Essays on financial incentives in the secondary healthcare sector

Katja Grašič

PhD

University of York

Economics and Related Studies

August 2022

DECLARATION

I declare that this thesis represents original work, of which I am the main author. Chapter 2 “Incentivising Hospital Quality through Care Bundling” is co-authored with James Gaughan, Nils Gutacker and Luigi Siciliani. Chapter 3 “Can financial incentives shift health care from an inpatient to an outpatient setting?” is co-authored with Luigi Siciliani. I am the sole author of Chapter 4 “The effect of DRG classification reform on coding intensity and reported costs: Evidence from England”.

In both co-authored papers I was the principal author. I have contributed to the research questions, developed the original aspects of the identification strategy and performed the empirical analyses. I have also written the first draft of the chapters. My co-authors advised on refinement of the research questions, the empirical strategy, and made comments on drafts of the chapter.

ACKNOWLEDGMENTS

My PhD was funded by the National Institute for Health Research (NIHR) Doctoral Research Fellowship Program (DRF-2016-09-097). Chapter 2 was further funded by NHS England. The views expressed are those of the authors and not necessarily those of the NIHR, NHS England or the Department of Health and Social Care. NHFD data (used in Chapter 2) are re-used with permission of Royal College of Physicians (Ref: FFFAP/2016/002). Hospital Episode Statistics (used in Chapters 2, 3 and 4) are copyright 2006-2015, re-used with the permission of NHS Digital (Ref: NIC-84254-J2G1Q). Welsh Admitted Patient Care Data Set (used in Chapter 4) are copyright 2006-2013, re-used with the permission of NHS Wales. All rights reserved.

First and foremost I would like to express my sincere gratitude to my supervisors, Luigi Siciliani and Martin Chalkley, who have given me continuous support and guidance throughout the entire PhD process. I would like to thank my Thesis Advisory Group member Andrew Street, for his advice and suggestions at the meetings and between them, and my co-authors James Gaughan, Nils Gutacker and Luigi Siciliani for their help and assistance. Special thanks goes to the admin and IT team at Centre for Health Economics for their help with data access and to Berry Puyk from NHS Wales for his help with the access to Welsh data. Finally, I would like to extend my thanks to all of the staff members at the Centre for Health Economics, and the Health Policy Team in particular, for their kindness, help and collegiality.

This thesis greatly benefited from invaluable comments and feedback. I would like to particularly thank Mathilde Peron, Ricarda Milstein, Anne Sophie Oxholm, Brendan McElroy, Jonas Schreyögg and participants at the the following events: EUHEA 2020 Autumn Series, Virtual iHEA 2021 conference, Winter HESG 2021, and seminar participants at the Hamburg Centre for Health Economics.

Last but not least, I wish to thank my close friends and family for all of their love and support. I could not have finished this without you.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

This thesis consists of three empirical essays, contributing to the understanding of key policy issues related to financial incentives in the healthcare setting. Chapters 2 and 3 contribute to a growing literature on the Pay for Performance (P4P) schemes in the secondary care, while Chapter 4 focuses on the design of the prospective payment reimbursement system.

Chapter 2 evaluates the impact of a financial incentive, designed to improve care of fragility hip fracture patients in England and implemented in 2010. The scheme adopts a bundled approach, by which nine different criteria related to quality of care must be met in order for the hospitals to receive bonus payment. Analysis is based on the difference-in-difference framework, with Wales as a control group. Results show large and statistically significant effect of the scheme on the uptake of the incentivised quality measures. Effects on patients mortality are small and mostly insignificant.

Chapter 3 considers a financial incentive scheme designed to shift inpatient activity to outpatient setting and implemented in England in 2012. The scheme targets three conditions and operates by overpaying the outpatient activity while concurrently underpaying inpatient activity. Using difference-in-difference approach, the results indicate large and significant effects of the policy on increasing the proportion of patients treated in the outpatient setting, without harming the quality of care or increasing the overall volume of activity.

Chapter 4 estimates the effect of a Diagnosis Related Groups (DRG) classification reform on hospitals' coding behaviour. The chapter considers a major reform in English DRG system in 2009 which highlighted the role of reported comorbidities in the reimbursement process. The analysis is based on the difference-in-difference framework. Results indicate significant effect of the reform on coding intensity, increasing the probability of being coded to a severe HRG and, consequently, the overall treatment cost.

CHAPTER 1

Introduction

Across the Organisation for Economic Cooperation and Development (OECD) countries, the government expenditure for healthcare has been steadily rising in the last decade by around 3.3% each year. The COVID-19 pandemic has further increased the health spending with an estimated growth of 4.9% in 2020 (OECD,2022). With the associated share of the GDP spending on healthcare rising from an average of 8.8% to 9.9% within a short time span, countries are increasingly seeking measures to improve the efficiency of the healthcare systems while maintaining or improving the quality of care. An important tool for policy makers to achieve these goals are various financial incentives, which are designed to motivate healthcare providers to deliver high quality care in an efficient way. The principal aim of this thesis is to gain better understanding of the effects of these incentives in the secondary care setting. This can assist in optimising the design of the policies in the future.

An increasingly popular stream of financial incentives, considered in Chapters 2 and 3, are the Pay-for-Performance (P4P) schemes. These aim to encourage the adoption of the clinical best practice, thereby reducing unwarranted practice variation and improving the overall quality and efficiency of the healthcare system. Typically targeting specific medical conditions, P4Ps vary greatly in their design and implementation. However, despite the increase in their popularity, there is lack of consensus regarding their effectiveness, with existing research studies reporting conflicting results (Milstein and Schreyoegg, 2016; Markovitz and Ryan, 2017). These differences might be partly explained by specific de-

sign features of P4P schemes, including the the size of the associated financial bonus and the particulars related to the selection and implementation of the incentivised measures (Mendelson et al., 2017; Emmert et al., 2012; Van Herck et al., 2010). This makes research on the specific elements of the P4P design especially relevant for policy makers.

One of the design issues, arising when concurrently incentivising several measures of quality, relates to whether the payment should be linked to the performance of each measure separately or whether it should be conditional on adequate performance across all measures, which we refer to as 'care bundling'. This issue is explored in Chapter 2, which focuses on a financial incentive designed to improve the care of fragility hip fracture patients in England. The Best Practice Tariff for hip fracture was introduced in all English hospitals in April 2010. It rewards providers based on a care bundle that consists of nine process measures (i.e. measures that focus on delivering a specific process rather than on patient outcomes). If the measures are jointly achieved, hospitals receive a bonus on a per-patient basis, which is paid on top of the base price. The design of the scheme and selection of the process measures was evidence-based. The size of the bonus is significant, and amounts up to 20% of the baseline tariff. This is considerably larger than the bonus typically rewarded in the P4P schemes, which is around 3-5% (Milstein and Schreyoegg, 2016).

The specific research questions considered in Chapter 2 are addressing the impact of the hip fracture BPT on i) providing care according to the quality indicators; ii) outcomes, measured as 30/90/365 days survival rates. This chapter further explores the 'care bundling' element of the scheme, including estimation of the effect by individual criteria. The analysis uses the National Fragility Hip Fracture data (NFHD) for England and Wales. The sample includes patients over the age of 60 who were treated for fragility hip fracture in the period from April 2008 to March 2015. The empirical approach in Chapter 2 is based on the difference-in-difference regression framework, by which we compare changes in outcomes between England and Wales, where the BPT was not implemented.

The results suggest that the policy was successful in increasing the proportion of patients

for whom all of the criteria were met by 52 percentage points. Results further suggest large heterogeneity across the measures, with the effect being smaller in areas in which the achievement was already high prior to the implementation of the policy. However, while the absolute achievement rates varied across criteria in both England and Wales in the pre-policy period, by 2014/15 English providers achieved comparable achievement rates across all of the criteria, regardless of the initial performance. In contrast, Welsh providers improved individual care processes in a less systematic way. The difference in response across countries indicates that the bundled element of the scheme focused the attention of English providers on all of the care processes, rather than on individual tasks. The scheme is further associated with a small reduction of 30/80/365 day mortality, albeit only the coefficient measuring the mortality 90-days post admission is statistically significant. Overall, Chapter 2 finds that a scheme based on care bundle, which is evidence based, coupled with a sizable bonus, can be effective in improving hospital performance.

While most P4P incentives focus on improving the quality of care, a small number of schemes directly target efficiency. This is often done by encouraging a shift from providing a more expensive treatment to a less costly alternative. However, this might come at an expense of lowering the quality of care (Chalkley and Malcomson, 1998). Chapter 3 contributes to the limited evidence on the effects of the P4P schemes that specifically target efficiency (Gaughan et al., 2019). Specifically, the chapter considers the BPT for outpatient services, which encourages providers to shift the care from the more expensive inpatient setting towards outpatient setting. The BPT focuses on three procedures: diagnostic cystoscopy, diagnostic hysteroscopy and hysteroscopic sterillisation. It operates by simultaneously increasing the price paid to hospitals for performing the procedure in the outpatient setting and decreasing the price for the inpatient procedures. This chapter specifically investigates the effect of the BPT on: (i) the choice of treatment setting (intensive margin); (ii) the quality of care; and (iii) total patient volume (extensive margin). It further considers spill-over effect of the BPT scheme on very closely related, but non-incentivised

procedures.

The analysis in Chapter 3 utilises patient level Hospital Episode Statistics (HES) data for Inpatient and Outpatient settings for the period from April 2009 to March 2016. The empirical strategy is based on the difference-in-difference framework, by which we compare the change in outcomes between the incentivised and the control procedures across the pre- and post- policy period. The sample includes patients who, during the study period, either had an incentivised BPT procedures or any of the procedures used to construct control groups (sigmodoscopy, lower genital procedures, vacuum aspiration with cannula). The latter were chosen based on the suitability to be performed in both, inpatient and outpatient setting, clinical relevance and pre-policy time trends. Sample further includes patients who had any of the two procedures used to test the spill-over effect of the BPT scheme (urethra endoscopic procedures, hysteroscopy with insertion of inuterine device). These two procedures were selected based on clinical similarity to the incentivised procedures.

The results show that a targeted incentive scheme can result in a swift and substantial change in the choice of the treatment setting. Estimates suggest a positive and significant effect of the policy on the probability to have the procedure performed in the outpatient setting for all three incentivised conditions, with the largest effect observed for cystoscopy and hysteroscopy (36.1 percentage points (pp) and 16.3 pp, respectively). The observed policy effect is smaller for sterilisation (3.8 pp). The BPT had no effect on quality of care, measured as the probability to have the procedure repeated within 60/90 days. Furthermore, the policy did not significantly increase the total patient volume. Estimates also suggest that the policy had a positive and statistically significant effect on shifting the treatment setting for closely related, but non-incentivised conditions. Results of Chapter 3 demonstrate that a financial incentive can be effective in shifting patients from inpatient to outpatient setting, without having a negative impact on either quality of care or patient volume. This gives policy makers a strong tool to increase healthcare efficiency and reduce costs.

While P4P schemes typically target specific treatments and have explicitly defined out-

comes measures, policy makers also try to improve the efficiency by implementing changes to the overall reimbursement system. In order to improve quality and efficiency of healthcare delivery, most OECD countries have implemented the Diagnosis Related Groups (DRG) classification system to standardise hospitals' reimbursement and encourage cost-containment (OECD, 2010). The system works by grouping patients into one of many DRG groups based on their clinical and demographic characteristics. Each DRG group then attracts a fixed payment, regardless of the patient's specific care pathway or the actual treatment cost. The assumption behind the DRG based reimbursement system is resource homogeneity within the groups. Where there is large variation in the resource use within a DRG, the payment is either too high or too low for many patients, penalizing hospitals with unfavourable patient casemix. Countries typically respond to this issue by refining the DRG system and creating new groups to better account for differences across patients. This increases the role of reporting complications/co-morbidities and may create an incentive for hospitals to upcode. There is a lack of evidence on the extend to which a classification reform changes the coding practice across hospitals. Chapter 4 contributes to filling this gap in knowledge.

To improve resource homogeneity across the English Healthcare Resource Groups (HRGs), in 2009 all of the existing 600 HRG groups were replaced with 1500 new groups. The aim of the Chapter 4 is to estimate the causal effect of this classification revision on hospitals' coding behaviour. Chapter employs difference-in-difference modelling approach, comparing changes in coding and treatment intensity across two distinct healthcare systems with similar population cohorts. In particular, chapter compares changes in coding for patients treated in hospitals in England (treatment group) to those treated in Wales (control group), where the classification revision did not affect hospital reimbursement. Analysis relies on two main data sources. For information on patients treated in England, the patient-level Hospital Episode Statistics (HES) dataset is used. This contains comprehensive data on patient's care pathway, including their socio-demographic and clinical

details. Information on Welsh patients comes from Admitted Patient Care (APC) dataset.

The results indicate a significant effect of the HRG classification revision on the intensity of clinical coding. Refinement of the HRG groups increases the probability of being coded to a 'severe' HRG (indicating complications and comorbidities) by 3.3 percentage points, while increasing the average price paid to hospitals by £58.5. Reform significantly increased the number of reported diagnosis codes by 0.56, with no effect on procedures. Chapter also analyses the severity of the reported procedure and diagnosis codes, based on the expected resource use for each reported procedure and diagnosis code. Results suggest there was no increase in the severity for either diagnoses or procedures. This indicates that the change in HRG composition is mainly driven by more extensive coding of diseases, rather than changes in the treatment pathway. Results further indicate that the effect is largest for hospitals with perceived easier casemix in the pre-policy period, suggesting that these hospitals were *catching up* in coding to increase their marginal utility following the reform. Overall, results of Chapter 4 suggest that policy makers must balance an increase in quality of coding and fairness across providers with associated increase in healthcare expenditure when considering a reform of the DRG classification system.

Taken together, all three chapters contribute to the literature on the impact of financial incentives on hospital care. This thesis demonstrates that financial incentives can be a powerful tool to change behaviour and increase the efficiency and quality of care, in particular when the associated financial reward is large. Importantly, as shown in Chapters 2 and 3, this can be achieved without increasing the overall healthcare spend.

CHAPTER 2

Incentivising Hospital Quality through Care Bundling

2.1 Introduction

Policymakers are increasingly implementing pay-for-performance (P4P) schemes to incentivise the adoption of best practice, thereby reducing unwarranted practice variation and improving the overall quality and efficiency of the healthcare system.[1] Despite the international movement towards P4P schemes, the evidence about their effectiveness remains inconclusive with some studies reporting substantial improvements in quality whereas others fail to identify them (Milstein and Schreyoegg, 2016; Markovitz and Ryan, 2017). Design features of the P4P schemes, such as the size of the financial incentives or the modalities by which payments are determined, may drive some of this heterogeneity (Mendelson et al., 2017; Emmert et al., 2012; Van Herck et al., 2010).

Most P4P schemes incentivise improvements in process measures of quality, while others incentivise health outcomes (Milstein and Schreyoegg, 2016). A key design issue in P4P schemes with multiple incentivised performance measures is whether separate bonus payments should be made for each measure provided, or whether a single bonus payment should be made conditional on all measures being provided jointly, which we refer to as

[1] Recent examples within secondary care include the Hospital Readmission Reduction Program in the US, and Advancing Quality in the UK.

"care bundling". Conditioning payment on a bundle of processes can give strong incentives to providers to deliver them all, but may discourage some providers to deliver any at all if they find at least one of the processes to be particularly costly.

This study contributes to the literature on P4P and its design features by analysing the effectiveness of a national P4P scheme that incentivises hospitals through a single additional payment for every patient that receives a care bundle of nine process measures. The Best Practice Tariff (BPT) for fragility hip fracture, introduced in the English NHS in 2010, seeks to ensure timely access to surgery, involvement of geriatricians throughout the entire care pathway, and tertiary prevention of fractures. These process measures reflect best practice standards developed by the British Geriatric Society and the British Orthopaedic Association (2007) on the basis of clinical evidence and professional consensus. BPT payments are conditional on the delivery of the *entire* care bundle, i.e. hospitals do not receive the bonus payment for patients for whom one or more process measures are not achieved. A second distinctive feature of this P4P scheme is that the size of the financial incentive is economically significant and amounts to up to 20% of the baseline episodic payment to hospitals. This is important because one of the reasons for the lack of provider response is the relatively small payment (typically around 5% of the revenues (Cashin et al., 2014b)).

After developing a theoretical model of provider incentives under care bundling, we implement a difference-in-difference (DID) strategy to identify the causal effect of the BPT policy on care provision in England using Wales as a control group. Both countries have similar healthcare systems and share key institutional features such as training and regulation of healthcare professionals, free care at the point of use, and population demographics. Furthermore, hospitals in both countries report to the same clinical audit, the National Hip Fracture Database (NHFD), which ensures that achievements of the incentivised care standards are recorded and disseminated to the public in a consistent way. We then estimate the relationship between changes in care provision and patient health outcomes using a two-way fixed effects models at hospital level. Both sets of results are combined to estab-

lish the overall effect of the BPT payment policy on patient survival that operates through measurable improvements in care quality.

Our results suggest that a P4P scheme based on care bundling, which is evidence based and awards a sizable bonus payment, is effective in improving hospital performance to a large extent. Specifically, we find that the BPT increases the proportion of patients who receive the complete care bundle by 51.7 percentage points (pp). There is considerable heterogeneity in the impact of the BPT on the set of process measures that are incentivised with the largest improvements occurring in the involvement of geriatricians in the care process (+20 to 65 pp) and much smaller effects in e.g. falls prevention (+6 pp). The size of the improvement across process measures is inversely related to pre-policy achievement levels as English hospital seek to establish a similar level of achievement across all process measures to maximise pay-out. We do not find evidence that English providers continue to exert efforts to deliver targeted process measures once they have failed to meet at least one measure, i.e. when the financial incentive is removed. Based on our results, we estimate that the introduction of the BPT policy may have helped to improve 30-days survival rates in hip fracture patients by 0.3 percentage points.[2]

Our analysis makes several contributions to the literature and offers new insights into the optimal design of P4P schemes and the behavioural response of providers. First, it provides evidence on the use of a P4P payment rule which incentivises a care bundle to incentivise quality of care. Our empirical results show that the care bundle payment can be effective in stimulating provider effort. Unlike the more common P4P arrangements with a separate payment for each of the incentivised measures, care bundles have only one payment which is conditional on satisfying all the incentivised measures in an all-or-

[2]A previous study by Metcalfe et al. (2019) found that the introduction of the hip fracture BPT in the English NHS led to a 1.7 percentage point reduction in 30-day mortality compared to a control group of patients treated in Scotland. However, these effects appear to be driven by a worsening of outcomes in Scotland, rather than marked improvements in outcomes in England, which suggests that the estimated effect may not be due to the BPT alone. Metcalfe et al. (2019) did not examine how the BPT policy affects hospital achievements of incentivised process measures, a prerequisite for a causal effect on mortality, nor how the care bundle approach affects hospital decision-making. We extend previous work in these directions.

nothing approach. This creates a distinct decision problem for the provider, who needs to balance the cost of delivering all the measures against the prospect of a single payment. We characterise these incentive issues in the theoretical model in Section 2.3. The key insight is that for some providers the bundled price will increase the scope of providing care to patients as this is the only way to gain the payment. Instead, other providers with a relatively high cost of one of the process measures of quality may not respond at all under care bundling, while they would have partially responded under a scheme incentivising each individual process. The comparison again highlights the financial incentive given by the bundled payment to provide *all* or *nothing*. We also show that this insight holds even in the presence of synergies on costs across processes and that, as intuitively expected, the presence of cost synergies increases the scope of providing bundled care, for a given incentive scheme. We also briefly show that the case for a bundled payment is reinforced by the presence of synergies on health benefits across care processes.

Second, our study provides new evidence on the marginal contribution of P4P over and above other common policy leavers, such as the dissemination of clinical guidelines, increased monitoring of care processed, and public reporting of comparative performance information. Existing P4P schemes have often been implemented alongside other policy leavers thereby making it difficult to isolate the effect of financial incentives. For example, in order to operationalise the UK Quality Outcome Framework, the largest and most widely studied P4P scheme in primary care internationally, family doctors were given new quality standards and were required to improve their data recording and monitoring systems. Data on the quality of care of each practice were reported to the payer on a regular basis and were published in the public domain to inform patient choices. In contrast, the fragility hip fracture BPT draws on an existing data collection and incentivises care standards that had been agreed upon previously and where provider performance was already reported publicly. Hence, both the treatment (England) and the control (Wales) group in our study experience the same non-financial stimuli. This increases our confidence that any

difference in post-policy behaviours can be attributed to the BPT policy.

Third, our study is one of few to evaluate a high-powered quality improvement scheme with considerable bonus payments of up to 20% of baseline payments. As mentioned above, the lack of response to previous P4P schemes may have been due to the small size of the financial bonuses (Milstein and Schreyoegg, 2016), which typically do not exceed 5% of the base price (Cashin et al., 2014b) and, thus, may be insufficient to compensate providers for their additional costs. The size of the BPT bonus increased throughout the study period from 7% to 20% of the baseline payment, which permits us to test empirically how the size of financial incentives affects provider behaviour.

Fourth, we contribute to a sparse literature on the effect of P4P in the clinical area of hip fracture care. Fragility hip fractures are common in elderly people and are a lead cause of mortality and morbidity, with associated disability, need for long-term institutional care, and high medical costs (Tajeu et al., 2014). In 2000, an estimated 1.6 million hip fractures occurred worldwide, and this number is expected to increase to 6.3 million by 2050 (Cooper et al., 2011). While P4P has been implemented in various settings covering a range of conditions, we are not aware of other P4P schemes for hip fracture outside of the English NHS, despite its high health burden. In this study we show that a successfully implemented P4P scheme can improve process quality and outcomes for hip fracture patients.

The study is organised as follows. Section 2.1.1 reviews the related literature. Section 2.2 discusses the institutional setting in which providers operate. Section 2.3 outlines the theoretical implications of the bundled payment arrangement. Sections 2.4 and 2.5 describe the data and the details of the empirical approach. Section 2.6 presents the empirical findings. Section 2.7 is devoted to discussion and concluding remarks.

2.1.1 Related literature

Our research contributes to the literature within the broader area of hospital incentive schemes. Milstein and Schreyoegg (2016) reviewed P4P programs covering the inpatient

setting across the OECD countries and found that, out of 34 programs in the sample, approximately half lacked any statistical evaluation. The existing evaluations often experienced design issues, including lack of a suitable control group, which may lead to potential bias. While the review uncovered large heterogeneity across the programmes in respect of incentive design and clinical areas covered, they were all typically associated with small size of the P4P bonus and generated only limited improvement in performance. Additional P4P reviews confirm the modest effect to P4P schemes on changing providers' behaviour (Mendelson et al., 2017; Eijkenaar, 2012).

Existing reviews also commonly suggest that the reason for the limited success of the financial incentives lies in the particular design features (Eijkenaar, 2013; Ogundeji et al., 2016b; Milstein and Schreyoegg, 2016; Scott et al., 2016). A meta-analysis of P4P effects estimates shows that P4P schemes incentivising process measures generate larger responses than those incentivising health outcomes measures (Ogundeji et al., 2016b). Eijkenaar (2013) compares different remuneration methods, including separate payment for each P4P incentive and the *"all-or-nothing"* arrangement where providers receive bonus payments only once a certain threshold across patients is met[3]. The study finds advantages and limitations of different payment arrangements, noting that the optimal financial bonus structure depends on the specific incentivised program. Milstein and Schreyoegg (2016) finds that the payment based on absolute scores is usually preferred over the relative rankings, while Scott et al. (2016) finds that schemes which base reward on improvements over time have lower probability of being effective compared to those that reward performance at a single time point.

In contrast to the existing literature, the care bundle arrangement used in the BPT for hip fracture in this study, by which a provider must to meet several criteria on a per-patient basis to receive the bonus, does not typically feature in the P4P payment design. This is

[3]This approach differs from the care bundling as it is not based on several quality measures. Instead, it requires a single criterion/quality measure to be met for a defined proportion of all patients for the hospital to receive the payment

despite evidence based care bundles being often considered to deliver best clinical care and hence promoted as best practice. A study of very-low birth-weight babies from 32 neonatal departments in Germany found that an intervention bundle is feasible and can reduce blood-stream infections in neonatal departments (Salm et al., 2016). López-Cortés et al. (2013) found that and evidence based bundle of six adjunctive measures improved the management of patients with bacterial blood infections and reduced mortality. Similar results were found for by Takesue et al. (2015). Care bundles were further shown to be effective in the ICU to reduce ventilator-associated pneumonia and improve outcomes (Resar et al., 2005). Our study shows that P4P can be an effective tool for policy makers to promote adherence to the evidence based care bundles, hereby improving care processes and, subsequently, patient outcomes.

More specifically, our analysis extends and improves previous studies evaluating BPT for hip fracture. The initial assessment of the BPT hip fracture scheme (McDonald et al., 2012) suggests a positive effect of the policy on the uptake of four criteria that were included in the study. The analysis was based on aggregated hospital level data covering one year pre- and post-policy. Our study extends this study by using a longer pre- and post-policy period and providing causal estimates of the effect of the BPT policy, as well as exploring heterogeneous effects and potential mechanisms at work. While the evidence of the effect of hip fracture BPT on adherence is limited, there is some research on its effect on patients outcomes, measured by mortality. Metcalfe et al. (2019) compared the outcomes of patients in England to those in Scotland (which is not part of NHFD) and found a 1.7 percentage point reduction in mortality between 2010 and 2016 as a result of the BPT in England. Similarly, Neuburger et al. (2015) identified a statistically significant fall in 30 day mortality in England by 1.8% one year after the introduction of the BPT.

2.2 Institutional background

Public healthcare in the UK is funded through general taxation and is free to patients at the point of use. The delivery of healthcare is decentralised, with each of the four countries of the UK (England, Scotland, Wales, and Northern Ireland) operating their own National Health Service (NHS) to provide primary and secondary care services to their resident populations. From the founding of the UK public healthcare system in 1948 until the political devolution in 1999, England and Wales operated a common NHS with shared resources and policies. While priorities and policies in both countries have begun to diverge since then, the organisation of the health services still remains broadly comparable to this date. For example, both systems have similar healthcare expenditures per capita[4] and they continue to share the same professional regulation (e.g. on clinical training, conduct, and fitness to practice) and similar pay structure for their doctors and nurses (OECD, 2016). Care pathways for hip fracture patients are also similar, with patients in both countries accessing emergency care either by presenting at an hospital emergency department (ED) (e.g. arriving by ambulance, self-referral) or by urgent referral from their family doctor. Clinical guidelines for hip fracture care are issued by the National Institute for Health and Care Excellence (NICE) and apply to both countries equally.[5]

One important aspect in which the English and Welsh NHS differ is in how they reimbursement hospital providers for the care they deliver. Welsh hospitals are paid via a capitation system, where each hospital receives a lump sum that is linked to the size of local population they serve; not to the actual volume or quality of service provided. Reimbursement for the hip fracture patients is included in this sum and there is no further bonus paid to hospitals that meet best practice standards. Conversely, hospitals in England are reimbursed via a prospective payment system that was introduced in 2003 and now covers

[4] In 2014/15, the English NHS spent £2,055 per capita on healthcare, whereas the Welsh NHS spent £2,083 per capita (Harker, 2014).

[5] The current clinical guideline CG124 details the appropriate management of hip fracture and has been in place since 2011. The fragility hip fracture BPT incentivises many elements of this guideline.

more than 60% of total hospital activity (Grasic et al., 2013). Patients are categorised into distinct healthcare resource groups (HRGs; similar to DRGs in other countries) according to their age, severity and the care that was provided. Hip fracture patients fall mostly into one of 10 orthopaedic HRGs, typically related to major hip procedures, although patients may also be grouped to other HRGs if there are significant concomitant medical conditions (e.g. a stroke).

Until March 2010, English hospitals were paid a base price for each hip fracture patient, where the price reflected the historical average costs of treating patients in this particular HRG. In April 2010, the BPT for fragility hip fracture was introduced to incentivise hospitals to deliver best practice care processes according to the definitions set out by the relevant medical societies[6] (Department of Health, 2010). Under this system, hospitals now receive a lower base price (P_0) for all patients, irrespective of the quality of care provided. In addition, they can earn a relatively large bonus payment (P_b) on top of this base payment for each patient for whom the full set of BPT criteria are met. In the financial year[7] 2010/11, this bonus payment amounted to £412, which was subsequently increased to £800 in 2011/12 and to £1,350 in 2012/13.[8]. During our study period[9], hospitals received $P_0 + P_b$ for each patient for whom all of the following criteria were met, and P_0 otherwise:

- BPT 1: Surgery within 36 hours;
- BPT 2: Shared care by surgeon and geriatrician;
- BPT 3: Care protocol agreed by geriatrician, surgeon and anaesthetist;

[6]The BPT for hip fracture was one of the original four BPTs introduced in 2010; the other three incentivised conditions included cholecystectomy, stroke and cataracts. BPT has since expanded and as of 2022 covers more than 80 different conditions. The reimbursement rules and bonus size differ across the conditions.

[7]Financial years run from 1st of April to 31st of March of the following year.

[8]The average HRG base price for hip fracture patients in 2014/15 was £6,369 (England, 2014). With the average hospital treating 450 eligible patients every year, this represents an average potential bonus of £0.6M or approximately 0.15% of the average total hospital budget.

[9]In 2017, the list of BPT criteria was revised and some new criteria were introduced whereas some old ones were removed.

- BPT 4: Pre-operative cognitive function assessment (introduced in 2012);
- BPT 5: Post-operative cognitive function assessment (introduced in 2012);
- BPT 6: Perioperative assessment by geriatrician;
- BPT 7: Geriatrician-led multidisciplinary rehabilitation;
- BPT 8: Secondary prevention including falls;
- BPT 9: Bone health assessment.

In line with these BPT criteria, patients should be operated on within 36h from the time they present at the ED or - if the patient was not admitted via the ED - the time of diagnosis. This reflects empirical evidence that timely surgery can improve survival, decrease length of stay and the incidence of pressure ulcers, and facilitate a return to independent living (Lee and Elfar, 2014). The BPT also greatly emphasises the role of ortho-geriatricians in the treatment of hip fracture patients, with four of the nine criteria requiring their direct involvement (BPT2, BPT3, BPT6, BPT7). Geriatricians are expected to see each patient in the perioperative period, i.e. within 72h of admission (BPT6), to ensure their fitness for surgery. They also should coordinate with the orthopaedic surgeon and agree on the type of care the patient should receive (BPT2). Furthermore, they should be involved in the development of care protocol for patients with hip fracture (BPT3) and coordinate their activities with the rehabilitation team (BPT7), which has been shown to reduce length of inpatient stay (Cameron, 2005; Kalmet et al., 2016; Lau et al., 2017). The latter two criteria are not necessarily achieved for each patient separately, but serve as a general set of rules for patients treated for hip fracture. Since April 2012, all patients must also receive a simple "memory test" type of assessment at two time points (BPT4, BPT5). Finally, patients with hip fracture are at increased risk of falls, and the BPT therefore incentivises preventive activities such as medication review, physiotherapy work to improve strength and balance, and an assessment of the home environment (BPT8). Because many hip fracture patients

have osteoporosis, they should undergo bone strengthening treatment and/or bone density scans (BPT9) to reduce the risk of future fractures.

These criteria follow closely national clinical standards for hip fracture care set by the British Orthopaedic Association and British Geriatrics Society and monitored through a collaborative clinician-led audit, the National Hip Fracture Database (NHFD)[10], which was launched in April 2007 (Neuburger et al., 2015). The aim of the audit is to comprehensively describe the quality of care delivered to fragility hip fracture patients in the four UK countries, and to facilitate benchmarking among hospital providers and health care systems (British Orthopaedic Association, 2007). All participating hospitals benefit from regular regional and national meetings for peer support. Performance data has been published in annual reports since 2009 (covering the period from October 2007 to September 2008), creating non-financial incentives to improve in both countries. The hip fracture BPT relies on the data from the NHFD to assess compliance against the incentivised criteria. It's important to note that both England and Wales follow the same NHFD guidelines since 2007, with both countries incentivised to perform well on the BPT scores due to benchmarking and public reporting of results.

2.3 Theoretical framework

To fix ideas, we provide a simple model of provider behaviour under a payment which rewards bundled care, and then compare to a system where the payment is not bundled. Our model contributes to the health economics theoretical literature on P4P, which has focused on multi-tasking (Eggleston, 2005; Kaarboe and Siciliani, 2011; Mak, 2018), gaming (Kuhn and Siciliani, 2009) and selection effects (Lisi et al., 2020), and more broadly to the literature on provider incentives (Ellis and McGuire, 1986a; Ma, 1994b; Chalkley and Malcomson, 1998). However, none of these studies focuses on bundling.

[10]The NHFD is now commissioned by the Healthcare Quality Improvement Partnership and managed by the Royal College of Physicians as part of the Falls and Fragility Fracture Audit Programme.

Consider a hospital treating an emergency patient. The hospital can treat the patient with some basic care, or can provide the patient with two additional care processes, 1 and 2, that generate additional patient benefits (and costs for the provider). For simplicity, we assume that all the patients are the same and do not differ in severity. The hospital has four different treatment options: i) a basic treatment, ii) the provision basic treatment and process 1, but not care process 2; iii) the provision of basic treatment and care process 2, but not care process 1; and iv) the provision of basic treatment and care processes 1 and 2.

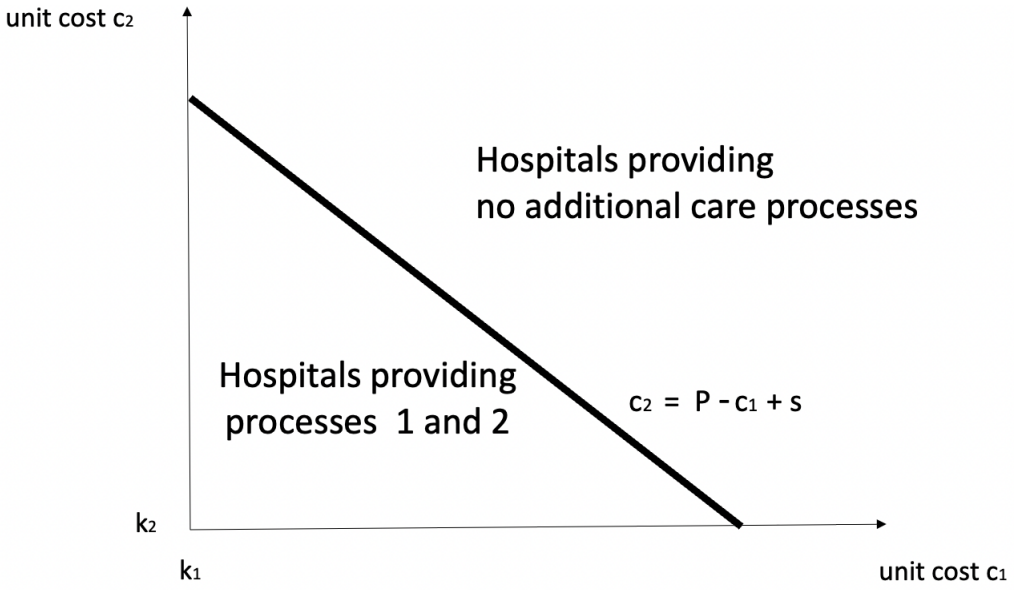
The cost of process 1 and 2 is respectively defined with c_1 and c_2. If both care processes are provided, the hospital sustains costs c_{12}. We allow for the possibility of cost synergies so that $c_{12} = c_1 + c_2 - s$, where $s \geq 0$ measures the extent of the cost synergies. Similarly, the patient benefit from process 1 and 2 is respectively defined with b_1 and b_2, and b_{12} if the processes are jointly provided.

We assume that providers maximise their financial surplus, defined with π_i, where $i = 1, 2, 12$ denoting the cases when care processes 1, 2 or both are provided. We further assume that providers differ in costs, e.g. to reflect different degree of efficiency in providing processes 1 and 2, so that c_1 and c_2 are distributed with joint density function $f(c_1, c_2)$ over the support $c_1 \in [k_1, +\infty)$, $c_2 \in [k_2, +\infty)$.

Under bundled payment, we assume that the funder pays a price P if both care processes are provided, and zero otherwise (and this price is above the sum of minimum provider costs, $P > k_1 + k_2$). Under these arrangements, the provider has an incentive to provide both processes if $\pi_{12} > 0$, or, more explicitly, if $P > c_1 + c_2 - s$. The solution is described in Figure 2.1. It shows that only providers with relatively lower costs on care processes 1 and 2 have an incentive to provide them. Providers with relatively high costs provide neither processes. It is also immediate to see that higher cost synergies or a higher bundled payment will induce more hospitals to provide both care processes (the diagonal line shifts upwards).

Notice that each provider has never an incentive to provide only one of the two pro-

Figure 2.1: Bundled payment

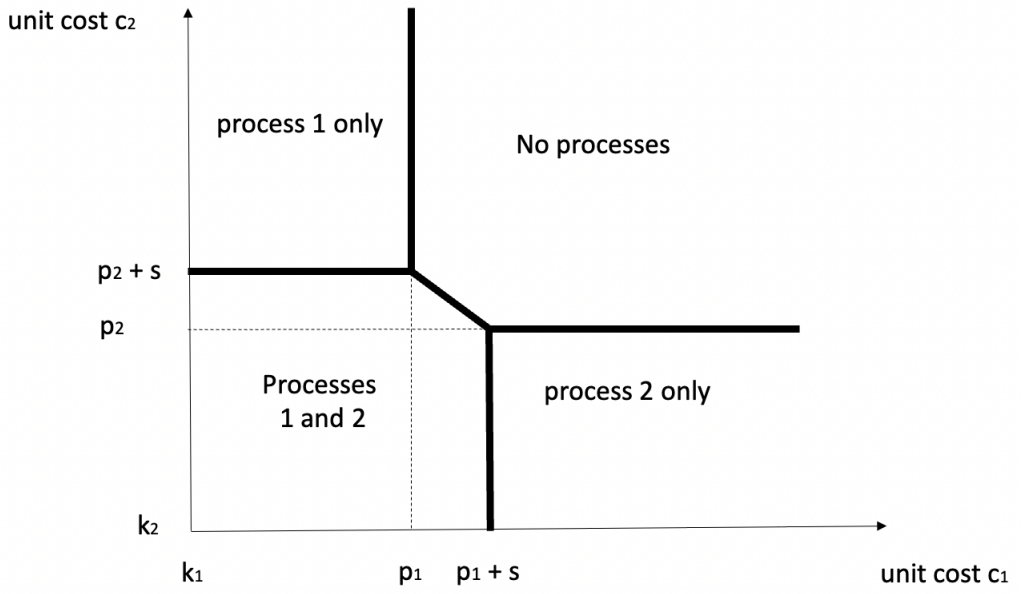


Notes: The image shows the incentives for providing processes 1 and 2 under bundled payment arrangement.

cesses based on financial ground. This would instead *not* be the case in a system where the payment is not bundled. To see this, suppose that the purchaser pays a price p_1 every time process 1 is provided, and p_2 every time process 2 is provided, where each of these prices are above the minimum provider costs ($p_1 > k_1, p_2 > k_2$). Under these payment arrangements, three possible scenarios arise. In the first one, the hospital provides only process 1 if providing process 1 is profitable, $\pi_1 > 0$, and providing process 1 is more profitable than providing both processes, $\pi_1 > \pi_{12}$. These two inequalities reduce to $c_1 < p_1$ and $c_2 > p_2 + s$. In the second scenario, the hospital provides only process 2 if it is profitable to do so, $\pi_2 > 0$, and if it is more profitable than providing both processes, $\pi_1 > \pi_{12}$. These two inequalities reduce to $c_2 < p_2$ and $c_1 > p_1 + s$. Finally, the hospital provides both processes if it is more profitable to provide both processes than just one of the two processes, i.e. $\pi_{12} > \pi_1$ and $\pi_{12} > \pi_2$, or $c_2 < p_2 + s$ and $c_1 < p_1 + s$, and the profits of providing both processes are positive, $\pi_{12} > 0$. Figure 2.2 illustrates the solution. As intuitively expected, hospitals with relatively low cost of process 1 and high cost of process 2 provide

only process 1. Hospitals with high cost for both processes provide none, and providers with relatively low costs provide both. It is still the case that higher cost synergies increase the number of providers who respond to the financial scheme, for a given pair of prices.

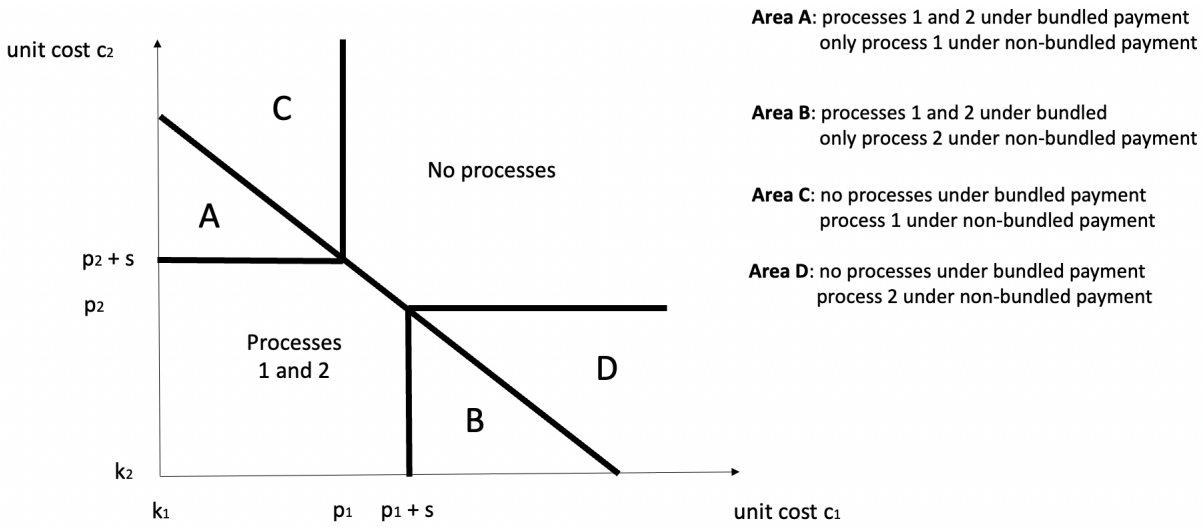
Figure 2.2: Non-bundled payment



Notes: The image shows the incentives for providing processes 1 and 2 under non-bundled payment arrangement.

To further compare the solutions under a bundled payment with a system when payment is not bundled, in Figure 2.3 we set $P = p_1 + p_2$, so that the bundled price is the sum of the prices when the payment is not bundled and the condition for $\pi_{12} > 0$ under bundled payment is $c_2 < p_1 + p_2 - c_1 + s$. Figure 2.3 illustrates that providers that fall in areas A and B provide both processes under bundled payment, while only one of the two processes when the payment is not bundled. Therefore, in some cases, the bundled price increases the scope of providing additional care to patients. Conversely, providers in areas C and D provide no processes under bundled payment, but provide one of the two processes when one of the two bundles is not provided. Therefore, in other cases, the bundled price reduces the scope of providing care to patients. The comparison again highlights the strong financial incentive given by the bundled payment to provide "all" or "nothing".

Figure 2.3: Comparison of bundled and non-bundled payment



Notes: The image shows the difference in incentives for providing processes 1 and 2 under bundled versus non-bundled payment arrangement.

We conclude the comparison by discussing possible effects on patient health benefits under different payment schemes. To illustrate the key effects at work, consider the special case where the number of providers under area A is the same as under area C, while the number of providers under area B is the same under area D. In this special case, the total payment of the funder to providers involves the same spending given the assumption that $P = p_1 + p_2$. It is straightforward to show that if patient benefits present no synergies, so that $b_{12} = b_1 + b_2$, then total patient benefit is also the same. It is only if patients benefits present synergies across care processes, so that $b_{12} > b_1 + b_2$, then total patient benefit will be higher under bundled payment. This illustrates one a possible benefit of bundled payment, where synergies in benefits (in addition to costs) are advocated as a rationale for providing different care processes under the same payment.[11]

[11]The discussion relies on the strong assumption that the total number of providers is constant under the two schemes. This is a strong assumption, as the number of providers are likely to differ. However, the additional effects when the number of providers in areas A and C or B and D differ are predictable and depend on the price levels under each scheme, and the shape of the joint density function. A full welfare analysis is beyond the scope of this section, which mostly aims at illustrating provider incentives under different schemes.

2.4 Data

Our empirical analysis is based on audit data from the NHFD for England and Wales for the period from April 2008 to March 2015. The NHFD collects information on care quality for patients aged 60 or over who are admitted to a public NHS hospital with a fragility hip fracture.[12] We use the recorded information to derive binary indicators of BPT achievement (overall, and by criterion) for each patient in accordance with the BPT payment rules. We define the overall BPT achievement measure as meeting seven criteria in the period from April 2008 to March 2012 and eight criteria in the period from April 2012 to March 2015. The indicator for the pre-operative cognitive function assessment (BPT4) is included in the overall measure from April 2012 onwards, when it was first used for BPT payment purposes. The indicator for post-operative cognitive function assessment (BPT5) was not collected in the NHFD prior to April 2011 and is, therefore, excluded from all analyses.

The NHFD provides information on patients' socio-demographic characteristics including age (recoded to 5-year age brackets), sex, and admission source. The latter gives an indication on patients' location at the time of the fracture: already in the hospital, in a care home, in a rehabilitation centre, or at home. NHFD further provides information on patients' clinical characteristics, including their predicted operative risk, as measured by the American Association of Anaesthesiologists (ASA) physical status classification system. Values range from one for a healthy patient to five for a patient who is not expected to survive. Values two to four detail the progressive severity of systemic disease. Further variables describing patients' medical condition include fracture type (based on the anatomical location of the fracture), which is coded as either intracapsular or extracapsular, and patient-reported level of mobility before the fracture. The latter can take four levels: freely mobile, some mobility, no outdoor mobility and no functional mobility. The socio-demographic and medical characteristics are used in our empirical analysis to adjust

[12] Private provision of hip fracture treatment is not recorded but is likely to be close to zero given the emergency nature of the condition. Private hospitals in the UK focus on selected elective procedures and do not normally have the required capabilities to treat emergency patients.

for differences in case-mix across providers following the risk-adjustment methodology developed by the Royal College of Surgeons (Tsang and Cromwell, 2014).

The initial sample includes 334,850 observations, out of which 317,574 (95%) are in England and 17,276 (5%) in Wales[13]. In the main analysis we exclude observations with any of the variables recorded as 'unknown' or missing. This applies to 55,697 (17.5%) observations in England, and 3,255 (18.8%) in Wales, reducing the sample to 275,898 patients, of which 14,021 were treated in Wales. Appendix Table A1 reports the number of hospitals and patients in our estimation sample by year and country. Participation in the audit is voluntary and has increased steadily since its inception (from 5 hospitals to 13 in Wales; from 111 to 166 in England), reaching full coverage of eligible hospitals in both countries by 2012/13.

Table 2.1 provides descriptive statistics of patient characteristics for both countries. In the pooled sample across years and countries, the average age of patients is 83 years, with the vast majority (83%) being female. The bulk of patients (78%) are admitted from residential homes. More than half of all patients are classified as having severe systematic disease (ASA score 3) with approximately 12% facing constant threat to life (ASA score 4 or 5). 36% of patients report being freely mobile before the fracture, with 2% reporting no functional mobility.

Additionally, to investigate the association between BPT achievement and patients' health, we extract information on whether patients are alive at 30/90/365 days following admission. Mortality information was not included within our NHFD data extract and Welsh mortality data from other sources were similarly not available. We therefore conduct this analysis for England only, by combining the provider level BPT achievement rates with provider level mortality information. The latter are derived by linking admission data from Hospital Episode Statistics (HES) dataset with official death records from the Office for National Statistics (ONS). HES is an admission-level dataset that provides

[13]Hip fracture incidence rates are comparable across England and Wales (Curtis et al., 2016). Differences in the number of cases reflect difference in the size of the population of both countries

Table 2.1: Descriptive statistics of patient characteristics

Patient characteristic	Mean / %	
	Wales (N=14,021)	England (N=261,877)
Age (in years)	82.2	82.6
Male sex	27.9%	26.9%
Admission source		
Hospital	4.8%	3.3%
Long-term / nursing care home	15.7%	18.0%
Residential home	79.0%	78.1%
Other	0.5%	0.7%
ASA grade		
1 (Normal healthy patient)	2.7%	2.6%
2 (Mild systemic disease)	31.3%	30.9%
3 (Severe systemic disease)	53.7%	54.8%
4 (Severe systemic disease, constant threat of life)	11.8%	11.3%
5 (Moribund, not expected to survive)	0.5%	0.4%
Pre-fracture mobility		
Freely mobile before operation	34.8%	35.8%
Full or partial outdoor mobility	25.4%	26.4%
Some indoor mobility	37.8%	35.7%
No functional mobility	2.0%	2.1%
Fracture type		
Intracapsular	56.9%	58.4%
Extracapsular	43.2%	41.6%

Notes: This table presents descriptive statistics for our full sample, reported separately for England and Wales. Data are pooled over the entire study period April 2008 to March 2015.

detailed information on patients' clinical and socio-demographic characteristics. For the mortality analysis, we limit our sample to patients aged 60 or over who are treated in English NHS hospitals in the period April 2010 to March 2015 and are eligible for inclusion in the BPT scheme based on their reported diagnosis, procedure and HRG codes (NHS England, 2016)[14]. We limit our sample to the post-policy period only, as the overall BPT achievement rates are zero in the pre-policy period. We first risk-adjust the mortality data and then aggregate them at a quarterly basis for each provider.[15] We merge these quarterly mortality rates with hospital-level BPT achievement figures (i.e. the proportion of patients for whom a criterion is met in the same calendar quarter) obtained from the NHFD. More details about the risk-adjustment procedure are provided in the Appendix.

Table 2.2 presents descriptive statistics for 30/90/365-day mortality rates in the post-policy period. We observe a reduction in over time across all mortality.

Table 2.2: Hospital mortality rates following hip fracture care in England

	30-day mortality		90-day mortality		365-day mortality	
Financial year	Mean	SD	Mean	SD	Mean	SD
2010/11	0.071	0.031	0.152	0.045	0.272	0.054
2011/12	0.068	0.029	0.140	0.047	0.264	0.058
2012/13	0.067	0.033	0.147	0.047	0.270	0.052
2013/14	0.055	0.024	0.120	0.036	0.240	0.047
2014/15	0.056	0.029	0.128	0.046	0.254	0.051
Overall	0.067	0.032	0.144	0.050	0.265	0.057

Notes: Statistics are calculated from quarterly hospital-level mortality data derived from HES. Each observation is a hospital-quarter combination. The data cover the period Q2/2010 to Q1/2015.

[14]Because some providers did not report to the NHFD at the beginning of our study period, the hip fracture sample from HES is larger than the NHFD sample and consists of 373,274 admissions.

[15]We use calendar quarters rather than months as a measure of time to avoid issues with small number of reported patient cases in the denominator of the mortality rates.

2.5 Methods

2.5.1 Effect of the policy on performance

The aim of our main analysis is to establish a causal link between the introduction of the fragility hip fracture BPT and subsequent changes in the delivery of the incentivised care bundle for hip fracture patients in England. Observed improvements in care quality following the start of the payment policy may be confounded by external effects such as changes in the healthcare production technology, preferences and beliefs among clinical staffs, or demography. We therefore employ a DID approach with patients admitted to English hospitals forming the treatment group and those admitted to Welsh hospitals being in the control group.[16] Both healthcare systems have similar organisational characteristics, serve similar patient populations, and are subject to the same clinical guidelines (see Section 2.2). Staffing levels of ortho-geriatricians, a key input required to meet several BPT criteria, were also comparable prior to the introduction of the payment policy (NHFD, 2010). Hence, observed care quality in Wales during the post-policy period can serve as a credible counterfactual estimate for England.

Our base model takes the following form:

$$Y_{iht} = \alpha + \theta(England_i \ D_t) + \boldsymbol{X_i}'\delta + \boldsymbol{v_t} + \boldsymbol{v_s} + \boldsymbol{v_h} + \epsilon_{iht} \tag{2.1}$$

where Y_{iht} is a dummy variable equal to 1 if the patient i in hospital h in month t fulfils all[17] BPT criteria and 0 otherwise. D_t is a dummy variable equal to 1 in the post-policy period (from April 2010 to March 2015), and equal to 0 in the pre-policy period (from April 2008 to March 2010). $England_i$ is a dummy variable equal to 1 if patient i is treated in England (the treatment group) and equal to 0 if treated in Wales (the control group). $\boldsymbol{v_t}$ is a

[16] The approach of comparing similar geographic regions to test for policy effects has been widely used in literature, for example, comparing England and Scotland (Propper et al., 2008; Farrar et al., 2009) or different US states (Kolstad and Kowalski, 2016).

[17] Seven criteria in financial year 2010/11, and eight criteria from 2011/12 onwards. See Section 2.4.

vector of indicators for each financial year in the study (2009/10 to 2014/15 with reference category 2008/9). $\boldsymbol{v_s}$ is a vector of calendar months (January to December, with reference category April) to adjust for seasonality. $\boldsymbol{v_h}$ is a vector of hospital fixed effects. $\boldsymbol{X_i}$ is a vector of patient characteristics, α is the intercept and ϵ_{iht} is the error term. We estimate (2.1) as a linear probability model with standard errors clustered at the hospital level. The key coefficient of interest is θ, which measures the average treatment effect on the treated (ATT) patient population over the post-policy period.

The size of the bonus payment increased in year 2 (2011/12), and then again in year 3 (2012/13). We implement a version of our base model that captures differential responses over time, i.e.

$$Y_{iht} = \alpha + \sum_{k=2010/11}^{2012/13} \theta_k(England_i \ Year_k) + \boldsymbol{X_i}'\delta + \boldsymbol{v_t} + \boldsymbol{v_s} + \boldsymbol{v_h} + \epsilon_{iht} \qquad (2.2)$$

where $Year_k$ are binary indicators for the three post-policy periods with differing incentive payments and the θ_k coefficients measure the corresponding ATTs in year k.

The validity of the DiD approach relies on the parallel trends assumption, i.e. the outcomes of interest would develop similarly in both groups in the absence of the policy intervention. We explore this assumption in two ways. First, we carry out a a visual inspection of the pre-policy trends (see Appendix Figure A1). Second, we test the assumption empirically by using pre-policy data (2008/9-2009/10) to estimate the following model that allows for country-specific trends:

$$Y_{iht} = \alpha + \sum_{k=1}^{7} \beta_k Quarter_k + \sum_{k=1}^{7} \tau_k(England_i \ Quarter_k) + \boldsymbol{X_i}'\delta + \boldsymbol{v_s} + \boldsymbol{v_h} + \epsilon_{iht} \quad (2.3)$$

where $Quarter_k$ are binary variables taking value of 1 if the patient was treated in quarter k ($k = 1, ..., 7$), and 0 otherwise. We use quarterly dummies rather than year dummies (as used in the main regression) to add granularity better control for short term changes in the pre-policy period. The reference group is the first quarter. The coefficients τ_k capture

the difference in the pre-policy trends between the two countries. The null hypothesis for the parallel trends assumption is $H_0 : \tau_k = 0$. Failure to reject this assumption provides reassurance that the parallel trends assumption holds.

We perform three sensitivity analyses of our base model. First, we control for the increasing number of hospitals reporting to the NHFD by estimating our main models (eqs. 2.1-2.2) on a balanced panel of hospitals that have participated in the audit in all financial years since 2008/9 and have treated at least 30 patients per year. Second, we estimate both models with a limited set of control variables (age, sex, and fracture type) which are recorded for all patients in the NHFD and drop other control variables where information is missing for some patients. This increases the estimation sample to 334,835 patients at the possible expense of a less comprehensive risk-adjustment. Finally, we use the seven initial BPT criteria that have been in place since April 2010 (i.e. excluding BPT4 and BPT5) to define achievement in all years of the sample.

2.5.2 Heterogeneity

We extend our analysis in two directions to explore heterogenous responses across different process measures of quality. First, the level of pre-policy achievement differs considerably across process measures which implies differential scope for improvement. English hospitals could then have an incentive to focus on improving those process measures where pre-policy achievements were lowest in order to deliver the full care bundle. Furthermore, process measures are likely to differ in marginal cost and patient benefit, which may also affect how providers prioritise these process measures. To explore heterogeneous effect of the policy across different process measures, we re-estimate the models in eqs. 2.1 - 2.3 for each of the BPT criteria separately. Building on this, we also test whether the policy had an impact on the number of criteria met, with the understanding that providers in Wales might still aim to improve on the care they provide, but with less emphasis on achieving all the criteria. We perform this analysis using the same regression framework as specified above.

Third, the payment is conditional on providing all of the measures. Therefore, the financial incentive to provide any additional processes drops to zero if at least one other process measure has been missed. We exploit the sequential nature of care processes within the hip fracture care pathway to study how missing one or more BPT criteria, and therefore forgoing the bonus payment, affects provider behaviour with respect to any remaining criteria that are yet to be met. Assuming that achieving BPT criteria is costly to the provider one would expect providers to exert less effort to meet criteria once at least one criterion has been missed. To investigate this empirically, we categorise BPT criteria into two groups: those where processes are expected to take place before or during surgery (BPT1, BPT3, BPT4, BPT6) and those were processes take place after surgery (BPT7, BPT8, BPT9).[18] We then estimate the effect of failing to meet at least one earlier criterion on the propensity of meeting all post-surgical criteria using the following model:

$$Post_{ith} = \alpha + \gamma Pre_{ith} + \theta(England_i \ Pre_{ith}) + \boldsymbol{X_i}'\delta + \boldsymbol{v_t} + \boldsymbol{v_s} + \boldsymbol{v_h} + \epsilon_{iht} \qquad (2.4)$$

where $Post_{ith}$ is a binary indicator taking the value of 1 if all BPT criteria related to post-surgical care processes were attained for patient i in hospital h at time t, and 0 otherwise. Pre_{ith} is a binary indicator equal to 1 if one or more of the pre/mid-surgery BPT criteria are missed and 0 otherwise. The coefficient θ captures the difference in response between England and Wales if a pre-surgery criteria is missed. A positive estimate of θ suggests that English hospitals are more likely than Welsh hospitals to achieve post-surgery criteria once an earlier criterion is missed, which, in turn, implies that the existence of the BPT incentive policy has positive spillovers on process measures even when these are no longer incentivised. We run eq. 2.4 as a linear model with standard errors clustered at the hospital level. The pre-surgery criteria rely heavily on the involvement of geriatricians and have attainment level close to zero in both countries prior to the introduction of the

[18]The allocation of BPT criteria to these two groups was informed clinical expert opinion.

BPT (see Appendix Table A2). We therefore limit this analysis to the period after the BPT introduction (i.e. April 2010 onwards). Furthermore, note that BPT2 (shared care across specialities) is excluded from analysis because achievement level are consistently very low in Wales, which markedly reduces the proportion of patients for whom all pre-surgery criteria are fulfilled ($Pre_{ith} = 1$).[19] This analysis relies on the assumption that providers do not decide to meet the pre-surgery criteria based on their expectations about the likelihood of meeting the post-surgery criteria. Considering the nature of the criteria and the uncertainty about meeting the subsequent criteria (as they are days/weeks away), this seems to be a reasonable assumption.

2.5.3 Effect of the policy on health outcomes

The process measures were selected for the incentive scheme based on their potential to improve health outcomes (Cameron, 2005; British Orthopaedic Association, 2007). In the last model, we estimate whether differential changes over time in BPT attainment rates across different hospitals affected hospital mortality rates in England, using the following two-way fixed effects panel approach:

$$\hat{Z}_{hq} = \alpha + \theta Y_{hq} + \boldsymbol{v_t} + \boldsymbol{v_q} + \boldsymbol{v_h} + \epsilon_{hq} \tag{2.5}$$

where $\hat{Z}_{hq}$ is the risk-adjusted 30/90/365-day mortality rate[20] of hospital h in quarter $q = 1, \ldots, 28$ (with 1 for the first quarter in the financial year 2008/9 and 28 for the last quarter in year 2014/15), Y_{hq} is the proportion of patients receiving BPT incentivised care, $\boldsymbol{v_t}$ is a vector of indicator variables for financial years (from 2009/10 to 2014/15 with reference category 2008/9), and $\boldsymbol{v_q}$ is a vector of indicators for calendar quarters (April-June, July-Sept, Oct-Dec, with reference category Jan-Mar) to capture seasonal effects[21]. The key

[19] The achievement of all pre-surgery BPT criteria is 2.9% when including BPT2 and 15.3% when excluding this criterion. See Table A2 for further details.

[20] Further details on these computations are presented in the Appendix.

[21] While the analysis on the effect of the policy on performance uses monthly dummies, the mortality analysis uses quarterly data. This is due to the fact that the latter is performed on hospital level data, with a

coefficient of interest θ captures the marginal contribution of the BPT achievement rate on mortality. If $\theta < 0$ then this would suggest that BPT achievement is associated with an improvement in patient health and reduces mortality risk).

As a further sensitivity analysis we estimate eq. 2.5 replacing the proportion of patients who meet all of the BPT criteria (i.e. the condition that triggers BPT payment) with the average number of criteria achieved. In this case, the coefficient θ captures the marginal effect of one additional BPT criterion being achieved on mortality risk. Note, that we do not run the sensitivity analysis including all eight measures separately. This is due to high correlation across the measures causing unstable regression estimates.

2.6 Results

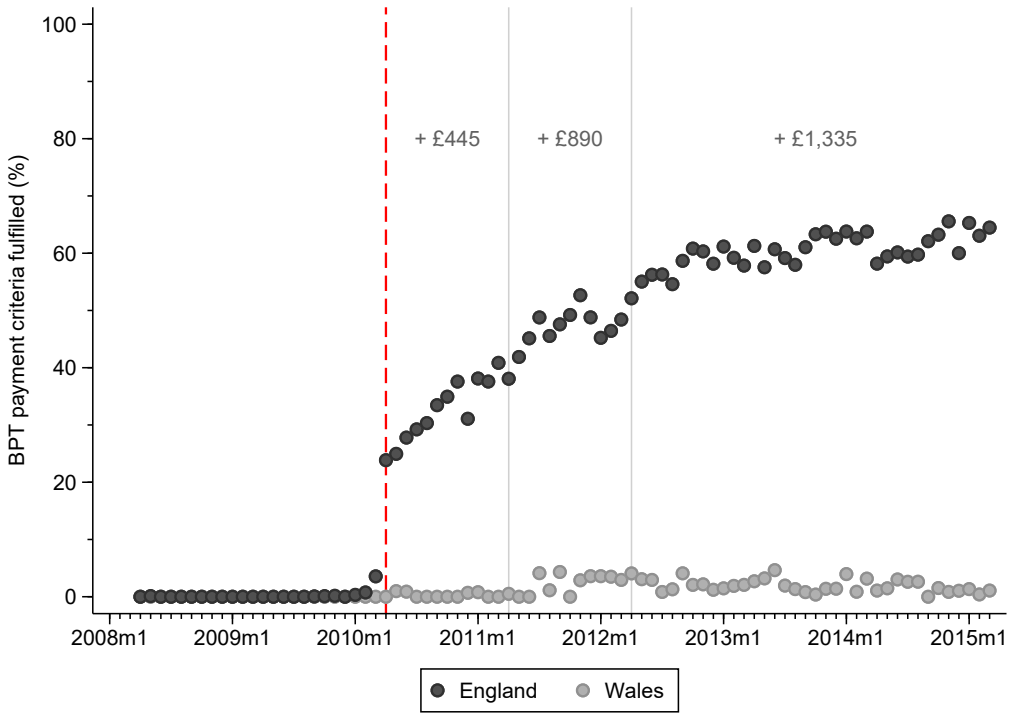
2.6.1 Policy effect

Figure 2.4 shows the proportion of patients receiving the full care bundle in England and Wales over time, where achievement is defined as having met 7 criteria in the period up to March 2011 and as having met 8 criteria in the subsequent period. Vertical lines indicate changes in the size of the bonus paid to providers in England that deliver the full care bundle. Achievement is very low in both countries during the pre-policy period. After the introduction of the BPT payment policy in April 2010, care processes improve rapidly in England but remain low in Wales.

Table 2.3 presents results of the main DID analyses both in terms of the average effect of the policy (eq. 2.1) and for each of the three years in the post-policy period (eq. 2.2). Our key results indicate a large and statistically significant increase in the probability of receiving the incentivised care bundle of 51.7 percentage points (pp) ($p<0.001$) over the post-policy period; 30.3 pp in the first period (April 2010 to March 2011), 44.7 pp in the

typical hospital treating less than 100 patient per month. Quarterly data was used to reduce the error associates with the small sample size.

Figure 2.4: Time-series of BPT achievement in England and Wales



Notes: The figure shows the BPT achievement in England and Wales from April 2008 to March 2015. The red line marks the start of the BPT incentive (April 2010), while the gray lines indicate the changes to the bonus size. The numbers above the plot indicate the size of the bonus in the respective period.

second period (April 2011 to March 12) and 57.4 pp in the third period (April 2012 to March 15) (all $p<0.1$).

The effect of patient characteristics on the likelihood of receiving the full care bundle are also detailed in Table 2.3. Male patients are 1.9pp less likely to receive all incentivised BPT criteria, whereas older patients are increasingly more likely to do so. The level of predicted operative risk, as measured by the ASA grade, is inversely related to the likelihood of BPT pay-out: compared to healthy patient (ASA Grade 1), the probability of receiving the full care bundle decreases by 1.4 pp for patients with mild systemic disease (ASA grade 2), by 5.2 pp and 14.7 pp for patients with increasingly severe systemic disease (ASA Grade 3 and 4, respectively) and by 27.9 pp for moribund patients (ASA grade 5). The effect of pre-fracture mobility is less pronounced (between 0.6 and -2.0pp) and does not follow a clear

Table 2.3: Regression estimate of the effect of the BPT payment policy on the probability of delivering the incentivised care bundle

	(1) Single post-policy period		(2) Multiple post-policy periods	
	Estimate	SE	Estimate	SE
Policy effect				
Average	0.517***	0.027		
April 2010 - March 2011			0.303***	0.036
April 2011 - March 2012			0.447***	0.041
April 2012 - March 2015			0.574***	0.029
Age groups				
60-64 (reference)				
65-69	0.008	0.006	0.008	0.006
70-74	0.022***	0.006	0.022***	0.006
75-79	0.033***	0.005	0.033***	0.005
80-84	0.040***	0.006	0.040***	0.006
85-89	0.040***	0.006	0.040***	0.006
90+	0.052***	0.006	0.052***	0.006
Gender				
Female (reference)				
Male	-0.019***	0.002	-0.019***	0.002
Admission source				
Hospital (reference)				
Care home	0.138***	0.009	0.137***	0.009
Residential home	0.124***	0.009	0.123***	0.009
Other	0.127***	0.013	0.126***	0.013
ASA Grade				
ASA Grade 1 (reference)				
ASA Grade 2	-0.014**	0.006	-0.014**	0.005
ASA Grade 3	-0.052***	0.006	-0.052***	0.006
ASA Grade 4	-0.147***	0.008	-0.148***	0.008
ASA Grade 5	-0.279***	0.024	-0.279***	0.024
Mobility				
Full mobility (reference)				
Some outdoor mobility	-0.004	0.003	-0.004	0.003
Some indoor mobility	0.006**	0.003	0.006**	0.003
No functional mobility	-0.020***	0.007	-0.020***	0.007
Fracture type				
Extracapsular (reference)				
Intracapsular	-0.018***	0.002	-0.018***	0.002
Hospital fixed effects	X		X	
Time (month) fixed effects	X		X	
Patient characteristics	X		X	
Number of hospitals	185		185	
Number of patients	275,898		275,898	

Notes: Model 1 estimates the average effect of the BPT policy over the first five years after the policy introduction (based on eq. 2.1). Model 2 estimates separate effects for the three post-policy periods that coincide with changes in the size of the bonus payment (based on eq. 2.2). Note that the third period covers three financial years, whereas the other periods cover one financial year each.

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

pattern. Patients who are admitted from residential home or nursing home are between 12-14 pp more likely to meet all of BPT criteria than those who are already in the hospital at the time of fracture.

We do not find evidence of a violation of the parallel trends assumption that underpin these DID analyses. All of the estimates for the difference in trends in the pre-policy period, presented in Table 2.4 are small (between -0.001 and 0.016 pp/quarter). With the exception of the last quarter, all of the estimates are also statistically insignificant. This finding is confirmed by visual inspection of the data (Figure 2.4). The coefficient in the last calendar quarter (Jan10-Mar10) prior the start of the policy is statistically significant, indicating a potential anticipation effect. However, the effect is still small in comparison to the overall effect of the policy, and unlikely to bias results.

Table 2.4: Differences in achievement of care bundle between England vs Wales by quarter of pre-policy period

Quarter	Estimate	SE
Apr08 - June08	*Reference*	
July08 - Sept08	-0.001	0.001
Oct08 - Dec08	0.000	0.001
Jan09 - Mar09	0.000	0.001
Apr09 - June09	-0.001	0.001
July09 - Sept09	0.000	0.001
Oct09 - Dec09	0.001	0.001
Jan10 - Mar10	0.016***	0.004

Notes: We test the parallel trends assumption using data from the pre-policy period (April 2008 to March 2010) and quarterly dummies. Standard errors (SEs) are clustered at hospital level.

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

Our sensitivity analyses, presented in Table 2.5, confirm the robustness of our main results. The estimate of the policy effect increases to 55.2 pp when we base our analysis on a balanced panel of hospitals that have participated in the clinical audit in all years since the start of our dataset. Although this suggests that early contributors to the NHFD performed slightly better than average, any selection bias is likely to be of little practical importance. Furthermore, the overall achievement rate of the early contributors was close

to zero in the pre-policy period, which is likely to be the case for hospitals joining the NHFD at a later point as well. We also find that adjusting for a limited set of non-missing patient characteristics has limited effect on the estimated policy effect (50.9pp), which suggests that hospitals may face a relatively similar case-mix of patients.[22] Exclusion of one criterion (BPT4: 'pre-operative mental health assessment') from the overall measure does not change the estimated effect of the policy (51.7 pp), mainly as the achievement rates for this criterion were already high in both countries prior to its inclusion in the BPT incentive.

Table 2.5: Results of sensitivity analyses

	(1) Balanced panel		(2) Limited case-mix adjustment		(3) Subset of 7 BPT criteria	
Policy effect	Estimate	SE	Estimate	SE	Estimate	SE
Average	0.552***	0.018	0.509***	0.019	0.517***	0.027
April 2010 - March 2011	0.389***	0.029	0.303***	0.029	0.303***	0.037
April 2011 - March 2012	0.501***	0.024	0.443***	0.033	0.447***	0.041
April 2012 - March 2015	0.552***	0.018	0.579***	0.021	0.575***	0.029
Hospital fixed effects	X		X		X	
Time (month) fixed effects	X		X		X	
Patient characteristics	X		Limited		X	
Number of hospitals	78		185		185	
Number of observations	146,845		334,835		275,898	

Notes: Model 1 uses a balanced panel of hospitals who reported to the NHFD in all years during our study period. Model 2 adjusts for a limited set of case-mix variables (age, sex, fracture type) for which information was available for the full sample. Model 3 excludes BPT4 ('Pre-operative cognitive function assessment') from the definition of the care bundle. Model 4 excludes observations in the six months prior to policy start, when providers may have been aware of impeding changes to payment modalities ('anticipation period'). Standard errors are clustered on hospital level.

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

2.6.2 Heterogeneity

The analyses at the level of care bundles hide some important heterogeneity across the criteria that constitute the care bundle (Table 2.6). The BPT has the largest impact on process measures related to geriatrician involvement (BPT2, BPT3, BPT6, BPT7), ranging from

[22]Patients sustaining a fragility hip fracture are unlikely to bypass their local hospital (Gutacker et al., 2016), which implies that concerns over endogenous selection are unlikely to apply.

64.0 pp for shared care across specialities to 19.8 pp for geriatrician-led multidisciplinary rehabilitation. These results resonate with Neuburger et al. (2017), who reported a large increase in the number of full-time equivalent geriatricians working in NHS hospitals in England following the start of the BPT payment policy. The policy also increased the proportion of patients with bone health assessment carried out by 24.0 pp, while the policy impact is generally less pronounced for the remaining BPT criteria. This is due to either i) similar improvements being observed in both countries, or ii) initial rates being already high prior to the policy (See Table A3 in the Appendix).

The BPT policy increased the probability of receiving surgery within 36h (BPT1) by 13 pp. The effects on secondary prevention (BPT8) and pre-operative cognitive assessment (BPT 4) are both small (6 pp and -2.4pp, respectively) and not statistically significant. Taken together, the policy increased the number of criteria met by 2.0.

Table 2.6: Estimated policy effects by BPT criterion

	(1)		(2)					
	Full post-policy period		April 2010 to March 2011		April 2011 to March 2012		April 2012 to March 2015	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Criterion met								
BPT1: Surgery within 36 hours	0.134***	0.032	0.109***	0.036	0.119***	0.025	0.143***	0.037
BPT2: Shared care across specialties	0.640***	0.143	0.517***	0.130	0.616***	0.149	0.670***	0.149
BPT3: Multidisciplinary care protocol	0.276**	0.131	0.254*	0.136	0.234*	0.139	0.291**	0.138
BPT4: Pre-op cognitive function test	-0.024	0.093	n/a		n/a		-0.024	0.093
BPT6: Peri-op geriatric asses.	0.411***	0.059	0.334***	0.078	0.373***	0.081	0.435***	0.056
BPT7: Multidisciplinary rehab	0.198***	0.069	0.274***	0.105	0.170*	0.068	0.192***	0.073
BPT8: Falls prevention	0.060	0.133	0.119	0.090	0.011	0.138	0.062	0.147
BPT9: Bone health assessment	0.240***	0.077	0.160**	0.068	0.216***	0.065	0.261***	0.081
Total number of criteria met								
Count of criteria	2.032***	0.264	1.754***	0.323	1.742***	0.267	2.158***	0.295

Notes: Model 1 estimates the average effect of the BPT policy over the first five years after the policy introduction (based on eq. 2.1). Model 2 estimates separate effects for the three post-policy periods that coincide with changes in the size of the bonus payment (based on eq. 2.2). Note that the third period covers three financial years, whereas the other periods cover one financial year each. Standard errors (SEs) are clustered at the hospital level.

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

As can bee seen in Table 2.6, for most criteria the magnitude of the effect is increasing over time, with the largest effects observed during the period from 2012/3 to 2014/15 for 5 out of 8 criteria. However, as in the case of the overall measure, the rate of change is slowing over time, with the largest absolute increase observed in the first year of the policy for all the criteria. This suggest that hospitals responded strongly at the start of the policy, despite the lower BPT bonus in the first year relative to the following years.

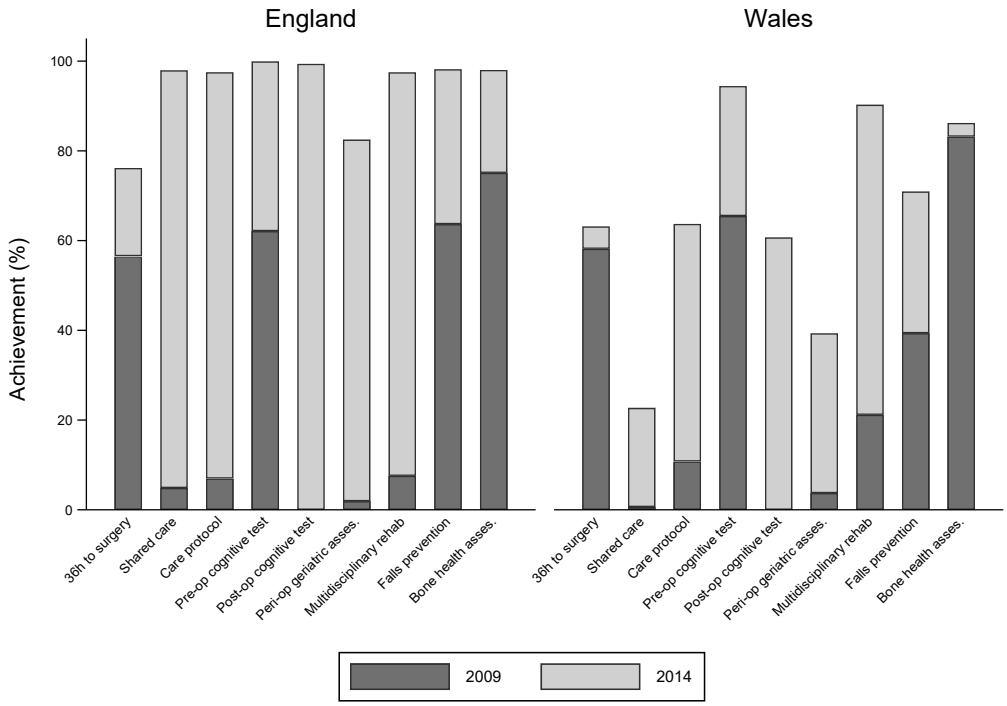
Figure 2.5 shows that in the England all BPT measures reached similar levels by 2014/15, regardless of the initial values before the start of the policy in 2009/10[23]. Conversely, despite similar starting levels as in England, Wales shows much larger dispersion between the BPT achievement levels in 2014/15. The wide-ranging improvement in England is likely due to the bundled element of the BPT, as it requires the hospitals to improve on all segments to receive the additional payment, as also highlighted by our theoretical model in Section 2.3.

Appendix Table A4 reports the results of the parallel trends test for individual criteria (see also Appendix Figure A1 for a visual representation). For most individual criteria (BPT1-BPT8) there is no evidence of systematic deviations in trends between England and Wales prior to the start of the BPT payment policy. However, the test of parallel trends rejects the null hypothesis for bone health assessment (BPT9). Four of the corresponding coefficients are positive and statistically significant, which implies that English hospitals were, on average, adopting these care processes more rapidly than than Welsh hospitals prior to the start of the BPT policy. As a result, the corresponding policy effect estimates for these criteria are likely to be over-estimated. Table A4 also reports the estimates of the pre-trends analysis for the number of criteria achieved. In this case two out of seven criteria are significant, however, the magnitude of the estimates is small.

We next look at how providers respond to the loss of financial incentive once at least

[23]While for majority of criteria providers reached very high achievement rates, BPT1 (36h to surgery), seems to be lagging behind. However, this is most likely due to clinical guidelines that state that it's beneficial to postpone the surgery for some groups of patients (predominantly younger ones), as for them further stabilisation results in better overall clinical outcome.

Figure 2.5: Achievement of BPT criteria in England and Wales, 2009/10 and 2014/15



Notes: The figure shows the achievement of the BPT criteria in England in Wales in the year before the implementation of the BPT (2009/10) and in the last year of our study period (2014/15).

one previous BPT criterion has been missed for a given patient. As previously shown, English hospitals increased their achievement across nearly all BPT criteria following the start of the BPT payment policy. However, our results in Table 2.7 suggest that once at least one of the pre-surgical BPT criteria has been missed, English hospitals are no more likely than Welsh hospitals to achieve subsequent criteria. This suggests that English hospitals are indeed responding to the financial incentive created by the BPT rather than unobserved time-varying factors.

The difference in the achievement rates across England and Wales further suggests that some measures are easier to achieve/are less resource intensive. In particular, measures related to geriatrician involvement, which are likely associated with higher resource use, only improved in England and not in Wales. On the other hand, nurse-led measures like mul-

Table 2.7: The effect of failing one or more pre-operative criteria on the probability of meeting the set of post-surgical criteria

	(1) Observed		(2) Without FE		(3) With FE	
Period	Estimate	SE	Estimate	SE	Estimate	SE
Pre-criteria failed	-0.259**	0.093	-0.245**	0.097	-0.131***	0.021
England	0.149*	0.075	0.152*	0.074	n/a	
Pre failed × England	0.095	0.094	0.115	0.095	0.043	0.022
Hospital fixed effects					X	
Time (month) fixed effects			X		X	
Patient characteristics			X		*X*	
Number of hospitals	182		182		182	
Number of observations	237,614		237,614		237,614	

Notes: Estimated during post-policy period April 2010 - March 2015. Analysis excludes BPT2 from pre-surgery measures due to low achievement during first year. Standard errors are clustered at hospital level.

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

tidisciplinary rehabilitation (BPT7), falls prevention (BPT8) and bone health assessment (BPT8) are mostly met by hospital providers in both countries.

2.6.3 Effect of process measures on patient health outcomes

Table 2.8 presents the results of a two-way fixed effects model that estimates how changes in hospitals' BPT achievement rates over time affect hospital mortality rates. We find that a 100 pp increase in the achievement of the care bundle (i.e. going from 0% to 100%) is associated with, on average, a 1.2 pp reductions in mortality at 90 days but not at one year after admission. Given that the BPT payment policy improved the delivery of the care bundle by 51.7 pp, this implies an overall effect of the policy on 90-day mortality of -0.62 pp ($= 0.517 \times 0.012$). However, this analysis does not distinguish between patients for whom none of the criteria was met and those for whom all but one criteria was met. The latter group would have received most, but not all, beneficial care processes incentives under the BPT, which would lead to downward biased estimates of the effect of care bundles on mortality. We therefore also estimate how achieving one additional BPT criterion affects the probability of mortality within a given time frame. Our results show that each extra

BPT criterion met is associated with a reduction in 30-day mortality of 0.1pp, and this effect increases to 0.2pp by the end of the first year after admission. The estimated increase in the in number of criteria met is 2.0 (see Table 2.6), which suggests the policy effect on 1-year mortality equals to -0.4pp ($= 2.0 \times -0.2$).

While the effect of BPT on mortality is relatively small, it's important to note the improvement in survival was not one of the explicit goals of the BPT. Rather, the BPT aimed to improve the patient experience and rehabilitation process. Unfortunately, we do not have measures available to capture patient well-being, mobility and rehabilitation, which would provide a more comprehensive measure of the effect of BPT on patient outcomes.

Table 2.8: Association between mortality and BPT achievement

	Care bundle is met (yes/no)		Number of BPT criteria met	
Death within	Estimate	SE	Estimate	SE
30 days	-0.004	0.004	-0.001	0.001
90 days	-0.012*	0.006	-0.002	0.001
365 days	-0.005	0.007	-0.002	0.002
Number of hospitals	142		142	
Number of hospital-quarters	2,527		2,527	

Notes: The estimation is performed on quarterly hospital data and is restricted to English hospitals that reported data for at least 20 patients to NHFD in each quarters of our dataset. The presented estimate of the effects of achieving all BPT measures on mortality is based on 100% BPT achievement (compared to 0% BPT achievement). Standard errors (SE) are clustered at hospital level.

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

2.7 Discussion

We have investigated whether the BPT financial incentive improved the care pathway for hip fracture patients. Our estimates show that the policy increased the overall adherence to the criteria by 51.7pp. While this suggest that the scheme was very successful in changing provider's behaviour, the overall result hides considerable heterogeneity in the response across different criteria. The policy effect was largest for the four criteria associated with geriatrician involvement in the care, with smaller improvement observed in other areas.

Our results are robust to different specification in respect to patient characteristics, balanced sample of hospitals and subset of BPT criteria.

The sizeable difference between the overall policy effect and the effect of individual criteria suggests the English providers responded strategically to the scheme, by improving across all dimensions to the level required for the bonus payment. While the absolute achievement rates varied across criteria in both England and Wales prior to the introduction of the BPT, by 2014/15 English providers achieved comparable achievement rates across all of the criteria, regardless of the initial performance. In contrast, Welsh providers improved individual care processes in a less systematic way. The difference in response across countries indicates that the bundled element of the scheme focused the attention of English providers on all of the care processes, rather than individual tasks.

Our results for the effect of the BPT on mortality are lower that the ones reported in Metcalfe et al. (2019), who studied the effect of BPT on mortality comparing achievement rated between England and Scotland. The discrepancy might be due to the difference in the research objective: while we study the effect of meeting the BPT criteria on mortality, they compare mortality scores without linking them to the BPT achievement rates. The improvement in mortality might stem from wider changes to behaviour after the implementation of the policy and it is not directly linked to individual criteria.

Our results suggest the BPT scheme for hip fracture yields better results compared to several other P4P programs Milstein and Schreyoegg (2016). While it is difficult to pinpoint the reasons for this relative success of the scheme, it may be attributed to specific design features. The bonus size in P4P schemes within the hospital setting is often relatively small (less than 2-3% of total reimbursement) (Milstein and Schreyoegg, 2016), whereas this scheme operates with a substantially larger bonus, reaching 21% of the total reimbursement of care. Furthermore, the selection of the incentivised criteria relied heavily on clinical input, mirroring the clinical guidance for the optimal treatment of hip fracture patients. Hence the scheme may have given stronger motivation and resources to the physicians to

treat the patients according to their own best practice standards.

In summary, we believe there is sufficient evidence to conclude that the BPT policy for hip fracture is overall successful in improving care pathway for hip fracture patient and reducing mortality.

2.8 Conclusions

Our study shows that P4P can be effective in increasing the provision and, more broadly, the quality of health care. Previous studies have shown that P4P has had limited impact within the context of secondary care, and has only occasionally been more successful within the context of ambulatory or primary care. We have shown that P4P can be successful also in the context of secondary care and highlight different distinct economic features of the scheme. First, the scheme incentivised process measures of care as opposed to health outcomes and these process measures were chosen to reflect best practice standards based of clinical evidence and professional consensus. Second, there was just one single payment which was conditioned on a bundle of processes. The simplicity of the scheme combined with the strong financial incentive to provide the bundle could be a contributing factor to greater provider attention and focus to increase efforts. Third, the size of the bonus was significant, potentially suggesting that the small effects from previous schemes may be driven by the small bonus. Overall, the study supports policies that gradually move away from activity-based financing in favour of payment models that reward quality directly.

CHAPTER 3

Can financial incentives shift health care from an inpatient to an outpatient setting?

3.1 Introduction

Driven by the rapid growth of healthcare spending, policymakers across OECD countries are under renewed pressure to develop policies that contain costs while preserving quality of care (OECD, 2010; Cashin et al., 2014a). One policy lever to induce healthcare providers to reduce costs and increase efficiency is the use of financial incentives (Ellis and McGuire, 1993; Hollingsworth, 2008; Hussey et al., 2009).

For hospital care, which accounts for about half of health spending, one area that has been targeted to reduce costs is the substitution of the more expensive inpatient care with the less expensive outpatient (ambulatory) care[24] (Davis and Russell, 1972; Elnicki, 1976; Vitikainen et al., 2010). In the 70s, ambulatory care was initially limited to few selected treatments (Vitikainen et al., 2010), but has since gathered momentum thanks to advances in medical technology. Many more conditions are treated in an ambulatory (office-based) setting, including diagnostic procedures, chemotherapy and dialysis (Dobson et al., 2013; Esparza et al., 1989; Wellenstein et al., 2017). This has resulted in lower costs, quicker discharge and faster patients' recovery (Ancona-Berk and Chalmers, 1986; Castells et al.,

[24]The definition of outpatient care varies across countries. We refer to outpatient care as the ambulatory/office based care. Day procedures that are theatre based are considered to be performed in the inpatient setting.

2001; Robinson and Beyer, 2010; Hayes et al., 2015; Gordan et al., 2019).

Despite the availability of the relevant technology and the recognised benefits, the move towards the outpatient setting for suitable procedures has been slow. In the UK, less than a third of these procedures is performed on an outpatient basis (Department of Health, 2012). This can be explained by the existing financial incentives, with higher revenues and profit margins for the inpatient relative to the outpatient setting (Higgins et al., 2016; Bach and Jain, 2017; Fisher et al., 2017), or provider reluctance to change, given that changing established practices can be costly, making providers reluctant to invest time and resources in the reorganisation of their services (Gaughan et al., 2019).

Strategic pricing or pay for performance in the form of a financial incentive, which explicitly incentivises outpatient care, can potentially encourage hospitals to change their modus operandi and shift the provision of care towards the incentivised setting (Ellis and McGuire, 1986b; Ma, 1994a; Hodgkin and McGuire, 1994). There have been several policy initiatives aimed at changing providers' behaviour through financial incentives (OECD, 2010), and their limited effects have been documented (Milstein and Schreyoegg, 2016). However, the changes in financial incentives have been triggered by a wider overhaul of hospital pricing structure and were generally small in size, limiting the ability to translate these findings in the context of targeted interventions explicitly aimed at boosting outpatient care (Bach and Jain, 2017; Milstein and Schreyoegg, 2016; Eijkenaar, 2012; OECD, 2010; He and Mellor, 2013). This study contributes to filling this gap in knowledge (van Hoof et al., 2019).

This paper examines the effect of a financial scheme that incentivises the shift in provision from a high-cost to a low-cost setting. We study the effect of the introduction of a Best Practice Tariff (BPT) for outpatient care that rewards providers for treating patients in an office-based ambulatory setting rather than in an inpatient setting. The scheme was introduced across all hospitals in England in April 2012 and affected three procedures. Two are high volume diagnostic procedures (diagnostic cystoscopy, diagnostic hyteroscopy) and

the third is a form of sterilisation for women (hysteroscopic sterilisation). The scheme operates by simultaneously increasing the price paid for the office-based outpatient procedure by a significant amount (up to 470% increase[25]), and, in the case of the two diagnostic procedures, by also lowering the price paid in the inpatient setting.

We employ difference-in-difference methods to assess whether the introduction of incentive scheme was successful in increasing the probability of being treated in the outpatient setting (the intensive margin). We also test if the incentive scheme affected volume (the extensive margin), quality of care, as measured by repeated procedures, and whether it had positive or negative spillover effects for closely-related unincentivised procedures, which could be driven by cost synergies or effort diversion, respectively. Our control group includes procedures that were chosen based on clinical relevance, suitability of treatment in both the inpatient and the outpatient setting, and parallel trends in the probability of being treated in the outpatient setting in the pre-policy period.

Our results show that a targeted financial stimulus can result in a swift and substantial shift in the choice of the treatment setting. We find a positive and significant effect of the policy on the probability of having the procedure performed in the outpatient setting for all three incentivised procedures, with the largest effect being observed for cystoscopy and hysteroscopy (36.1 percentage points (pp) and 16.3 pp, respectively). The effect is smaller for sterilisation (3.8 pp), possibly due to the inpatient price remaining the same for this procedure (the inpatient price was instead lowered for cystoscopy and hysteroscopy).

Our results suggest that there was instead no effect of the scheme on quality, as measured by the probability of having the same procedure repeated within 60 or 90 days. We also find no evidence that the policy affected the overall volume (the extensive margin): the effect on patient volume was quantitatively small and statistically insignificant. We instead find evidence of positive spillover effects, with the BPT increasing the probability of being

[25]The increase in the price is calculated by comparing the outpatient price in the pre-policy year to the outpatient price in the first year post-BPT introduction. The highest increase is observed for sterilisation, where the pre-policy price for the outpatient procedure is £242, increasing to £1,137 after the introduction of the BPT. See Table 3.1 for details.

treated in the outpatient setting for the non-incentivised procedures by 4.8 to 12.3 pp.

We contribute to the limited literature on the effect of financial incentives specifically targeting behavioural changes across hospital settings. This is in contrast to previous studies (see review in section 3.2) that look at the hospital response across settings following a major pricing overhaul, typically driven by the introduction of the DRG reimbursement system. Moreover, the size of incentive scheme arising from such policy changes have generally been small, and associated with small or no changes in behaviour, making difficult to assess the scope of more direct and sizeable financial incentives. Unlike most studies of financial incentives that focus exclusively on the incentivised outcome measure (Milstein and Schreyoegg, 2016), we also explore a number of secondary outcomes, related to the quality of care, volume, and spillover effects. Several reviews of pay for performance schemes note the lack of research on the effects of financial rewards on non-incentivised dimensions of care, despite its importance when implementing a P4P program. We are therefore able to uncover some of the mechanisms by which the scheme works and provide additional evidence on the overall effectiveness of the scheme.

The study is organised as follows. Section 3.2 presents the related literature. Section 3.3 provides the institutional background. Section 3.4 presents the empirical approach. Section 3.5 describes the data. Section 3.6 presents the results. Section 3.7 discusses implications for providers' revenues and costs, and spending for the funder. Section 3.8 concludes.

3.2 Related literature

Our paper relates to four strands of the literature on the effect of financial incentives in the hospital setting. The first focuses on the effect of changes in the DRG price on volume and quality of care (Gruber and Owings, 1996; Yip, 1998; Dafny, 2005; Gruber et al., 1999; Grant, 2009; Clemens and Gottlieb, 2013; Papanicolas and McGuire, 2015; Januleviciute et al., 2016; Verzulli et al., 2016; Batty and Ippolito, 2017; Shin, 2019). Exploiting price

shocks in reimbursement, these studies generally find that hospitals supply more treatment when the reimbursed price increases or the price of alternative (competing) treatments decreases. A second strand of the literature studies the effect of the relative price change in the presence of multiple treatments for a given health condition (Foo et al., 2017; Papanicolas and McGuire, 2015). These studies find that hospitals increase the proportion of the more profitable treatment, even when not in accordance with clinical guidance (Papanicolas and McGuire, 2015). These findings are is in line with ours. However, in most of the studies the price shock is relatively small (around 3-5%), corresponding to a relatively small price increase. In our study, the price increase is more dramatic ($>$50% for all procedures).

The third strand of the literature relates specifically to the effect of financial incentives on the substitution between healthcare settings, with an emphasis on (ambulatory) outpatient and inpatient care, which is close to our focus. Leader and Moon (1989) found that the move towards prospective payment in the US led to a large decline in inpatient visits, coinciding with the surge in outpatient appointments[26]. Similar findings were reported by Hadley and Swartz (1989) and Hadley et al. (1989). More recently, He and Mellor (2013) examined whether a change in Medicare outpatient payment rates under the Outpatient Prospective Payment System (OPPS) caused outpatient/day case care to shift towards the inpatient setting[27]. The study found a reduction in the number of hernia repair procedures performed in an outpatient setting, while the number of inpatients procedures remained the same. In these studies the shift across settings was the result of a by-product of a broader change in the payment system, such as the introduction of a DRG prospective payment. Our study differs because the price was targeted at selected procedures with the only aim of strongly incentivising care in the outpatient setting.

Fourth, our paper contributes to the broader literature on the effects of Pay for Perfor-

[26]Note that the classification of outpatient activities differs between the USA and UK, with the US definition being broader. Day case procedures that are performed in a theatre setting are classified as inpatient in the UK, while in the USA they are classified as outpatient procedures.

[27]In this case the shift is between day-case and overnight stay, rather than between theatre and non-theatre based settings

mance (P4P) schemes that are typically aimed at improving quality of care by financially rewarding the implementation of specific care processes, generally linked to best practice, or improvements in health outcomes. Only few P4P schemes incentivise measures of performance related to efficiency (Gaughan et al., 2019). P4P schemes are heterogeneous across countries and conditions and evidence regarding their effectiveness remains limited (Milstein and Schreyoegg, 2016; Mendelson et al., 2017; Eijkenaar, 2012; Ogundeji et al., 2016a; Vlaanderen et al., 2019), with many programs lacking proper evaluation (Milstein and Schreyoegg, 2016). Our study provides a comprehensive evaluation of one of the few schemes focusing on efficiency.

3.3 Institutional background

The English National Health Service (NHS) is funded by general taxation and it is free at the point of consumption. Patients are registered with a general practitioner (GP), who acts as gatekeeper and patients require a GP referral to access a hospital specialist. Patients have a legal right to select the hospital, nonetheless choice remains limited (NHS England and Monitor, 2015). When recommending a hospital treatment for a given medical condition, providers have to inform the patient about the risks associated with the treatment and offer them possible alternatives[28] (Citizen Advice , 2020).

Hospitals provide elective care in the inpatient and the outpatient setting. Inpatient care refers to patients who are admitted to a hospital either with an overnight stay or as a day case whereby the patient is provided with a hospital bed for tests or surgery, but will not stay overnight. Instead, outpatient care refers to ambulatory type of care carried out by specialists in an office based setting. Outpatient care represents the largest share of NHS contacts in the hospital setting (Royal College of Physicians, 2018). In 2018/19, there were over 17 million inpatient admissions (NHS Digital, 2019a), and over 123 million outpatient

[28]The patient is only entitled to treatment deemed appropriate by the consultant (Citizen Advice , 2020).

attendances (NHS Digital, 2019b).

For inpatient care, hospitals are paid per patient treated, with the tariff based on the national average reported costs adjusted for local input prices (NHS England and NHS Improvement, 2019). The tariff varies by healthcare resource groups (HRGs), the English version of Diagnosis Related Groups (DRGs). Patients are grouped into HRG based on the recorded diagnosis and procedures codes, clinical setting as well as demographic characteristics (e.g. age).

For outpatient care, the payment is made per attendance, with the tariff typically based on the clinical specialty (e.g. urology) and attendance type (first or follow-up attendance). For selected procedures performed in the outpatient setting the tariffs are based on HRGs, using the same HRG codes used for inpatients. However, in most cases the outpatient tariffs are lower compared to those in the inpatient setting (NHS England and NHS Improvement, 2019).

3.3.1 BPT policy for outpatient care

The *BPT Outpatient scheme* was introduced in 2012/13 in a move to shift some activity from the theatre based inpatient setting to the outpatient office based setting. The BPT includes three procedures: diagnostic cystoscopy, diagnostic hysteroscopy and hysteroscopic sterilisation. These procedures were selected by the Department of Health based on expert clinical advise, further supported by high outpatient rates achieved by a small number of providers before the start of the scheme (Department of Health, 2012).

For cystoscopy and hysteroscopy the BPT consisted of two tariffs, one for the outpatient setting and one for inpatient setting. As shown in Table 3.1, before 2012/13 the tariffs reflected the expected cost, with the inpatient tariff being about three times the outpatient tariff. Instead, since 2012/13 the outpatient tariff was at least 60% higher than the inpatient tariff, and therefore set at level which was significantly higher than the expected cost of the outpatient procedure while the inpatient setting was priced below the expected cost. Only

providers who sustained a high proportion[29] of outpatient care could break even under this arrangement. For hysteroscopic sterilisation, the BPT substantially increased the price for the outpatient setting by about four times in 2012/13, with the inpatient price remaining part of the conventional national price setting increasing over the years at a slower rate, by up to 26% in a given year (Department of Health, 2012).

To qualify for the outpatient BPT tariff, the procedure must be coded to the outpatient department and be performed in a non-theatre based setting with local or no anaesthetic. Procedures in the inpatient department are instead performed in a theatre-based setting, typically under general anaesthetic.

Table 3.1: National tariff, representing the price [in £] paid to providers for performing the BPT procedures over time.

	Cystoscopy		Hysteroscopy		Sterilisation	
	Inpatient	Outpatient	Inpatient	Outpatient	Inpatient	Outpatient
2010/11	687	231	771	231	771	274
2011/12	714	257	733	242	733	242
2012/13	260	403	260	457	928	1,137
2013/14	251	444	268	472	1,034	1,174
2014/15	246	436	264	465	1,018	1,156
2015/16	248	438	250	498	1,123	1,238

Notes: National tariff represents the price [in £] paid to providers for performing the BPT procedures over time.

Source: NHS England Tariff Workbooks, for years 2010/11-2015/16; available online https://improvement.nhs.uk/resources/national-tariff/. Tariff is not readily available for financial year 2009/10, although the tariff structure is analogous to the other years in our sample.

Hysteroscopy and cystoscopy are both established diagnostic tests that are in widespread use across the UK. Cystoscopy is a diagnostic endoscopic procedure involving a telescopic examination of the bladder and urethra using a cystoscope. It is used to check for common problems such as frequent urinary tract infections, long lasting pelvic pain as well as to remove tissue for biopsy and help with the diagnosis and follow up of urogenital cancers (NHS Direct, 2020). The procedure is routinely used in both male and female patients.

[29]The threshold to break-even in 2012/13 is set at 60% outpatient rate for diagnostic hysteroscopy and 50% outpatient rate for diagnostic cystoscopy. The estimated achievable rate is 80% for hysteroscopy and 50% for cystoscopy (Department of Health, 2012).

While in many countries cystoscopy is performed in an office based setting (Casteleijn et al., 2017), it was mainly performed as an inpatient procedure in the beginning of our time series with only 12% of all cystoscopies performed in the outpatient setting in 2009/10.

Hysterescopy is also classed as a endoscopic procedure and involves the use of miniaturised endoscopic equipment to directly visualise and examine the uterine cavity. It is primarily used for assessment of abnormal uterine bleeding and investigation of reproductive problems (NHS Direct, 2018). While it was historically performed in the inpatient setting, advances in endoscopic technology and ancillary instrumentation have facilitated the move to outpatient setting with or without the use of local anaesthesia (Yen et al., 2019).

Unlike cystoscopy and hysteroscopy, hysteroscopic sterilisation is a treatment rather than a diagnostic tool. It is one of two main forms of sterilisation procedures for women and it is primarily performed in the outpatient setting. The technique is relatively novel, as it was first introduced in 2001. It involves insertion of titanium (nitinol) metal device into the fallopian tube (Murthy et al., 2017). The alternative method is the inpatient laprascopic form of sterilisation, historically regarded as the gold standard by which female sterilization techniques are measured (Greenberg, 2008).

All three incentivised procedures are deemed safe when performed in the outpatient setting, with the general anaesthesia in the inpatient setting typically presenting bigger risk to patients than complications arising form the outpatient procedures [30]. Nevertheless, while generally successful, safe and well-tolerated, outpatient hysteroscopy can be associated with significant pain in up to 35% of women (Iaco et al., 2000; Ahmad et al., 2017), which is also the primary reason for early abandonment of procedure (Ahmad et al., 2017). In the UK, there are several patient groups actively advocating for better pain control and more choice in the selection of suitable hysteroscopy technique. The largest one is the Campaign Against Painful Hysteroscopies (CAPH), which started in 2012/13. Their campaign included notable presence in media; in response to their action, the pain management issue

[30]Most patients can be treated in both settings and thus the two settings can be considered as almost perfect substitutes.

was discussed three times in the national Parliament (CAPH, 2018).

While cystoscopy is also associated with discomfort and anxiety, patients generally do not experience extreme pain during the procedure (Greenstein et al., 2014; Falavolti et al., 2017). In a prospective UK study the success rate of outpatient cystoscopy was more than 96%, accompanied by high levels of tolerability and patient satisfaction (Lee et al., 2009).

Hysteroscopic sterilisation is generally associated with less pain and shorter recovery time compared to the inpatient laproscopic procedure and it is suitable for patients with increased anaesthetic risks associated with laparoscopic technique (Royal College of Obstetricians and Gynaecologists , 2016). However there is evidence of higher rates of post-operative complications. While patients undergoing hysteroscopic sterilization have a similar risk of unintended pregnancy, they have 10-fold higher risk of undergoing the operation a second time compared with patients undergoing laproscopic sterilization (Mao et al., 2015). In addition, compared to the laprascopic technique, hysteroscopic sterilisation is irreversible (Royal College of Obstetricians and Gynaecologists , 2016). According to Royal College of Obstetricians and Gynaecologists (RCOG), women should be presented with a comprehensive description of benefits and risks of both techniques (Royal College of Obstetricians and Gynaecologists , 2016).

3.4 Empirical strategy

We employ a difference-in-difference approach to estimate the causal effect of the introduction of the *BPT Outpatient scheme* on providers' behaviour. We are interested in assessing whether providers responded by increasing the proportion of patients treated in an outpatient versus the inpatient setting, and whether in turn this affected the quality of care received by patients, the total volume of patient treated either in an inpatient or an outpatient setting, or diverted effort from non-incentivised procedures.

First, we estimate the effect of the BPT policy on the probability of patients being treated

in the outpatient setting using the following regression specification:

$$Y_{iht} = \alpha + \beta\ BPT_i + \theta\ (D_t\ \ BPT_i) + \boldsymbol{X_i'}\boldsymbol{\delta}_1 + (\boldsymbol{X_i'}\ \ BPT_i)\boldsymbol{\delta}_2 + \boldsymbol{v_t} + \boldsymbol{v_s} + \boldsymbol{v_h} + \epsilon_{iht} \quad (3.1)$$

where Y_{iht} is a binary variable taking value of 1 if the patient i in hospital h in month t (ranging from 1, indicating April 2009, to 84, indicating March 2016) is treated in an outpatient setting and 0 if the patient is treated in an inpatient setting. D_t is a dummy variable equal to 1 in the post-policy period (from April 2012 to March 2016), and equal to 0 in the pre-policy period (from April 2009 to March 2012). BPT_i is a dummy variable equal to 1 if the patient receives an incentivised procedure and equal to 0 if the patient receives a non-incentivised (control) procedure. $\boldsymbol{v_t}$ is a vector of binary year indicators for each financial year in the study (2010/11 to 2015/16 with reference category 2009/10). $\boldsymbol{v_s}$ is a vector of calendar months (January to December, with reference category April) to adjust for seasonality. $\boldsymbol{v_h}$ is a vector of hospital fixed effects to control for time-invariant hospital factors. $\boldsymbol{X_i}$ is a vector of patient characteristics, described in more detail in section 3.5. We allow the effect of patient characteristics to differ across the incentivised and the non-incentivised procedure (the treatment and the control group). α is the intercept and ϵ_{iht} is the error term. We estimate (3.1) as a linear probability model with standard errors clustered at the hospital level. The key coefficient of interest is θ, which gives the average treatment effect on the treated over the post-policy period. In economic terms, it provides an estimate of the effect of the financial scheme on the probability of patients being treated in the outpatient rather than in the inpatient setting.

Since the BPT scheme was rolled out across all hospitals in the English NHS at the same time, we construct control groups that include a selection of comparable non-incentivised procedures that can be performed in both an inpatient and outpatient setting and show similar pre-policy trends to the incentivised BPT procedures. In more detail, to select the control groups we proceed as follows:

- We first select procedures (HRGs) that can be performed in both the inpatient and outpatient setting and hence have a separate HRG tariff in 2015/16 (242 conditions meet this criterion).

- We only keep procedures performed in either the urology or gynecology department (as the BPT conditions are performed in these two departments). However, as cystoscopy and hysteroscopy are endoscopic procedures, we keep diagnostic endoscopies that are performed in other departments (i.e. colonoscopy)[31]. This restriction reduces the number of potential control groups to 32.

- We visually inspect the parallel trend assumption for each of these 32 potential control groups for our primary outcome, the probability of being treated in the outpatient setting. We exclude control groups for whom the trend is not parallel[32]. This restriction reduces the number of potential control groups to seven.

- For these seven control groups we check if they are clinically relevant. In particular, we seek medical advice regarding the adverse effects (including severe pain) of performing the procedure in the outpatient setting, whether the patient had choice associated with the selection of the treatment setting and information on any potential new technologies or incentives that might change the probability of treatment in the outpatient setting during our study period.

- As a result of the above process, we identify three final control groups, one for each incentivised procedure. The control groups are sigmoidoscopy (for cystoscopy), lower genital procedures (for hysteroscopy) and vacuum aspiration with cannula (for sterilisation). See section 2.4 for more details.

[31]The reasoning for including endoscopies is that while they target different body parts, endoscopies share many similarities in the way they are delivered, including the use of endoscope and similar pain relief techniques.

[32]While we inspect the parallel trends assumption separately for each outcome, this does not inform the selection of the control group. Only the probability of being treated as outpatient is used for this purpose.

Equation (3.1) above estimates the average effect of the policy across the entire post-policy period. To test if the effect differs across years, we also use the following flexible specification:

$$Y_{iht} = \alpha + \beta BPT_i + \boldsymbol{\theta}\,(\boldsymbol{v_t}\ BPT_i) + \boldsymbol{X_i'\delta}_1 + (\boldsymbol{X_i'}\ BPT_i)\boldsymbol{\delta}_2 + \boldsymbol{v_t} + \boldsymbol{v_s} + \boldsymbol{v_h} + \epsilon_{iht} \quad (3.2)$$

where $\boldsymbol{v_t}$ is a vector of binary year indicators for each financial year across both the pre- and post-policy period (2010/11 to 2015/16, with reference year 2009/10). In this case $\boldsymbol{\theta}$ is a vector of policy estimates for each financial year in the study, where we expect these to be not significantly different from zero in the pre-policy years. As before, we estimate (3.2) as a linear probability model with standard errors clustered at the hospital level.

While the main objective of the *BPT Outpatient scheme* is to shift procedures from the outpatient to the inpatient setting, incentive programs may influence other aspects of patient care. To better understand the overall impact of the BPT, we explore three more outcome measures. For this part, we focus only on the two diagnostic procedures, cystoscopy and hysteroscopy, as the selected outcome measures are less relevant or appropriate for sterilisation[33]. First, we analyse the impact of the scheme on quality of care. In particular, diagnostic endoscopic procedures (including hysteroscopy and cystoscopy) are frequently associated with severe pain when performed without anaesthesia (as it is generally the case in the outpatient setting), resulting in a failure to complete the clinical investigation. In this case the procedure has to be repeated at a later date[34], causing additional stress and inconvenience for the patient and extra demand on the healthcare system. The rate of failed procedures is a common health outcome measure for endoscopic procedures (Relph et al.,

[33]Sterilisation differs from the two diagnostic procedures in many aspects; unlike hysterosopy and cystoscopy, it is considered a treatment rather than a diagnostic procedure; it's not performed in high volumes and it's associated with much greater patient involvement in decision making. Therefore, we don't consider the selected additional outcomes measures as appropriate for this particular condition

[34]It has been shown that the pain experienced during the procedure is the leading cause of repeated treatments, accounting for over 80% of all cases (Ahmad et al., 2017). However, occasionally procedures have to be repeated when they are originally performed in the inpatient setting, for a variety of reasons including incomplete view of the tissue.

2016; Genovese et al., 2020). Since we cannot observe this metric directly in our data, we measure instead whether the procedure was repeated at a later date. This relies on the assumption that the per patient rate of other factors for failed operation (anatomical factors and structural abnormalities) remains constant throughout the period. We estimate the effect of the policy on the probability to have a repeated procedure (in any setting) within the following 60 or 90 days of the initial procedure[35]. In this case the dependent variable Y_{iht} takes the value 1 if the patient had the repeated procedure within the next 60/90 days and 0 otherwise. As pain is likely to be more prominent for hyteroscopy than cystoscopy, a priori we expect larger effect of the policy on repeated procedures for hysteroscopy. We estimate the effects of the BPT on patient outcomes using the same regression specification as in (3.1), with the same independent regressors.

Second, we investigate the effect of the financial scheme on total patient volume. The BPT significantly changes the profitability of the incentivised outpatient procedures. While this motivates providers to shift patients from the inpatient to the outpatient setting and hence the *intensive* margin, it could also affect the *extensive* margin, by inducing providers to offer the procedure to patients who would not otherwise be eligible according to their medical need. These effects can thus contribute to an increase in the total volume of procedures and affect the overall NHS healthcare costs. We analyse the effect of the BPT on total volume (inpatient and outpatient cases combined) using quarterly data reported at the provider level using the following specification:

$$Y_{jhq} = \alpha + \beta BPT_j + \theta(D_t \ BPT_j) + \boldsymbol{X'_{jhq}}\boldsymbol{\delta_1} + (\boldsymbol{X'_{jhq}} \ BPT_j)\boldsymbol{\delta_2} + \boldsymbol{v_t} + \boldsymbol{v_s} + \boldsymbol{v_h} + \epsilon_{jhq} \quad (3.3)$$

where Y_{jhq} is the total number of procedures j performed in hospital h in quarter $q = 1, \ldots, 21$, where the quarter 1 corresponds for the period April-June 2009 whereas the

[35]We use different time points for two reasons. First, several cancer treatments specify that patients should have recurrent endoscopic investigations every three months. We therefore include procedures done within 60-days, as patient's are unlikely to need repeated investigation in this period on medical grounds. And second, we include repetition within 90-days to take into account potential waiting time for the procedure.

last quarter 21 corresponds to the period Jan-March 2016. BPT_j equals to 1 if j is an incentivised procedure and 0 if it is a control procedure. D_t is a dummy variable equal to 1 in the post-policy period (from April 2012 to March 2016), and equal to 0 in the pre-policy period (from April 2009 to March 2012). X_{jhq} is a vector of patient characteristics for condition j averaged on a quarterly basis by provider. $\boldsymbol{v_t}$ is a vector of binary year indicators for each financial year in the study (2010/11 to 2015/16 with reference category 2009/10). $\boldsymbol{v_h}$ is a vector of hospital fixed effects and $\boldsymbol{v_s}$ is a vector of four seasonal dummies[36] (July-September, October-December, January-March, with reference category April-June). Our key coefficient θ measures whether volume increased more quickly for the incentivised procedure than for the control one. We estimate (3.3) using OLS.

Third, we investigate whether the BPT policy had negative or positive spillover effects on closely related unincentivised procedures, and therefore affected the probability of selecting the outpatient setting for procedures that are clinically similar to the BPT procedures but were not incentivised, a form of unintended consequence. On one hand, limited capacity to treat patients in the outpatient setting could reduce and crowd out the ability of the provider to perform outpatient activity for other procedures, a form of negative spillover. On the other hand, the incentivised BPT could induce providers to invest in outpatient capacity, therefore reducing the marginal cost of treating patients in an outpatient setting even for non-incentivised procedured, a form of positive spillover. For this estimation we select two procedures that are very similar to the incentivised hysteorsopy and cystoscopy, but do not attract the bonus: we use hysteroscopy with insertion of uterine device to test spillover effects of the BPT for hysteroscopy, and we use endoscopic urethra procedures to test the spillover effects for cystoscopy. In our estimation we use the same empirical approach as in (3.1) and the same control groups. A negative (positive) coefficient θ here would provide evidence of negative (positive) spillover effects, implying that the BPT reduced (increased) the probability of providing the procedure in an outpatient setting for a

[36] As the analysis is performed on the provider level, quarterly data was used to avoid the issue of small numbers of patients treated in hospitals each month.

closely-related procedure that was not incentivised by the BPT.

All the analyses described in this section use a difference-in-difference approach and rely on the parallel trends assumption. We first use visual inspection of the pre-parallel trends to select the control procedures. We then also test the plausibility of the parallel trend assumption empirically by estimating model (3.2) only for the period prior to the BPT introduction (2009/10-2011/12), and testing if the coefficients associated with the year dummies interacted with the treatment group are statistically significantly different from zero. We perform this test separately for each outcome and procedure.

3.5 Data

We employ data from the Hospital Episode Statistics (*HES*), which separately collects information on inpatient and outpatient care. *HES Outpatients* is a dataset comprised of all office-based consultations and procedures, detailing information about the patient's hospital visit, including their socio-demographic data (ie. age, gender, income deprivation), whether the visit was patient's first attendance, and OPCS[37] codes of any procedures[38] carried out during the appointment. *HES Inpatients* includes all inpatient and day-case admissions, detailing information on patient's care pathway, including admission and discharge date, type of admission (elective or non-elective), patient's diagnosis (ICD) and procedure (OPCS) codes, and socio-demographic data.

Our study period is from April 2009 to March 2016, with the pre-policy period running from April 2009 to March 2012. Our sample consists of all patients aged 19 or older who, during this period, had either a BPT procedure (our treatment group) or any of the

[37]OPCS Classification of Interventions and Procedures, first published by Office of Population Censuses and Surveys, is the classification of procedures used by clinical coders within the NHS. It provides codes for operations, procedures and interventions performed during inpatient stays, day case surgery and outpatient treatments in the NHS hospitals. While the codes themselves are different, the code set is comparable to the American Medical Association's Current Procedural Terminology (NHS Digital, 2020).

[38]Diagnosis code fields are included in the dataset, however they are only available for less than 1% of all consultation visits and are hence not used in our analysis.

procedures that we use to construct the control groups. Our sample across the three BPTs and the corresponding control groups consists of 5,723,343 observations[39]. Additionally, we have 235,227 observations for the two procedures used to test for possible spillover effect[40]. Including those, the full sample consists of 5,958,570 observations, of which 3,747,670 (62.9%) are in the inpatient setting and 2,210,900 (37.1%) in the outpatient setting[41].

Patients are placed into the treatment and control group based on their assigned HRG codes (English equivalent to the DRG codes), mirroring the method used by the BPT scheme for payment purposes. The English HRG system is frequently updated, with groups added and removed on a yearly basis. To create consistent series throughout the study period, we use the NHS Digital Grouper software[42] to assign coherent set of HRG groups across all years. The grouper software is developed by the NHS for payment and benchmarking purposes and uses patient's clinical and demographic characteristics to assign the HRG group.

3.5.1 Outcome measures

We estimate the effect of the BPT on three outcome variables. First, our main dependent variable is a categorical variable equal to one if the patient was treated in the outpatient setting and equal to zero if treated in the inpatient setting. Second, we construct a measure of *hospital quality*, proxied with the probability of the patient receiving the same procedure

[39]Out of 5,723,343 observations, just over 3 million correspond to the BPT incentivised procedures: 2,359,964 are for cystoscopy, 558,618 for hysteroscopy and 116,216 for sterilisation. The rest of the observations make up the control groups: 1,499,406 are for sigmoidoscopy, 1,002,097 for lower genital procedures and 187,042 for vacuum aspiration with cannula.

[40]The two control groups for spillover effect are urethra procedure (121,286 observations) and hysteroscopy with insertion of uterine device (113,941 observations)

[41]The 63% inpatient and 37% outpatient proportions refer to the inpatient/outpatient cases across all conditions (including control conditions), and across the entire time period. This includes the pre-policy period, when the inpatient cases were dominant.

[42]We use Payment Grouper version 2016, freely available to download from the following link https://webarchive.nationalarchives.gov.uk/20171011074955/http://content.digital.nhs.uk/article/2063/Archive-payment

again within the next 60/90 days (”repeated procedure” for short). We construct a categorical variable equal to one if the patient received the same procedure (at least once) within 60 and 90 days from the original procedure; we assign a value of 1 if the number of past procedures is one or more and 0 otherwise. Third, we measure *volume* as the total number of procedures performed across both the inpatient and outpatient setting. This variable is measured at provider level for each quarter and we only include providers present in all quarters of the study period[43] (131 out 167 of for cystoscopy and 132 out of 167 for hysteroscopy). The provider exclusion only applies to the estimation of volume; in all other estimations we use the full sample.

3.5.2 Control variables

We control for patients' clinical and socio-demographic characteristics, including age (measured as a categorical variable with 5-year bands and two separate categories for 19 to 24 and 90+), sex (male=1) and the number of past emergency hospital visits in the year prior to the procedure (measured as a categorical variable with values from 0 to 4 and 5+) as a proxy of patient severity[44]. As a proxy of socio-economic status, we use the income deprivation score of the English Indices of Deprivation 2010 for patient local area of residence (coded in quintiles).

3.5.3 Control group procedures

Procedures in the control group were selected based on their comparability to procedures in the treatment group in both inpatient and outpatient settings, their clinical relevance and whether they exhibited a parallel trend with the treatment group in the pre-policy period.

[43]There are two main reasons why hospital do not report in all quarters: (i) hospital mergers; (ii) very small numbers of cases, with frequent zeroes. We are more concerned with the former. Several providers merged during the period, which affects the total volume; to avoid this issue, we exclude these providers from the analysis.

[44]Because the diagnosis information is not available for outpatient attendances, we cannot construct the usual Elixhauser/Charlson index to control for severity. As an alternative we use the past number of emergency admissions (in the year prior) as a proxy for patient severity.

The three control groups for the main outcomes are flexible sigmoidoscopy, lower genital procedure and vacuum aspiration with cannula.

Flexible Sigmoidoscopy is used as a control procedure for the BPT procedure diagnostic cystoscopy and is used to evaluate the lower part of the large intestine. Like cystoscopy, it is an endoscopic procedure (with a sigmoidoscope) used as a screening tool to detect polyps and cancerous cell. It can be safely performed in an office based setting with or without the use of pain relief. Similar to cystoscopy, pain is considered one of the main reasons for a failed procedure (Doria-Rose et al., 2005). It is mainly performed in the gastroenterology department, while the diagnostic cystoscopy is performed in the urology department. Thus the chance of a spillover effect across the two procedures is minimal (Kelly et al., 2008).

Lower genital procedure is used as a control procedure for hysteroscopy. It includes an array of procedures grouped to a common HRG group, including procedures of the Bartholian gland[45] (drainage/balloon catheter insertion) and procedures of the vulva. These procedures can typically be performed in the office based setting using local anaesthetic. While all of procedures in this group are performed in the gynaecology department, they tackle unrelated gynaecological problems.

Vacuum aspiration with cannula is used as a control procedure for hysteroscopic sterilisation. This is a safe and effective alternative method for surgical management of miscarriage. It can be performed in the outpatient setting under local anaesthesia. While vacuum aspiration and sterilisation are very distinct procedures, they share several similarities: (i) both procedures were typically performed in the inpatient setting with effective and safe alternatives for surgical management emerging relatively recently (Sharma, 2015); (ii) both procedures require strong patient involvement in the clinical process, including giving the relevant information and support[46]; (iii) patient demographic is similar across two procedures (women of childbearing age).

[45]Bartholin's cyst usually appears as a lump in the genital area; it can become painful and infected, in which case it needs treatment (drainage in the first instance).

[46]NICE guidance requires providers to provide all the necessary information and give support to women who experience miscarriage: https://tinyurl.com/yxpc5ed3

3.5.4 Spillover effects

We test whether the policy affected the probability of being treated in the outpatient setting for procedures that are clinically very similar to the BPT procedures but are not part of the the BPT scheme. One possible concern is that by incentivising an increase in the outpatient setting for one procedure, this may contribute to crowding out and reducing the probability of being treated in an outpatient setting for the non-incetivised procedures, a form of negative spillover effect[47]. To test for spillover effects for cystoscopy we use *Endoscopic urethra procedure*. This is a collection of procedures that are typically performed alongside cystoscopy and are grouped to a single HRG. Examples include retrograde pyelogram, which involves introducing contrast dye into the urinary system during cystoscopy and endoscopic urine sampling, by which a urine sample is taken during cystoscopy to check for tumors and infections. To test for spillover effects for hysteroscopy we use *Hysteroscopy with insertion of inuterine device*. In this case hysteroscopy is followed by insertion of inuterine device, which is generally a straightforward, office based procedure.

For this analysis the dependent variable is again a binary variable equal to one if the patient was treated in an outpatient rather than an inpatient setting. Because of the similarity between the incentivised conditions and the non-incentivised conditions where spillover may be present (in both, trends and clinical similarity) we use the same control groups as in the main analysis (i.e. sigmoidoscopy as the control group of endoscopic urethra procedure, lower genital procedure as the control group of hysteroscopy with insertion of inuterine device).

3.5.5 Descriptive statistics

Table 3.2 presents the sample mean and standard deviation (SD) for each of the outcomes in the pre- and post- policy period (2009/10-2011/12 and 2012/13-2015/16, respectively).

[47]The effect could also go in the opposite direction and hasten the more from an inpatient to an outpatient setting; in this case we would see a positive spill-over effect.

For the three incentivised conditions, we observe sizeable difference across the two periods in the proportion of patients treated in the outpatient setting. The outpatient rates for cystoscopy increased from 13% in the pre-policy period to 52%. Similarly, for hysterescopy it increased from 41% to 62%.[48] The outpatient rate for sterilisation increased from 0% to 5%. Instead, for the control groups the increases in outpatient rates are much smaller. For sigmoidoscopy, the control group of cystoscopy, it increased from 13% to 15%. For lower genital procedure, the control group of hysteroscopy, it increased from 74% to 82%. For vacuum aspiration, the control group for sterilisation, the outpatient rate increased from 1% to 2%.

[48]Department of Health/NHS estimated maximum achievable rate is considerably lower for cystoscopy (50%) than for hysteroscopy (80%) (Department of Health, 2012).

Table 3.2: Descriptive statistics. Outcomes

(a) Cystoscopy, corresponding control group & procedure to test for spillover effects

	Cystoscopy				Sigmoidoscopy				Urethra procedure			
	Pre-policy		Post-policy		Pre-policy		Post-policy		Pre-policy		Post-policy	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Proportion treated as outpatient	0.13	0.33	0.52	0.50	0.13	0.34	0.15	0.36	0.03	0.18	0.13	0.34
Volume (on quarterly basis)	77,292	2,973	89,529	4,526	46,622	3,864	58,746	4,268	4,029	225	4,558	213
Reoperation within 60 days	0.05	0.21	0.06	0.23	0.03	0.17	0.04	0.20	0.03	0.17	0.03	0.17
Reoperation within 90 days	0.07	0.25	0.08	0.27	0.04	0.20	0.05	0.22	0.05	0.21	0.05	0.21

(a) Hysteroscopy, corresponding control group & procedure to test for spillover effects

	Hysteroscopy				Lower genital procedures				Hysteroscopy with ID			
	Pre-policy		Post-policy		Pre-policy		Post-policy		Pre-policy		Post-policy	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Proportion treated as outpatient	0.41	0.49	0.62	0.48	0.74	0.44	0.82	0.38	0.20	0.40	0.38	0.49
Volume (on quarterly basis)	18,781	435	20,828	1,211	35,196	1,040	36,234	1,355	4,340	186	3,866	113
Repeated procedure within 60 days	0.04	0.19	0.05	0.21	0.07	0.25	0.07	0.26	0.00	0.05	0.00	0.05
Repeated procedure within 90 days	0.05	0.21	0.05	0.23	0.09	0.29	0.10	0.30	0.00	0.06	0.00	0.05

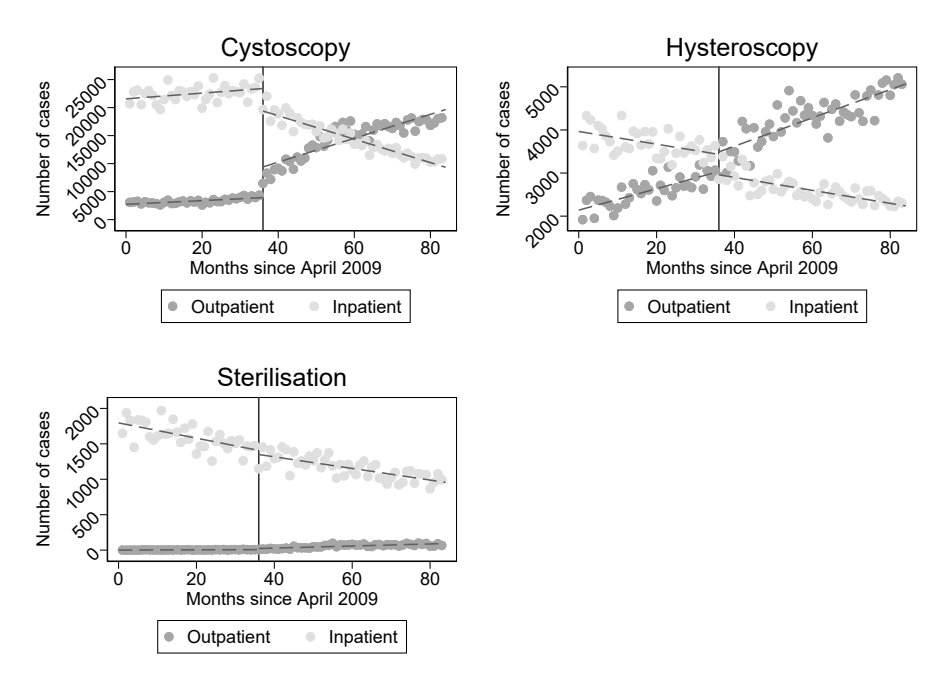
(a) Sterilisation & the corresponding control group

	Sterilisation				Vacuum aspiration			
	Pre-policy		Post-policy		Pre-policy		Post-policy	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Proportion treated as outpatient	0.00	0.05	0.05	0.21	0.01	0.09	0.02	0.13
Volume (on quarterly basis)	4,822	355	3,647	301	6,934	238	6,489	303

Notes: This table shows the sample mean and standard deviation (SD) for the outcomes across treatment/control procedures in the pre-and post-policy period (2009/10-2011/12 and 2012/13-2015/16, respectively). The table shows the proportion of patients treated in the outpatient setting and of having a repeated procedure within 60 or 90 days. For volume, the table presents the average number of patients each quarter in the pre/post-policy period.

Figure 3.1 displays the monthly volume of the BPT procedures across the inpatient and outpatient setting. It provides suggestive evidence of substitution between the two settings. For cystoscopy, there is a rapid decline in inpatient activity after the start of the policy in April 2012 and concurrent increase in the number of outpatient cases. This is also the case for hysteroscopy, though it is less pronounced, with the volume in the inpatient setting already trending downwards prior to the start of the policy. The volume of outpatient hysteroscopic sterilisation increases slowly after the start of the policy and remains low throughout.

Figure 3.1: Volume of BPT procedures over time in an outpatient and inpatient setting



Notes: Number of inpatient and outpatient procedures measured on a monthly basis for the three incentivised BPT procedures: diagnostic cystoscopy, diagnostic hysteroscopy and sterilisation. Time period is from April 2009 to March 2016. The vertical line indicates the start of the BPT policy in April 2012. The scale (y-axis) differs between the three graphs, hence the figures are not directly comparable.

Table 3.2 also suggests that the proportion of patients who require the same procedure again within 60 or 90 days is relatively stable over time both in the treatment and the control groups. About 7% (5%) of patients require additional cystoscopy (hysteroscopy) within 90 days in the pre-policy period. The total volume, across the inpatient and outpatient

setting, averaged across quarters and providers increases over time for both cystoscopy and hysteroscopy (and their control groups) but reduces for sterilisation (and its control group).

Table 3.3 presents descriptive statistics for the patients characteristics in our final sample. While we observe variations across procedures, the patients composition is relatively similar across the treatment group/control group pairs. We briefly comment on the patient characteristics in the treatment group. Patients treated for cystoscopy are on average 66 years old, 64% are male, and 83% did not have any emergency admissions in the previous year. The largest share of patients belongs to the most deprived quintile (28%). Patients treated for hysteroscpy are on average 53 years old, and large majority (92%) did not have and emergency admissions in the past year. Largest hare of patients belongs to the most deprived quintile. Patients treated for sterilisation were on average 35 years old, with majority of patients not having an emergency admission in the year prior to the procedure. Unlike for hysteroscopy and sterilisation, most patient belong to the least deprived group (28%).

Table 3.3: Descriptive statistics. Patient characteristics

(a) Cystoscopy, the corresponding control group & the procedure to test the spill-over effect

	Cystoscopy		Sigmoidoscopy		Urethra procedures	
	Mean	SD	Mean	SD	Mean	SD
Age	66.17	15.25	58.43	16.99	58.53	16.88
Male	0.64	0.48	0.48	0.50	0.55	0.50
Deprivation quintiles						
Most deprived	0.28	0.45	0.28	0.45	0.25	0.43
2nd quintile	0.18	0.39	0.18	0.38	0.17	0.38
3rd quintile	0.19	0.39	0.19	0.39	0.19	0.39
4th quintile	0.18	0.39	0.19	0.39	0.20	0.40
Least deprived	0.17	0.38	0.17	0.38	0.20	0.40
Past emergencies						
0	0.83	0.37	0.86	0.35	0.57	0.49
1	0.04	0.19	0.04	0.20	0.05	0.22
2	0.04	0.19	0.04	0.20	0.13	0.33
3	0.02	0.14	0.02	0.14	0.07	0.26
4	0.02	0.12	0.01	0.11	0.05	0.22
5+	0.05	0.23	0.03	0.16	0.12	0.33
Observations	2,359,964		1,499,406		121,286	

(c) Hysteroscopy, the corresponding control group & the procedure to test the spill over-effect

	Hysteroscopy		Lower genital procedures		Hysteroscopy with ID	
	Mean	SD	Mean	SD	Mean	SD
Age	52.71	13.01	55.20	20.87	44.17	7.69
Male	-		-		-	
Deprivation quintiles						
Most deprived	0.27	0.44	0.23	0.42	0.25	0.43
2nd quintile	0.17	0.38	0.17	0.38	0.17	0.37
3rd quintile	0.18	0.38	0.19	0.39	0.18	0.39
4th quintile	0.19	0.39	0.20	0.40	0.20	0.40
Least deprived	0.19	0.40	0.21	0.41	0.21	0.40
Past emergencies						
0	0.92	0.27	0.90	0.30	0.93	0.25
1	0.03	0.16	0.02	0.14	0.03	0.16
2	0.02	0.15	0.02	0.13	0.02	0.14
3	0.01	0.10	0.01	0.10	0.01	0.10
4	0.01	0.08	0.01	0.08	0.01	0.07
5+	0.01	0.10	0.04	0.20	0.01	0.08
Observations	558,618		1,002,097		113,941	

(c) Sterilisation & the corresponding control group

	Sterilisation		Vacuum aspiration with cannula	
	Mean	SD	Mean	SD
Age	34.66	8.37	29.75	6.71
Male	-		-	
Deprivation quintiles				
Most deprived	0.18	0.39	0.19	0.39
2nd quintile	0.14	0.35	0.14	0.35
3rd quintile	0.17	0.38	0.17	0.38
4th quintile	0.22	0.42	0.22	0.41
Least deprived	0.28	0.45	0.28	0.45
Past emergencies				
0	0.92	0.28	0.86	0.35
1	0.03	0.17	0.09	0.29
2	0.03	0.17	0.03	0.17
3	0.01	0.10	0.01	0.10
4	0.01	0.08	0.00	0.07
5+	0.01	0.09	0.01	0.07
Observations	116,216		187,042	

Notes: Table shows the descriptive statistics for the patients' characteristics in the regression sample across all study years (2009/10-2015/16) inclusing sample mean, and standard deviation (SD). Age is patient's age at the time of admission. Deprivation deciles are based on the continuous IMD index of income deprivation that takes value from 0 (least deprived) to 1 (most deprived); . "Past emergencies" measure the number of emergency admissions in the year prior to the procedure.

3.6 Results

3.6.1 Main effects

In this section, we present our findings on whether the introduction of the BPT policy increased the probability of being treated in an outpatient setting. The results of our difference-in-difference analysis are presented in Table 3.4 (see Table A1 for full regression results). The DiD estimates show a sizeable, positive and statistically significant effect (at 0.1% level) of the *BPT Outpatients Scheme* on the probability of being treated in the outpatient setting for all three BPT procedures. The effect is larger for cystoscopy, and equal to 36.1 percentage points (pp), while it is 16.3 pp for hyteroscopy. The effect is smaller and equal to 3.8 pp for hysteroscopic sterilisation.

Table 3.4: Difference-in-difference estimates of the impact of the BPT Outpatients scheme on the probability of treatment in the outpatient setting

Treatment group	Cystoscopy (1)	Hysteroscopy (2)	Sterilisation (3)
DiD coefficient	0.361***	0.163***	0.038***
	(0.031)	(0.022)	(0.010)
Adjusted R^2	0.385	0.285	0.092
Number of hospitals	168	167	158
Observations	3,859,365	1,560,714	303,256

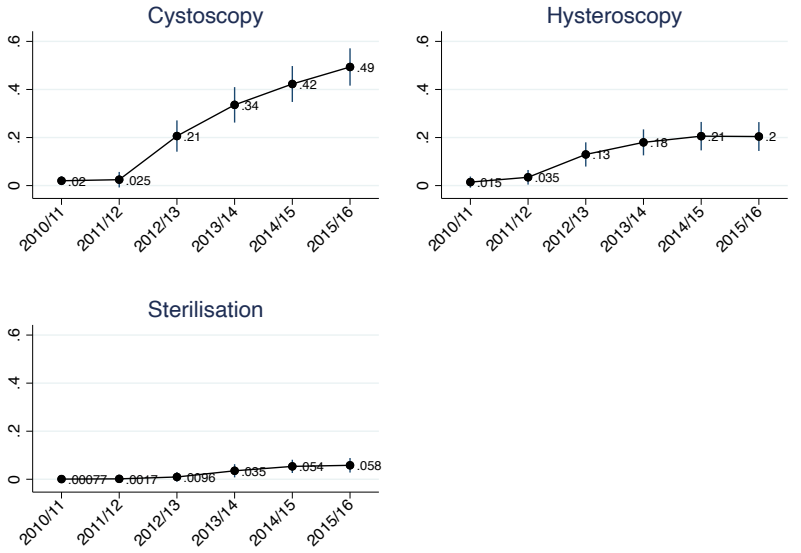
Notes: Dependent variable is the probability to be treated in the outpatient setting. Time period is from April 2009 to March 2016. Models are estimated by OLS with standard errors (presented in parenthesis under the coefficients), clustered at hospital level. Models are run separately for each treatment-control procedure pair (cystoscopy-sigmoidoscopy; hysteroscopy-lower genital procedures; sterilisation-vacuum aspiration). All models control for casemix and a set of month, year and hospital fixed effects.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Figure 3.2 shows the effect of the BPT scheme for each financial year both pre and post policy (using the first year of the pre-policy period as reference group) with 95% confidence intervals. The effect is close to zero in the pre-policy years (2010/11 and 2011/12) and is increasing over time for all three conditions in the post-policy years. For hysteroscopy we observe that most of the effect is concentrated in the first year of the BPT. For cystoscopy

and sterilisation the increase is more gradual over time.

Figure 3.2: Probability of being treated in an outpatient setting by year



Notes: Effect of the BPT across financial years relative to 2009/10 with 95% confidence interval.

The effect differs across the three conditions. While cystoscopy and hysteroscopy both had similar estimated achievable rates (Department of Health, 2012), the starting point is different across the two conditions: cystoscopy had a much lower probability of outpatient treatment in the pre-policy period compared to cystoscopy (13% vs 41%). Furthermore, the outpatient treatment for hysteroscopy received negative publicity (CAPH, 2018), which might negatively affect the uptake rates (for both, hospitals and patients). The effect is smallest for sterilisation. Compared to the other two procedures, for which the inpatient prices decreased after the introduction of the BPT, they inpatient prices increased for sterilisation. Smaller size of the effect might also be due to relative novelty of the outpatient sterilisation technique (Murthy et al., 2017) and increased involvement of the patient in the decision making (Royal College of Obstetricians and Gynaecologists , 2016).

3.6.1.1 Sensitivity analyses for the main outcome

We perform additional sensitivity analyses to confirm the robustness of our estimates. We first adjust for the transition period, by excluding from the analysis the period within 6 months on either side of the start of the policy in April 2012. This controls for possible anticipatory effect or delayed transition after the start of the policy. Using the same specification as in the main model, we find that the estimates are adjusted slightly upwards once the transition period is taken into account. As can be seen in table B2(a) in the Appendix, effect of the policy changes from 0.361 to 0.379 for cystoscopy, from 0.163 to 0.180 for hysteroscopy and from 0.038 to 0.043 for sterilisation. All of the coefficients are statistically significant.

We further control for changes in the number of hospitals in the sample, by excluding those hospital who did not report data in all quarters of our study period (therefore using a balanced panel). In this way, we remove any hospitals that merged from the sample, as this could affect their medical practice. In this case the estimates are adjusted slightly downwards: 0.357 for cystoscopy, 0.152 for hysteroscopy and 0.040 for sterilisation (see Table B2(b) in the Appendix). All of the coefficients are statistically significant.

3.6.1.2 Parallel trends assumption for the main outcome

The estimates from the difference-in-difference analyses rely on the parallel trends assumption. We test the plausibility of this assumption both visually and empirically. Figure B1 in the Appendix displays the trends in the proportion of patients treated in the outpatient setting which appear to be parallel for all treatment-control group pairs. We test this also empirically. For the analysis presented in Table 3.5, we restrict the sample to the pre-policy years, and test wither the probability of being treated in an outpatient setting differs between the treatment and the control group in 2010/11 and 2011/12 relative to the first year of the policy in 2009/10. The coefficients are very small, especially when compared to the effects of the policy (presented in Table 3.4) and mostly insignificant. This suggest the

Table 3.5: Empirical test for the parallel trends assumption for the primary outcome measure

Treatment group	Cystoscopy (1)	Hysteroscopy (2)	Sterilisation (3)
DiD coefficient for 2010/11	0.018*	0.015	0.000
	(0.009)	(0.011)	(0.002)
DiD coefficient for 2011/12	0.021	0.036*	0.001
	(0.016)	(0.015)	(0.006)
Adjusted R^2	0.464	0.373	0.122
Number of hospitals	159	157	153
Observations	1,488,249	650,416	141,140

Notes: Dependent variable is the probability to be treated in the outpatient setting. Time period is from April 2009 to March 2012. Standard errors (presented in parenthesis under the coefficients) are clustered at hospital level. Models are run separately for each treatment-control procedure pair (cystoscopy-sigmoidoscopy; hysteroscopy-lower genital procedures; sterilisation-vacuum aspiration). All models include a constant, case mix variables and a full set of month and hospital dummies. The null hypothesis for the parallel trends assumption is that the DiD coefficients are jointly zero.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

parallel trend assumption is likely to hold.

3.6.2 Effect on quality of care and volume

Table 3.6 reports the effect of the policy on hospital quality showing little impact. The results suggests that the BPT policy slightly reduced the risk of having a repeated cystoscopy within 60 and 90 days by 0.2 pp and 0.4 pp, respectively. However, the coefficients are not statistically significant. For hysteroscopy, the estimates show instead an increased risk of repeated hysterescopy within 60 and 90 days by 0.7 and 0.9 pp, though again the coefficients are not statistically significant.

While all of the coefficients are small and insignificant, the results are nevertheless in line with our a priori expectation that increasing the proportion of patients in an outpatient setting could potentially affect more hysteroscopy patients relative to cystoscopy patients. This is due to the fact that outpatient hysteroscopy is associated with severe pain when performed without anaesthesia, resulting in a failure to complete the clinical investigation and need for repeated procedures (Iaco et al., 2000). While cystoscopy is also associated

Table 3.6: Difference-in-difference estimates of the impact of the BPT Outpatients scheme on the probability of having a repeated procedure within 60/90-days

	Repeated procedure: 60 days		Repeated procedure: 90 days	
	Cystoscopy	Hysteroscopy	Cystoscopy	Hysteroscopy
	(1)	(2)	(3)	(4)
DiD coefficient	-0.002	0.006	-0.004	0.006
	(0.002)	(0.004)	(0.002)	(0.005)
Adjusted R^2	0.008	0.022	0.012	0.029
Number of hospitals	168	167	168	167
Observations	3,859,365	1,560,714	3,859,365	1,560,714

Notes: Dependent variable is the probability to be treated in the outpatient setting. Time period is from April 2009 to December 2015. Models are estimated by OLS with standard errors (presented in parenthesis under the coefficients) clustered at hospital level. Models are run separately for each treatment-control procedure pair (cystoscopy-sigmoidoscopy; hysteroscopy-lower genital procedures; sterilisation-vacuum aspiration). All models control for casemix and include a set of month, year and hospital dummies.

*** p<0.001, ** p<0.01, * p<0.05

with discomfort and anxiety, patients generally do not experience such extreme pain during the procedure. Overall, these results suggests the the BPT manages to improve efficiency without negatively affecting quality of care.

Table 3.7 reports the effect of the BPT policy on volume. Although the coefficient is positive for both cystoscopy (by 17.8 patients, less then 3%, per hospital and quarter) and hysteroscopy (by 23.0 patients, less than 5%), it is not statistically significant. Therefore, although the policy significantly affected the intensive margin (by increasing substitutions across settings), it did not affect the extensive margin.

We further test for the plausibility of the parallel trends assumption for the additional analyses on quality and volume. Figures B2 and B3 in the Appendix shows the trends over time for volume and for the probability of having repeated procedure within 60/90 days. The pre-trends appear parallel for both treatment-control pairs. The results of the empirical test are presented in Tables B3(a)-(c) in the Appendix. All coefficients are small and/or insignificant, confirming that the parallel trend assumption is likely to hold.

Table 3.7: Difference-in-difference estimates of the impact of the BPT Outpatients scheme on volume

	Cystoscopy (1)	Hysteroscopy (2)
DiD coefficient	17.471	22.954
	(13.753)	(12.269)
Adjusted R^2	0.706	0.546
Number of hospitals	131	132
Observations	7,336	7,392

Notes: Dependent variable is the probability to be treated in the outpatient setting. Time period is from April 2009 to March 2016. Models are estimated by OLS with standard errors (presented in parenthesis under the coefficients) clustered at hospital level. Models are run separately for each treatment-control procedure pair (cystoscopy-sigmoidoscopy; hysteroscopy-lower genital procedures). Both models control for casemix and include a set of month, year and hospital dummies. Only hospitals who report to all quarters are included in the analysis.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

3.6.3 Spillover effects

We have also tested whether the BPT policy had an effect on closely related procedures. Table 3.8 suggests that the BPT policy had a sizeable impact on the proportion of patients treated in an outpatient setting for selected closely-related procedures that were not incentivised by the BPT. For endoscopic urethra procedures, the policy increased the probability of outpatient treatment by 4.5 pp. While this estimate is statistically significant, the effect is much lower than for cystoscopy (35.0 pp). For hysteroscopy with insertion of inuterine device, the effect of the policy on outpatient treatment is 12.4 pp and significant. This is more in line with the effect observed for hysteroscopy (16.4 pp), suggesting a considerable positive spillover effect.

The results are in line with figure 3.3, which shows the trends in proportion of outpatient cases for the treatment, control and spill-over conditions. For cystocsopy the spillover procedure is endoscopic urethra procedures for which we observe a modest shift in the proportion of cases after the introduction of the policy. For hysteroscopy, the spill-over procedure is hysteroscopy with insertion of a ID. For this pair, we observe almost identi-

Table 3.8: Difference in difference results of the main effects for the spill-over procedures

Treatment group	Endoscopic urethra procedures (1)	Hysteroscopy with ID (2)
DiD coefficient	0.045***	0.124***
	(0.020)	(0.021)
Adjusted R^2	0.445	0.363
Number of hospitals	167	167
Observations	1,084,180	1,076,988

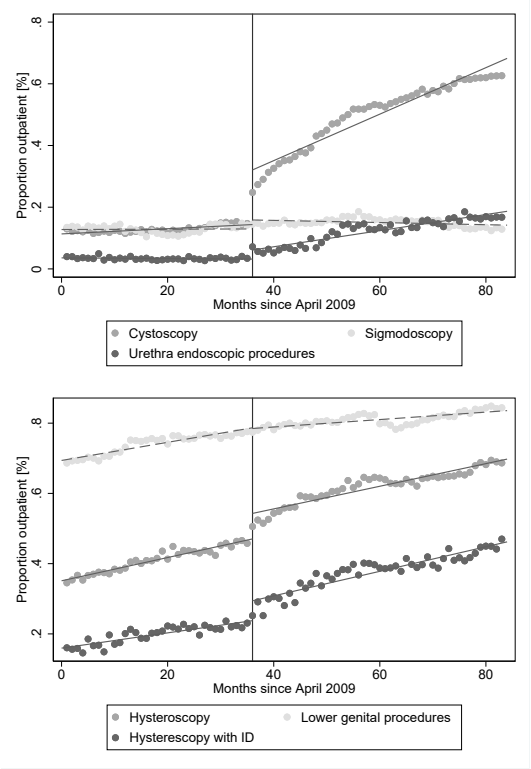
Notes: Dependent variable is the probability to be treated in the outpatient setting. Time period is from April 2009 to March 2016. Models are estimated by OLS with standard errors (presented in parenthesis under the coefficients), clustered at hospital level. Models are run separately for each spill over-control procedure pair (endoscopic urethra procedure-sigmoidoscopy; hysteroscopy with ID-lower genital procedures). All models include a constant, case mix variables and a full set of month and hospital dummies.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

cal patterns over time both pre- and post- policy in the treatment (hysteroscopy) and the spillover (hysteroscopy with ID) condition.

We also test for the parallel trends assumption for the spillover effect. Figure 3.3 shows the trends in the probability of being treated in the outpatient setting, which appear parallel for both across both spill-over/control pairs. The results of the empirical test are presented in Tables B3(d) in the Appendix. All coefficients are small and/or insignificant, confirming again that the parallel trend assumption is likely to hold.

Figure 3.3: Proportion of patients treated in the outpatient setting over time for the treatment, control and spillover procedures



Notes: The figure shows the trends in the proportion of outpatient procedures for the two diagnostic BPT procedures (cystoscopy and hysteroscopy), the corresponding control groups (sigmoidoscopy and lower genital procedures) and the applicable spillover procedures (endoscopic urethra procedure and hysteroscopy with ID).

3.7 Effect on revenues and profits

We further explore the effect of the BPT on hospitals' revenues and profits. Our results show that the BPT increased the probability of being treated in the outpatient setting compared to the inpatient setting, while also having a small effect on the total volume, which affect both revenues and costs[49].

[49] At the same time, the inpatient prices (for cystoscopy and hysteroscopy) are set in a way that a hospital only *breaks-even* if big enough proportion of patients (50% and 60%, respectively) is treated in the outpatient setting.

3.7.1 Effect on revenues

Define p_O^1 and p_I^1 as the prices paid to hospitals in the post-policy period *(1)* for treating the patient in an outpatient and inpatient setting, respectively; and p_O^0, p_I^0 as the outpatient and inpatient prices in the pre-policy period *(0)*. Similarly, define V_O^1 and V_I^1 as the volumes of patients treated in an outpatient and inpatient setting in the post-policy period *(1)*, and V_O^0 and V_I^0 as the outpatient and inpatient volumes in the pre-policy period *(0)*. The hospital revenues in the post-policy (R_{Post}) and pre-policy period (R_{Pre}) are then given by:

$$R_{Post} = p_O^1 V_O^1 + p_I^1 V_I^1 \tag{3.4}$$

$$R_{Pre} = p_O^0 V_O^0 + p_I^0 V_I^0, \tag{3.5}$$

with the difference in revenues corresponding to:

$$\Delta R = R_{Post} - R_{Pre} \tag{3.6}$$

Table 3.9 presents the pre- and post-policy prices and volumes for the three conditions. The prices are those used to pay hospitals in 2011/12, which is the year prior to the introduction of the policy, and in 2012/13, which is the year the BPT was introduced. The pre-policy volumes are based also on 2011/12. The volumes for 2012/13 are estimated using our regression results, which show that the policy increased to a great extent the proportion of patients treated in the outpatient setting, and to a small extent the volume. More details on the calculations are available in the appendix B2. We can therefore use the data from table 3.9 to compute the hospital revenues for 2011/12 both before the policy and post policy, as if the policy had been in place for that year.

Table 3.9: Coefficients used to calculate the effect of the BPT on hospital's revenue and costs

(a) Prices

	Cystoscopy	Hysteroscopy	Sterilisation
Pre-policy price			
Outpatient p_O^0	257	242	242
Inpatient p_I^0	714	733	733
Post-policy price			
Outpatient p_O^1	403	457	1,137
Inpatientp_I^1	260	260	928

(b) Volumes

	Cystoscopy	Hysteroscopy	Sterilisation
Pre-policy volume			
Outpatient V_O^0	45,010	34,227	348
Inpatient V_I^0	276,487	41,834	11,268
Post-policy volume			
Outpatient V_O^1	208,972	57,698	1,022
Inpatient V_I^1	121,680	30,391	10,594

(c) Costs

	Cystoscopy	Hysteroscopy	Sterilisation
Pre-policy costs			
Outpatient c_O^0	159	197	269
Inpatient c_1^0	422	775	1,155
Post-policy costs			
Outpatient c_O^1	154	179	274
Inpatient c_I^1	453	733	1,181

Notes: The post-policy and pre-policy prices are based on the prices paid to the hospitals in 2012/13 (year of the BPT introduction) and 2011/12 (one year prior the introduction). The pre-policy volume is based on the 2011/12 volume of inpatient and outpatient attendances. The post-policy volume is based on the estimated effect of the policy, including the change in overall volume and in the proportion of patients treated in the outpatient setting.

For cystoscopy the revenues pre-policy and post-policy are respectively equal to:

$$\begin{aligned} R_{Pre} &= p_O^0 V_O^0 + p_I^0 V_I^0 \\ &= 257 * 45,010 + 714 * 276,487 \\ &= 208.98M \end{aligned}$$

$$\begin{aligned} R_{Post} &= p_O^1 V_O^1 + p_I^1 V_I^1 \\ &= 403 * 208,972 + 260 * 121,680 \\ &= 115.85M \end{aligned}$$

The results show a sharp reduction in revenues for cystoscopy, reducing from £208.98 million in the pre-policy period to £115.85 million in the post-policy period - a reduction of £93.12 million. This also implies that the payer had large savings following this policy, about 45 per cent of the original hospital reimbursement. Given that the outpatient tariff post policy is higher than the inpatient one, hospitals could have increased revenues by further increasing the number of patients treated in an outpatient setting. But even in the extreme case where 100 per cent of patients were treated in an outpatient setting, the revenues would be at most 133.25M, which would still imply a 36 per cent savings for the funder.

A qualitatively similar picture arises for hysteroscopy. The revenues pre policy were £38.95 million, and reduced to £34.27 million post policy, therefore dropping by £4.68 million, or 12 per cent. The effect on revenues is therefore less dramatic than for cystoscopy. Note that prices (pre-policy and post-policy) and outpatient volumes for hysteroscopy are comparable to those to cystoscopy, but volumes in an inpatient setting are much smaller pre policy. This explains why the reduction in revenues is much more contained, as the reduction in revenues is mostly driven by the lower tariff in the inpatient setting. Also, note

that in this scenario the hospital could have increased the revenues to £40.25 million under the extreme scenario where 100 per cent of patients were treated in the outpatient setting in the post-policy period (while we estimate only 65.5 per cent being treated in the outpatient setting in the post-policy period).

Differently from cystoscopy and hysteroscopy, for sterilisation both the inpatient and outpatient tariff increased post policy, though the outpatient one increased much more. Therefore, by construction hospital revenues increased, and went from 8.34 million in the pre-policy period to £10.99 million in the post-policy period, with an increase in revenues of £2.65 million or 32 per cent, and an equivalent increase in reimbursement for the payer. The number of patients treated in an outpatient setting increased only from 348 to 1,022, which represents 9.6 per cent of the post-policy volume. Moreover, the tariff in the outpatient setting was higher than than in the inpatient setting (again differently from the other two procedures), and therefore hospital revenues would have increased even more post policy if a higher proportion of patients were treated in an outpatient setting.

These results indicate substantial savings for the payer for cystoscopy and hysteroscopy, with a relatively smaller increase in reimbursement for sterilisation.

3.7.2 Profits

Even if the BPT reduced hospital revenues for cystoscopy and hysteroscopy, this does not necessarily imply that profits also reduced because the higher proportion of patients treated in an outpatient setting also reduced provider total cost of provision across the two settings. Hospital profit for the pre- and post-policy period is:

$$\pi_{Post} = \left(p_O^1 - c_O^1\right) V_O^1 + \left(p_I^1 - c_I^1\right) V_I^1 \tag{3.7}$$

$$\pi_{Pre} = \left(p_O^0 - c_O^0\right) V_O^0 + \left(p_I^0 - c_I^0\right) V_I^0 \tag{3.8}$$

The effect of the BPT on the change in profit is then as follows:

$$\Delta\pi = \pi_{Post} - \pi_{Pre} \tag{3.9}$$

We obtain the pre- and post- policy cost information, presented in table 3.9, from the yearly Reference Costs publication, which provides treatment information across hospitals and HRGs. Combining this with information on prices and volumes in the post-policy period from table 3.9, we can calculate the post policy profit for cystoscopy from equation (3.7).

For cystoscopy, the pre-policy profit was £85.15 million, which was reduced to £28.55 million in the post-policy period, resulting in a reduction in profit of £56.60 million. Therefore, the BPT reduced both revenues and profits. Although the positive price mark-up and volume of patients treated in the outpatient setting increased, the price mark-up in the inpatient setting went from positive to negative which more than offsets the first effect given the high volume in the inpatient setting.

The change in profit is instead positive for cystoscopy. Profit was £0.22 pre policy and increased to £1.88 post policy, an increase of £1.66 million. Therefore, in this case profits increased despite the reduction in revenues. In this scenario, the volume of inpatients is smaller before pre- and post policy, and therefore the reduction in profits driven by the reduction in the price mark-up in the inpatient setting is more than offset by the increase in the price mark-up and volume in the outpatient setting. Therefore, in this scenario both the funder and the provider were better off following the policy.

For sterilisation, hospitals were making a loss in both periods. In the pre-policy period the loss was -£4.76 million, decreasing to -£1.80 million in the post-policy period (the loss was smaller by £2.97 million). In this scenario, prices and price mark-ups increased in both settings, therefore increasing revenues and reduces losses.

3.8 Conclusion

Our study shows that a financial incentive can be successful in shifting patients from inpatient to outpatient setting, without pronounced negative consequences. Our results show positive effect for all three incentivised conditions (diagnostic hysteroscopy, diagnostic cystoscopy and sterilisation), with the biggest effect seen for the two diagnostic conditions (between 16 and 36 percentage points). This is not surprising, as the financial incentive is less pronounced for sterilisation. Namely, the prices for inpatient procedure did not reduce for this condition, hence eliminating the *penalty* element of the scheme.

The BPT manages to improve efficiency without reducing patient benefit. With hospitals and policy makers often having to balance the cost and quality dimensions of care, this is a particularly positive finding. Furthermore, hospitals did not respond to the scheme by increasing the overall volume of patients, which means the scheme did not motivate supplier induced demand. We further observe positive spill-over effect on the related conditions. This suggest the providers changed their working patterns in a way that accommodates further shift across settings. Despite the hospitals high take-up of the scheme, because of the way the bonus is structured hospitals made lower profit after the introduction of the policy. At the same time, the purchaser lowered their costs.

These findings have important policy implications in light of rapid growth in medical spending across most healthcare systems. While most existing literature shows limited effect of small financial incentives, our study shows that the effect is large when the scheme is targeted and the bonus very large.

CHAPTER 4

The effect of the DRG classification reform on coding intensity and healthcare expenditure: Evidence from England

4.1 Introduction

In order to improve efficiency and productivity of healthcare systems most OECD countries have implemented the Diagnosis Related Groups (DRG) classification system to standardise hospitals' reimbursement and encourage cost-containment (Davis and Rhodes, 1988; Busse et al., 2013b; Mihailovic et al., 2016). The system works by grouping patients into one of many DRG groups based on their clinical and demographic characteristics. Each DRG group then attracts a fixed payment, regardless of the patient's specific care pathway or the actual treatment cost.

The main assumption behind the DRG based system is resource homogeneity within the groups (Horn et al., 1986; Busse et al., 2013a). This means that the actual treatment cost should be similar across all patients who are grouped to the same DRG, as the price paid to hospitals typically corresponds to the average treatment cost within that group. Where there is large variation in the resource use within a DRG, the payment is either too high or too low for many patients, which can create financial hardship for hospitals with unfavourable casemix, while creating large profits for others (Bojke et al., 2016; Stephani et al., 2017).

For example, if all complicated (and hence more costly) hip replacement surgeries are performed by a single specialist hospital, this hospital is disadvantaged[50] compared to other hospitals in the case where a single price is applied to all hip replacements (complicated and non-complicated).

Many countries respond to this issue by refining the DRG classification to better account for differences in treatment costs across patients (Busse et al., 2013a). This is usually done by splitting a DRG into two or more new groups, based on the reported co-morbidities and complications. Increasing the number of groups improves homogeneity, as it ensures greater similarity of patients within a group. However, it might also create an incentive for hospitals to upcode. This includes increasing either the coding or treatment intensity (number and invasiveness of the procedures) to boost their revenue (Busse et al., 2013a; Cook and Averett, 2020), which can lead to an overall increase in healthcare expenditure. Despite the widespread use of the DRGs and frequent updates of the classification across countries, there remains a lack of evidence on the effects of these reforms on the intensive margin by increasing hospital's coding and treatment intensity (OECD, 2010).

The aim of this study is to fill this knowledge gap by investigating the effect of the DRG classification change on hospitals' behaviour, focusing on a major DRG reform which was implemented in England in April 2009. England uses it's own DRG version called Healthcare Resource Groups (HRGs), first used for payment purposes in 2003. Due to the perception that the existing classification did not capture well the treatment complexity of a sizeable proportion of patients (PA, 2008), the HRG system was completely reformed in 2009. This included a large expansion of the number of groups, increasing from 550 to 1,500 (Grasic et al., 2013), heightening the role of patient's recorded complications and co-morbidities.

Our analysis is based on the difference-in-difference approach using Wales as a control group. We compare changes in cost, coding and treatment intensity after the HRG revision

[50]In this case the hospital will incur higher per-patient cost (assuming that complicated surgery is more expensive), but receive the same payment as other hospitals.

in England, to changes in Wales. The two countries have similar demographics and share the same coding rules and HRG classification. However, unlike England, Wales utilises the HRGs for reporting purposes only. This means that a change in classification does not impact the reimbursement of Welsh hospitals, which are paid by a lump sum payment. Unlike in many other health care systems where DRGs are considered to be a black-box (Herwaarden et al., 2020), the rules behind the HRG construction are fully transparent. This allows us to retrospectively apply the new, post-reform HRG classification to the historic datasets from England as well as to the data from Wales. Hence we are able to construct a unique, coherent HRG data series across time and countries.

We focus on patients who were treated for respiratory conditions, as they are treated in most regional hospitals, rather than in specialised centres. Furthermore, unlike for some other medical areas, the revised HRG system allows for respiratory patients to be coded across different severity levels, based on the reported comorbidities. For example, there are separate HRG codes for pneumonia with and without comorbidities. This makes respiratory conditions susceptible to changes in coding behaviour. We first show that the reform increased the probability of being coded to a *severe* HRG - defined as an HRG that requires presence of additional complications and co-morbidities[51] - by 3.3 percentage points (pp). This is further reflected in the HRG price increase, with the average price effect of £58.5 or 4.3% of the total payment. These results suggest that the upcoding associated with the reform increased the overall healthcare expenditure for respiratory patients by approximately £57 million per year. Considering that respiratory patients represent a small proportion of all inpatient activity (around 5%), the overall cost to the system is likely considerably higher[52].

While the above results indicate a change in hospitals' behaviour, this could stem from an increase in coding of diseases, or from the change in the patients' treatment pathway. We

[51] An example of a more severe HRG is Pneumonia with complications and co-morbidities, while pneumonia without complications would be the non-severe HRG in this case.

[52] We can't directly apply our estimates to other patient groups as respiratory patients are relatively costly compared to some other disease areas, for example diagnostic procedures.

first show that the reform significantly increased the number of reported diagnosis codes by 0.56, with no effect on procedures. Next, we analyse the severity of the reported procedure and diagnosis codes, based on the expected resource use for each reported procedure and diagnosis code[53]. Our results suggest there was no increase in the severity for neither diagnosis nor procedures. This indicates that the change in HRG composition is mainly driven by more extensive coding of diseases, rather than changes in the treatment pathway.

We further explore heterogeneity in the response to the reform across hospitals. We find that hospitals with the lowest average rate of reported complications and co-morbidities in the pre-reform period respond more strongly, by increasing their coding at a faster pace compared to hospitals for which the coding/treatment intensity was already high before the reform. This indicates that hospitals might have been *catching-up* in coding to increase their marginal utility following the reform. There are two possible interpretation of these results: (i) either the reform encouraged hospitals to improve the coding quality, or (ii) some hospitals are fraudulently upcoding to increase their revenue. Considering the large reduction in apparent casemix differences across hospital after the reform, as well as trends in the number and severity of the reported diagnosis codes, it's likely the reform increased the coding quality. These results suggest that policy makers must balance an increase in quality of coding with associated increase in healthcare expenditure when considering a reform of the DRG classification system.

The rest of the paper is structured as follows. Section 4.2 provides a review of related literature. Section 4.3 provides a short theoretical motivation for DRG split and the implications of upcoding on the overall healthcare budget. Section 4.4 provides the institutional background on the English HRG system. Section 4.5 outlines the empirical methods. Section 4.6 describes the data and section 4.7 describes the results. Section 4.8 is devoted to discussion and conclusion.

[53]We measure procedure and diagnoses severity using the *Procedure and Diagnosis Hierarchy List*, which is published alongside the HRG classification and measures the expected resource use of each procedure.

4.2 Related literature

This paper contributes to our understanding of the financial incentives created by the DRG payment in publicly funded health system. Research on upcoding typically links changes in DRG prices to changes in activity levels. The influential paper by Dafny (2005) exploits the 1988 policy change in Medicare in USA, which increased the price[54] of approximately 43% of all groups. The author utilises the structure of the DRGs, with most coming in pairs (with and without complications/co-morbidities) and compares the share of patients with/without co-morbidities before and after the reform. Her results provide evidence of upcoding towards the higher priced (weighted) DRGs, with the argument that there is nothing to suggest that the patients were becoming sicker over the time frame or that the coding was more accurate.

Several papers adopt Dafny's approach to estimate the effect of price changes on upcoding. Barros and Braun (2017) use politically driven change in DRG prices in Portugal in 2006 as an exogenous source of variation to estimate changes in the distribution of patients across DRGs with and without co-morbidities. While their results indicate presence of upcoding, the effects are quantitatively small. Similarly, Verzulli et al. (2016) studied the effect of a 1-year increase in DRG prices in Emilia-Romagna region of Italy on their volume. They find an increase in volume of the affected surgical DRGs (procedure based), but there is no effect for the medical DRGs (diagnosis based). These estimates differ from the results from the study by Januleviciute et al. (2016), that exploits the variations in prices created by the changes in the national average treatment cost in Norway. In this case the authors find that 10% increase in price leads to about 0.8-1.3% increase in the number of patients treated for medical DRGs, while they find no effect for surgical DRGs.

Using somewhat different approach, Jürges and Köberlein (2015) study upcoding in German neonatal departments, following the introduction of the DRG system in 2003. In

[54] In this case the price is characterised by the DRG-weight which is the relative price compared to the baseline. Weight of 1 typically corresponds to the average cost.

this setting the DRG weights are higher for the care of neonates who are just below a threshold birth-weight. Authors use this variation to estimate the effect of the DRGs on the reported birth-wight, focusing on neonates who are just below and just above the threshold. They find a substantial upcoding of birth-weight, particularly for infants with higher expected treatment costs. Similar results were found in subsequent studies on upcoding birth-weight in neonatal departments (Hochuli, 2020; Hennig-Schmidt et al., 2018; Groß et al., 2021).

Recently, more papers focused directly on studying the implication of classification change on change in coding behaviour. Cook and Averett (2020) explore the effect of the 2008 MS-DRG classification change in the US on upcoding, measured as an increase in the DRG cost-weight. Authors use a combination of methods, including difference-in-difference regression framework. In this case they compare the change in cost weights between the DRGs that changed to those that remained the same pre- and -post reform. Authors find evidence of upcoding in government, non-profit, and for-profit hospitals, with their most conservative results suggesting the classification change increased the cost-weight by 3 percent.

Similarly, Milcent (2020) explores the introduction of DRG severity levels in France in 2009, resulting in an increase in the number of groups from 800 to 2,200. Using interrupted time series design, the author found a decrease of 2.1% in the probability of being recorded as a non-severe patient. The estimated decrease was smaller for the public compared to the for-profit hospitals.

This paper extends the existing literature in several ways. First, our study is the first to our knowledge that compares the changes in coding across two very similar health care systems, with the reform affecting only one of them. Most other research focuses on comparing changes across DRGs within the same setting, which might suffer from spill-over effect. In particular, hospitals might increase coding across all medical conditions, including the ones that are not part of any reform. Second, we exploit the fact that the DRG

grouping in UK is transparent, allowing us to re-group the data to provide a consistent series of *new* (post-reform) DRG codes over time in both settings. This means that we are able to obtain the DRG codes for patients in the pre-policy period had the patient been admitted after the change in classification. This gives us unique opportunity to detect changes in coding and the related hospital reimbursement. Third, we extend the existing literature by exploring different mechanisms of upcoding, including changes in the volume and intensity of coding of diagnosis codes and procedure codes. The latter provides us with insight on whether the increase in HRG severity stems from the change in treatment (changes in the recorded procedure codes) or from increase of coding the diagnosis codes. Finally, we explore heterogeneity in response across hospitals, based on their coding patterns prior to the classification reform. This gives us an opportunity to study response to the reform across hospitals and identify the effect of the reform on the apparent casemix differences across hospitals.

4.3 Motivating framework

4.3.1 The rationale for finer categorisation of DRGs

The patient classification system assumes that patients in a particular DRG have certain diagnostic characteristics $X_1, \ldots, X_N$. It further assumes that the average cost of treatment is predictable and is the same across providers after taking into account any differences in efficiency. However, some patients grouped to this DRG might have additional co-morbidities and complications which do not affect the assigned DRG, although they might increase the cost of care.

Suppose the average cost of treating a patient who does not have any additional co-morbidities is c_0, that co-morbidities and complications increase the cost to c_1. We assume there are $N = n + m$ patients treated across all hospitals where m patients have additional

co-morbidities. The resulting per-patient cost for providing care is therefore:

$$C = \frac{mc_1 + nc_0}{m+n}. \tag{4.1}$$

In many DRG systems the price P paid to hospitals equals to the average cost of treatment, so that $P = C$.

For a specific hospital h with m_h patient with complications and n_h patients without, this price may be either greater or smaller than its average cost depending on whether $\frac{m_h}{n_h}$ is greater or smaller than $\frac{m}{n}$ The price equals the hospital's cost only if the proportion of patients with complications in the hospital is the same as the national proportion.

This can have severe financial consequences for hospitals with a high concentration of patients with complications and co-morbidities, as their cost is higher than the reimbursement they receive. On the other hand, hospitals with less severe casemix might profit. To prevent inequality in reimbursement, many health care systems split the DRGs to better account for varying levels of resource use across patients.

4.3.2 The risk of finer coding for expenditure

If the classification system is changed to take into account complications it is possible to specify two prices (P_1 and P_0) which reflect the cost of treating the different kinds of patients – with and without complications respectively. As before, we have $n + m$ patients, out of which m are with complication and n without. We assume that the price of the DRG equals to the cost of treatment, so that $c_0 = P_0$ and $c_1 = P_1$. The full healthcare expenditure (both the actual and the one paid to providers) E_1 for this treatment is hence:

$$E_1 = mP_1 + nP_0 \tag{4.2}$$

In this case the overall expenditure is the same as in previous case when we did not have the DRG split. However, if hospitals increase their coding, so that the number of patients

grouped to the more severe HRG increases by k. The total expenditure E_2 now increases:

$$E_2 = (m + k)P_1 + (n - k)P_0 = E_1 + k(P_1 - P_0) \tag{4.3}$$

With k patients additionally grouped to the more severe HRG, the total expenditure increases by $k(P_1 - P_0)$. Unlike in previous case with a single HRG, hospitals now have incentive to increase the coding as it directly impact their profit (increasing the intensity of coding would not increase the profit in previous case).

The above suggests that it is important - from the perspective of oversight of the health-care system - to measure the sensitivity of reported case-mix coding to changes in the classification system. This constitutes a key motivation for our empirical study.

4.4 Institutional setting

Healthcare in the UK is primarily funded through general taxation and is free at the point of use. Patients are registered with a general practitioner (GP), who acts as a gatekeeper and provides referrals to access the secondary care (no referral is needed for emergency care). While access to medical care is similar regardless of the geographical location, the UK does not have a common healthcare system. Instead, each of the four constituent countries of the UK (England, Scotland, Wales, and Northern Ireland) is responsible for their own National Healthcare Service (NHS).

In this paper we focus on the comparison of medical coding between England and Wales. There are several reasons why Wales is a suitable control group. England and Wales operated a common NHS from 1948 (founding year of the NHS) until the political devolution in 1999. Under this arrangement, all polices and resources were shared across the two countries. While these have begun to diverge since then, the organisation of the health services still remains broadly comparable to this date, including similar access to care and

comparable healthcare expenditures per capita[55]. The two countries continue to share the same professional regulation (e.g. on clinical training, conduct, and fitness to practice) and have comparable pay structure in place for their doctors and nurses (OECD, 2016). Population health profile is also similar across England and Wales (ONS, 2013). Furthermore, both countries have adopted the same classification and reporting of diseases (using International Classification of Diseases (ICD) codes) as well as procedures (using Office of Population Censuses and Surveys (OPCS) codes).

One important aspect in which the English and Welsh NHS differ is in their approach to reimbursement of hospital activity. Since 2003 English hospitals are reimbursed via prospective payment system. Initially limited to selected condition, it now covers majority of inpatient and outpatient hospital activity (Grasic et al., 2013). Patients are categorised into distinct Healthcare Resource Groups (HRGs) according to their age, sex, co-morbidities and the medical care received. The latter two are recorded using the ICD and OPCS codes. Payment is then based on the allocated group, regardless of any additional medical services provided. Conversely, Welsh hospitals are paid via a capitation system. In this case each hospital receives a lump sum, which is linked to the size and demographic characteristics of the local population they serve, not to the actual volume or type of service provided. Wales still uses HRGs (the same version as England), however, their use is limited to reporting purposes only. Hence, the coding of co-morbidities is of lesser importance in Wales as it does not directly impact reimbursement.

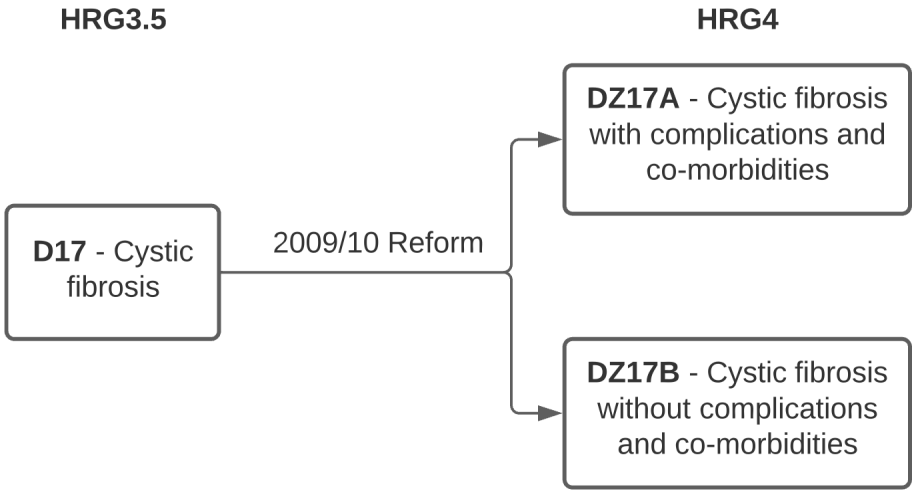
4.4.1 2009 HRG classification reform

HRGs were introduced in 1991 to facilitate reporting and benchmarking. They were first used for payment purposes in 2003/4. At start, this only covered a small number of selected procedures, however, by 2006/7 most hospital admissions and attendances were reimbursed

[55] In 2014/15, the English NHS spent £2,055 per capita on healthcare, whereas the Welsh NHS spent £2,083 per capita (Harker, 2014).

via the HRGs (Grasic et al., 2013). Initially, hospitals were reimbursed using the HRG3.5[56] classification system, consisting of 550 unique HRGs. However, there was perception that the system did not capture well the treatment complexity of some patients, resulting in a loss of revenue for many hospital providers (PA, 2008). This led to an overhaul of the classification system and the introduction of a new version (HRG4) in 2009/10, which better describes the care that a patient receives in terms of treatment and resources. This study explores the effects of this change on hospitals' coding behaviours.

Figure 4.1: Classification reform



Notes: The image displays an example of a classification change between the HRG3.5 (pre 2009/10) and HRG4 (2009/10 and after) classification systems for Cystic Fibrosis. There was a single code for Cystic Fibrosis under the HRG3.5 system (D17); under the HRG4 classification system there are two codes, one for the base case (DZ17B) and one that includes complications and co-morbidities (DZ17A).

The HRG4 design represented a major development from HRG3.5 with the number of groups increasing from 550 to 1,400. In particular, the number of groups for respiratory illnesses increased from 35 in HRG3.5 system to 95 in HRG4 system. While part of the increase is due to expansion of the HRGs to new treatment settings (for example outpatient setting), a substantial part of the rise in the number of groups is driven by the incorporation of complications and co-morbidities into the classification logic. As can be seen in Figure 4.2, HRGs split into 2 or more new groups, better reflecting patient's complexity.

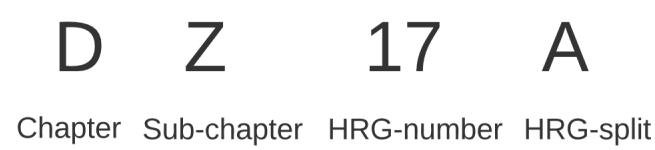
[56]While there were earlier versions of the HRG system, those were not used for payment purposes.

Patients are grouped into the HRG groups based on the reported procedure and diagnosis codes; hence only those patients with recorded relevant complications and co-morbidities (required codes are different for each HRG) are grouped into the *severe* HRG.

4.4.2 HRG4 Structure

The HRGs are classed into 21 chapters that indicate the clinical area in which the patient is treated (see Table C1 for the description of each chapter). This is symbolised by the first letter of the HRG group, for example, the letter *D* stands for *Respiratory System*. Chapters can be further grouped into sub-chapters, giving more information about the area of treatment[57]. First for characters of the HRG comprise the so called *core* HRG, which corresponds to the general treatment area (for example pneumonia). The last character of the code is a letter, called a *split*, that represents the complexity of the case and indicates any co-morbidities present. The higher up the alphabet, the more complicated complicated the HRG; the letter *Z* stands for no-split present. Figure 4.2 shows an example of an HRG for Cystic Fibrosis with complications and co-morbidities. The main feature of the HRG classification reform is the addition of the splits - while previously all patients would be grouped to the same core HRG, the addition of the split heightens the role of co-morbidities and differentiates across patient severity.

Figure 4.2: Example of an HRG4



Notes: The image shows an example of an HRG4 group for Cystic Fibrosis with complications and comorbidities; first four characters represent the core HRG (DZ17) while the last character (A) defined the severity of the HRG.

[57]For example, AA and AB are both sub-chapters within the Nervous System chapter, with *AA* standing for *Nervous System Procedures and Disorders* and *AB* standing for *Pain Management*. Not all chapters have multiple sub-chapters; for example, all HRGs within chapter D are grouped to the DZ sub-chapter.

The grouping logic is published on a yearly basis in the *Code-to-Group* workbook, which details the mechanism behind coding and makes it possible to analyse which diagnosis and procedure groups need to be listed for a patient to be grouped to a particular HRG. Classification logic and documentation are in public domain in several countries, such as Australia and England, while some other countries treat is as a *black box* (van Herwaarden, 2018). Access to the classification logic and documentation enables us to know explicitly which co-morbidities are required to group a patient to a higher value HRG split.

4.5 Empirical strategy

We employ a difference-in-difference (DiD) modelling approach, comparing changes in coding and treatment intensity between patients treated in hospitals in England (treatment group) and those treated in Wales (control group), where the classification reform did not affect hospital reimbursement.

We first focus on understanding the effect of the reform on the severity of the HRG groups and the corresponding effect in the average prices paid to the hospital[58]. Our main outcomes of interest are (i) the probability of being coded a *severe* HRG that includes complications and co-morbidities and (ii) the average price of the treatment paid to hospitals, corresponding to the increase in health expenditure for the payer. The latter takes into account the fact that the prices for a particular treatment differ depending on whether the coded HRG is with or without complications and co-morbidities; as the proportion of patients coded to HRG with CC increases, so does the average cost of this particular treatment.

For the analysis we exploit the fact that the English HRG grouping system is transparent, meaning we directly observe the classification logic and the underlying rules. This allows us to *go back in time* and construct the HRG codes using the new revised classification for

[58]Once the proportion of HRGs with/without complications and co-morbidities (CC) changes, the average price for the treatment changes accordingly as the HRG with CC typically attracts a higher price

all the years and for both countries. We model the change in outcomes using the following regression framework:

$$Y_{iht} = \alpha + \theta(England_i\ D_t) + \boldsymbol{X_i'\delta_1} + \boldsymbol{H_i'\delta_2} + \boldsymbol{v_s} + \boldsymbol{v_t} + \boldsymbol{v_h} + \epsilon_{iht} \tag{4.4}$$

Here Y_{iht} is the outcome of patient i in hospital h in month t. For the probability of being grouped to severe HRG the Y_{iht} is a binary variable taking value 1 if the patient is grouped to an HRG that required presence of additional complications and co-morbidities and 0 for the base case. For the treatment cost the outcome variable is the national tariff corresponding to the associated HRG. D_t is a dummy variable equal to 0 in the pre-reform period (April 2006 - March 2009) and 1 in the post-reform period (from April 2009-March 2013). $England_i$ is a dummy variable equal to 1 for patients treated in England (treatment group) and 0 for patients treated in Wales (control group). $\boldsymbol{X_i}$ is a vector of patient characteristics, further described in the Data section. The coefficient vectors $\boldsymbol{v_s}$ and $\boldsymbol{v_t}$ are calendar month (February to December, with January as the reference category) and financial year (2007/8-2012/13 with 2006/7 as reference category) dummies, respectively. $\boldsymbol{v_h}$ are hospital fixed effects, α is the intercept and ϵ_{iht} is the error term. We further include a vector of core HRG dummies[59], $\boldsymbol{H_i}$, indicating the main part of the HRG before splitting for possible complications and co-morbidities. Including these dummies allows us to isolate the effect of the HRG split on the outcomes in question. We estimate a linear model with standard errors clustered at hospital level. The key coefficient of interest is θ, which measures the average treatment effect on the treated (ATT) patient population over the post-policy period.

We validate the analytical approach of our main analysis by testing whether the pre-intervention trends are parallel. The assumptions underpinning difference-in-difference analysis require that in the absence of treatment, the difference between the treatment and control group is constant over time. We test this assumption in two ways: by visual in-

[59]Core HRG refers to the main part of the HRG before splitting for possible complications and co-morbidities. Each treatment is associated with very distinct medical intervention and we would not expect them to serve as substitutes.

spection and by empirical analysis. For the latter, we limit the data to the pre-policy period (2006/7-2008/9) and estimate the following model:

$$Y_{iht} = \alpha + +(England_i \, \boldsymbol{v_t}')\boldsymbol{\theta} + \boldsymbol{X_i'\delta_1} + \boldsymbol{H_i'\delta_2} + \boldsymbol{v_s} + \boldsymbol{v_t} + \boldsymbol{v_h} + \epsilon_{iht} \qquad (4.5)$$

The vector of coefficients $\boldsymbol{\theta}$ captures the difference in the pre-intervention trends between the treatment and control group. The null hypothesis for the parallel trends assumption is that the $\boldsymbol{\theta}$ coefficients are jointly zero.

We perform additional sensitivity analyses/robustness checks. First, we run the regression excluding 6 months on either side of the time of the reform. This adjust for any anticipatory effect of the reform, as well as any post-reform adjustment period. Second, we perform the analysis on a subset of hospitals that report observations in every year of the study period. This removes from the sample any hospitals that merged during the study period, which might affect their reporting processes. Our third sensitivity analysis concerns the selection of the treatment time point. In the main analysis, we base the treatment time on the admission date, as this is typically better coded that the discharge date[60]. However, some patients might have been admitted before the reform, but discharged after the implementation of the reform. To investigate the robustness of our results we re-estimate the models setting the discharge date as the treatment date.

4.5.1 Mechanisms

To better understand the mechanisms underpinning the effect of the reform on the main outcomes, we carry out two additional analyses. First, we analyse the effect of the reform on coding intensity, using the number of recorded procedure/diagnosis codes as the outcome variable. An increase in diagnosis codes might suggest a change in coding behaviour, while an increase in procedure codes could reflect either increased coding or additional

[60] In case patient receives further rehabilitation after the main treatment there is no discharge date recorded

treatments.

Second, to disentangle the increase in coding from the increase in treatment intensity, we study the underlying *severity* of the reported procedure and diagnosis codes. The *Diagnosis and Procedure Hierarchy List*, developed by NHS England, provides us with a severity score of each reported procedure or diagnosis code, based on the expected resource use. For the analysis, we use this list to obtain the severity score for each reported diagnosis and procedure code and calculate the mean and maximum score across all reported diagnosis/procedure codes for each observation. This allows us to observe whether any changes in coding after the reform was due to increased reporting of low-value co-morbidities (low-value diagnosis codes), upcoding of diseases (high-value diagnosis codes) or increase in the treatment intensity (high-value procedure codes). We estimate the effect of the reform on the severity using the regression framework presented in equation 4.4, with the dependent variable measuring the maximum/mean reported severity score for each observation (measured across all of the reported codes for each observations). Typically, maximum severity score drives the HRG, and is therefore better indicator of upcoding. However, the mean severity score gives us a better overview of the coding direction (volume vs severity). We estimate this effect separately for diagnoses and procedures.

4.5.2 Heterogeneous treatment effects across hospitals

In the second part of our empirical analysis we estimate the heterogeneous treatment effects of the classification reform across hospitals. In particular, we are interested in whether behavioural change was larger for hospitals with lower coding/treatment intensity at the start of our study period. The hypothesis is that these hospitals might need to *catch-up* in coding once the HRG revision increased the marginal utility of additional coding.

To estimate heterogeneous treatment effects we first construct three distinct subgroups of English hospital (terciles), based on the proportion of patients coded to severe HRG at the start of our series in 2006/7. English hospitals with the lowest scores in 2006/7 are

grouped to the first tercile, with the highest scoring hospital comprising the third tercile. Welsh hospitals serve as a control group. We are interested in three outcomes: price od the HRG, probability of being coded to a severe HRG and the number of reported diagnosis codes. The regression has the following form:

$$Y_{iht} = \alpha + (D_t \, \boldsymbol{Q_i}')\boldsymbol{\theta} + \boldsymbol{X_i'\delta_1} + \boldsymbol{H_i'\delta_2} + \boldsymbol{v_s} + \boldsymbol{v_t} + \boldsymbol{v_h} + \epsilon_{iht} \tag{4.6}$$

$\boldsymbol{Q_i}$ is the vector of tercile dummies of English hospital. Analogous to regression equation (4.4), Y_{iht} is the outcome (probability of severe HRG, treatment price, number of diagnoses) of patient i in hospital h in month t. All other terms are defined as in equation (4.4). Vector of coefficient $\boldsymbol{\theta}$ captures the differences in the effect of the reform across the three hospital terciles.

4.6 Data

Our analysis employs two main data sources. For information on patients treated in England, we use patient-level Hospital Episode Statistics (HES) inpatient dataset. This provides comprehensive information on patient's hospital care, including their socio-demographic and clinical details. Information on patients treated in Wales comes from the Welsh Admitted Patient Care (APC) dataset. Like HES, this is a patient-level source, with similar layout and variables coverage. Importantly, both datasets utilize the same coding system to report clinical information: ICD-10[61] classification for diagnosis codes and OPCS[62] classification for procedure codes[63].

The study period runs from April 2006 to March 2012; the first three years (up to March 2009) comprise the pre-policy period, with the post-policy period running from April 2009 to March 2012. We focus on patients who were treated for respiratory conditions, includ-

[61] International Classification of Diseases codes.
[62] OPCS Classification of Interventions and Procedures.
[63] Both systems also use the same ICD/OPCS version.

ing (among others) pneumonia, asthma and COPD. We identify our sample by including all patients who were grouped to an HRG beginning with D (Respiratory/Thoraric illnesses chapter). This disease area was chosen for several reasons: (i) HRGs codes in this chapter do not undergo further changes during our study period; (ii) respiratory/thoracic illnesses are fairly common and are treated in most hospitals, rather than in specialised centres; (iii) England and Wales are both members of the British Thoraric Society that issues guidelines for treatment of respiratory illnesses and thus follow the same standard treatment protocols. By concentrating on a specific subgroup we ensure that the treatment and control groups are comparable and can better identify the mechanisms driving changes in coding of the activity. Our sample comprises of all finished consultant episodes (observations) for patients aged 19 or over. We exclude Welsh patients treated in England and English patient treated in Wales from the sample, as hospitals are reimbursed separately for the cross border patients. Our full sample includes 6.7 million observations, of which 6.3 million are from England and 0.4 million from Wales.

To assign consistent HRG codes across all years and countries, we use special software called the *Grouper*, which is freely available to download from the NHS Digital website and is updated on a yearly basis. We use the 2009/10 grouper version on both, the English and the Welsh datasets. Furthermore, to uncover the severity of the coded diagnosis and procedure codes, we use the Procedure and Diagnosis Hierarchy lists. Published by NHS Digital, the two lists provide for each diagnosis (ICD) and procedure (OPCS) code a value on a scale from 1 to 40 based on their expected resource use. A higher value represents a more costly (severe) diagnosis/procedure. Hence this data enables us to track changes in coding severity over time.

We attach a price to each observation, using the 2009/10 English National Tariff Data. This corresponds to the price paid for the medical care to English hospitals in 2009/10. Price differs by HRG and by whether the treatment is elective or an emergency. Using the same price across both countries and over time allows us to detect the effect of the

classification reform on the price of treatment, independent of the price inflation.

4.6.1 Outcome measures

We construct several dependent variables to estimate the effect of classification change on hospitals' coding behaviour. Our main outcome of interest is the probability of being coded to an HRG with complications and co-morbidities. We exploit the fact that most HRGs come in pairs, with patients without relevant co-morbidities grouped to the base HRG and patients with co-morbidities grouped to the *severe* HRG. Our variable takes the value 0 for the base case, with severe HRGs[64] coded as 1.

The assigned HRG directly affects the price paid to the hospital for the medical treatment. To quantify this relationship, we estimate the effect of the reform on the average price of a treatment, using the national tariff[65] for each HRG as the outcome variable. This is analogous to the DRG-weight[66] used in related studies (Dafny, 2005; Barros and Braun, 2017).

We construct a comprehensive set of additional outcome measures to better understand the mechanisms by which the classification change affects the coding. This includes the number of recorded diagnosis/procedure codes, which ranges from 0-12 for procedures and 0-14 for diagnoses. We further study whether the classification change resulted in an increase of *severity* (expected resource use) of the recorded medical codes. This is done by attaching the relevant hierarchy codes, that serve as a resource use proxy (described in section 4.6). We are interested in both, the maximum recorded value for each patient (maximum severity across all diagnosis codes recorded for a particular patient), as well as

[64] While most HRGs come in pairs, in some cases there are further splits of the relevant complications and co-morbidities. We code as 1 all of the HRGs that require recording of additional co-morbidites.

[65] Providers need to report average cost of treating patients by HRG in a yearly cost collection exercise. The capture the hospital level average cost of treating patients in each HRG. These so called Reference Costs are then used to calculate National Tariff, upon which the providers are paid. There is a three year gap between Reference costs submission and the National Tariff.

[66] In our context, the DRG-weight corresponds to the standardised price, with the mean price having the weight 0.

the mean value across all of the codes recorded for a particular patient.

4.6.2 Control variables

In our models we control for patient's characteristics that are independent from the reported clinical information (ie. reported diagnosis and procedure codes). Socio-demographic control variables include patient's age (coded as a categorical variable in 5 year bands with separate categories for 18 to 24 and 90+), sex (male=1) and socio-economic status (coded in quintiles). For the latter we use the Welsh Index of Multiple Deprivation (for patients treated in Wales) and the English Index of Multiple Deprivation (for patients treated in England). Deprivation index definition is the same in both cases and measures deprivation across multiple domains (such as education, income and environment) in the patient's immediate local area.

4.6.3 Descriptive statistics

Table 4.1 presents descriptive statistics for patient characteristics for the English and Welsh sample. Patients are comparable across the two countries by age (average age 68.8 in England vs 69.9 in Wales) and sex (proportion of male is 0.5 in England vs 0.49 in Wales). Deprivation among patients is somewhat higher in Wales compared to England. (21% in most deprived group in England vs 26% in Wales). Proportion of emergency admissions is slightly higher in England compared to Wales (89% vs 86% emergencies).

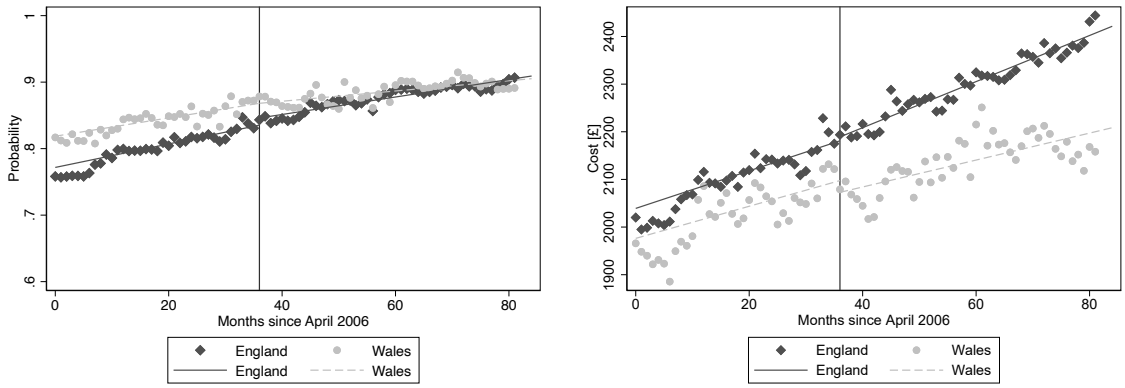
Table 4.2 presents the mean and standard deviation for the main outcomes in the pre and post reform period across both countries. In the pre-reform period Wales has slightly higher proportion of cases coded to severe HRG (0.84 vs 0.80), higher number of recorded diagnosis codes (4.87 vs 4.36) and higher maximum diagnosis severity (13.14 vs 12.90). On the other hand, the average treatment price in the pre-policy period is higher in England (£2,102 vs £2,024), likely related to higher number of recorded procedures with higher severity. The gap between England and Wales is smaller in the post-policy period across

Table 4.1: Patient's characteristics (mean and standard deviation) for England and Wales

	England		Wales	
	Mean	SD	Mean	SD
Age	68.80	17.13	69.86	15.90
Male [proportion]	0.50	0.50	0.49	0.50
Deprivation				
Most deprived	0.21	0.42	0.26	0.44
2nd quintile	0.20	0.41	0.23	0.42
3rd quintile	0.20	0.40	0.19	0.39
4th quintile	0.20	0.39	0.16	0.36
Least deprived	0.20	0.38	0.14	0.34
Total Observations	6,300,335		416,412	
proportion emergency	0.89	0.35	0.87	0.35

Notes: This table presents the mean and standard deviation [SD] for the patient characteristics used as explanatory variables in the regression models. The estimates include average age [measured in years], proportion of male patients [using binary sex variable] and the proportion of patients in each of 5 deprivation quintiles, based on the English and Welsh indices of multiple deprivation. We further report the number of observations (patients) in the sample and the proportion of the observations that were admitted as emergency. Estimates are measured across the entire study period (2006/7-2012/13).

Figure 4.3: Probability of being coded to a severe HRG and treatment cost over time



(a) Probability of severe HRG

(b) Treatment price

Notes: Probability of being coded to a severe HRG (a) and treatment cost (b) for respiratory/thoracic patients in England and Wales over time from April 2006 to March 2013. The vertical line represents the time of the reform of the HRG classification in England in 2009.

most outcomes, with the exception of treatment cost for which we observe higher increase in England compared to Wales.

Figure (4.3a) presents the number of patients grouped to severe HRGs over time. We

Table 4.2: Mean and standard deviation (SD) for the main outcomes in the pre and post reform periods across England and Wales

	England			
	Pre-reform		Post-reform	
	Mean	SD	Mean	SD
Proportion coded to severe HRG	0.80	0.40	0.88	0.34
Average treatment price [in £]	2,101.60	1,098.17	2,306.84	1,107.92
Number of coded diagnosis codes	4.36	2.61	5.90	3.18
Number of coded procedure codes	1.37	1.10	1.56	1.45
Max diagnosis severity	12.90	1.72	13.15	1.64
Max procedure severity	9.12	6.81	8.52	7.00
Mean diagnosis severity	7.86	4.47	7.57	4.22
Mean procedure severity	1.34	3.62	1.27	3.45
Observations	2,384,032		3,916,303	
	Wales			
	Pre-reform		Post-reform	
	Mean	SD	Mean	SD
Proportion coded to severe HRG	0.84	0.37	0.89	0.32
Cost	2,023.94	1,093.64	2,135.47	1,147.45
Number of reported diagnosis codes	4.87	2.64	5.71	3.09
Number of reported procedure codes	0.58	1.34	0.69	1.60
Max diagnosis severity	13.14	1.44	13.23	1.41
Max procedure severity	8.48	6.16	7.73	6.41
Mean diagnosis severity	9.32	2.43	9.23	2.27
Mean procedure severity	1.02	2.89	0.79	2.61
Observations	173,958		242,454	

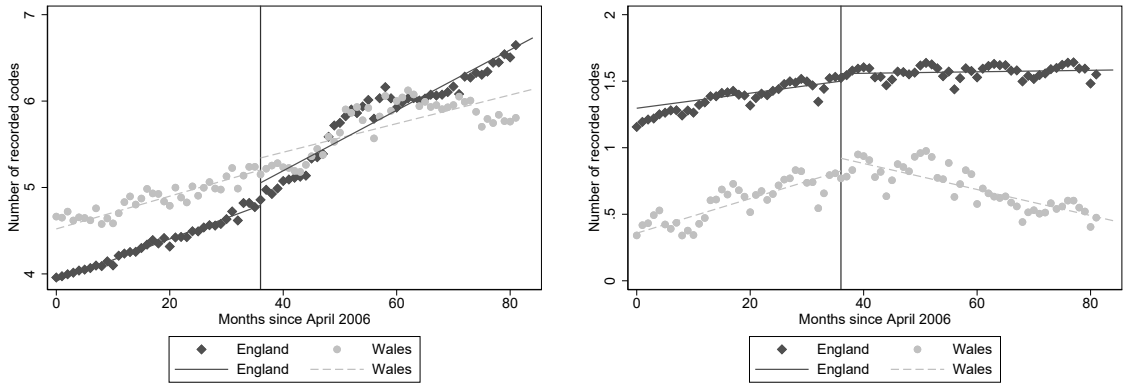
Notes: This table presents the mean and standard deviation [SD] for the main outcome measures: proportion of patients coded to severe HRG, average treatment price (including both severe and non-severe cases), measured in £, number of recorded diagnosis and procedure codes and the maximum and mean severity of the reported diagnosis and procedure codes, based on the expected resource use. Estimates are reported separately for the pre-reform period (2006/7-2008/9) and post-reform period (2009/10-2012/13) across both the treatment (England) and control group (Wales).

can can observe similar trends across England in Wales in the pre-reform period, with the difference among the two countries shrinking over time in the post-reform period. Similarly, figure (4.3b) presents the price of treatment over time across the two countries. We see that treatment price is very similar in the pre-policy period across both England and Wales, while the trend is slightly steeper for England in the post-policy period.

In Figure (4.4a) we can observe the number of recorded diagnosis codes over time in England in Wales. While initially Wales recorded more diagnoses per observation, English

hospitals rapidly increased coding after the reform. The pattern for the number of recorded procedure codes, seen in Figure (4.4b), is more stable over time.

Figure 4.4: Number of recorded diagnosis and procedure codes over time



(a) Number of recorded diagnosis codes

(b) Number of recorded procedure codes

Notes: Number of recorded diagnosis (a) and procedure codes (a) for England and Wales over time from April 2006 to March 2013. The vertical line represents the time of the reform of the HRG classification in England in 2009.

4.7 Results

4.7.1 Effect of the reform on the main outcomes

Our results, presented in table 4.3, suggest that the classification reform increased the probability of being coded to a severe HRG by 3.3 percentage points (pp). The average treatment price increased by £58.54 pounds. Using logarithm of the cost as dependant variable, our estimates suggest the treatment price increased by 4.30 percent. Estimates are statistically significant to five percent level. These results provide evidence of upcoding following the classification reform, with the size of the effect in line with existing literature (for example, the most conservative results by (Cook and Averett, 2020) suggest the cost increased by around 3 percent following a DRG reform in the US). Combining the estimates on the effect of reform on treatment price with the number of patients treated in each post reform year, we estimate that the total cost of the reform amount to around £57 million per year for patients with respiratory disease.

As can be seen in table 4.3, the probability of being coded to a severe HRG increases with age and is 48.7 pp higher for patients aged 90+ compared to the youngest group in the sample (18-24 year old). Accordingly, the cost of care is also highest for the oldest group (£727 higher compared to the youngest group). There is small difference in probability of severe HRG/cost between male and female (0.4 pp and £10.37, respectively). Compared to the most deprived, the least deprived are 3.2 pp less likely to have recorded complications and co-morbidities, with their care £62.6 cheaper. Cost of care is considerably higher for emergency patients (by £932). This is expected, as the there is a separate tariff for elective and emergency activity for the same HRG. Emergency patients are also more likely to be grouped to a severe HRG (by 17.3 pp).

Validity of the difference-in-difference estimator relies on the pre-reform trends being parallel across both countries. While we first confirm this visually (see Figures 4.3 and 4.4), the graphs reveal strong seasonal patterns and are not adjusted for casemix. We therefore

Table 4.3: Main regression results: probability of being grouped to a severe HRG, treatment cost and log of treatment cost

	Probability - severe HRG (1)		Treatment price (2)		Log price (3)	
	Estimate	SE	Estimate	SE	Estimate	SE
Post reform X England	0.033***	0.006	58.535*	23.809	0.043*	0.017
Age						
(reference=18-24)						
25-29	0.034***	0.003	27.236***	4.669	0.019***	0.003
30-34	0.077***	0.004	69.843***	5.054	0.050***	0.003
35-39	0.136***	0.004	132.260***	4.962	0.095***	0.004
40-44	0.202***	0.004	205.333***	5.008	0.144***	0.004
45-49	0.263***	0.005	281.345***	5.474	0.196***	0.004
50-54	0.319***	0.005	356.697***	5.89	0.244***	0.005
55-59	0.362***	0.005	432.399***	6.551	0.293***	0.005
60-64	0.393***	0.005	490.841***	7.383	0.329***	0.006
65-69	0.421***	0.005	542.823***	8.085	0.360***	0.006
70-74	0.444***	0.005	590.666***	8.221	0.388***	0.006
75-79	0.464***	0.005	640.721***	8.125	0.416***	0.006
80-84	0.478***	0.006	691.607***	7.918	0.445***	0.006
85-89	0.487***	0.006	726.709***	8.08	0.464***	0.006
90+	0.487***	0.005	726.779***	8.331	0.465***	0.006
Sex						
(reference=female)						
Male	0.004***	0.001	10.373***	1.261	0.004***	0.001
Deprivation						
(reference=most deprived)						
2nd quintile	-0.008***	0.001	-16.707***	1.793	-0.008***	0.001
3rd quintile	-0.017***	0.001	-31.324***	2.234	-0.016***	0.001
4th quintile	-0.024***	0.001	-46.747***	2.502	-0.024***	0.001
Least deprived	-0.032***	0.001	-62.552***	2.868	-0.032***	0.002
Emergency						
(reference=elective)						
Emergency	0.173***	0.011	932.269***	30.617	0.541***	0.022
Constant	0.300***	0.017	5959.362***	39.328	8.370***	0.011
Observations	6,640,515		6,640,515		6,640,515	
Number of hospitals	257		257		257	
Hospital Fixed Effects	✓		✓		✓	
Season dummies	✓		✓		✓	
Year dummies	✓		✓		✓	
Core HRG dummies	✓		✓		✓	

Notes: This table presents estimated effects on the classification reform on the probability of grouping a patient to severe HRG (1), average treatment price in £ (2) and log of treatment price (3). The table presents the main effects as well as the relevant estimates for the patients characteristics. All models include season dummies (calendar months), hospital fixed effects, year dummies (2006/7-2012/13) as well as dummies representing the core HRG before the split into severe and non-severe. Standard errors are clustered by hospital.

$*$ p$<$0.05 $**$ p$<$0.01 $***$ p$<$0.001

further test this assumption empirically, by estimating the change in the difference across both countries in the pre-reform period. Table 4.4 presents the coefficients for the years prior to the reform (2007/8 and 2008/9, with 2006/7 being the reference year) across the three main outcomes (grouped to HRG with CC, treatment cost and log of treatment cost). All of the relevant coefficients are small and insignificant, suggesting there is no difference in trends across the two countries over time in the pre-reform period.

Table 4.4: Placebo regression results for the main outcomes: probability of being grouped to a severe HRG, treatment cost and log of treatment cost

	Probability of severe HRG	Treatment price	Log price
	(1)	(2)	(3)
(Reference=2006/7 x England)			
2007/8 x England	0.008	-14.284	-0.008
	(0.007)	(11.383)	(0.006)
2008/9 x England	0.019	-1.999	-0.002
	(0.011)	(11.657)	(0.007)
Observations	2,508,636	2,508,636	2,508,636
Number of hospitals	216	216	216
Hospital fixed effects	✓	✓	✓
Season dummies	✓	✓	✓
Year dummies	✓	✓	✓
Core HRG dummies	✓	✓	✓

Notes: This table presents the results of the empirical test for the parallel pre-trends for the three main outcomes: probability of grouping a patient to severe HRG (1), average treatment price in £ (2) and log of treatment price (3). The models are estimated on the pre-reform data (2006/7-2008/9). The table include the estimates for the interaction term of the pre-policy years and the treatment group (England). All models include season dummies (calendar months), hospital fixed effects, year dummies (2006/7-2008/9) as well as dummies representing the core HRG before the split into severe and non-severe. Standard errors are clustered by hospital.

$*$ p$<$0.05 $**$ p$<$0.01 $***$ p$<$0.001

We perform several sensitivity analyses to substantiate the validity of the main results. First, we exclude from the analysis observations that occurred within six months (on either side) of the reform. This way we control for any anticipatory/adjustment effect. As can be seen in Table 4.5, the estimates for cost and for the probability of a severe HRG are slightly larger once the adjustment period is taken into account (estimated effects are 3.8 pp increase in probability, with the £61.79 increase in treatment cost). This suggest the either (i) hospitals started to prepare for the reform already in the period before it was officially implemented; or (ii) the adjustment is gradual in the post-policy period. The estimates are slightly smaller once we perform the analysis on a balanced panel of hospitals who reported

the data in all years (estimated effects 3.1 pp increase in probability and an increase of £56.99 in treatment cost). Once we base the time of treatment on the discharge data rather than the admission date, the sample size is reduced (from 6.6 million to 4.8 million). The estimates of the probability of a severe HRG remain unchanged when using the discharge date (3.3 pp), however, the effect of the reform on treatment price is smaller and non-significant (£41.618). This is likely due to the sample selection: the severe (and hence most expensive) cases are typically moved to other wards and departments for rehabilitation rather then discharged, resulting in missing value of the discharge date.

Table 4.5: Sensitivity analyses

	Including adjustment period (1)		Balanced panel (2)		Discharge based date (3)	
	Probability of severe HRG	Treatment Cost	Probability of severe HRG	Treatment Cost	Probability of severe HRG	Treatment Cost
Post reform X England	0.038***	61.794*	0.031***	56.988*	0.033***	41.618*
	(0.007)	(25.693)	(0.006)	(24.097)	(0.006)	(23.696)
Observations	5,708,747		5,792,515		4,822,412	
Number of hospitals	257		153		256	
Hospital Fixed Effects	✓		✓		✓	
Season dummies	✓		✓		✓	
Year dummies	✓		✓		✓	
Core HRG dummies	✓		✓		✓	

Notes: This table presents the results of the sensitivity analyses for the two outcomes: probability of grouping a patient to severe HRG and average treatment price in £. Model (1) excludes 6 months of either side of the implementation time of the reform to adjust for possible anticipatory/adjustment period. Model (2) is run on a balanced panel of hospitals that treated patients in each year of the study period (2006/7-2012/13). Model (3) uses discharge date rather than admission date to assign the time of the treatment. All models include season dummies (calendar months), hospital fixed effects, year dummies (2006/7-2008/9) as well as dummies representing the core HRG before the split into severe and non-severe. Standard errors are clustered by hospital.

$*$ p$<$0.05 $**$ p$<$0.01 $***$ p$<$0.001

4.7.2 Mechanisms

We are further interested in understanding what drives the upcoding and the increase in the treatment cost. Our estimates, presented in Table 4.6, show that the HRG reform significantly increased the number of diagnosis codes recorded on patient's records by 0.560 codes, while having no effect on the number of recorded procedure codes (estimated increase is 0.021 and insignificant). While our results suggest there is a small increase in the probability of having any procedure recorded (by 2.0 pp), the result is statistically insignificant. Assuming hospitals would report any additional procedures they perform as to optimise their revenue, these results suggest the reform increased the coding of comorbidities, without causing an increase in the treatment intensity by performing more procedures.

This is further confirmed by analysing the severity of the reported diagnosis and procedure codes - measured as the expected resource use for each code. As can be observed in Table 4.6, the maximum severity score for the reported diagnoses increased by 0.051, with the mean severity actually decreasing by 0.145. The latter is likely a result of an overall increase in coding intensity, with more low-severity diagnoses recorded. For procedures the maximum severity increased by 0.149, with the increase in mean of 0.018. With the severity score ranging from 1 to 40, the estimates suggest the effect of the reform on the severity of reported procedures is very small.

The empirical analysis of the pre-trends confirms that there is no difference in trends across England in Wales for all of the above outcome measures before the implementation of the reform. As can be seen in table C2 in the Appendix, all of the relevant coefficients are small and insignificant, confirming the validity of our results.

Table 4.6: Results of the regressions analyses on the effects of the reform on coding diagnoses and procedure codes

	Number of Diagnoses (1)	Number of Procedures (2)	Probability to record a procedure (3)	Max severity Diagnosis (4)	Max severity Procedures (5)	Mean severity Diagnoses (6)	Mean severity Procedures (7)
Post reform X England	0.560***	0.021	0.020	0.051	0.149	-0.145	0.018
	(0.112)	(0.057)	(0.014)	(0.018)	(0.080)	(0.231)	(0.067)
Observations	6,640,515	6,640,515	6,640,515	6,640,515	6,640,515	6,640,515	6,640,515
Number of hospitals	257	257	257	257	257	257	257
Hospital Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Season dummies	✓	✓	✓	✓	✓	✓	✓
Year dummies	✓	✓	✓	✓	✓	✓	✓
Core HRG dummies	✓	✓	✓	✓	✓	✓	✓

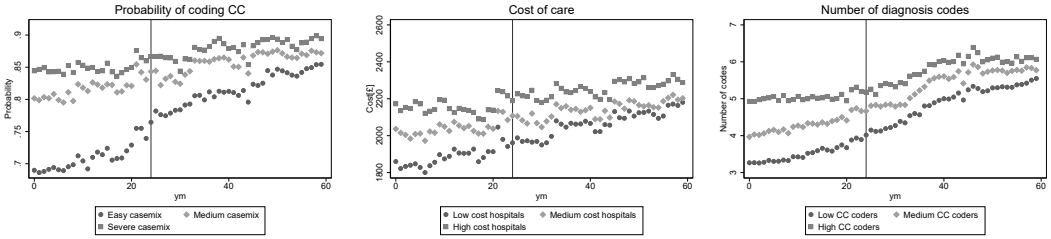
Notes: This table presents the results of the analyses on the mechanisms of upcoding; specifically the coefficients measuring the effect of the classification reform (interaction term between post reform period and treatment group (England)). The outcome variables are as follows: the number of recorded diagnosis codes (1), the number of recorded procedure codes (2), probability of having any procedure reported in the medical record (3), maximum severity across all recorded diagnosis codes (4), maximum severity across all recorded procedure codes (5), mean severity across all recorded diagnosis codes (6), and mean severity across all recorded procedure codes (7). All models use data from the full study period (2006/7-2012/13). All models include season dummies (calendar months), hospital fixed effects, year dummies (2006/7-2008/9) as well as dummies representing the core HRG before the split into severe and non-severe. Standard errors are clustered by hospital.

$*$ $p<0.05$ $**$ $p<0.01$ $***$ $p<0.001$

4.7.3 Heterogeneous Treatment Effect across Hospitals

Figure 4.5 presents the trends in probability of being coded to severe HRG, average treatment price, and the number of recorded diagnosis codes across hospital terciles based on the their coding intensity in the pre-reform period. This is measured as the proportion of severe HRGs in the second year of the study period (2006/7). We observe large differences across hospitals in the pre-reform period, however, these dissipate in the post-reform period. In particular, the figures suggest that hospitals with a low proportion of severe at the start of the study period *catch-up* once the reform is implemented.

Figure 4.5: Trends for the main outcomes for England and Wales over time (from April 2007 to March 2012).



(a) Probability of being coded to severe HRG

(b) Treatment price

(c) Number of recorded diagnosis codes

Notes: This graphs shows the trends for selected outcomes for England and Wales over time (from April 2007 to March 2012). The vertical line represents the time of the reform of the HRG classification in England in 2009. We present trends for three outcome measures: probability of grouping a patient to severe HRG (a), average treatment price in £ (b) and the number of recorded diagnosis codes (c). We group hospitals in terciles, with the first tercile (low coders) comprised of hospitals with the lowest proportion of severe HRGs at the start of the period (2006/7). Second tercile are medium coders (medium proportion of severe HRGs) and the third terciles is comprised of high coders. These are the hospitals with the highest proportion of severe HRGs at the start of the policy.

Table 4.7 presents the estimates for the main outcomes across hospitals, with *low-coders* referring to hospitals with the smallest proportion of severe HRGs at the start of the period, medium coders referring to the middle group and high coders referring to hospitals with the largest share of severe HRGs. Across all three outcome measures, the effect is largest for the low coders, with the reform increasing the probability of being grouped to HRG with CC by 6.8 pp for this group. For the middle group of hospitals the estimated increase

Table 4.7: The effect of the classification reform across hospitals based on their coding intensity in 2006/7 (start of the study period)

	Grouped to severe HRG	Treatment cost	Number of diagnoses
	(1)	(2)	(3)
Post reform X Low coders	0.068***	99.888***	0.871***
	(0.118)	(23.895)	(0.006)
Post reform X Medium coders	0.016***	52.344*	0.525***
	(0.004)	(23.273)	(0.087)
Post reform X High coders	0.003	31.273	0.154
	(0.004)	(23.195)	(0.105)
Observations	6,383,460	6,383,460	6,383,460
Number of hospitals	198	198	198
Hospital fixed effects	✓	✓	✓
Season dummies	✓	✓	✓
Year dummies	✓	✓	✓
Core HRG dummies	✓	✓	✓

Notes: This table presents the results of the analyses analysis on heterogeneous treatment affects across hospitals, based on the proportion of patients grouped to severe HRG at the start of the period. Estimates are reported for three outcome measures: probability of grouping a patient to severe HRG (1), average treatment price in £ (2) and the number of recorded diagnosis codes (3). We group hospitals in terciles, with the first tercile (low coders) comprised of hospitals with the lowest proportion of severe HRGs at the start of the period (2007/8). Second tercile are medium coders (medium proportion of severe HRGs) and the third terciles is comprised of high coders. These are the hospitals with the highest proportion of severe HRGs at the start of the policy. The reported estimates are the interaction between each of the tercile and the post-reform policy. Control group consists of hospitals from Wales. All models use data from the study period 2007/8-2012/13. All models include season dummies (calendar months), hospital fixed effects, year dummies as well as dummies representing the core HRG before the split into severe and non-severe. Standard errors are clustered by hospital.

$*$ p$<$0.05 $**$ p$<$0.01 $***$ p$<$0.001

in 1.6 pp. For hospitals that were already coding well the complications and co-morbidities in the pre-policy period (high coders) the effect of the reform is negligible (0.3 pp and insignificant). We observe similar pattern for treatment cost, which increased by £99.89 in low-coding hospitals, £52.34 in the middle group and £31.27 in the high coding group. The number of reported diagnosis codes also increased more for the low coders group (by 0.871 codes) compared to high coders (0.154). These results suggest some hospitals were *catching-up* with coding once the reform was implemented. The difference in casemix appears thus much greater in the pre-reform period, compared to the post-reform period.

There are two possible interpretations of these results: (i) the low coders have improved the quality of their reporting, or (ii) the low coders are upcoding the patients to severe HRGs by either by changing the treatment pathway or are participating fraudulent coding. Results of our analysis of mechanics of upcoding indicate that the higher proportion of severe HRG is mainly due to increased reporting of lower severity diagnosis codes, without much

change to the reporting of procedures, suggesting comprehensive rather than fraudulent coding. Furthermore, the apparent casemix is much more similar across hospitals post the HRG reform (see Figure 4.6). While we do not have information on the true underlying casemix across hospitals, considering our sample consists of respiratory patients, we do not expect large underlying differences in the presenting cohort of patients once we adjust for age, sex and deprivation (Vaughan et al., 2021). This suggests there was an improvement in the underlying coding quality.

4.7.4 Interpretation and policy application

The main motivation behind the classification reform is to provide a fair reimbursement system across hospitals. While our results show an increase in the overall treatment cost after the reform, this might be justified by creating a level playing field across hospital.

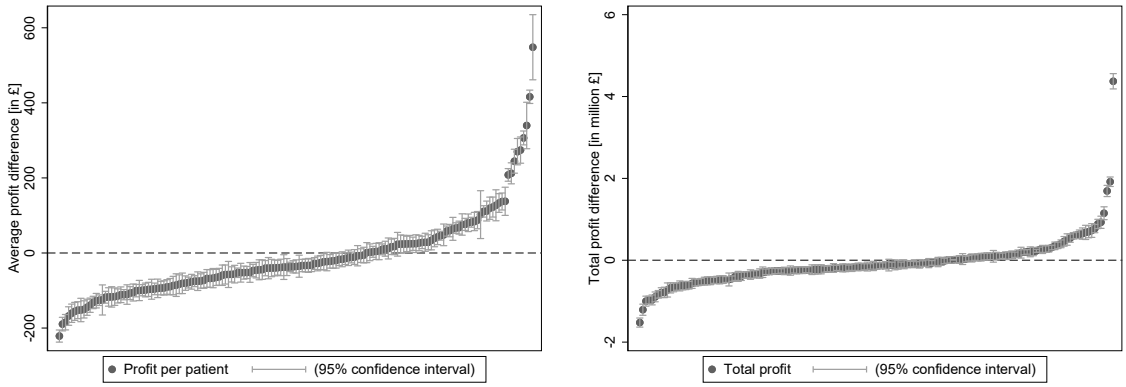
To better understand the implications of the reform on hospitals' profits, we look at the cost and revenue distribution before the reform. We assume that the price of the post-reform HRG corresponds to the actual cost of treatment in both, the pre- and post- reform period. This suggest that the hospitals profit post-reform is 0. We further assume the hospitals get paid a weighted average of severe and non-severe prices in the pre-reform period, so that the price paid to the hospital is the same across patients who are grouped to the same core HRG (regardless of reported complications and co-morbidities). This is used as a proxy for the HRG3.5 groups, which are typically not split by co-morbidities. We calculate the change in profit as the price minus the cost.

Figure 4.6 shows that over half of hospitals were making a loss in the pre-policy period for treating respiratory patient, with the per patient profit ranging from a negative loss of £221.29 per person to a positive profit of £548.20 pounds per person, with a mean profit per person of -£13.71. Similarly, when looking at the total hospitals' profit, it ranges from -£1.52 million per year to £4.37 million per year, with the average of -£0.07 million. The total loss across hospitals with negative profit equals to £34.55 million pounds. These results suggest that while many hospitals were disadvantaged before the start of the policy, the total loss accrued is less than the cost of upcoding after the classification reform was implemented (£57 million).

This has important policy implications. Namely, in the NHS type of healthcare system, hospitals are mostly publicly owned; this means that any hospitals' deficit is typically covered by the state. In this particular case, the deficit due to the changes in casemix is lower than the increase in healthcare expenditure due to the reform, rendering the reform as less optimal choice for the payer. However, as seen in section 4.7.3, the policy makers must

take into account both, the cost of the reform, fairness across the hospitals, as well as the likely improvements in the coding of diseases and procedures.

Figure 4.6: Mean and total change in profit



(a) Mean change in profit

(b) Total change in profit

Notes: The figure present the per patient (mean) (a) and total (across all patients) (b) change in profit across hospitals when the HRG system moves from the verion HRG3.5 to HRG4. Graphs are based on the data from the pre-policy period. The total profit is calculated on a per-year basis.

4.8 Conclusion

In order to improve efficiency and productivity of healthcare systems most OECD countries have implemented the Diagnosis Related Groups (DRG) classification system to standardise hospitals' reimbursement and encourage cost-containment. Our study shows that expansion of the number of DRGs can prompt an increase in coding of comorbidities, leading to an increase of hospital expenditure. In particular, results indicate that hospitals mainly respond to the classification reform by more comprehensive coding of diagnosis codes, without changing the patients' care pathway via performing more procedures or changing their severity.

Our results further show that the effect of the reform is the largest for hospitals who were coding fewer comorbidities in the period prior to the implementation of the reform. This suggest some hospitals were *catching up*, with the casemix across hospitals being much more similar after the reform compared to before the reform.

There are two main limitations to this study. First, it is limited by it's focus on a single disease area (respiratory illnesses). In particular, while we show that the upcoding is driven mainly by an increase in the coding of diagnoses, the effect on treatment intensity might be larger in more procedure (rather than diagnosis) driven clinical areas (ie. orthopaedics). For these clinical areas our results show the lower bound with the actual upcoding effects possibly bigger. A further limitation relates to the calculation of the financial effects of the scheme. All of the calculations on cost, revenue and profit are based on the revised (post-reform) version of HRGs (HRG4), with the assumption that the pre-reform HRGs correspond to the core HRG before splitting them based on complications and comorbidities. While this largely corresponds to the actual 3.5 version of the HRGs used in the pre-reform period, there might nevertheless be differences across both versions that we cannot capture with our approach. However, due to the way complications an co-morbidities were integrated into the revised HRG version, we expect those differences to be small.

CHAPTER 5

Conclusion

This thesis considers three different types of financial incentives in the secondary care setting. In particular, it analyses (i) a P4P scheme targeting an improvement in the quality of care, (ii) a P4P incentive aimed at increasing efficiency of care delivery, and (iii) wider changes in the hospital reimbursement model, with a particular focus on the modification of the underlying DRG classification. Overall, the thesis finds large and statistically significant effects in all three cases, suggesting that providers strongly react to financial stimulus. This has important policy implications in light of rapid growth in medical spending across most healthcare systems, as it demonstrates that policy makers have effective tools at their disposal to improve efficiency and quality of healthcare delivery in the secondary care sector.

Chapter 2 investigates whether the financial incentive in a form of a Best Practice Tariff (BPT) for hip fracture improved the quality of care for hip fracture patients. The scheme simultaneously incentivises nine quality measures by only rewarding providers when they meet all of them (on patient level). Estimates show that the policy increased the overall adherence to the criteria by 51.7pp. This suggests that the scheme was very effective in changing provider's behaviour, which contrasts many of the existing studies on P4P schemes that find small or no effect of financial policies on the quality of care (Milstein and Schreyoegg, 2016; Eijkenaar, 2012). However, the overall result hides substantial heterogeneity in the response across different criteria, with the largest effects observed for the

criteria related to geriatrician involvement in the care, with smaller improvement observed in other areas. The sizeable difference between the overall policy effect and the effect of individual criteria suggests the providers responded strategically to the scheme, by improving across all dimensions to the level required to obtain the bonus payment.

This has important policy implications as it suggests that incentivising several measures simultaneously can be an effective way to improve a wide range of performance indicators at the same time. In particular, it shows that P4P can be used to improve care across the entire care pathway, rather than just focusing on particular aspects of patient's experience. The results presented in this Chapter further imply that focusing on process rather than outcome measures can be an efficient way to improve quality of care, provided that the incentivised measures are in line with the best clinical practice.

A main limitation of Chapter 2 is the relatively short pre-policy period (2 years). This is due to data availability and the limited data quality in the first year of the data collection. However, we find consistency in coding across the pre-period across indicators, suggesting clear trends in achievement rates across both England and Wales. Due to data limitations, this chapter only considers mortality as a patient's outcome, despite the scheme being designed to improve the overall patient's experience. Future research could consider wider impacts of the scheme on patient's outcomes.

While Chapter 2 studies how P4P programs can be used to improve quality of care, the key focus of Chapter 3 is the use of financial schemes to encourage greater efficiency in the care delivery. Focusing on the BPT for Outpatients, this chapter illustrates how changes in the reimbursement levels - and associated marginal profit - can be successful in shifting patients from inpatient to outpatient setting, without harming the quality of care. Our estimates show a positive effect of the BPT on the proportion of patients treated in the outpatient setting for all three incentivised conditions (diagnostic hysteroscopy, diagnostic cystoscopy and sterilisation), with the biggest effect seen for the two diagnostic conditions (between 16 and 36 percentage points), and a smaller effect observed for sterilisation (4

percentage points).

The BPT for Outpatients manages to improve efficiency without reducing the quality of care, measured as the probability of having the same procedure repeated at a later date. The scheme did not increase the overall volume of patients. We also observe positive spill-over effect of the BPT on the probability of being treated in the outpatient setting on the related conditions. Despite the hospitals high take-up of the scheme, hospitals made lower profit after the introduction of the policy due to the particular structure of the scheme. At the same time, the purchaser lowered their costs.

These results have considerable policy implications. First, they show that a targeted financial policy can be successful in shifting care from the outpatient to the inpatient, without harming the quality of the care. This means that the policy makers do not have to trade off quality for efficiency, as it's possible to address one dimension without harming the other. Second, in case where the bonus for outpatient services is accompanied by a reduction in the price paid for the inpatient care, the health care purchaser reduces their overall healthcare expenditure. This is particularly relevant, as policy makers increasingly explore ways to limit the rising trends in healthcare spending. However, price settings warrants a careful consideration, as the increase in profit for the purchaser might come at an expense of individual providers creating deficit for providing the care.

A limitation of Chapter 3 is that it focuses only on the three incentivised treatments as the same type of financial incentive might not work across all of the procedures. However, a larger majority of procedures that are deemed suitable for both the inpatient and the outpatient setting - and could thus be included in the incentive scheme - have very similar characteristics to the ones that considered in the study. Most of these procedures are diagnostic tests, which can be performed either under the general anaesthesia or in the outpatient setting using alternative (often novel) techniques. This similarity suggest that the results of this Chapter are likely to be generalised to other contexts. Further limitation of Chapter 3 relates to the measurement of quality of care. The chapter uses the probabil-

ity of having a repeated procedure within 60/90 days as a proxy for premature stop to the original procedure. While this measures the clinical outcome of the procedure, it does not fully capture patient experience, which might be affected by the shift from inpatient to the outpatient setting. This remains an area for future research.

While Chapters 2 & 3 deal with incentives that target specific conditions and are associated with pre-defined measures of success, Chapter 4 explores a whole-system change in reimbursement. This chapter focuses on a reform of the DRG classification which saw a substantial increase in the number of DRG groups by heightening the role of co-morbidities and complications. Results show that the expansion of the number of DRGs can prompt an increase in coding co-morbidities, which in consequences leads to an increase of hospital expenditure. Furthermore, hospitals respond to the DRG expansion by increasing the coding of diagnoses, without performing more procedures or altering the care pathway. The effect of the reform is largest for hospitals that were not coding the comorbidities in the period prior to the implementation of the reform. This might suggest some hospitals were catching up in coding, with the casemix across hospitals being much more similar after the reform compared to before the reform.

Results of Chapter 4 suggest that the reform likely improves quality of coding, by closing the coding gap across hospitals. Namely, while the average treatment payment differs substantially across hospitals in the pre-reform period, the difference is negligible in the last year of our study period. This has two possible interpretations: (i) the casemix across hospitals might be similar, but some hospitals were not fully reporting co-morbidities; (ii) some hospitals fraudulently upcode and hence profit in the post-reform period. The result on the coding patters across diagnosis and procedure codes indicate that (i) is more likely, suggesting an increase in the quality of coding.

We derive several policy implications from the results of Chapter 4. First, hospitals will likely respond to classification reform by changing their coding behaviour, as this will typically improve their financial position. However, hospitals will likely focus on the improve-

ment in the coding the diagnoses, rather than changing patients care pathway. While change in classification will likely increase the total health expenditure, policy makers should balance this against the likely increase in the quality of coding and better understating of the population health. Furthermore, our results suggest that the casemix differences across hospitals are relatively small once we take the improvement in coding into account. This suggests that even a DRG system with small number of codes is mainly fair across hospitals.

Chapter 4 is limited by its focus on respiratory illnesses, which are more likely to be treated without surgical procedures. This means that the effects on treatment intensity (coding of procedures) might be larger in more procedure driven clinical areas (ie. orthopaedics). For these clinical areas our results show the lower bound with the actual upcoding effects possibly larger. Future research is needed to estimate the effect of classification change for procedure driven medical areas. A further limitation of Chapter 4 relates to the calculation of the financial effects of the scheme, which are based on the revised (post-reform) version of HRGs (HRG4), with the assumption that the pre-reform HRGs correspond to the core HRG before splitting them based on complications and comorbidities. While this largely holds for majority of cases, there might nevertheless be small differences across both versions that we cannot capture with our approach. This could be addressed in future research.

Taken together, all three Chapters provide evidence that financial incentives can be effective in changing provider's behaviour. Well designed schemes can improve both quality of care and efficiency, without increasing the overall healthcare expenditure. However, as evidenced in Chapter 4, hospitals also respond by upcoding to maximise their financial reimbursement. While this can be a sign of an improvement in the coding quality, policy makers nevertheless need to consider upcoding when implementing a new financial policy in the secondary care sector.

Appendix A

Figure A1: BPT attainment over time for England and Wales

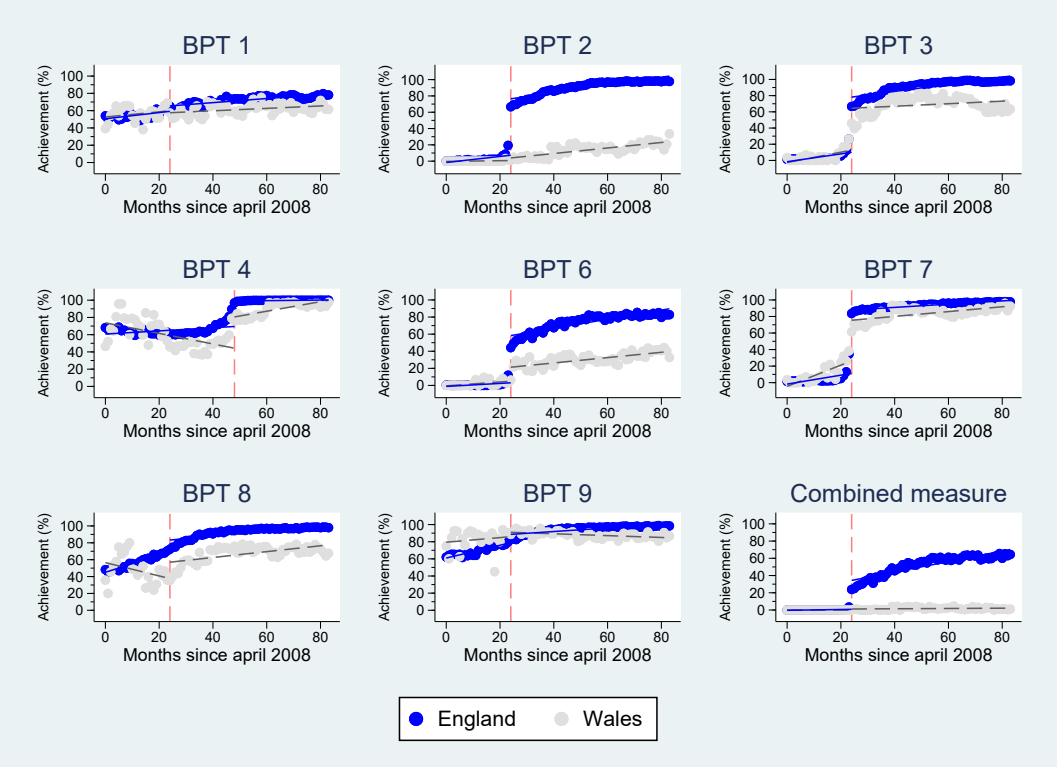


Table A1: Number of hospitals and patients included in the estimation sample, 2008/9 to 2014/15.

Financial year	Hospitals		Patients	
	Wales	England	Wales	England
2008/09	5	111	399	14,850
2009/10	10	154	808	22,227
2010/11	11	166	1,401	33,120
2011/12	12	166	2,176	41,299
2012/13	13	164	2,904	45,960
2013/14	13	164	3,147	50,824
2014/15	13	163	3,186	53,597

Notes: The table shows the number of hospitals reporting to the NHFD and the number of patients recorded within NHFD by year and country. The dashed line represents the start of the BPT policy in April 2010.

Table A2: Number of patients (proportion of total) for which the all pre-/during & post-surgical BPT criteria were achieved

	Pre-surgery		Pre-surgery excl. BPT2		Post-surgery	
Period	England	Wales	England	Wales	England	Wales
Pre-policy (April 08 to March 10)	129	0	184	0	1,422	77
	(0.3%)	(0%)	(0.5%)	(0%)	(3.4%)	(6.4%)
Post-policy (April 10 to March 15)	124,726	375	125,620	1,956	201,660	7,757
	(55.5%)	(2.9%)	(55.9%)	(15.3%)	(89.7%)	(60.5%)

Notes: 'Pre-surgery' refers to care process that are expected to take place prior to or during the surgery (BPT1, BPT3, BPT4, BPT6). 'Post-surgery' refers to care processes that are expected to take place after the surgery (BPT7, BPT8, BPT9).

Table A3: Proportion of patients receiving BPT care

	Financial year						
Criterion	2008	2009	2010	2011	2012	2013	2014
England							
Care bundle: All criteria met	0.00	0.00	0.24	0.38	0.58	0.61	0.62
1: Surgery within 36 hours	0.54	0.57	0.65	0.71	0.74	0.76	0.76
2: Shared care across specialties	0.01	0.05	0.75	0.88	0.95	0.97	0.98
3: Multidisciplinary care protocol	0.02	0.07	0.75	0.90	0.95	0.98	0.98
4: Pre-op cognitive function assessment	0.64	0.62	0.61	0.72	0.99	1.00	1.00
6: Peri-op geriatric assessment	0.00	0.02	0.56	0.71	0.79	0.82	0.83
7: Geriatrician-led multidisciplinary rehab	0.02	0.08	0.88	0.92	0.94	0.96	0.98
8: Secondary prevention including falls	0.51	0.64	0.81	0.92	0.96	0.97	0.98
9: Bone health assessment	0.64	0.75	0.87	0.94	0.96	0.98	0.98
Wales							
Care bundle: All criteria met	0.00	0.00	0.00	0.02	0.02	0.02	0.01
1: Surgery within 36 hours	0.56	0.58	0.56	0.62	0.63	0.64	0.63
2: Shared care across specialties	0.00	0.01	0.04	0.12	0.17	0.15	0.23
3: Multidisciplinary care protocol	0.02	0.11	0.55	0.75	0.78	0.74	0.64
4: Pre-op cognitive function assessment	0.73	0.65	0.46	0.52	0.80	0.96	0.94
6: Peri-op geriatric assessment	0.01	0.04	0.23	0.30	0.29	0.32	0.39
7: Geriatrician-led multidisciplinary rehab	0.03	0.21	0.73	0.87	0.87	0.84	0.90
8: Secondary prevention including falls	0.57	0.39	0.53	0.69	0.73	0.71	0.71
9: Bone health assessment	0.84	0.83	0.91	0.90	0.85	0.86	0.86

Notes: This table presents the proportion of patients receiving BPT care according to the 9 separate incentivised measures. Scores are reported separately for England and Wales. BPT scores are reported separately for each financial year, which 1st April to 31st March of the following calendar year.

Table A4: Empirical test of the parallel trends assumption: individual criteria

	BPT1		BPT2		BPT3		BPT4		BPT6		BPT7		BPT8		BPT8		Number of criteria	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE		
Quarter																		
July08 - Sept08	-0.074	0.073	0.004	0.003	- 0.103	0.072	- 0.089	0.080	0.003	0.007	- 0.088	0.058	- 0.19	0.12	-0.078	0.050	-0.525***	0.094
Oct08 - Dec08	-0.107**	0.048	0.009	0.009	- 0.251	0.215	- 0.126	0.090	0.057	0.058	- 0.163	0.158	- 0.17	0.14	-0.015	0.030	-0.640*	0.272
Jan09 - Mar09	-0.022	0.091	0.002	0.011	- 0.288	0.228	- 0.082	0.093	0.026	0.032	- 0.196	0.179	- 0.11	0.13	-0.006	0.031	- 0.593	0.316
Apr09 - June09	-0.024	0.106	0.010	0.016	- 0.306	0.240	- 0.053	0.088	- 0.003	0.015	- 0.261	0.190	- 0.05	0.11	0.116***	0.036	- 0.521	0.439
July09-Sept09	-0.052	0.089	0.013	0.021	- 0.325	0.231	- 0.079	0.098	- 0.028	0.026	- 0.349*	0.186	- 0.07	0.12	0.139***	0.039	- 0.671	0.392
Oct09 - Dec09	0.011	0.094	0.010	0.025	- 0.362	0.255	- 0.169	0.098	- 0.031	0.036	- 0.339	0.206	0.14	0.12	0.265***	0.073	- 0.310	0.445
Jan10 - Mar10	-0.080	0.090	0.078*	0.036	- 0.304	0.259	- 0.087	0.105	0.006	0.046	- 0.301	0.218	0.08	0.12	0.106**	0.051	- 0.413	0.454
April10 - June10	n/a	n/a	n/a	n/a	n/a	n/a	- 0.048	0.088	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
July10 - Sept11	n/a	n/a	n/a	n/a	n/a	n/a	- 0.002	0.102	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Oct10 - Dec10	n/a	n/a	n/a	n/a	n/a	n/a	- 0.017	0.092	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Jan11 - March11	n/a	n/a	n/a	n/a	n/a	n/a	- 0.018	0.088	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
April11 - June11	n/a	n/a	n/a	n/a	n/a	n/a	0.029	0.096	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
July11 - Sept11	n/a	n/a	n/a	n/a	n/a	n/a	- 0.002	0.089	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Oct11 - Dec11	n/a	n/a	n/a	n/a	n/a	n/a	0.049	0.088	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Jan12 - March12	n/a	n/a	n/a	n/a	n/a	n/a	0.076	0.096	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Notes: We perform the analysis of to test the pre-parallel trends for each individual BPT criteria and for the number of criteria met, using the data from the period April 2008-March 2010 and quarterly dummies. We perform the analysis of to test the pre-parallel trends for BPT4, using the data from the period April 2008-March 2012 and quarterly dummies. The reference category for quarters is Apil 2008 - June 2008. Standard errors are clustered on hospital level.

Risk-adjustment of hospital mortality rates

We follow standard methodology (e.g. Ash and Ellis, 2012) and apply indirect standardisation to calculate quarterly hospital mortality rates that are adjusted for case-mix differences. To do so, we first estimate a logistic regression model of mortality within a given time window controlling for patients' age (in 5-year brackets), gender, number of conditions as defined by the Elixhauser algorithm (from 0-31), and the location from which the patient was admitted (own home or an institution, such as care home), i.e.

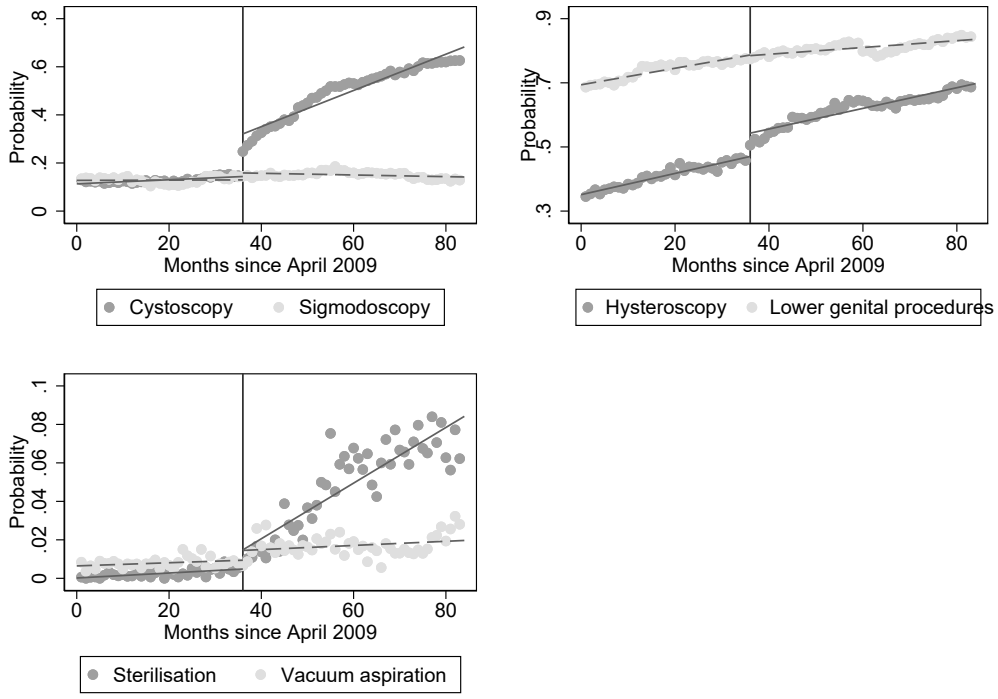
$$Pr[Z_i = 1|\boldsymbol{X_i}] = \frac{e^{\alpha+\boldsymbol{X_i'\delta}}}{1+e^{\alpha+\boldsymbol{X_i'\delta}}} \tag{5.1}$$

where Z_i equals to 1 if patient i died within 7/30/90/365 of admission and 0 otherwise, and X_i is a vector of patient characteristics. We then use this model to predict the probability of death, $\hat{Z}_{ihq}$, for patient i treated in hospital h in quarter $q = 1, \ldots, 28$ (with 1 for the first quarter in year 2008/9 and 28 for the last quarter in year 2014/15). The indirectly standardised hospital mortality rate $\hat{Z}_{hq}$ is then calculated as

$$\hat{Z}_{hq} = \frac{\sum_{i=1}^{N_{hq}} Z_{ihq}}{\sum_{i=1}^{N_{hq}} \hat{Z}_{ihq}} \times \frac{1}{N_q} \sum_{i=1}^{N_q} Z_{iq}. \tag{5.2}$$

Appendix B

Figure B1: Probability of treatment in the outpatient setting - trends over time



Notes: Figure shows pre- and post- trends of the treatment/control group pairs for the primary outcome: probability of treatment in the outpatient setting

Figure B2: Volume over time

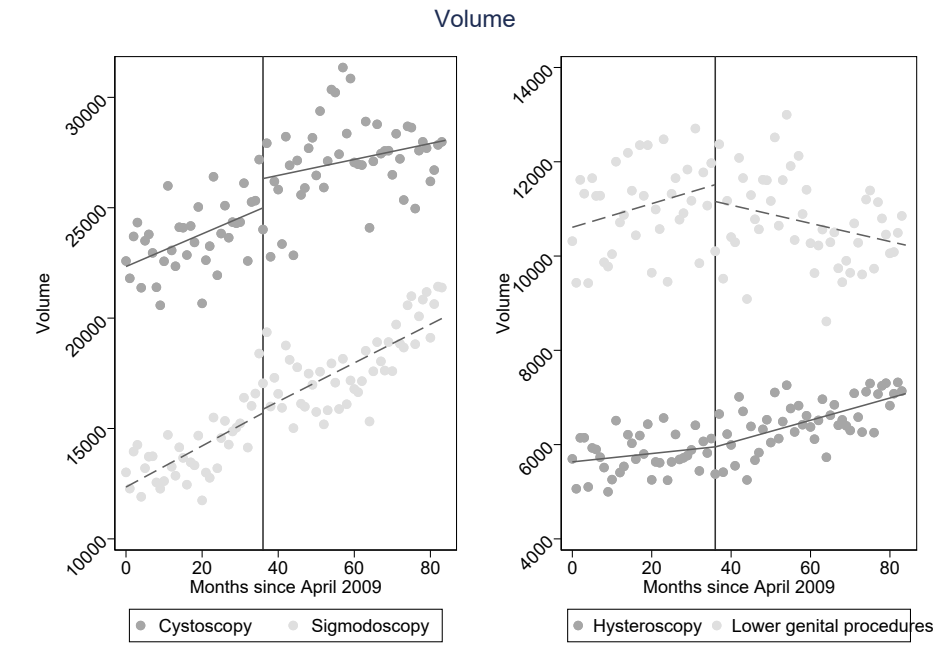


Figure B3: Re-operation within 60/90 days

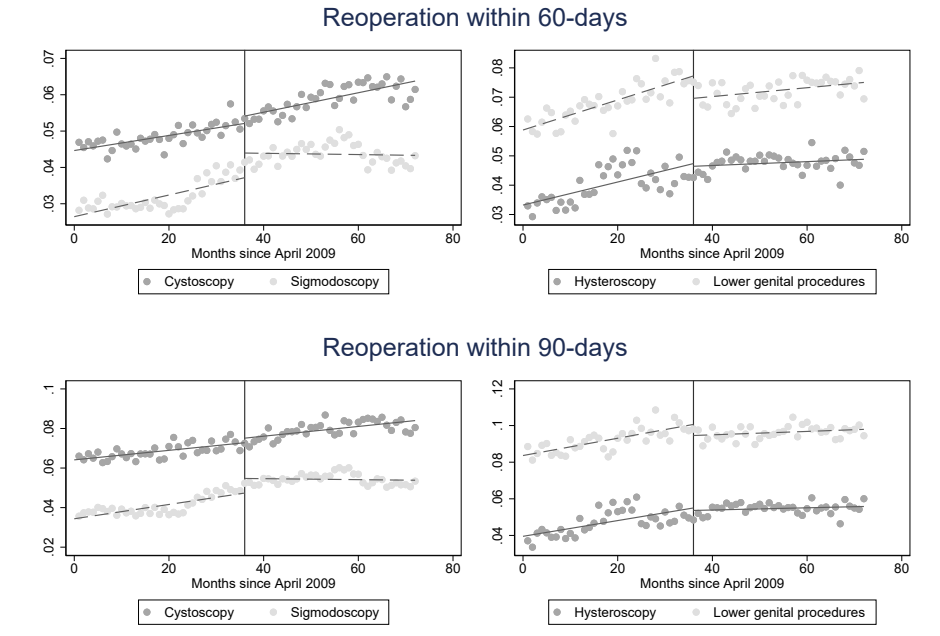


Table B1: Main regression results: coefficients of patient characteristics for the main analysis

	Cystoscopy		Hysteroscopy		Sterilisation	
	Estimate	SE	Estimate	SE	Estimate	SE
Diff-in-diff estimates						
Treatment (BPT group)	-0.051	0.027	-0.270***	0.033	-0.034***	0.010
Diff-in-diff coefficient	0.361***	0.031	0.163***	0.022	0.038***	0.010
Age group						
Reference: 19-24						
26-30	0.015***	0.003	-0.001	0.008	0.002	0.002
31-35	0.021***	0.003	0.008	0.006	0.002	0.003
36-40	0.026***	0.003	0.034***	0.008	0.003	0.004
41-45	0.031***	0.005	0.062***	0.010	0.006	0.004
46-50	0.033***	0.005	0.092***	0.013	0.032**	0.012
51-55	-0.050**	0.015	0.124***	0.015	0.211	0.108
56-60	0.026***	0.007	0.165***	0.019	0.630***	0.168
61-65	0.027**	0.008	0.216***	0.023	0.722***	0.117
66-70	0.017*	0.008	0.257***	0.025	0.630**	0.196
71-75	0.016	0.008	0.286***	0.026	0.708***	0.004
76-80	0.017	0.009	0.305***	0.027	0.040*	0.018
81-85	0.014	0.009	0.319***	0.027	0.091	0.046
86-89	0.015	0.010	0.321***	0.027	0.152*	0.071
90+	0.028*	0.012	0.326***	0.027	0.001	0.005
BPT x 26-30	0.010	0.005	0.081***	0.013	0.012***	0.004
BPT x 31-35	0.014*	0.007	0.100***	0.015	0.020***	0.005
BPT x 36-40	0.018*	0.007	0.090***	0.016	0.021***	0.005
BPT x 41-45	0.019*	0.008	0.087***	0.021	0.018**	0.006
BPT x 46-50	0.023**	0.009	0.059**	0.022	-0.014	0.012
BPT x 51-55	0.113***	0.018	0.026	0.023	-0.197	0.108
BPT x 56-60	0.035**	0.011	-0.016	0.026	-0.599***	0.168
BPT x 61-65	0.032**	0.012	-0.076**	0.029	-0.671***	0.117
BPT x 66-70	0.042***	0.012	-0.124***	0.031	-0.614**	0.196
BPT x 71-75	0.043**	0.013	-0.155***	0.032	-0.662***	0.017
BPT x 76-80	0.044***	0.013	-0.183***	0.032		
BPT x 81-85	0.048***	0.014	-0.191***	0.033		
BPT x 86-89	0.053***	0.015	-0.195***	0.034		
BPT x 90+	0.049**	0.017	-0.126**	0.038	0.061	0.055
Sex						
Reference: Female						
Male	0.002	0.002				
BPT x Male	0.021***	0.004				
Past emergency visits						
Reference: no visits						
1 visit	-0.020***	0.004	-0.052***	0.006	0.000	0.002
2 visits	-0.026***	0.005	-0.008	0.006	0.001	0.003
3 visits	-0.030***	0.006	0.004	0.006	0.008*	0.004
4 visits	-0.037***	0.007	0.017**	0.006	0.003	0.007
5+ visits	-0.044***	0.008	0.030***	0.004	0.003	0.003
1 visit	-0.001	0.005	0.034***	0.008	-0.002	0.004
2 visits	0.039***	0.006	0.017*	0.007	-0.004	0.004
3 visits	0.050***	0.008	0.018*	0.008	-0.014*	0.006
4 visits	0.043***	0.009	-0.006	0.010	-0.012	0.008
5+ visits	0.060***	0.010	-0.016	0.010	-0.011*	0.005
Deprivation						
Reference: least deprived						
2nd decile	0.006	0.010	0.000	0.004	-0.002	0.002
3rd decile	0.007	0.009	0.004	0.005	-0.002	0.003
4th decile	0.011	0.011	0.007	0.006	-0.002	0.003
Most deprived	0.006	0.015	0.007	0.008	-0.003	0.003
BPT x 2nd decile	-0.008	0.017	-0.014	0.010	0.003	0.004
BPT x 3rd decile	-0.012	0.015	-0.027*	0.012	0.007	0.006
BPT x 4th decile	-0.016	0.017	-0.034*	0.014	0.008	0.007
BPT x Most deprived	-0.007	0.024	-0.030	0.023	0.016*	0.007
Constant	0.126***	0.015	0.613***	0.018	0.014***	0.003
Observations	3,859,365		1,560,714		303,256	

Notes: Dependent variable is the probability to be treated in the outpatient setting. Time period is from April 2009 to March 2016. Models are estimated by OLS with standard errors (presented in parenthesis under the coefficients), clustered at hospital level. Models are run separately for each treatment-control procedure pair (cystoscopy-sigmoidoscopy; hysteroscopy-lower genital procedures; sterilisation-vacuum aspiration). All models control for casemix and a set of month, year and hospital dummies.
*** p<0.001, ** p<0.01, * p<0.05

Table B2: Sensitivity analyses of the difference-in-difference estimates of the impact of the BPT Outpatients scheme on the probability of treatment in the outpatient setting

Treatment group	Cystoscopy (1)	Hysteroscopy (2)	Sterilisation (3)
(a) Anticipatory/adjustment period			
DiD coefficient	0.379***	0.180***	0.043***
	(0.032)	(0.027)	(0.012)
Adjusted R^2	0.382	0.287	0.098
Number of hospitals	167	165	159
Observations	3,266,040	1,304,678	265,268
(b) Balanced panel			
DiD coefficient	0.357***	0.152***	0.040***
	(0.032)	(0.023)	(0.011)
Adjusted R^2	0.374	0.270	0.092
Number of hospitals	131	132	135
Observations	3,528,821	1,417,426	279,392

Notes: Dependent variable is the probability to be treated in the outpatient setting. Time period is from April 2009 to March 2016, with the model (a) excluding period from October 2011 to October 2012 (6 month on either side of the start of the policy in April 2012). Model (b) excludes providers that did not report in all quarters of our study period. Models are estimated by OLS with standard errors (presented in parenthesis under the coefficients), clustered at hospital level. Models are run separately for each treatment-control procedure pair (cystoscopy-sigmoidoscopy; hysteroscopy-lower genital procedures; sterilisation-vacuum aspiration). All models control for casemix and a set of month, year and hospital dummies.

*** p<0.001, ** p<0.01, * p<0.05

Table B3: Empirical test for the parallel trends assumption for the primary outcome measure

Treatment group	Cystoscopy (1)	Hysteroscopy (2)
(a) Repeated operation within 60-days		
DiD coefficient for 2010/11	0.002	0.004
	(0.002)	(0.006)
DiD coefficient for 2011/12	-0.002	-0.003
	(0.003)	(0.003)
Adjusted R^2	0.012	0.025
Number of hospitals	161	157
Observations	1,346,109	591,394
(b) Repeated operation within 90-days		
DiD coefficient for 2010/11	0.001	0.006
	(0.001)	(0.006)
DiD coefficient for 2011/12	-0.002	-0.001
	(0.006)	(0.004)
Adjusted R^2	0.008	0.032
Number of hospitals	161	157
Observations	1,346,109	591,394
(c) Volume		
DiD coefficient for 2010/11	9.804	0.896
	(7.071)	(9.611)
DiD coefficient for 2011/12	-10.930	12.463
	(10.884)	(13.116)
Adjusted R^2	0.752	0.591
Number of hospitals	131	132
Observations	3,144	3,168
(d) Spill-over effect		
DiD coefficient for 2010/11	0.018*	0.015
	(0.009)	(0.011)
DiD coefficient for 2011/12	0.021	0.036
	(0.016)	(0.015)
Adjusted R^2	0.464	0.373
Number of hospitals	159	157
Observations	1,488,249	650,416

Notes: Time period is from April 2009 to March 2012. Standard errors (presented in parenthesis under the coefficients) are clustered at hospital level. Models are run seperately for each treatment-control procedure pair (cystoscopy-sigmoidoscopy; hysteroscopy-lower genital procedures; sterilisation-vacuum aspiration) and each outcome. All models include a constant, case mix variables and a full set of month and hospital dummies. The null hypothesis for the parallel trends assumption is that the DiD coefficients are jointly zero.

*** p<0.001, ** p<0.01, * p<0.05

B2 Calculation of the volume

The post-policy volume is calculated by taking into account the change in the proportion of patients treated in the outpatient setting β_P^1 as well as an increase in the overall volume β_V^1. With the pre-policy volumes of inpatient and outpatient appointments V_O^0 and V_I^0, respectively, the new post policy outpatient and inpatient volumes equal to:

$$\begin{aligned} V_O^1 &= [V_O^0 + V_I^0 + \beta_V^1] \times [\beta_P^1 + \beta_P^0] \\ V_I^1 &= [V_O^0 + V_I^0 + \beta_V^1] \times [1 - \beta_P^1 - \beta_P^0] \end{aligned}$$

where β_P^0 is the proportion of patients treated in the outpatient setting in the pre-policy period:

$$\beta_P^0 = \frac{V_O^0}{V_O^0 + V_I^0}$$

Table B4 shows the coefficients used in the calculation of the post-policy volume. The policy effects corresponds to the effect in the last post-policy year. The volume increase corresponds to the volume effect, multiplied by the number of quarters *(4)* and hospitals *(122)*.

Table B4: The effect of the BPT on hospital's revenue - coefficients

	Cystoscopy	Hysteroscopy	Sterilisation
Policy Effect β_P	0.490	0.209	0.058
Volume increase β_V	9,155	12,028	-
Pre-policy volume			
Outpatient V_O^0	45,010	34,227	348
Inpatient V_I^0	276,487	41,834	11,268
Post-policy volume			
Outpatient V_O^1	208,972	57,698	1,022
Inpatient V_I^1	121,680	30,391	10,594

Notes: The pre-policy volume is based on the 2011/12 volume of inpatient and outpatient attendances. The post-policy volume is based on the estimated effect of the policy, including the change in overall volume and in the proportion of patients treated in the outpatient setting.

Appendix C

Table C1: HRG Chapters and Chapter Description

HRG Chapter	HRG Chapter Description
A	Nervous System
B	Eyes and Periorbita
C	Mouth Head Neck and Ears
D	Respiratory System
E	Cardiac Surgery and Primary Cardiac Conditions
F	Digestive System
G	Hepatobiliary and Pancreatic System
H	Musculoskeletal System
J	Skin, Breast and Burns
K	Endocrine and Metabolic System
L	Urinary Tract and Male Reproductive System
M	Female Reproductive System and Assisted Reproduction
N	Obstetrics
P	Diseases of Childhood and Neonates
Q	Vascular System
R	Radiology and Nuclear Medicine
S	Haematology, Chemotherapy, Radiotherapy and Specialist Palliative Care
U	Undefined Groups
V	Multiple Trauma, Emergency Medicine and Rehabilitation
W	Immunology, Infectious Diseases and other contacts with Health Services
X	Critical Care and High Cost Drugs

Notes: This table shows the description of the HRG chapters, which are distinguished by the first letter of the HRG. This research analyses the effects of the classification reform on changes to coding in chapter D: Respiratory system.

Table C2: Placebo analysis for the secondary outcomes

	Number recorded Diagnoses	Number recorded Procedures	Probability Procedure HRG	Max severity Diagnosis	Max severity Procedures	Mean severity Diangoses	Mean severity Procedures
2007/8 X England	0.087	-0.085	0.005	0.231	-0.002	0.054	0.928**
	(0.095)	(0.107)	(0.007)	(0.167)	(0.013)	(0.140)	0.345
2008/9 X England	0.181	-0.115	0.004	0.170	-0.012	0.022	0.928**
	(0.163)	(0.147)	(0.007)	(0.301)	(0.022)	(0.018)	0.345
Observations	2,508,636	2,508,636	2,508,636	2,508,636	2,508,636	2,508,636	2,508,636
Number of hospitals	216	216	216	216	216	216	216
Hospital Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Season dummies	✓	✓	✓	✓	✓	✓	✓
Year dummies	✓	✓	✓	✓	✓	✓	✓
Core HRG dummies	✓	✓	✓	✓	✓	✓	✓

Notes: This table presents the results of the empirical test for the parallel pre-trends for the outcomes related to coding procedures and diagnoses: the number of recorded diagnosis codes (1), the number of recorded procedure codes (2), probability of having any procedure reported in the medical record (3), maximum severity across all recorded diagnosis codes (4), maximum severity across all recorded procedure codes (5), mean severity across all recorded diagnosis codes (6), and mean severity across all recorded procedure codes (7). The models are estimated on the pre-reform data (2006/7-2008/9). The table include the estimates for the interaction term of the pre-policy years and the treatment group (England).

$*$ p$<$0.05 $**$ p$<$0.01 $***$ p$<$0.001

BIBLIOGRAPHY

G. Ahmad, S. Saluja, H. O'Flynn, A. Sorrentino, D. Leach, and A. Watson. Pain relief for outpatient hysteroscopy. *Cochrane Database of Systematic Reviews*, 2017. URL https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD007710.pub3/abstract.

V. A. Ancona-Berk and T. Chalmers. An analysis of the costs of ambulatory and inpatient care. *American Journal of Public Health*, 76:1102–1104, 1986.

A. S. Ash and R. P. Ellis. Risk-adjusted payment and performance assessment for primary care. *Medical Care*, 50:643–653, 2012. doi: 10.1097/MLR.0b013e3182549c74.

P. B. Bach and R. H. Jain. Physician's office and hospital outpatient setting in oncology: It's about prices, not use. *Journal of Oncology Practice*, 13:4–5, 2017.

P. Barros and G. Braun. Upcoding in a national health service: the evidence from portugal. *Health Economics*, 26:600–618, 2017. doi: 10.1007/s10754-020-09287-x.

M. Batty and B. Ippolito. Financial incentives, hospital care, and health outcomes: Evidence from fair pricing laws. *American Economic Journal: Economic Policy*, 9:28–56, 2017.

C. Bojke, K. Grasic, and A. Street. How should hospital reimbursement be refined to support concentration of complex care services? *Health Economics*, 27:e26–e38, 2016. doi: 10.1002/hec.3525.

British Orthopaedic Association. The care of patients with fragility fracture. 2007. URL https://www.bgs.org.uk/sites/default/files/content/attachment/2018-05-02/Blue%20Book%20on%20fragility%20fracture%20care.pdf.

R. Busse, A. Geissler, A. Aaviksoo, A. Cots, U. Hakkinent, and et al. Diagnosis-related groups in europe moving towards transparency, efficiency and quality in hospitals. 347: 1–7, 2013a. doi: https://doi.org/10.1136/bmj.f3197.

R. Busse, A. Geissler, W. Quentin, and M. Wiley. Diagnosis related groups in europe: moving towards transparency, efficiency, and quality in hospitals? *BMJ (Clinical research ed.)*, 347:1–7, 2013b. doi: https://doi.org/10.1136/bmj.f3197.

I. D. Cameron. Coordinated multidisciplinary rehabilitation after hip fracture. *Disability and Rehabilitation*, 27(18-19):1081–1090, 2005. doi: https://doi.org/10.1080/09638280500061261.

CAPH. 2018. URL https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD007710.pub3/abstract. Last accessed on 30th of July 2020.

C. Cashin, Y. Chi, P. Smith, M. Borowith, and S. Thomson. Paying for performance in healthcare. *European Observatory on Health Systems and Policies Series*, 2014a.

C. Cashin, Y. L. Chi, P. Smith, M. Borowitz, and S. Thomson. Paying for performance in health care. implications for health system performance and accountability. *European Observatory on Health Systems and Policies Series*, 2014b. URL https://www.euro.who.int/__data/assets/pdf_file/0020/271073/Paying-for-Performance-in-Health-Care.pdf.

N. Casteleijn, J. Vriesema, S. Stomps, O. van Balen, and E. Cornel. The effect of office based flexible and rigid cystoscopy on pain experience in female patients. *Investigative and Clinical Urology*, 58:48–53, 2017.

X. Castells, J. Alonso, M. Castilla, C. Ribó, F. Cots, and J. M. Antó. Outcomes and costs of outpatient and inpatient cataract surgery: a randomised clinical trial. *Journal of Clinical Epidemiology*, 54:23–29, 2001.

M. Chalkley and J. Malcomson. Contracting for health services with unmonitored quality. *The Economic Journal*, 108:1093–1110, 1998. doi: https://doi.org/10.1111/1468-0297.00331.

Citizen Advice . Nhs patients rights. 2020. URL https://www.citizensadvice.org.uk/health/nhs-healthcare/nhs-patients-rights/. Last accessed 4th of August 2020.

J. Clemens and J. D. Gottlieb. Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review*, 104, 2013.

A. Cook and S. Averett. Do hospitals respond to changing incentive structures? evidence from medicare's 2007 drg restructuring. *Journal of Health Economics*, 73, 2020. doi: https://doi.org/10.1016/j.jhealeco.2020.102319.

C. Cooper, Z. A. Cole, C. R. Holroyd, S. C. Earl, N. C. Harvey, E. M. Dennison, L. J. Melton, S. R. Cummings, J. A. Kanis, and IOF CSA Working Group on Fracture Epidemiology. Secular trends in the incidence of hip and other osteoporotic fractures. *Osteoporosis International*, 22:1277–1288, 2011. doi: 10.1007/s00198-011-1601-6.

E. M. Curtis, R. van der Velde, R. J. Moon, J. P. W. van den Bergh, P. Geusens, F. de Vries, T. P. van Staa, C. Cooper, and N. C. Harvey. Epidemiology of fractures in the united kingdom 1988-2012: Variation with age, sex, geography, ethnicity and socioeconomic status. *Bone*, 87, 2016. doi: 10.1016/j.bone.2016.03.006.

L. S. Dafny. How do hospitals respond to price changes? *American Economic Review*, 95: 1525–1547, 2005.

C. Davis and D. J. Rhodes. The impact of drgs on the cost and quality of health care in the united states. *Health Policy*, 9, 1988. doi: https://doi.org/10.1016/j.jhealeco.2022.102581.

K. Davis and L. Russell. The substitution of hospital outpatient care for inpatient care. *The Review of Economics and Statistics*, 54:109–120, 1972.

Department of Health. A simple guide to payment by results. 2010. URL http://data.parliament.uk/DepositedPapers/Files/DEP2010-2028/DEP2010-2028.pdf.

Department of Health. Payment by results guidance for 2012-13. 2012. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/216212/dh_133585.pdf.

A. Dobson, A. M. El-Gamil, M. T. Shimer, J. E. DaVanzo, A. Q. Urbanes, G. A. Beathard, and T. F. Litchfield. Clinical and economic value of performing dialysis vascular access procedures in a freestanding office-based center as compared with the hospital outpatient department among medicare esrd beneficiaries. *Seminars in dialysis*, 26, 2013. doi: 10.1111/sdi.12120.

V. Doria-Rose, P. A. Newcomb, and T. R. Levin. Incomplete screening flexible sigmoidoscopy associated with female sex, age, and increased risk of colorectal cancer. *Gut*, 54:1273–1278, 2005. doi: 10.1136/gut.2005.064030.

K. Eggleston. Multitasking and mixed systems for provider payment. *Journal of health economics*, 24:211–223, 2005. doi: 10.1016/j.jhealeco.2004.09.001.

F. Eijkenaar. Pay for performance in health care: an international overview of initiatives. *Medical Care Research and Review*, 69, 2012.

F. Eijkenaar. Key issues in the design of pay for performance programs. *The European Journal of Health Economics*, 110:115–130, 2013. doi: https://doi.org/10.1007/s10198-011-0347-6.

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

R. P. Ellis and T. G. McGuire. Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of health economics*, 5:129–151, 1986a. doi: https://doi.org/10.1016/0167-6296(86)90002-0.

R. P. Ellis and T. G. McGuire. Provider behavior under prospective reimbursement. cost sharing and supply. *Journal of Health Economics*, 5:129–151, 1986b.

R. A. Elnicki. Substitution of outpatient for inpatient hospital care: a cost analysis. *Inquiry*, 13:245–261, 1976.

M. Emmert, F. Eijkenaar, H. Kemter, A. S. Esslinger, and O. Schöffski. Economic evaluation of pay-for-performance in health care: a systematic review. *The European Journal of Health Economics*, 13:755–767, 2012. doi: 10.1007/s10198-011-0329-8.

N. England. National tariff system 2014/15. 2014. URL https://www.gov.uk/government/publications/national-tariff-payment-system-2014-to-2015.

D. M. Esparza, N. Young, and J. A. Luongo. Effective planning for office and outpatient chemotherapy administration. *Seminars in oncology nursing*, 5, 1989. doi: https://doi.org/10.1016/0749-2081(89)90077-6.

C. Falavolti, F. Sergi, T. Petitti, and M. Buscarini. Does listening to music during flexible cystoscopy and bladder biopsy decrease patient's pain? *Journal of Clinical Urology*, 2017. doi: 10.1177/2051415816666689.

S. Farrar, D. Yi, M. Sutton, M. Chalkley, J. Sussex, and A. Scott. Has payment by results affected the way that english hospitals provide care? difference-in-differences analysis. *BMJ*, 339, 2009. doi: https://doi.org/10.1136/bmj.b3047.

M. D. Fisher, R. Punekar, and Y. M. Yim. Differences in health care use and costs among patients with cancer receiving intravenous chemotherapy in physician offices versus in hospital outpatient settings. *Journal of Oncology Practice*, 13:e37–e44, 2017.

P. K. Foo, R. S. Lee, and K. Fong. Physician prices, hospital prices, and treatment choice in labor and delivery. *American Journal of Health Economics*, 3:422–453, 2017.

J. Gaughan, N. Gutacker, K. Grasic, N. Kreif, L. Siciliani, and A. Street. Paying for efficiency: Incentivising same-day discharges in the english nhs. *Journal of Health Economics*, 68, 2019.

F. Genovese, G. D'Urso, F. D. Guardo, G. Insalaco, A. Tuscano, L. Ciotta, A. Carbonaro, V. Leanza, and M. Palumbo. Failed diagnostic hysteroscopy: Analysis of 62 cases. *European journal of obstetrics, gynecology, and reproductive biology*, 245:193–197, 2020. doi: 10.1016/j.ejogrb.2019.10.031.

L. Gordan, M. Blazer, V. Saundankar, D. Kazzaz, S. Weidner, and M. Eaddy. Cost differential of immuno-oncology therapy delivered at community versus hospital clinics. *The American Journal of Managed Care*, 25:e66–e70, 2019.

D. Grant. Physician financial incentives and cesarean delivery: New conclusions from the healthcare cost and utilization project. *Journal of Health Economics*, 28:244–250, 2009.

K. Grasic, A. Mason, and A. Street. Paying for the quantity and quality of hospital care: The foundations and evolution of payment policy in england. *Health Economics Review*, 5:1–10, 2013.

J. A. Greenberg. Hysteroscopic sterilization: History and current methods. *Reviews in Obstetrics and Gynecology*, 1:113–121, 2008.

A. Greenstein, I. Greenstein, S. Senderovich, and N. J. Mabjeesh. Is diagnostic cystoscopy painful? analysis of 1,320 consecutive procedures. *International Brazilian Journal of Urology*, 40, 2014. doi: 10.1590/S1677-5538.IBJU.2014.04.13.

M. Groß, H. Jürges, and D. Wiesen. The effects of audits and fines on upcoding in neonatology. *Health Economics*, 30, 2021. doi: https://doi.org/10.1002/hec.4272.

J. Gruber and M. Owings. Physician financial incentives and cesarean section delivery. *RAND Journal of Economics*, 27:99–123, 1996.

J. Gruber, J. Kim, and D. Mayzlin. Physician fees and procedure intensity: The case of cesarean delivery. *Journal of Health Economics*, 18:473–490, 1999.

N. Gutacker, L. Siciliani, G. Moscelli, and H. Gravelle. Choice of hospital: Which type of quality matters? *Journal of Health Economics*, 50, 2016. doi: 10.1016/j.jhealeco.2016.08.001.

J. Hadley and K. Swartz. The impacts on hospital costs between 1980 and 1984 of hospital rate regulation, competition, and changes in health insurance. *Inquiry*, 26:35–47, 1989.

J. Hadley, S. Zuckerman, and J. Feder. Profits and fiscal pressure in the prospective payment system: Their impacts on hospitals. *Inquiry*, 26:354–365, 1989.

R. Harker. Nhs funding and expenditure. 2014. URL https://commonslibrary.parliament.uk/research-briefings/sn00724/.

J. Hayes, J. R. Hoverman, M. E. Brow, D. C. Dilbeck, D. K. Verrilli, J. Garey, J. L. Espirito, J. Cardona, and R. Beveridge. Cost differential by site of service for cancer patients receiving chemotherapy. *The American Journal of Managed Care*, 21:189–196, 2015.

D. He and J. M. Mellor. Do changes in hospital outpatient payments affect the setting of care? *Health Services Research*, 48, 2013.

H. Hennig-Schmidt, H. Jürges, and D. Wiesen. Dishonesty in health care practice: A behavioral experiment on upcoding in neonatology. *Health Economics*, 28, 2018. doi: https://doi.org/10.1002/hec.3842.

S. V. Herwaarden, I. Wallenburg, J. Messelink, and R. Bal. Opening the black box of diagnosis-related groups (drgs): unpacking the technical remuneration structure of the dutch drg system. *Health Economics, Policy and Law*, 12:196–209, 2020. doi: 10.1017/S1744133118000324.

A. Higgins, G. Veselovskiy, and J. Schinkel. National estimates of price variation by site of care. *American Journal of Managed Care*, 22:e116–e121, 2016.

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

D. Hodgkin and T. G. McGuire. Payment levels and hospital response to prospective payment. *Journal of Health Economics*, 13:1–29, 1994.

B. Hollingsworth. The measurement of efficiency and productivity of health care delivery. *Health Economics*, 17:1107–1128, 2008.

S. D. Horn, R. A. Horn, P. D. Sharkey, and A. F. Chambers. Severity of illness within drgs. homogeneity study. *Medical Care*, 24, 1986. doi: 10.1097/00005650-198603000-00005.

P. Hussey, H. de Vries, J. Romley, M. Wang, S. Chen, P. Shekelle, and E. McGlynn. A systematic review of health care efficiency measures. *Health Services Research*, 44: 784–805, 2009.

P. D. Iaco, A. Marabini, M. Stefanetti, C. D. Vecchio, and L. Bovicelli. Acceptability and pain of outpatient hysteroscopy. *Journal of the American Association of Gynecologic Laparoscopists*, 7:71–75, 2000.

J. Januleviciute, J. E. Askildsen, O. Kaarboe, L. Siciliani, and M. Sutton. How do hospitals respond to price changes? evidence from norway. *Health Economics*, 25:620–636, 2016. doi: 0.1002/hec.3179.

H. Jürges and J. Köberlein. What explains drg upcoding in neonatology? the roles of financial incentives and infant health. *Journal of Health Economics*, 43:13–26, 2015. doi: 10.1016/j.jhealeco.2015.06.001.

C. Kaarboe and L. Siciliani. Multi-tasking, quality and pay for performance. *Health Economics*, 20, 2011. doi: https://doi.org/10.1002/hec.1582.

P. H. S. Kalmet, B. B. Koc, B. Hemmes, R. H. M. ten Broeke, G. Dekkers, P. Hustinx, M. G. Schotanus, P. Tilman, H. M. J. Janzing, J. M. A. Verkeyn, P. R. G. Brink, and M. Poeze. Effectiveness of a multidisciplinary clinical pathway for elderly patients with hip fracture: A multicenter comparative cohort study. *Geriatric Orthopaedic Surgery and Rehabilitation*, 7:81–85, 2016. doi: 10.1177/2151458516645633.

S. B. Kelly, J. Murphy, A. Smith, H. Watson, S. Gibb, C. Walker, and R. Reddy. Nurse specialist led flexible sigmoidoscopy in an outpatient setting. *Colorectal Disease*, 10(4): 390–393, 2008. doi: 10.1111/j.1463-1318.2007.01271.x.

J. T. Kolstad and A. E. Kowalski. Mandate-based health reform and the labor market: Evidence from the massachusetts reform. *Journal of Health Economics*, 47:81–106, 2016. doi: https://doi.org/10.1016/j.jhealeco.2016.01.010.

M. Kuhn and L. Siciliani. Performance indicators for quality with costly falsification. *Journal of Economics & Management Strategy*, 18:1137–1154, 2009. doi: https://doi.org/10.1111/j.1530-9134.2009.00240.x.

T. W. Lau, C. Fang, and F. Leung. The effectiveness of a multidisciplinary hip fracture care model in improving the clinical outcome and the average cost of manpower. *Osteoporsis International*, 28:791–798, 2017. doi: 10.1007/s00198-016-3845-7.

S. Leader and M. Moon. Medicare trends in ambulatory surgery. *Health Affairs*, 8, 1989. doi: https://doi.org/10.1377/hlthaff.8.1.158.

D. J. Lee and J. C. Elfar. Timing of hip fracture surgery in the elderly. *Geriatric Orthopaedic Surgery and Rehabilitation*, 5:138–140, 2014. doi: 10.1177/2151458514537273.

J. Lee, S. Doumouchtsis, S. Jeffery, and M. Fynes. Evaluation of outpatient cystoscopy in urogynaecology. *Archives of Gynecology and Obstetrics*, 279, 2009. doi: 10.1007/s00404-008-0773-6.

D. Lisi, L. Siciliani, and O. R. Straume. Hospital competition under pay-for-performance: Quality, mortality, and readmissions. *Journal of Economics & Management Strategy*, 29:289–314, 2020. doi: https://doi.org/10.1111/jems.12345.

L. E. López-Cortés, M. D. del Toro, J. Gálvez-Acebal, E. Bereciartua-Bastarrica, M. C. Fariñas, M. Sanz-Franco, C. Natera, J. E. Corzo, J. M. Lomas, J. Pasquau, A. del Arco, M. P. Martínez, A. Romero, M. A. Muniain, M. de Cueto, Pascual, J. Rodríguez-Baño, for the REIPI/SAB group, C. Velasco, F. J. Caballero, M. Montejo, J. Calvo, M. Aller-Fernández, L. Martínez, M. D. Rojo, and V. Manzano-Gamero. Impact of an Evidence-Based Bundle Intervention in the Quality-of-Care Management and Outcome of Staphylococcus aureus Bacteremia. *Clinical Infectious Diseases*, 57(9):1225–1233, 08 2013. ISSN 1058-4838. doi: 10.1093/cid/cit499. URL https://doi.org/10.1093/cid/cit499.

A. Ma. Health care payment systems: Cost and quality incentives. *Journal of Economics and Management Strategy*, 3:93–112, 1994a.

C. T. A. Ma. Health care payment systems: cost and quality incentives. *Journal of Economics & Management Strategy*, 3:93–112, 1994b. doi: https://doi.org/10.1111/j.1430-9134.1994.00093.x.

H. Y. Mak. Managing imperfect competition by pay for performance and reference pricing. *Journal of health economics*, 57:131–146, 2018. doi: 10.1016/j.jhealeco.2017.11.002.

J. Mao, S. Pfeifer, P. Schlegel, and A. Sedrakyan. Safety and efficacy of hysteroscopic sterilization compared with laparoscopic sterilization: an observational cohort study. *BMJ*, 2015. doi: https://doi.org/10.1136/bmj.h5162.

A. A. Markovitz and A. M. Ryan. Pay-for-performance: Disappointing results or masked heterogeneity? *Medical Care Research and Review*, 74, 2017. doi: 10.1177/1077558715619282.

R. McDonald, S. Zaidi, S. Todd, F. Konteh, K. Hussain, J. Roe, and T. Allen. A qualitative and quantitative evaluation of the introduction of best practice tariffs: An evaluation report commissioned by the department of health. 2012.

A. Mendelson, K. Kondo, C. Damberg, A. Low, M. Motúapuaka, M. Freeman, M. O'Neil, R. Relevo, and D. Kansagara. The effects of pay-for-performance programs on health, health care use, and processes of care: A systematic review. *Annals of Internal Medicine*, 166:341–353, 2017. doi: 10.7326/M16-1881.

D. Metcalfe, C. K. Zogg, A. Judge, D. C. Perry, B. Gabbe, K. Willett, and M. L. Costa. Pay for performance and hip fracture outcomes. *The Bone & Joint Journal*, 101-B: 1015–1023, 2019. doi: https://doi.org/10.1302/0301-620X.101B8.BJJ-2019-0173.R1.

N. Mihailovic, S. Kocic, and M. Jakovljevic. Review of diagnosis-related group-based financing of hospital care. *Health Services Research and Managerial Epidemiology*, 3, 2016. doi: https://doi.org/10.1177/2333392816647892.

C. Milcent. From downcoding to upcoding: Drg based payment in hospitals. *International Journal of Health Economics and Management*, 21:1–26, 2020. doi: 10.1007/s10754-020-09287-x.

R. Milstein and J. Schreyoegg. Pay for performance in the inpatient sector: A review of 34 p4p programs in 14 oecd countries. *Health Policy*, 120:1125–1140, 2016.

P. Murthy, J. Edwards, and M. Pathak. Update on hysteroscopic sterilisation. *The Obstetrician adn Gynaecologist*, 22:227–235, 2017.

J. Neuburger, C. Currie, R. Wakeman, C. Tsang, F. Plant, B. D. Stavola, D. Cromwell, and J. van der Meulen. The impact of a national clinician-led audit initiative on care and mortality after hip fracture in england. *Medical care*, 53:686–691, 2015. doi: 10.1097/MLR.0000000000000383.

J. Neuburger, C. Currie, R. Wakeman, A. J. adn C. Tsang, F. Plant, H. Wilson, D. A. Cromwell, J. van der Meulen, and B. D. Stavola. Increased orthogeriatrician involvement in hip fracture care and its impact on mortality in england. *Age and ageing*, 46:187–192, 2017. doi: 10.1093/ageing/afw201.

NHFD. The national hip fracture database national report 2010. 2010. URL https://www.nhfd.co.uk/20/hipfractureR.nsf/945b5efcb3f9117580257ebb0069c820/7de8dac5ec3b468980257d4f005188f2/FILE/NHFD2010Report.pdf.

NHS Digital. Hospital admitted patient care activity 2018-19. 2019a. URL https://digital.nhs.uk/data-and-information/publications/

statistical/hospital-admitted-patient-care-activity/2018-19. Last accessed 4th of August 2020.

NHS Digital. Hospital outpatient activity 2018-19. 2019b. URL ttps://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/2018-19. Last accessed 4th of August 2020.

NHS Direct. 2018. URL https://www.nhs.uk/conditions/hysteroscopy/. Acessed on 28th of July 2020.

NHS Direct. 2020. URL https://www.nhs.uk/conditions/cystoscopy/. Acessed on 28th of July 2020.

NHS England and NHS Improvement. 2019/20 national tariff payment system. 2019. URL https://improvement.nhs.uk/documents/4980/1920_National_Tariff_Payment_System.pdf.

NHS England and Monitor. Populus summary: Outpatient appointment referrals. 2015. URL https://www.england.nhs.uk/wp-content/uploads/2015/09/monitor-nhse-outpatient-appointments-summary.pdf. Last accessed 4th of August 2020.

OECD. Health care systems: Getting more value for money. *OECD Economics Department Policy Notes*, No. 2, 2010.

OECD. Oecd reviews of health care quality: United kingdom 2016. 2016. doi: https://dx.doi.org/10.1787/9789264239487-en.

Y. Ogundeji, J. Bland, and T. Sheldon. The effectiveness of payment for performance in healthcare: a meta-analysis and exploration of variation in outcomes. *Health Policy*, 120:1141–1150, 2016a.

Y. K. Ogundeji, J. M. Bland, and T. A. Sheldon. The effectiveness of payment for performance in health care: A meta-analysis and exploration of variation in outcomes. *Health Policy*, 120, 2016b. doi: 10.1016/j.healthpol.2016.09.002.

ONS. General health in england and wales: 2011 and comparison with 2001. 2013. URL https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/articles/generalhealthinenglandandwales/2013-01-30.

S. PA. Ensuring pbr supports delivery of effective cancer services. 2008. URL http://data.parliament.uk/DepositedPapers/Files/DEP2008-2386/DEP2008-2386.pdf.

I. Papanicolas and A. McGuire. Do financial incentives trump clinical guidance? hip replacement in england and scotland. *Journal of Health Economics*, 44:25–36, 2015.

C. Propper, M. Sutton, C. Whitnall, and F. Windmeijer. Did 'targets and terror' reduce waiting times in england for hospital care? *B.E. Journal of Economic Analysis and Policy*, 8, 2008.

S. Relph, T. Lawton, M. Broadbent, and M. Karoshi. Failed hysteroscopy and further management strategies. *The Obstetrician & Gynaecologist*, 18(1):65–68, 2016. doi: 10.1111/tog.12261.

R. Resar, P. Pronovost, C. Haraden, T. Simmonds, T. Rainey, and T. Nolan. Using a bundle approach to improve ventilator care processes and reduce ventilator-associated pneumonia. *The Joint Commission Journal on Quality and Patient Safety*, 31(5):243 – 248, 2005. ISSN 1553-7250. doi: https://doi.org/10.1016/S1553-7250(05)31031-2. URL http://www.sciencedirect.com/science/article/pii/S1553725005310312.

W. R. Robinson and J. Beyer. Impact of shifting from office- to hospital-based treatment facilities on the administration of intraperitoneal chemotherapy for ovarian cancer. *Journal of Oncology Practice*, 6:232–235, 2010.

Royal College of Obstetricians and Gynaecologists . Femele sterilisation; consent advice no. 3. *Annals of Internal Medicine*, 2016. URL https://www.rcog.org.uk/globalassets/documents/guidelines/consent-advice/consent-advice-3-2016.pdf.

Royal College of Physicians. *Outpatients: The future*. 2018. URL https://www.rcplondon.ac.uk/file/11479/download.

F. Salm, F. Schwab, C. Geffers, P. Gastmeier, and B. Piening. The implementation of an evidence-based bundle for bloodstream infections in neonatal intensive care units in germany: A controlled intervention study to improve patient safety. *Infection Control and Hospital Epidemiology*, 37:798–804, 2016. doi: 10.1017/ice.2016.72.

A. Scott, M. Liu, and J. Yong. Financial incentives to encourage value-based health care. *Medical Care Research and Review*, 75:3–32, 2016. doi: https://doi.org/10.1177/1077558716676594.

M. Sharma. Manual vacuum aspiration: an outpatient alternative for surgical management of miscarriage. *The Obstetrician & Gynaecologist*, 17, 2015. doi: https://doi.org/10.1111/tog.12198.

E. Shin. Hospital responses to price shocks under the prospective payment system. *Health Economics*, 28:245–260, 2019.

V. Stephani, W. Quentin, and A. Geissler. Beyond drg-based hospital payment: How countries pay for variable, specialized and low volume care. *European Journal of Public Health*, 27, 2017. doi: https://doi.org/10.1093/eurpub/ckx187.188.

G. S. Tajeu, E. Delzell, W. Smith, T. Arora, J. R. Curtis, K. G. Saag, M. A. Morrisey, H. Yun, and M. L. Kilgore. Death, debility, and destitution following hip fracture. *The journals of gerontology: Series A, Biological sciences and medical sciences.*, 69:346–353, 2014. doi: 10.1093/gerona/glt105.

Y. Takesue, T. Ueda, H. Mikamo, S. O. nad S. Takakura, Y. Kitagawa, and S. Kohnu. Management bundles for candidaemia: the impact of compliance on clinical outcomes. *Journal of Antimicrobial Chemotherapy*, 70:587–593, 2015. doi: 10.1093/jac/dku414.

C. Tsang and D. Cromwell. Statistical methods developed for the national hip fracture database annual report, 2014. 2014. URL https://www.nhfd.co.uk/files/2014ReportPDFs/NHFD2014CEUTechnicalReport.pdf.

P. Van Herck, D. De Smedt, L. Annemans, R. Remmen, M. B. Rosenthal, and W. Sermeus. Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Services Research*, 2010. doi: 10.1186/1472-6963-10-247.

S. van Hoof, T. Quanjel, M. Kroese, M. Spreeuwenberg, and D. Ruwaard. Substitution of outpatient hospital care with specialist care in the primary care setting: A systematic review on quality of care, health and costs. *PLoS One*, 14, 2019.

L. Vaughan, M. Bardsley, D. B. D, and et al. Models of generalist and specialist care in smaller hospitals in england: a mixed-methods study. *Health Services and Delivery Research*, 9, 2021. URL https://www.ncbi.nlm.nih.gov/books/NBK568031/.

R. Verzulli, G. Fiorentini, M. L. Bruni, and C. Ugolini. Price changes in regulated healthcare markets: Do public hospitals respond and how? *Health Economics*, 26:1429–1446, 2016. doi: https://doi.org/10.1002/hec.3435.

K. Vitikainen, M. Linna, and A. Street. Substituting inpatient for outpatient care: what is the impact on hospital costs and efficiency? *The European Journal of Health Economics*, 11:395–404, 2010.

F. Vlaanderen, M. Tanke, B. Bloem, M. Faber, F. Eijkenaar, F. Schut, and et al. Design and effects of outcome-based payment models in healthcare: a systematic review. *European Journal of Health Economics*, 20:217–232, 2019.

D. J. Wellenstein, H. W. Schutte, H. M., J. Honings, P. Belafsky, G. Postma, R. Takes, and G. van den Broek. Office-based procedures for diagnosis and treatment of esophageal pathology. *Head & Neck*, 39, 2017. doi: 10.1002/hed.24819.

C. F. Yen, H. H. Chou, H. M. Wu, C. L. Lee, and T. C. Chang. Effectiveness and appropriateness in the application of office hysteroscopy. *Journal of the Formosan Medical Association*, 118:1480–1487, 2019.

W. C. Yip. Physician response to medicare fee reductions: Changes in volume of coronary artery bypass graft (cabg) surgeries in the medicare and private sectors. *Journal of Health Economics*, 17:675–696, 1998.