Three Essays on Hospital Doctor Incentives

Luis Cardoso Fernandes

PhD

University of York

Economics and Related Studies

October 2021

Abstract

This thesis comprises three essays on the effect of financial and non-financial incentives on the clinical behaviour and labour supply decisions of hospital doctors working in the English National Health Service (NHS).

Chapter 1 examines how the hospital environment in which surgeons practice affects their treatment decisions between cemented and cementless hip replacements. Using quasi-random changes in hospital environment brought about by surgeons' job moves, the analysis shows robust evidence of large effects of the clinical environment on surgeons' treatment choices.

Chapter 2 analyses whether the clinical activity of hospital doctors is affected by the receipt of performance-related financial awards known as the Clinical Excellence Awards. Using panel data information on NHS doctors' characteristics and award status, it finds limited empirical evidence of a negative effect on clinical activity measures.

Chapter 3 studies the responsiveness of doctors' labour supply to changes in take-home pay from NHS work. It exploits the 2016 UK pension reform which reduced take-home pay disproportionately for high earners. The analysis shows that hospital doctors affected by the reform responded by reducing their clinical activity.

Contents

List of Figures

List of Tables

Acknowledgements

First of all I would like to thank my supervisors Nils Gutacker and Martin Chalkley for all their invaluable support and expert advice throughout my time as a PhD student. They are exceptional mentors and incredible human beings. I would also like to express my gratitude to the Centre for Health Economics (CHE) for providing an incredible environment for my studies. My thanks also go to all organisers, supervisors and fellows of the Marie Curie PhD programme for their feedback and the extensive network I was able to build across all Europe. I thank the members of my advisory panel, Luigi Siciliani and Karen Bloor, for their guidance and comments on my work.

I would also like to thank my PhD buddy, Laurie, for all the good times in this journey and also for her friendship. Big thanks to my friend Dominic for the emotional support in the last weeks of writing up this thesis.

I dedicate this thesis to my loving partner, Taner, and to my mom, dad, brother, and sister, without whom this thesis would not be possible.

Declaration

I declare that this thesis is a presentation of original work and that I am the sole author, except where co-authorship is explicitly acknowledged. Funding for my studies was provided by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721402. Data for all chapters come from the Hospital Episode Statistics and are re-used with the permission of NHS Digital (Data Sharing Agreement DARS-NIC-84254-J2G1Q). All rights reserved. Data on consultant characteristics were obtained from the General Medical Council. No ethical approval was required for the work presented in this thesis.

Chapter 1 "*The Impact of the Hospital Environment on Surgeons' Treatment Choice*" is co-authored with Dr Nils Gutacker and Prof Martin Chalkley. I am the main author of this essay, having defined the theoretical and empirical model, assembled the data, constructed the variables, carried out the empirical analysis and written up the paper. All co-authors provided advice and comments during the development of the work and were involved in editing the paper. Previous versions of this paper were presented and discussed at the European Training Network Research in Progress workshop in Rotterdam (the Netherlands), the European Health Economics Association (EuHEA) conference in Maastricht (the Netherlands), the EuHEA PhD student-supervisor conference in Catania (Italy), and the Health Economists' Study Group meeting in York.

Chapter 2 "*The Effect of Winning Financial Awards on Consultant Activity Rates in the English NHS*" is single-authored. Previous versions of this paper were presented and discussed at the European Training Network research in progress workshop in Odense (Denmark), and at the "Methodological Workshop Concerning Working with Administrative Data" in Rotenburg (Germany).

Chapter 3 "*Doctors' Take-Home Pay and NHS Activity: Evidence from a UK Pension Reform*" is co-authored with Dr Nils Gutacker and Prof Martin Chalkley. I am the main author of this essay, having defined the theoretical and empirical model, assembled the data, constructed the variables, carried out the empirical analysis and written up the paper. All co-authors provided advice and comments during the development of the work and were involved in editing the paper. Previous versions of this paper were presented and discussed at the Health Economists' Study Group meeting in Cambridge (virtual meeting).

I affirm that this thesis has not previously been presented for an award at this or any other university or educational institution. Any views expressed in this document are exclusive responsibility of the author. All sources are acknowledged as References.

Introduction

Economic framework

Many markets are characterised by information asymmetries in which one of the parties has more or better information than the other. In health care markets, doctors are generally better informed than patients[1] about both the nature of illnesses (i.e. diagnostic information) and the effectiveness of alternative prevailing technologies to produce health improvements (i.e. treatment information) (Arrow, 1963). As the less-informed party in the interaction, patients frequently delegate some decision-making authority to a doctor to act on their behalf[2]. In return, the doctor receives payment — either directly or indirectly through third parties — for the information and services provided to the patient (Hurley, 2000). Health economists perceive this agreement between the doctor and the patient as one of *agency* (McGuire, 2000). Agency relationships[3] arise whenever a principal (the patient) delegates a task to an agent (the doctor) in return for compensation (Sappington, 1991; Laffont and Tirole, 1993).

The principal's expectation in an agency relationship is that the agent selects the course of action that is in the best interests of the principal. However, such perfect agency relationship between the patient and the doctor is unlikely to arise in health care markets (Pauly, 1980). The doctor manages the care and preferences of the patient with her own economic motives in mind, particularly with regards to income (which allows consumption), effort and leisure that contribute to own utility (McGuire, 2000). The *agency problem* arises when there is a

[1] The context in which doctor-patient interactions occur further accentuates the information asymmetry between both parties (Hurley, 2000). First, because consumption of health care is typically unpredictable and takes place at times of physical and mental vulnerability (Arrow, 1963), patients are likely to face higher costs of search (e.g. opportunity cost of time) and negotiation (i.e. cost of bargaining and making decisions) in health care exchanges than in other markets (Ferguson and Keen, 1996). Second, the scope to learn from experience is also limited as patients often purchase medical goods and services on a one-off basis, i.e. the consumption is irregular and occurs too infrequently to learn from past experiences (Hurley, 2000).

[2] This arrangement is legally required in some situations. For example, the access to certain medical goods (e.g., prescription drugs) is restricted to licensed health care professionals (Pauly, 1980).

[3] Agency relationships are ubiquitous in many areas of economic activity. Examples include the arrangements between corporate managers and shareholders or lawyers and their clients.

conflict of interests between the doctor and the patient (Hurley, 2000). And because doctors enjoy an information advantage in the relationship, they may exploit it for their own benefit by persuading patients to purchase medical services that they would not otherwise have purchased if fully informed.

Traditional economic models of doctor behaviour assume that doctors act as utility-maximising agents who seek income and leisure (McGuire, 2000). This assumption has led to the belief that under fee-for-service payment systems doctors may exploit their information advantage to induce services which fully informed patients would not choose — physician-induced demand (Evans, 1974). In fact, the two other traditional payment systems also have embedded financial incentives that may influence doctors' clinical decisions; briefly, capitation encourages undertreatment, and salary carries no incentive to exert effort (McGuire, 2000). The likely relationship between doctors' pay and clinical behaviour sprout the idea that specifying remuneration systems with financial incentives that elicit actions compatible with the objectives of patients has the potential to reduce agency costs (i.e. the difference between the ideal outcome and the outcome attained by the doctor) and to encourage doctors to act as better agents for their patients. As a result, a strong culture of performance-related contracting has emerged in many health care systems (Maynard, 2012), fueled by advancements in methods of measuring doctor performance and the quality of care (Fisher, 2006; Institute of Medicine, 2007). Parallel to the event of performance-related pay, commonly known as pay-for-performance schemes (P4P), an empirical literature grew to study the effect of such contracts (Armour et al., 2001; Town et al., 2005; Rosenthal and Frank, 2006; Petersen et al., 2006; Christianson et al., 2008; Scott et al., 2011; Li et al., 2014). However, findings from these studies are ambiguous and inconclusive. Moreover, a few studies also show that when poorly designed, explicit financial effects may give rise to undesirable doctor behaviour with unintended consequences for patient care (Rosenthal and Frank, 2006; Doran et al., 2008; Hutchison, 2008; Gravelle et al., 2010).

Another vexing information problem in doctor-patient interactions is incompleteness. Doctors may be better informed than patients, but they are not necessary perfectly informed. They do not know with certainty both the preferences of their patients and the medical production function (Arrow, 1963). While information asymmetry underlies the phenomenon of physician-induced demand, incomplete information is the motivation for the literature on small area variations (Wennberg and Gittelsohn, 1973). Variations in medical practice, particularly "unwarranted variations", are likely to stem from doctors' uncertainty about the true production function, and the irregular diffusion of clinical information in doctors' networks (Phelps, 2000). When there is uncertainty about "what works" in health care, and an array of alternative treatments is available, Wennberg (1984) argued that clinical choices are driven by doctors' own beliefs and habits — or their "practice style". The theory that doctors develop a "style" spawned a literature trying to identify which factors other than financial incentives drive doctors' behaviour. A few recent studies using innovative identification strategies suggest that factors related to doctors' practice environment (Molitor, 2018), and in particular their peers (Epstein and Nicholson, 2009; Avdic et al., 2021), are likely to matter more for doctors' treatment choices than their medical training. However, further research is needed to assert these claims.

Structure of this thesis

This thesis presents three essays on the role of financial and non-financial incentives on hospital doctors' clinical behaviour and labour supply. Each essay contributes to and extends the current literature. In particular, it offers significant advancements in the understanding of the role of the hospital environment in determining doctors' practice style (chapter 1); the potential unintended consequences of financial incentives on doctors' behaviour (chapter 2), and the labour supply responses of doctors to changes in remuneration (chapter 3).

In Chapter 1, we examine how the hospital environment in which surgeons practice affects their treatment decisions. Despite a large literature exposing the existence of profound and potentially unwarranted variations in medical and surgical rates both within and across geographical areas, the mechanisms that give rise to such variations are poorly understood (Chandra et al., 2011). Previous work shows that traditional patient characteristics (e.g., income, preferences, and health needs) and supply-side (e.g., hospital size, management practice, peer and organisational pressure, and the experience of doctors) are both equally important and have economically significant effects (Finkelstein et al., 2016). Less is, however, known about the relative importance of specific supply-side characteristics.

In this chapter, we investigate surgeons' choice of cemented vs. cementless hip replacements, which are two common methods of fixation in this surgical procedure. This clinical setting has a number of attractive features which improve the findings of previous work (Molitor, 2018; Avdic et al., 2021): i) there is no clear superiority in effectiveness of one procedure over the other, ii) both procedures are equally reimbursement to hospitals, iii) surgeons are trained in both approaches.

Our empirical approach follows a recent literature in health economics that contrasts the behaviour of individuals who change environments ('movers') to those that remain in their usual environment ('stayers'). Using quasi-random changes in hospital environment brought about by job moves of orthopaedic surgeons between April 2010 and March 2017, we identify the effect of the hospital environment on surgeons' treatment choices. We find that after the move, a surgeon's treatment choice changes on average by 6.4 percentage points for each 10-percentage point change in the clinical environment. We also show that surgeons adapt their behaviour fairly quickly following the move, with most of the adaptation occurring within the first year. Reassuringly, the change in practice style has no negative effects on three common quality metrics: 28-day emergency readmissions, one-year revisions, and change in patient-reported outcome measures (PROMs). Finally, we find no evidence of positive sorting of surgeons to hospitals, which could bias our findings.

Chapter 2 analyses how the clinical activity of consultants is affected by the receipt of a Clinical Excellence Award (CEA). CEAs are the main performance-related pay scheme available to doctors working in the English NHS. There are four levels of awards with associated increasing pay: Bronze, Silver, Gold, and Platinum. The scheme rewards doctors for achievements in the quality of care provided as well as for certain non-clinical activities (e.g. research and teaching). However, they do not directly relate to clinical activity rates, which may suffer if doctors pursue other incentivised activities to generate additional income through CEAs.

Using panel data information on NHS doctors between 2009 and 2015, I estimate a series of Poisson fixed-effects panel models that relate clinical activity to doctor characteristics and current award status. I measure clinical activity using the count of completed inpatient episodes of care for each consultant, and an adjusted measure I develop which weights each episode of care using the national average length of stay (LOS) for the respective Healthcare Resource Group (HRG). To complement may analysis, I also examine the effect of loosing an award on the clinical activity of consultants.

To account for the dynamic effects of the awards on clinical activity, I also extend the models to include leads and lags of the award in my analysis.

I find some empirical evidence of a negative effect of the two lowest level awards (Bronze and Silver) on clinical activity. Furthermore, I show that consultants do not change their clinical activity in pursuit of the awards, which could bias my results. In conclusion, the awards may have unintended consequences for the activity rates of consultants.

In Chapter 3, we study the responsiveness of doctors' labour supply to changes in effective income from NHS work. We exploit the 2016 UK pension reform which introduced tighter annual allowances for pension contributions of individuals with annual incomes in excess of £110,000. This policy affected many hospital consultants working in the English NHS and created a quasi-random shock in their take-home pay. We link three data sources to build a unique and rich panel data set with information on doctors' NHS inpatient activity, pay, and individual characteristics covering the years 2013 to 2018.

We find that consultants with estimated income derived from NHS work in excess of £110,000 experienced a 8.4 percent drop in activity rates post-reform as well as a similar relative decline in days of patient care - relative to doctors unaffected by the reform. A series of robustness checks, including a placebo test, confirm the findings. The UK pension reform, therefore, may have had detrimental consequences for patient treatment in the English NHS.

The last chapter of this thesis summarises key findings, discusses policy implications, and presents suggestions for future research.

Chapter 1

The Impact of the Hospital Environment on Surgeons' Treatment Choices

1 Introduction

Geographic variations in medical treatments are widely documented in the UK and elsewhere[4], but their causes remain poorly understood. A sizeable body of empirical literature has shown that these variations cannot be explained by traditional demand-side factors such as patient preferences, income, and health needs alone (see, for example, Barnato et al. (2007); Anthony et al. (2009); Song et al. (2010); O'Hare et al. (2010); Zuckerman et al. (2010); Chandra et al. (2011); Skinner (2011)). This finding has important economic implications. Variation that cannot be explained by the needs and preferences of patients may imply inefficiency (e.g., waste of scarce health care resources and sub-optimal health outcomes) if some patients are receiving excessively intensive treatment, and inequality, when patients with the same need receive different treatment (Wennberg, 2002; Appleby et al., 2011; OECD, 2014).

The importance of provider supply-side factors relative to patient demand-side factors has been demonstrated recently with the empirical work of Finkelstein et al. (2016). The authors suggest that almost half of the variation in health care utilisation amongst beneficiaries in the US Medicare system is driven by supply-side factors, whilst the remaining half is due to demand-side characteristics[5].

[4] Beginning with the seminal studies of Glover (1938) in England and Wales and Wennberg and Gittelsohn (1973) in the US, which uncovered large geographical variations in the rates of tonsillectomy, the literature on geographic variations grew quickly to show persistent differences in medical treatments across different institutional settings and patient populations (e.g., Skinner (2011); Corallo et al. (2014); OECD (2014))

[5] Similar results have been found in the Netherlands (Moura et al., 2019) and in Spain (Prieto and Lago-Peñas, 2012). Exceptionally, Salm and Wübker (2020) find that most of the variation in outpatient care in Germany is driven by demand-side factors, which they suggest to be due to constraints in the supply side.

A supply-side explanation for differences in healthcare delivery can arise from two distinct sources. One possibility is that the hospitals in which doctors work have an impact on their treatment decisions, and, therefore, moving a doctor across hospitals would change his or her practice style. The other possibility is that variations in medical care are due to systematic differences in the type of doctors practicing in each hospital. For example, doctors with aggressive treatment styles, as a result of beliefs or preferences, may flock to the same place, and this gives rise to hospitals with relatively more-aggressive treatment patterns. In this chapter, we assess the relative importance of these hypotheses and provide quasi-experimental evidence of hospital effects on surgeons' treatment choices.

Understanding whether variations in treatment choices are due to the institutional environment where doctors practice or doctors' own beliefs and preferences is a policy relevant question. If treatment choices are determined by doctors' beliefs and preferences alone, and environment is unimportant, policies affecting physician training and incentives (for example, the means of remuneration) will be useful. If the institutional environment is important then a focus on policies directed at institutions, such as hospital-level incentive schemes or tailored purchasing contracts, may be necessary.

We study how quasi-random changes in the clinical environment experienced by senior orthopaedic surgeons (known as consultants) moving across hospitals in England affects their choices between two substitute methods of fixation of replacement hips – cemented and cementless[6] - in hip replacement surgery, which is one of the most common procedures worldwide[7]. Our analysis focuses on the choice between these two methods of fixation for several reasons. First, although cemented and cementless hip replacements have been in use since 1970s, there is a lack of high-quality clinical evidence comparing the effectiveness of both techniques. The few existing studies, which suffer from short follow-up duration and small sample sizes, show no significant difference in patient outcomes as measured by revision, mortality and compli-

[6] Cemented prostheses use cement as a grout to hold the implant to the bone, whereas cementless prostheses achieve stability from being press-fit into the bone and from its porous coating that allows the bone to grow onto and into the component.

[7] Nearly 900,000 primary hip replacements were performed in the UK between 2003 and 2016 (Michael Green et al., 2017), and an estimated 2.5 million people were living with a hip implant in the US in 2010 (Maradit Kremers et al., 2015).

cation rates (see a meta-analysis by Abdulkarim et al. (2013)). Notwithstanding the weak evidence, the National Institute for Health and Care Excellence (NICE) favours cemented implants on the basis of long-term viability and relative cost-effectiveness (cemented prostheses are cheaper) (National Institute for Health and Care Excellence, 2011, 2014). NICE guidelines are non-binding; and the British Orthopaedic Association found in an assessment of 205 NHS hospitals that the choice of implant is not driven by evidence but rather by established local behaviour, surgeon location of original training and marketing by implant companies (Briggs, 2015). Second, NHS hospitals are paid equally for both procedures and surgeons have no financial incentive to choose either implant (Papanicolas and McGuire, 2015). Third, individual hospitals in the NHS control procurement practices with respect to hip prostheses. Managers are allowed to negotiate quantities and bargain prices with the suppliers, and one would therefore expect hospitals to play a significant role in the relative quantity of prosthesis types to be purchased (Davies and Lorgelly, 2013). Fourth, the propensity to perform a cemented versus cementless hip replacement varies substantially across hospitals in England (Briggs, 2015).

To measure the degree to which differences in doctors' treatment decisions are driven by the environment the consultants work in, we make use of patient-level administrative data from all publicly funded care in England, which includes information on age, sex, detailed diagnosis and procedure codes, comorbidities, and identifiers of hospitals and surgeons responsible for the care. We supplement these data with information on surgeons' personal characteristics and construct surgeons' employment histories to identify those who move their practice across hospitals. The exogenous shock in hospital environment that is provided by this quasi-random re-allocation of surgeons to hospitals is then used to estimate hospital effects on treatment choices. Since our interest lies in the treatment decision between cemented and cementless hip implants, we characterise the hospital environment by the rate of cemented hip replacements for all surgeons in the hospital, omitting the moving doctor's own cases. The critical identification assumption underlying our approach is that the timing of the move and the choice of the hospital of destination are independent of surgeons' treatment style. This independence condition would be violated if surgeons were to sort themselves into hospitals according to their preferred surgical approach. For example, surgeons who are more conservative may attract equally conservative surgeons to work in the same hospital. Additionally, we control for observed and unobserved time-invariant individual characteristics of surgeons such as their skills, innate ability, medical school of qualification or the setting of postgraduate training, through a fixed effects strategy.

Our first main finding is that the hospital in which surgeons practice matter for their observed treatment choices: surgeons moving from low to high-cemented environments increase the use of cemented implants, and vice versa. A 10 percentage points change in the hospital environment increases a surgeon's propensity to perform cemented hip replacements by 6.4 percentage points. We do not find evidence that a surgeon's practice style prior to the move varies systematically with the change in environment, suggesting that the timing and choice of destination is unlikely to be correlated with the practice style of surgeons before and after the move. We also show that surgeons adapt their treatment style fairly quickly following the move and with very limited further adjustments over time. Phelps and Mooney (1993) have long posited, in a model of physician learning about treatment efficacy, that physicians adjust to new environments through a learning process of adaptation in which prior beliefs are continuously updated. Our results do not offer support for this theory; however, we cannot rule out habit formation over longer time horizons than observed in our study. The second main result of our work is that the quick change in physician practice style has no negative effects on quality of care, measured by 28-day emergency readmission rates, one-year revision rates, or change in patient-reported outcome measures (PROMs). Finally, we find that surgeons respond symmetrically to changes in the environment, whether they move to high or low-cemented environments.

Our study is related to a scant but growing body of the literature that identifies differences in physician practice styles between and within regions and providers, and examines its causes and consequences (see, for example, Grytten and Sørensen (2003); Epstein and Nicholson (2009); Currie et al. (2016)). Most closely related to our work are the studies by Molitor (2018) and Avdic et al. (2021). Molitor (2018) uses cardiologists who migrate across states in the US to explore how their treatment choices for acute myocardial infarction change with changes in the location of practice. The author finds that the hospital environment largely explains the regional disparities in physician behaviour and puts an estimate on the change in practice style of 0.6-0.8 for every percentage point change in the clinical environment. Avdic et al. (2021) follows a similar approach to study the treatment style of Swedish doctors, and finds that they respond to changes in the clinical environment similarly to their US counterparts (Molitor, 2018).

We build on the work of Molitor (2018) and Avdic et al. (2021) to refine their findings in four ways. First, the clinical setting we examine has clear advantages in that unlike the treatment options in heart attacks, cemented and cementless hip implants are perfect substitutes with similar therapeutic success. Moreover, both implants are available in every hospital across the country, and surgeons are trained in both techniques. Our analysis is therefore purged from the effects of local resource constraints, lack of surgeon skill, and the possible clinical superiority of one method over the other. Second, both procedures are equally reimbursed to providers, and surgeons have no financial incentive driving their choice. Third, in the NHS patients have little choice of provider, much less of surgeon, and surgeons do not choose patients (Barrenho et al., 2021). Fourth, in contrast to previous work, we examine whether the change in practice style experienced by surgeons across their move between hospitals impacts patient health outcomes after surgery.

The rest of the paper is organised as follows. Section 2 presents the institutional background. Sections 3 and 4 discuss the data and empirical strategy. We present and describe our findings in Section 5. And Section 6 concludes.

2 Institutional Background

NHS care is funded by general taxation and free at the point of use. Patients have access to hospital care via General Practitioner (GP) referrals or in case of accident and need for emergency care. Although since 2006 NHS patients are legally entitled to choose the hospital provider (Gaynor et al., 2013), the choice of surgeon is potentially determined by a patient's position in waiting lists and surgeon availability (Barrenho et al., 2021). This eliminates any source of selection bias determined by patients selecting surgeons and vice-versa. Equally unlikely is patients' personal preferences for the type of implant, which seems to be mostly a function of surgeons' training and experience and local hospital norms (National Institute for Health and Care Excellence, 2014). Nevertheless, we control for patient characteristics in our models which may drive the choice of implant.

The focus of our analysis is on consultant surgeons, who are the most senior grade of NHS surgical staff, and who lead teams of more junior surgeons in the hospital and are responsible for medical cases, including discretion over the treatment provided. Consultants are usually employed by a single NHS hospital and work under salaried contracts. There is a single national salary scale, and terms and conditions resulting from negotiations between the UK government and the British Medical Association are reviewed and set out by an independent Review Body for Doctors' and Dentists' Remuneration. Because consultant pay is not tied to their clinical performance, their remuneration arrangements provide no incentive for the choice of method of fixation in hip replacement surgery.

Care for NHS-funded patients is equally paid across hospitals using a prospective activity-based payment system. The prices and tariffs paid by commissioners for hospital services are standardised on the basis of a national average of hospital costs for Health Care Resource Groups (HRGs). Similar to the US Diagnosis-Related Group (DRG) system, the HRG system classifies and groups services that share similar costs and patient characteristics so that hospital payment better reflects their workload composition. Additionally, tariffs are adjusted for regional exogenous variation input prices to account for the fact that provision is unavoidably more costly in certain places than others.

3 Data

3.1 Data Sources

As our primary data source, we use inpatient data from the Hospital Episode Statistics (HES) database, which collects administrative patient-level data for all NHS-funded care delivered in England by NHS hospitals and independent providers[8]. HES records contain detailed information on patient characteristics (e.g. age, sex, place of residence), clinical data (primary and secondary diagnoses using ICD-10 codes; and procedures using OPCS4 codes), details on the admission pathway (including date and mode of admission and discharge), and hospital identifiers and characteristics. Each episode of care in HES data is assigned to a consultant, who is in charge of delivering care. Consultants are uniquely identified across time and providers in the data, using a General Medical Council (GMC) registration number. We use these consultant identifiers to link the HES data to the GMC's List of Registered Medical Practitioners (LRMP) database. The LRMP holds information on all doctors qualified to practice in the UK, including their gender, year and place of primary medical qualification.

3.2 Sample Definition

Our raw data set was created by extracting patient episodes of primary elective (planned) hip replacement from HES, using the OPCS-4 codes[9] W37.1, for *"Primary total prosthetic replacement of hip joint using cement"*, and W38.1, for *"Primary total prosthetic replacement of hip joint not using cement"*. We focus on the universe of hip replacements performed between April 2010 and March 2017. Prior to 2010, the English NHS reimbursed both procedures differently, with cementless hip replacement being financially favoured over its cemented counterpart. From April 2010 onwards, both methods of fixation are paid equally. By focusing on cases treated from April 2010 onwards, we purge our analysis from the effect of provider financial incentives, which could be driving the treatment decision between both methods of fixation.

[8] Independent providers are private sector healthcare firms that are contracted by the NHS to provide care for NHS patients. They include private and charitable hospitals, as well as independent sector treatment centres.

[9] Similar to Papanicolas and McGuire (2015), we identify cases of cemented and cementless hip replacement using OPCS-4 instead of HRG codes because the HRG code systems are not consistent over time.

The raw data set, which excludes miscoded observations and duplicates, compromises 403,909 patients and 2,313 surgeons.

We then apply a series of restrictions to obtain our primary sample. First, we drop observations of patients treated in ISPs. These organisations treat both NHS and privately funded patients, but only the former are recorded in HES. Because we only observe a share of their clinical activity, we are unable to describe fully the clinical environment or the practice style of surgeons working in those settings. In our raw data set, one-quarter of the (NHS-funded) cases are performed in ISPs by nearly half of the surgeons in our sample. These surgeons simultaneously work in NHS hospitals, where most of their activity is concentrated - approximately three-quarters of the median surgeon's activity in our data is in an NHS hospital. This phenomenon, usually described as dual practice or *moonlighting*, is common practice in the English healthcare market, where surgeons and physicians employed by the NHS are allowed to undertake private work to supplement their NHS income. We recognise that the treatment decisions made by surgeons in those settings may impact their practice style in the NHS hospital, and test for the exclusion of surgeons working in ISPs as part of robustness checks. Second, we exclude hospitals performing fewer than 30 hip replacements to avoid problems of noisy estimates of hospital environment due to small denominators. We show that our findings are robust to changes in this definition. Finally, we exclude patients aged 0 to 17 because they represent a distinct group in which rates of hip replacement are particularly low. After imposing the sample restrictions described above, the primary full sample contains 302,410 hip replacements performed by 2,105 surgeons in 217 hospitals.

We define a subsample of patients that are treated by movers. We use patients' date of admission, surgeon codes, and hospital identifiers, to follow surgeons in our data set and construct individual employment histories. We define a practice period to be the time between the date of the first and last hip replacement performed by a surgeon in a given practice location (hospital). When a surgeon treats less than five cases in a given hospital, we exclude that practice period from our analysis, under the argument that such short-term arrangements represent clinical visits or temporary appointments (our results are robust to alternative assumptions).

Movers are surgeons with two sequential and non-overlapping practice periods, whereas non-movers are surgeons with one or more simultaneous practice locations. If a surgeon moves more than once in our sample (multi-movers), we include those moves independently in our analysis. We later exclude multi-movers to test for the robustness of our findings in the sample of one-time movers. For each move, we term the hospital of the first practice period as the hospital of origin, whilst the hospital of the second practice period is termed the hospital of destination. The year of the move is defined to be the year in which a mover performs his or her first hip replacement in the hospital of destination. Following this strategy, we have identified 74 movers, and a total of 84 moves. These surgeons treat 8,302 patients. For simplicity, we refer to this sample as the sample of movers.

3.3 Variable Definition

Measure of hospital environment

We follow Molitor (2018) to characterise the clinical environment wherein surgeons practice using the universe of decisions made by their peers in the same hospital and financial year. Specifically, we define hospital cemented rates for each surgeon j as the proportion of cemented cases to the total of cemented and cementless hip replacements, omitting his or her own cases. It has been shown that individual outcomes are strongly correlated with group average outcomes (Manski, 1993; Angrist and Pischke, 2008). Thus, by excluding surgeon j's cases from our measure of hospital environment, we mitigate any mechanical correlation of surgeon's treatment decisions and hospital cemented rates which could bias our estimates of environmental effects. We first formally define a raw cemented rate as:

$$P_{jkt} = \frac{1}{|i : i \in N, i \in K, i \in T, i \notin J|} \sum_{i \in N, i \in K, i \in T, i \notin J} (cem_{ijkt}) \tag{1.1}$$

where cem_{ijkt} denotes an indicator function for patient i treated by surgeon j in hospital k during year t receiving a cemented prosthesis. We then risk-adjust raw cemented rates by indirect standardisation to account for differences in patient case-mix between hospitals. This is completed in three steps.

First, we estimate a logit model of the choice of cemented hip replacement on observable patient characteristics[10]. Second, we use the fitted values from the logit model to compute average predicted hospital cemented rates, $\widehat{Pr_{jkt}}$. Finally, the risk-adjusted cemented rates are obtained by differencing out the predicted hospital cemented rates from the raw cemented rates, $\overline{cem_{jkt}} = Pr_{jkt} - \widehat{Pr_{jkt}}$.

One key simplification we make for ease of analysis is to use time-invariant risk-adjusted cemented rates, $\overline{cem_{jk}}$, to characterise the hospital environment. This is calculated for each hospital by averaging the yearly risk-adjusted cemented rates, $\overline{cem_{jkt}}$, across all years in the sample period, using the share of patients treated each year in the hospital as weights. We acknowledge and discuss in the conclusion that this is a limitation of our work.

Control variables

Whether the choice between both methods of fixation is driven by observed characteristics of patients is unclear. There is however a weak trend arising in the clinical literature which suggests that cemented prostheses may be favoured in women, older people and severe patients (Abdulkarim et al., 2013). We remain agnostic and estimate models with and without patient characteristics. We adjust our regression models for the following patient socio-demographic characteristics: age (in five-year bands with the separate category for <50 and >85), sex, age-sex interaction terms, ethnicity, and a proxy for patients' socioeconomic status based on the area-level Index of Multiple Deprivation of the neighbourhood in which they reside (McLennan et al., 2011). We further control for diagnosis of osteoarthritis, and the count of Elixhauser co-morbid conditions (grouped as 0, 1, 2–3, 4) (Elixhauser et al., 1998).

[10] We estimate the following logit model:

$$Pr(cem_{it} = 1) = G[X_i'\theta + \lambda_t + \epsilon_{it}] \tag{1.2}$$

where G[.] is the logit function, $Pr(cem_{it} = 1)$ is the probability that patient i treated in year t receives a cemented prosthesis; X_i is a row vector of K exogenous patient characteristics that include age, sex, age-sex interaction terms, ethnicity, elective surgery, diagnosis of osteoarthritis, number of Elixhauser comorbidities, and a proxy for the socioeconomic status; and,ϵ_{it} is the error term.

4 Empirical Strategy

In this section, we describe the empirical strategy that we adopt to identify hospital effects. We start by building intuition from a theoretical randomized experiment. We then introduce the estimation model, define estimands of interest and formalise a critical identification assumption.

4.1 Identification and Estimation of Hospital Effects

The ideal experiment to estimate whether the clinical environment impacts surgeons' practice patterns would be to randomly re-allocate surgeons to hospitals with different cemented rates and compare changes in surgeons' likelihood to perform cemented procedures by changes in hospital cemented rates. This experiment has a partial equilibrium interpretation, that is, it explores the effect of the exogenous shock in the clinical environment, while holding physician specific-factors fixed. In the long run, one would expect surgeons' practice styles to endogenously adjust to the new environment.

Building on this logic, we adopt a quasi-random experiment approach and focus our analysis on the subsample of movers defined above. Without random assignment, the critical identification assumption underlying our approach is that the timing of the move and the choice of the hospital of destination are orthogonal to surgeons' treatment decision between both methods of fixation. We revisit the test for this assumption in the following section. We characterise the change in hospital environment experienced by surgeon j, Δ_j, as the difference between the time-invariant risk-adjusted cemented rate in the hospital of destination, $d(j)$, and the hospital of origin, $o(j)$. Formally,

$$\Delta_j = \overline{cem}_{d(j)} - \overline{cem}_{o(j)}. \tag{1.3}$$

We then estimate

$$cem_{ijt} = \alpha POST_{jt} + \beta(\Delta_j \times POST_{jt}) + X_{it}\theta + \delta_t + \phi_j + \epsilon_{ijt} \tag{1.4}$$

where cem_{ijt} is an indicator variable that takes the value of one if patient $i = 1, \ldots, N$ treated by surgeon $j = 1, \ldots, J$ receives a cemented implant in year $t = 2010/11, \ldots, 2016/17$, and zero otherwise; $POST_{jt}$ is an indicator for the post-move period which takes the value one for episodes treated in the hospital of destination of surgeon j at time t; Δ_j is a vector of surgeon-specific time-invariant change in hospital environment experienced across the move; X_{it} is a row vector of K exogenous patient characteristics (those specified in Subsection 3.3); δ_t is a vector of year fixed effects that capture exogenous shocks common to all surgeons; ϕ_j is a vector of surgeon-specific fixed effects that account for time-invariant heterogeneity both observed (e.g., gender or postgraduate training) and unobserved (e.g., the skills and ability of surgeons); and $\epsilon_{ijt} \sim \mathcal{N}(0, \sigma^2)$ is a random-error term.

In this specification, the parameter of interest is β. It measures the estimated change in the propensity of a surgeon to perform a cemented hip replacement per unit difference in the clinical environment, as measured by Δ_j. Hypothetically, if surgeons' practice behaviour remains unchanged across the move, the estimated coefficient would be zero. On the other hand, if surgeons' propensity to perform a cemented hip replacement increases and matches exactly the average clinical behaviour in the hospital of destination, the coefficient would be one. When the estimated coefficient is between these two extreme scenarios, it measures the level of adjustment of surgeons' practice style to the new environment. The scalar parameter α measures the change in the propensity for a surgeon to perform a cemented hip replacement, which is purely due to the move. Finally, $\theta = [\theta_1, \ldots, \theta_k]$ is a $K \times 1$ vector of the effects of the corresponding K patient characteristics. We cluster standard errors at the surgeon and hospital level to account for potential serial and spatial correlation (Angrist and Pischke, 2008).

Equation 1.4 could be estimated using binary response models (e.g. probit or logit). However, we utilise linear probability models (LPM) for several reasons. First, unlike non-linear models, LPM estimators provide marginal effects estimates for coefficients, which allow for direct meaningful interpretation of treatment effects on the dependent variable (Angrist and Pischke, 2008). Second, our linear estimates of β are unbiased and consistent since there are few predictive values falling outside of the unit interval range (Horrace and Oaxaca, 2006). Finally, LPM estimates are equally unbiased in the presence of measurement error in dependent variable (Hausman, 2001).

4.2 Positive Sorting and Adaptation Process

We now revisit the key orthogonality assumption underlying our empirical strategy, namely that the timing of the move and the choice of hospital of destination are independent of surgeon-specific determinants of practice style and, thus, their pre-move behaviour. To this end, we test whether surgeons with higher propensity to perform cemented procedures in their hospital of origin choose hospitals of destination with higher cemented rates – in other words, whether there is positive sorting of surgeons to clinical environments. When this condition is violated, our estimate of coefficient β in equation 1.4 overstates[11] the effect of the environment on surgeons' treatment choice.

We are also interested in exploring how surgeons' practice styles evolve over time, especially in response to new hospital environments. Phelps and Mooney (1993) argue that physicians' practice styles develop in a "*learning-by-doing*" fashion. This means that beliefs formed during medical school and postgraduate training are continuously updated throughout a physician's career in response to new environments. These prior beliefs are modified in new environments by observation of colleagues trained in different schools of thought or by location-specific norms.

We test both hypotheses by adding flexibility to specification 1.4. This is achieved by substituting the pre and post-move linear trends in surgeons' practice style by a full set of interactions between Δ_j and periods of time relative to the move. We define these time intervals to be quarters of a year, k, and estimate the interaction for eight quarters before and after the move. We implement this by decomposing the first and second term in specification 1.4 as follows:

[11] The effect would be overestimated because consultants would be choosing hospitals more aligned with their clinical practice, and thus would be more responsive to the change in environment.

$$cem_{ijt} = \sum_{k=-8}^{7} [\alpha_k \mathbf{1}_{(t=k)j} + \beta_k \Delta_j \mathbf{1}_{(t=k)j}] + X'_{it}\theta + \delta_t + \phi_j + \epsilon_{ijt} \tag{1.5}$$

where $\mathbf{1}_{(t=k)j}$ is a surgeon-specific indicator function for each quarter relative to the move, k, and the other variables and coefficients are as described before. The coefficients of interest in equation 1.5 are the β_k. For $k < 0$, the coefficients provide evidence for our orthogonality condition. If there is no positive sorting of surgeons, their practice behaviours before the move should not differ systematically with Δ_j and the estimates of β_k should be close to zero and not statistically significant. When $k > 0$, the coefficients β_k characterise the change in surgeons' practice style over time relative to the move. If estimates become successively larger, surgeons progressively adjust to the new clinical environment as implied by Phelps and Mooney (1993).

5 Results

5.1 Summary Statistics and Descriptive Patterns

Summary statistics for patients, hospitals, and surgeons in our sample are presented in Table 1.1. A total of 302,410 hip replacements are performed between 2010 and 2017. The average cemented rate across the hospitals in our sample is 50.6%, or about half of all patients treated, and this split is relatively stable over time. We identify 74 movers in the entire universe of 2,105 surgeons.

Although our analysis only focuses on the set of movers, we show summary statistics separately for movers and non-movers in Table 1.2. Compared to non-movers, movers are more likely to be less experienced, hold a foreign medical qualification, and perform fewer hip replacements. About half of the movers' hospital of destination is within 50 kilometers of the hospital of origin.

Figure 1.1 shows NHS hospitals ranked in ascending order of cemented rates, in black and bold dots, and the respective hospital within (surgeon-level) variation, in light and grey dots, for 2013 and 2014. The hospital cemented rates vary across the full unit interval. A closer look at the tails of the distribution shows that some hospitals perform little to no

Table 1.1: Summary Statistics for Patients, Hospitals and Surgeons

	Patients	Hospitals				Surgeons	
			Cemented Rate				
	N	K	p25	Mean	p75	J	Number of Movers
2010	39,006	193	0.216	0.480	0.729	1,341	6
2011	40,046	196	0.227	0.494	0.785	1,292	9
2012	38,976	193	0.259	0.519	0.818	1,264	15
2013	39,900	195	0.245	0.513	0.801	1,237	6
2014	39,160	203	0.239	0.509	0.787	1,269	4
2015	36,571	199	0.221	0.512	0.827	1,266	8
2016	35,992	196	0.206	0.507	0.825	1,239	10
2017	32,759	194	0.243	0.516	0.827	1,209	16
2010-2017	302,410	217	0.235	0.506	0.801	2,105	74

Notes: The first column reports the number of primary total elective hip replacements. Second to fifth column describe the number of hospitals, and the mean, p25 and p75 for the distribution of cemented rates across hospitals. The last two columns report the total number of surgeons and movers. Statistics are for each year separately, except in the last row, where they are for the pooled sample 2010-2017. The sample is all primary elective hip replacements performed over the sample period (N=302,410).

Table 1.2: Characteristics of Movers and Non-movers

	Movers (1)	Non-Movers (2)
Male	0.97	0.97
UK Medical Qualification	0.43	0.61
Seniority, in years		
≤15	0.35	0.22
16-20	0.30	0.24
21-25	0.18	0.21
≥26	0.18	0.33
Annual Volume of Hip Replacements		
Mean	22.28	30.19
SD	19.65	30.40
Distance of the move, in Km		
≤50	0.45	
51-100	0.14	
101-150	0.14	
151-200	0.12	
≥201	0.14	
Surgeons, J	74	2,031

cemented hip replacements (on the left), whilst others almost exclusively perform cemented cases (on the right). The within variation is also substantial. Surgeons practicing in the same hospital have very different propensities to perform cemented hip replacements. These findings support three hypotheses in our analysis. First, both cemented and cementless implants are available in the majority of NHS hospitals, and thus the change in surgeons' practice style across the move is unlikely to be driven by local constraints in the availability of prostheses or purchasing arrangements. Second, most UK surgeons have training in both methods of fixation, i.e. individual surgeons' cemented rates do not cluster at zero or one. Finally, the existence of within-hospital variation suggests that the choice of implant is not determined solely by the organisational environment effects common to all surgeons in this organisation.

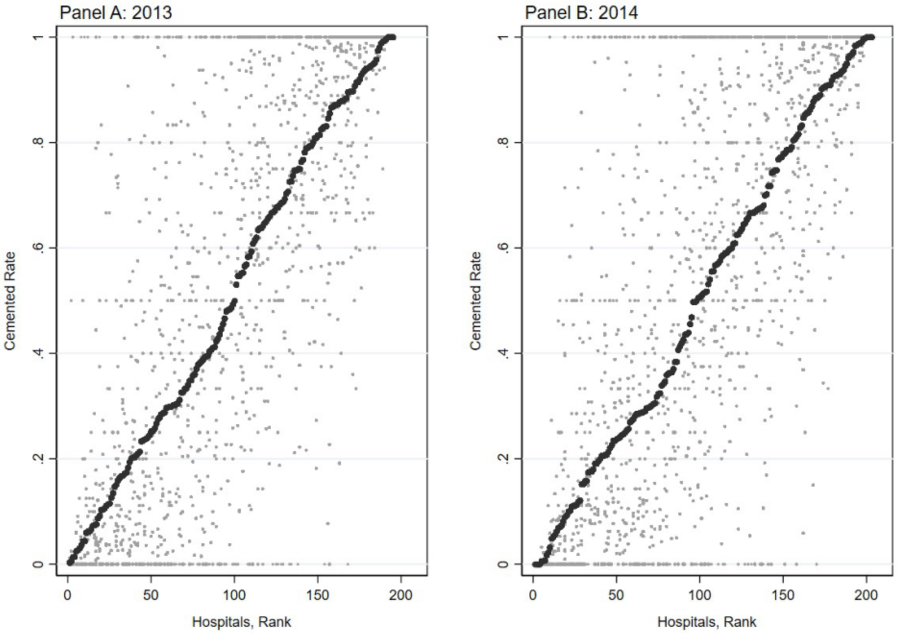

Figure 1.1: Distribution of Hospitals and Surgeons Cemented Rate

Notes: Hospitals (black and bold data points) are ranked in the x-axis in ascending order of the cemented rate in the y-axis, for the years of 2013 (panel A) and 2014 (Panel B). The light and grey data points are for individual surgeons practicing in that hospital. The sample is all primary elective hip replacements performed in 2013 (N=39,900) and 2014 (N=39,160).

In Figure 1.2, we show the spatial distribution of cemented rates across hospitals, divided into quintiles, for the year of 2014. Darker dots are for hospitals performing a higher proportion of cemented procedures. The figure does not suggest a particular spatial pattern in the distribution of cemented rates. It seems, however, that hospitals in the Greater London Area tend to perform, on average, fewer cemented cases.

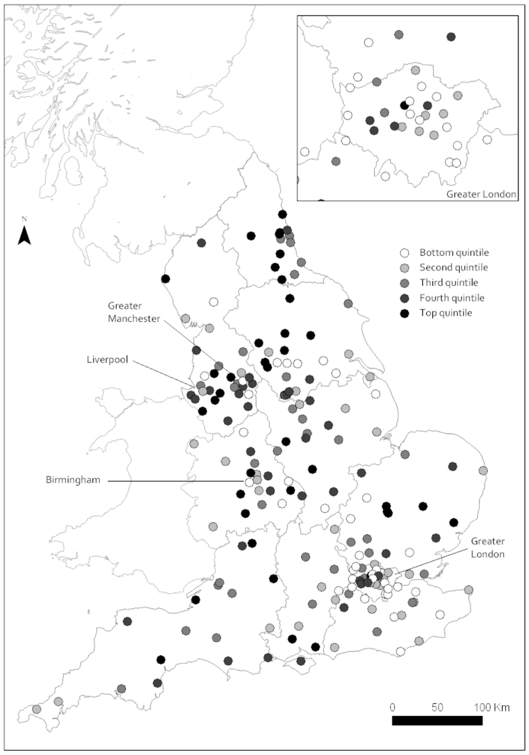

Figure 1.2: Geographical Distribution of Cemented Rate by Hospital

Notes: Map displays the distribution of cemented rates by hospital, divided into quintiles of cemented rate. Darker dots represent hospitals with higher cemented rates. The sample is all hospitals performing primary elective hip replacements performed 2014 (N=39,160).

To begin examining whether the clinical environment impacts surgeons' treatment choices, we first illustrate the nature of the variation that drives our analysis. Figure 1.3 shows the distribution of Δ_j: the change in the clinical environment between the hospital of destination and origin for each move (M=84), excluding surgeon's own cases. The average value of Δ_j is close to zero (mean = 0.005), with considerable variation around this average value (standard deviation = 0.40).

Figure 1.4 shows how surgeons respond to the new clinical environment. It portrays the change in surgeons' propensity to perform cemented procedures by the size of the change in hospital cemented rate experienced across the move. Three lines of best fit estimated by

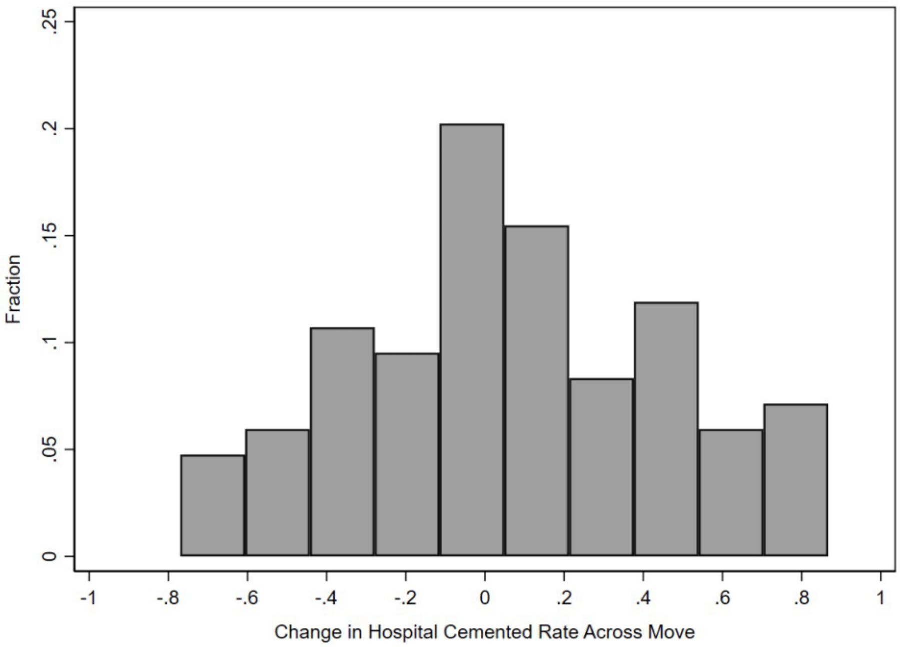

Figure 1.3: Change in Hospital Cemented Rate Across the Move

Notes: Distribution for the change in the risk-adjusted leave-out hospital cemented rate between the hospital of destination and origin (average = 0.005, std. dev. = 0.401, median = 0.000). The sample is all the moves (M=84).

OLS are also shown: the black solid line is for the entire sample of moves (M=84), while the short-dashed lines are fitted separately for surgeons moving to lower ($\Delta_j < 0$) and higher ($\Delta_j > 0$) cemented environments. The slope estimates have an interpretation similar to the coefficient of interest in Equation (1.4). If surgeons' decisions are purely determined by the clinical environment, one would expect a slope of one. Conversely, if surgeons are immune to changes in the clinical environment, their practice behaviour would not change, that is, the slope of the line would be zero. The OLS estimate for the slope of the line fitted to the entire sample of movers falls between both extreme scenarios. It yields a value of 0.704 (SE = 0.139), which is significant at the 1% level. Stated differently, a 10 percentage points increase in the hospital cemented rate, increases on average surgeons' propensity to do a cemented hip replacement by 7.0 percentage points. This result suggests the existence of large hospital effects. The estimate for surgeons moving to lower cemented environments ($\Delta_j < 0$) is smaller than the coefficient for those moving to higher cemented environments ($\Delta_j > 0$). The former produces a value of 0.678 (SE = 0.232) and the latter a value 0.924 (SE = 0.170), both being significant at the 1% level. This finding suggests that surgeons may respond asymmetrically to changes in the clinical environment. We turn to this issue in our analysis of heterogeneous treatment effects.

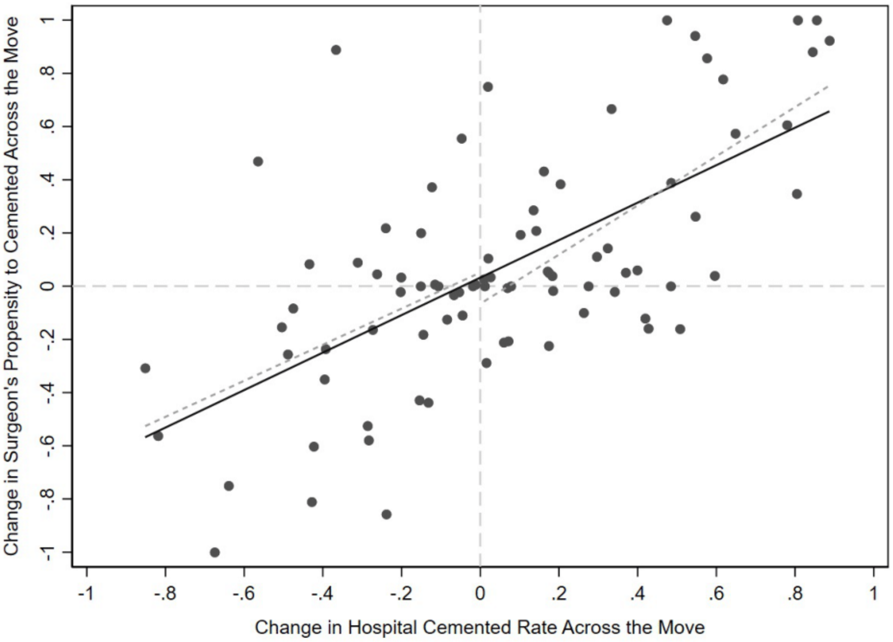

Figure 1.4: Change in Surgeon's Propensity to Cemented by Change in Hospital Cemented Rate Across the Move

Notes: The x-axis displays the change in the risk-adjusted leave-out cemented rate experienced by surgeons across the move. The y-axis shows the change in the risk-adjusted surgeons' propensity to perform a cemented hip replacement in the hospital of destination relative to the hospital of origin. Cemented rates at the hospital and surgeon level are risk-adjusted for patient characteristics, and computed over the period of two years before and after the move. Data points are for surgeons. The line of best fit for the entire sample of movers (solid line) is estimated by OLS (J=84, intercept = 0.032 (std. err.= 0.037), slope = 0.704 (std. err.= 0.092), R2 = 0.420). The two sort-dashed lines are lines of best fit for the sample of movers experiencing a negative change in the hospital cemented rate ($\Delta_j < 0$, J=38, intercept = 0.051 (std. err.= 0.102), slope = 0.678 (std. err.= 0.280), R^2 = 0.141) and a positive change in the hospital cemented rate ($\Delta_j > 0$, J=46, intercept = -0.066 (std. err. = 0.074), slope = 0.924 (std. err. = 0.170), R^2 = 0.400). The sample is all the moves (M=84).

5.2 Baseline Model

The estimate effects of the hospital environment on surgeons' treatment choices are set out in Table 1.3. In column 1, we adjust only for surgeon fixed effects to capture individual time-invariant heterogeneity. We then add year fixed effects in column 2, and add controls for patient characteristics in column 3. The estimates for hospital environment effects (Row 3) are all positive and statistically significant at the 1% level. The result in column 3 is of lower magnitude, suggesting that patient characteristics may determine some of the choice of hip implant fixation. A 10 percentage points change in the hospital cemented rate leads to an average change of 6.37 percentage points in a surgeon's propensity to perform a cemented hip replacement. This means that a surgeon moving from a hospital at the 25th percentile of cemented rate (P25 = 0.22) to a hospital at the 75th percentile (P75 = 0.84) would experience a positive change of roughly Δ_j=62 percentage points in the hospital environment, which would produce an increase in his own propensity to perform a cemented procedure of approximately 40 percentage points ($0.637 \times 0.62 = 0.395$). Furthermore, this estimate in column 3 is slightly smaller than the OLS estimate for the line of best fit in Figure 1.4. The coefficient of Δ_j in column 4 describes the degree of selective migration. If consultants with higher cemented rates in the hospital of origin (relative to their peers) move to hospitals with higher cemented rates, the coefficient would be positive. The estimated coefficient is small in magnitude and statistically insignificant suggesting lack of evidence of positive sorting.

5.3 Positive Sorting and Adaptation Process

Figure 1.5 plots the estimated coefficients β_k from equation 1.5; these are for the interaction of Δ_j with quarters of the year, k, relative to the move. We define quarter -1 to be the quarter before the move, whereas quarter 0 corresponds to the first quarter right after the move. We normalize the coefficient of the relative quarter before the move to zero, $\beta_{-1} = 0$. Upper and lower bounds of the 95% confidence interval are reported for each estimate. The figure shows a sharp, and discontinuous jump in surgeons' propensity to perform a cemented procedure across the move.

Our causal inference relies on the assumption that movers do not sort themselves into high and low-cemented environments according to unobserved determinants of treatment style. Such positive sorting of surgeons into hospitals would bias our estimates of hospital environment effects in Table 1.3. The pre-move coefficient estimates ($k < 0$) in Figure 1.5 show no

Table 1.3: Effect of Changes in Hospital Environment on Choice of Hip Implant Fixation

Dependent variable: $cem_{ijt} \in \{0,1\}$				
	(1)	(2)	(3)	(4)
Δ_j				0.048 (0.142)
$POST_t$	0.025 (0.028)	0.001 (0.030)	0.019 (0.027)	-0.022 (0.023)
$\Delta_j \times POST_t$	0.642*** (0.082)	0.682*** (0.083)	0.637*** (0.077)	0.672*** (0.071)
Surgeon FE	YES	YES	YES	NO
Hospital of origin FE	NO	NO	NO	YES
Year dummies	NO	YES	YES	YES
Patient characteristics	NO	NO	YES	YES
Patients, N	8,302	8,302	8,302	8,302
Surgeons, J	74	74	74	74
Moves	84	84	84	84

Notes: */**/*** indicate statistical significance at the 10%/5%/1% level. Standard errors are clustered at the physician and hospital level to allow for arbitrary serial correlation and heteroskedascity.

discernible trend: an F-test for the equality of coefficients fails to reject the null hypothesis of no trend at the 10% significance level ($F_{6,60} = 1.34$; p-value = 0.254). These are also insignificant and close to zero, implying that there are no pre-move trends and that the levels in surgeons' practice style are not a function of the change in hospital environment experienced across the move. The results of this auxiliary analysis therefore support the orthogonality assumption of our main analysis.

The post-move ($k > 0$) coefficients are all significant at the 1% level, and also show no discernible trend in the plot, levelling at around 0.6 – our estimate of hospital effects. An F-test for the equality of post-move coefficient estimates ($k > 0$) fails to reject the null hypothesis at the 10% significance level ($F_{7,60} = 1.28$; p-value = 0.278). This suggests that there is little to no adaptation process to the new clinical environment over time after the first year.

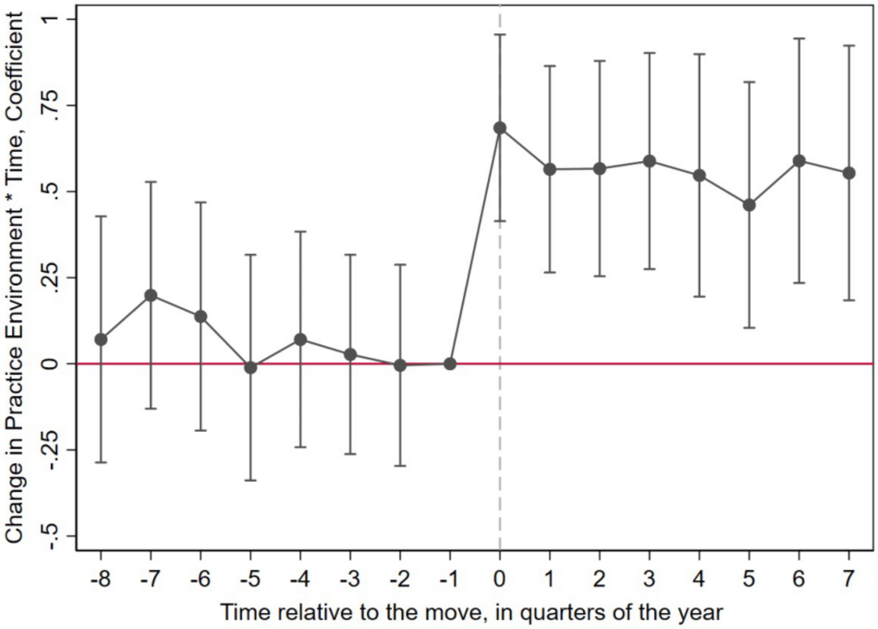

Figure 1.5: Event Study of Positive Sorting and Learning Process of Adaptation

Notes: Data points are estimates for the interaction of the change in the hospital risk-adjusted leave-out cemented rate with quarters of the year, k, relative to the move. Estimates are for eight quarters before and after the move. The x-axis displays the time relative to the move, in quarters of the year. The y-axis shows the value for the respective coefficient β_k. The coefficient for the quarter -1 relative to the move is normalized to 0. Standard errors are clustered at the physician and hospital level to allow for arbitrary serial correlation and heteroskedasticity. The analysis covers treatment decisions for all patients treated by movers during the eight quarters before and after the move (N=3,332).

5.4 Robustness Checks

We explore the robustness of our main finding to changes in the model specification and in the sample definition in Table 1.4. The first row presents the baseline results from Table 1.3. In row 2, we exclude surgeons who move more than once (multi-movers), as the effect of the second move may be contaminated by the first move. We therefore estimate the effect of the clinical environment only using the set of one-time movers in our sample, which consists of 61 surgeons. The estimate of the coefficient of interest is positive and statistically significant at the one percent level; however, of slightly smaller magnitude than our baseline result. In row 3, we extend our analysis and consider both elective and emergency THR. Elective procedures represent approximately 92 percent of all hip replacements in our sample. Yet the choice of method of fixation in elective cases may be correlated with the choice in emergency hip

Table 1.4: Robustness Checks

Dependent variable: $cem_{ijt} \in \{0,1\}$	$POST_{jt}$	$\Delta_j \times POST_{jt}$	Patients	Surgeons	Moves
	(1)	(2)	(3)	(4)	(5)
1. Baseline	0.019	0.637***	8,302	74	84
	(0.027)	(0.077)			
2. Excluding multi-movers	0.027	0.562***	6,504	61	61
	(0.032)	(0.097)			
3. Elective and emergency THR	0.025	0.643***	8,794	74	84
	(0.027)	(0.076)			
4. Overlap in Practice Periods <30 days	0.047	0551***	9,389	86	96
	(0.035)	(0.092)			
5. Surgeons without observed private practice in the two year before/after the move	0.017	0.614***	8,013	72	82
	(0.028)	(0.078)			
6. Minimum volume in a practice period, 10 THR	0.028	0.649***	9,139	58	64
	(0.024)	(0.083)			
7. Hospital minimum volume, 5 THR	0.019	0.637***	8,302	74	84
	(0.027)	(0.077)			

Notes: */**/*** indicate statistical significance at the 10%/5%/1% level. Standard errors are clustered at the physician and hospital level to allow for arbitrary serial correlation and heteroskedascity.

replacements. We find that our results are robust to including emergency THRs in our patient sample. In row 4, we show that our results are robust to relaxing our definition of a move to also include surgeons who happen to be practicing in both hospital of origin and destination, simultaneously, for up to 30 days. In row 5, we restrict our analysis to surgeons without any observed activity in an ISP in two years before and after the move. This suggests that the small amount of NHS-funded activity surgeons conduct in ISP is unlikely to contaminate our baseline results. Finally, in rows 6 and 7, we show that our findings are invariant to changes in the minimum volume cut-off to define a practice period or to include a hospital in the analysis.

5.5 Impact on Patient Health Outcomes

A rapid change in practice style following a move to a different hospital environment could have negative effects on patient safety, especially if surgeons are unfamiliar with the new treatment approach. We therefore examine whether a move and a subsequent change in treatment approach impacts on three common quality and safety metrics: probability of 28-day emergency readmission and one-year revision, and change in patient-reported outcome

Table 1.5: Effect of Changes in Hospital Environment on Patient Health Outcomes

Dependent variable:	28-day Emergency Readmission	1-Year Revision	PROMS
	(1)	(2)	(3)
$POST_t$	0.014	0.001	-0.141
	(0.010)	(0.003)	(0.572)
$\Delta_j \times POST_t$	-0.002	-0.001	1.981
	(0.019)	(0.004)	(1.249)
Surgeon FE	YES	YES	YES
Year Dummies	YES	YES	YES
Patient Characteristics	YES	YES	YES
Patients, N	8,302	8,302	4,218
Surgeons, J	74	74	74
Moves	84	84	84

Notes: */**/*** indicate statistical significance at the 10%/5%/1% level. Standard errors are clustered at the physician and hospital level to allow for arbitrary serial correlation and heteroskedascity.

measures (PROMs)[12]. To do so we re-estimate equation 1.4 using these safety metrics as outcome variables. Table 1.5 presents the results of this analysis. We do not find evidence that the move itself or the change in hospital environment have negative effects on any of the three safety measures analysed.

5.6 Surgeon Heterogeneous Effects

We also explore heterogeneity of hospital environment effects across subgroups of surgeons in the population of movers. In our model in specification (1.4), we assume that surgeons respond symmetrically to the change in environment. However, one may be concerned that surgeons respond differently depending on the direction of their move: from high- to low-cemented environments, and vice versa. We test for non-symmetric responses to hospital environments

[12] PROMs have clear advantages over the other two "*failure*" metrics because they measure patients' views on their health status and capture other dimensions of health (John Appleby and Devlin, 2005). We use the change in the Oxford Hip Score (OHS), a hip-specific measure of health status for patients with hip joint conditions (Dawson et al., 1996; Ostendorf et al., 2004), which is measured immediately before and six months after hip replacement surgery. The OHS is routinely collected in England since 2009 as part of the national PROMS programme (Department of Health, 2008). Data collection is mandatory for hospitals providing NHS-funded care, but survey participation is voluntary for patients. The OHS is based on 12 questions asking patients about the functional status (mobility) and pain, each of which is scaled from 0 to 4. The sum of the scores across items gives the OHS, with 0 being the worst and 48 the best health status.

by expanding equation (1.4) with a triple interaction of $\Delta_j \times POST_{jt}$ with an indicator variable H_j for surgeons moving to high-cemented environments (i.e. $\Delta_j > 0$). In a similar fashion, we test whether surgeons respond differently by (i) country of medical school qualification (UK vs. non-UK), (ii) distance of the move (in kilometers), and (iii) consultant seniority (in years). We estimate linear probability models of the form:

$$\begin{aligned} cem_{ijt} = \alpha_1 POST_{jt} + \alpha_2 H_j + \alpha_3 \Delta_j + \beta_1 (POST_{jt} \times H_j) + \beta_2 (\Delta_j \times POST_{jt}) \\ + \beta_3 (\Delta_j \times H_j) + \sigma_2 (\Delta_j \times POST_{jt} \times H_j) + X_{it}\theta + \delta_t + \phi_j + \epsilon_{ijt} \end{aligned} \tag{1.6}$$

where H_j is the variable of interest for the effects described above.

In Table 1.6, we look at heterogeneous effects across different subgroups of surgeons. The coefficient for the triple interaction ($\Delta_j \times POST_{jt} \times H_j$) in column 1 is not statistically significant, suggesting that surgeons respond symmetrically to changes in hospital environment. Next, we find no evidence that surgeons moving a longer distance between the hospitals of origin and destination are more responsive to the new environment (column 2). Finally, we do not find evidence that surgeons who qualified abroad (column 3) or who are more senior (column 4) respond differently compared to their locally trained or younger counterparts.

6 Conclusions

This study examines whether the hospital environment drives the treatment decisions made by NHS surgeons to shed new light on the sources of supply-side variation in medical care. Our analysis focuses on the universe of hip replacements performed in the English NHS to analyse surgeons' choice between cemented and cementless methods of fixation, which are used in this surgical procedure. Using eight years of NHS hospital discharge data we construct a panel of surgeons and trace their employment histories to identify job moves across hospitals. We use the variation in the clinical environment provided by the move to identify the effect of the hospital environment on the practice style of surgeons, where the hospital environment includes all factors that are not embedded in the surgeon.

Similar to Molitor (2018) and Avdic et al. (2021), we find evidence that the environment in which surgeons practice matters for their treatment decisions. Our results suggest that,

Table 1.6: Surgeon Heterogeneous Effects

Dependent variable: $cem_{ijt} \in \{0,1\}$				
$H =$	Asymmetric Response (1)	Distance of Move (2)	Medical Qualification (3)	Consultant Seniority (4)
$POST_t$	0.025 (0.028)	0.001 (0.030)	0.019 (0.027)	-0.022 (0.023)
$\Delta_j \times POST_t$	0.642*** (0.082)	0.682*** (0.083)	0.637*** (0.077)	0.672*** (0.071)
$\Delta_j \times POST_t \times (\Delta_j > 0)$	0.378 (0.228)			
$\Delta_j \times POST_t \times (51 - 100km)$		0.065 (0.211)		
$\Delta_j \times POST_t \times (101 - 150km)$		0.283 (0.218)		
$\Delta_j \times POST_t \times (151 - 200km)$		0.264* (0.154)		
$\Delta_j \times POST_t \times (201km)$		-0.237 (0.164)		
$\Delta_j \times POST_t \times (UK)$			-0.175 (0.163)	
$\Delta_j \times POST_t \times (16 - 20years)$				-0.028 (0.179)
$\Delta_j \times POST_t \times (16 - 20years)$				-0.314 (0.379)
$\Delta_j \times POST_t \times (26years)$				-0.175 (0.178)
Surgeon FE	YES	YES	YES	YES
Hospital of origin FE	YES	YES	YES	YES
Year dummies	YES	YES	YES	YES
Patient characteristics	YES	YES	YES	YES
Patients, N	8,302	8,302	8,302	8,302
Surgeons, J	74	74	74	74
Moves	84	84	84	84

Notes: */**/*** indicate statistical significance at the 10%/5%/1% level. Standard errors are clustered at the physician and hospital level to allow for arbitrary serial correlation and heteroskedascity.

on average, a 10 percentage points change in the hospital environment changes a surgeon's likelihood to perform a cemented procedure by 6.4 percentage points. This finding is in line with the 6.7 and 7.2 percentage points increase for a 10 percentage points change in environment reported by Molitor (2018) and Avdic et al. (2021), respectively. Furthermore, moving surgeons adapt immediately rather than gradually to the new hospital, indicating that policies targeting the organisations in which doctors work could have large effects in the short run. Simultaneously, the lack of post-move convergence suggests that policies aimed at changing physician-specific factors, such as the beliefs and preferences of consultants, may only have an effect in the long run.

We also show that practice style prior to the move does not vary systematically with the observed change in environment, implying that surgeons do not sort themselves into hospitals. Most importantly for patient welfare, our results show that the change in treatment style had no negative implications for the quality of care provided. This result is possibly explained by the lack of superiority in the effectiveness of one method of fixation over the other.

Our study has two main limitations. First, our approach relies on the assumption that our measure of hospital environment is time-invariant over our sample period. If hospitals follow different trends in cemented rates, then this key simplification introduces measurement error in our measure of change in the hospital environment. Reassuringly, our descriptive data shows that both the average and interquartile range of the cemented rate are stable across the period of analysis, and thus hospital differences in cemented rate are likely to be comparable across years. Second, the movement of consultants across hospitals is limited, and we show that movers are likely to be less experienced, hold a foreign medical qualification, and perform fewer hip replacements. Thus, if movers differ from non-movers in ways that change the relevance of hospital effects, our findings cannot be generalised to the non-mover population.

Our study contributes to recent work trying to open the "black box" of supply-side drivers of variation in medical care. Further research is required to identify specific mechanisms driving doctors' treatment decisions. In particular, our findings suggest that trying to disentangle and pin down specific factors related to the environment wherein doctors practice is an empirical task worth pursuing.

Chapter 2

The Effect of Winning Financial Awards on Consultant Activity Rates in the English NHS

1 Introduction

In the UK and other countries, policymakers have been experimenting with explicit financial incentives to influence doctors' behaviour in an attempt to improve efficiency in healthcare (Roland, 2004; Epstein, 2006; Scott, 2007; Greb et al., 2006). Recent examples include a range of performance-related pay schemes (i.e. pay-for-performance (P4P)) which conceptually entail linking doctors' pay to achievements on a set of performance targets in dimensions of health care such as processes of care (e.g. safety and quality), patient experience, and even health outcomes (Doran et al., 2017). If well-designed, P4P schemes may motivate doctors to direct efforts towards the delivery of cost-effective interventions, and, as a result, improve the quality of care provided, reduce variations in medical care, and curb health care expenditure (Maynard, 2012). However, if incentive schemes cannot fully specify all relevant aspects of doctors' work, critics contend that they may give rise to undesirable doctor behaviour with unintended consequences for patient care (Rosenthal and Frank, 2006; Doran et al., 2008; Hutchison, 2008; Gravelle et al., 2010).

In this chapter, I present evidence of a performance-related pay scheme, known as the

Clinical Excellence Awards (CEA)[13], which offers bonus payments to senior hospital doctors[14] in the English NHS (i.e. consultants) on the basis of their contributions to a high-quality service and several non-clinical activities, such as research and innovation, leadership roles, and teaching and training (see Table B.1 in the Appendix for further details on these activities). Every year, interested candidates nominate themselves and must provide clear evidence of their achievements — which should be significantly "over and above" the standard expected of their role — in the five years prior to the application (Advisory Committee on Clinical Excellence Awards, 2014). The national scheme[15] is designed so that consultants progress successively through four levels of awards with associated increasing pay — in ascending order of financial value, Bronze (worth £35,484 p.a. in 2013; see NHS Employers (2013)), Silver (£46,644 p.a.), Gold (£58,305 p.a.), and Platinum (£75,796 p.a.) — in increasingly competitive rank-order tournaments. Thus, new awards are allocated on the basis of a consultant's performance relative to her peers in the same award level. CEA payments are made to consultants on top of their salary and any additional clinical activity they undertake to increase that salary. To put into context, the awards can almost double a consultant's base salary at the highest level of the awards[16]. The CEA are given for a period of five years after which they are subject to review based on a new assessment of work performance (see Figure B.1 for a timeline of different periods associated with the awards).

The effect of each new CEA on the post-award work effort of consultants is *a priori* unknown. In the simple static labour-leisure framework, which lies at the heart of most prevail-

[13] CEAs came into force in 2004. They replaced the Distinction Awards scheme which was first introduced in 1948 at the formation of the NHS as a way to persuade doctors to accept and join the NHS (Mitchell et al., 2011). These awards have been contentious since their inception, mainly due to concerns over lack of transparency and possible discrimination in the allocation of awards against women, younger consultants, and those working in non-teaching hospitals and certain specialties (e.g. obstetrics and gynaecology, and dermatology) (Essex et al., 2021). In the past, CEAs were available in all countries of the UK. However, the scheme has been closed in Scotland since 2010, and suspended altogether in Northern Ireland since 2013 (Essex et al., 2021).

[14] CEAs are also awarded to senior academic general practitioners (GPs) and dentists. These are not the focus of this chapter.

[15] There are also nine local award levels, which I do not analyse in this chapter because these are structured differently. These are of lower financial value, employer-based rather than awarded by the Advisory Committee on Clinical Excellence Awards (ACCEA), and the respective lump sum is paid for only one year. Information on local awards is not publicly available.

[16] In April 2013, consultants in England were paid a base salary of £75,249 to £101,451, depending on tenure (NHS Employers, 2013).

ing theoretical models of doctor behaviour, doctors are assumed to act as utility-maximizing agents who choose their optimal level and mix of labour activities to trade off between consumption goods (enabled through their wages or income) and leisure (McGuire and Pauly, 1991; McGuire, 2000). Under suitable conditions on preferences, the labour supply function depends on income earned from work (for consultants, this is income derived from their salary and payments from additional activity undertaken) and non-labour income, which includes any source of income that is unrelated to work decisions at that single point in time. In this static form, it is reasonable to treat income derived from the awards as "non-labour" or "unearned" income because it is independent of consultants' current labour supply. Thus, if leisure is a normal good, the positive income shock provided by the awards increases demand for all goods, including leisure. As a result, the income effect may lead to a post-award reduction in work effort.

However, the preceding argument may not carry over when allowing for the intertemporal context of the CEA. As aforementioned, the scheme is designed so that the awards are subject to renewal every five years in light of new evidence of work performance from that period of time. Thus, reducing effort in the post-award period, as mentioned above, may result in the withdrawal of CEA payments in the future. If consultants value the income provided by the awards and want to keep it in the future[17], they must maintain work effort. In fact, consultants may even increase effort in the post-award period if they decide to pursue higher level awards, which demand evidence of more unique work achievements. That said, the awards are likely to induce "stepwise" increases in effort as consultants attempt to progress through the different levels of the awards.

Furthermore, consultants produce different types of services for the NHS, some of which are incentivised by the CEA whereas others are not. Volume of clinical activity is not directly incentivised, and thus may provide a weak signal of effort to the committees responsible for the allocation of the awards (Bloor et al., 2012). As a result, consultants may choose to substitute away from activity to dimensions of work explicitly incentivised by the scheme

[17] Theory of labor supply holds that workers' decisions over consumption and leisure are realistically made over time (Borjas, G. J., 2008). Workers are generally willing to trade some leisure today for some consumption tomorrow, if such transaction is of benefit to them (Borjas, G. J., 2008).

— a phenomenon known in the literature as *multitasking*[18] (Prendergast, 1999; Holmstrom and Milgrom, 1991; Baker, 1992; Eggleston, 2005) — and thus reduce the amount of time allocated to treat patients. That said, the "stepwise" increase in effort required in each new award level may come at the expense of a "stepwise" decrease in consultants' clinical activity, as consultants attempt to maintain the current award or pursue higher award levels.

In light of these concerns, determining the effect of the awards on clinical activity becomes a justifiable empirical investigation. It is equally one of interest to policymakers for two reasons. First, the scheme is costly for the NHS. In 2019, the total financial value of the national CEA was estimated to be nearly £130 million (Advisory Committee on Clinical Excellence Awards, 2019). To put into perspective, this cost equals, for example, the NHS total spending on outpatient child and adolescent mental health attendances in England (Essex et al., 2021). In a healthcare system that is concerned with efficiency gains and budget control, it is important to understand the added value and potential implications of the scheme for patient care. A second motive is that healthcare policymakers are increasingly concerned with NHS productivity (NHS, 2019), particularly with the activity of hospital consultants, in terms of treating NHS patients, which has been gradually falling in the last decade (Lafond et al., 2017).

In this chapter, I therefore examine whether being awarded a CEA impacts the activity rates of award winners. Specifically, I ask the question of whether consultants who are given a new award neglect patient care (volume) to possibly further focus on those dimensions explicitly incentivised by the scheme, in an attempt to keep their award or pursue higher levels of the awards. To complement my analysis, I also investigate whether loosing any of the awards, and hence the subsequent withdrawal of payments, is associated with the activity rates of consultants.

There is little prior evidence on the effect of the awards on consultant clinical activity. In two cross-sectional studies, Bloor et al. (2004, 2008) examine whether the activity rates

[18] The *multitasking problem* arises when only a subset of the relevant work activities are rewarded (Holmstrom and Milgrom, 1991). This theory says that while the desired behaviour may be induced for rewarded aspects of performance, improvements may come at the expense of activities that are not specified in workers' contracts (Prendergast, 1999).

of consultants are associated with holding any of the award levels. Adjusting for consultant individual characteristics (e.g. age, clinical specialty, type of contract, and gender) and organisational factors (e.g. whether they practice in teaching hospitals), they find no statistically significant differences in clinical activity between consultants with CEAs and those without. One obvious shortcoming of these studies is that the cross-sectional approach does not account for endogeneity issues due to unobserved factors likely to drive both the probability of receiving an award and the activity rates (e.g. individual ability, skill and motivation).

In my analysis, I follow NHS consultants working over a seven year period between 2009 and 2015. My data set includes detailed information on clinical activity, the awards, and consultant characteristics. I use two measures of consultant NHS activity: the count of completed inpatient episodes of care for each consultant and quarter of the year, and an adjusted measure I develop which weights each episode of care using the national average length of stay (LOS) for the respective Healthcare Resource Group (HRG). The motivation for the latter is that cases with longer expected LOS require more attention, time and effort from consultants. I begin my analysis with a simple model that includes time effects and consultant seniority to determine whether consultant activity differs after a new award is granted compared to before. Because I am concerned about the dynamic effects of the awards, I extend the model with a set of leads and lags to test for anticipation and adaptation effects in the five years leading to and after the awards. This empirical approach is similar to past work that examines the way in which life or work satisfaction evolves around the time of several life effects, including marital status (Clark et al., 2008; Frijters et al., 2008), unemployment (Clark et al., 2008), self-employment (Hanglberger and Merz, 2015), wage changes (Diriwaechter and Shvartsman, 2018), and trade union membership (Powdthavee, 2011).

I find weak evidence of a reduction in post-award activity rates in the two lowest levels of the awards (Bronze and Silver). This finding is sensitive to the econometric modelling approach applied. My results also suggest that being granted Gold and Platinum awards does not seem to have statistically significant impact on post-award consultant activity. Moreover, I show that loosing an award is not associated with the activity rates of consultants. Finally, I do not find clear evidence of anticipation effects for any of the award levels, which could bias the results.

This chapter is structured as follows: Section 2 reviews the literature on financial incentives and their unintended consequences. Sections 3 and 4 discuss the data and empirical strategy to assess the effect of the awards on post-award activity rates. I present and describe my findings in Section 5. And Section 6 concludes.

2 Literature Review

Health economists have long been interested in the role of financial incentives on doctors' behaviour. A sizeable body of the literature has assessed the implicit financial effects embedded in traditional remuneration systems — commonly fee-for-service, capitation, and salary — both theoretically (McGuire, 2000), and empirically (see Gosden et al. (2000) for a review) including in experimental settings (Lagarde and Blaauw, 2017; Brosig-Koch et al., 2017). The general conclusions of these studies are that fee-for-service induces overtreatment, capitation undertreatment, and salary carries no incentive to exert effort.

Evidence on the effects of explicit financial incentives in the form of performance payments to elicit specific doctors' behaviours is rapidly growing (Armour et al., 2001; Town et al., 2005; Rosenthal and Frank, 2006; Petersen et al., 2006; Christianson et al., 2008; Scott et al., 2011; Li et al., 2014). The findings from these empirical investigations are, however, ambiguous and inconclusive. Doctors seem to respond positively to some schemes but not others, and the estimated effects are at most modest. It is unclear whether the lack of meaningful findings is the result of methodological shortcomings (e.g. small sample sizes and ill-suited comparison groups) of existing studies, or in fact due to poorly designed incentive schemes (e.g. small size of payments) (Kantarevic and Kralj, 2013).

There is also a growing body of the literature suggesting that performance-related pay may give rise to strategic behaviour and unintended consequences. For example, there is evidence of doctors *gaming* performance indicators to attract more pay by wilfully miscoding diagnoses and selecting and excluding patients from indicator calculations (Rosenthal and

Frank, 2006; Doran et al., 2008; Hutchison, 2008; Gravelle et al., 2010). Another subset of studies, most relevant to this chapter, has highlighted the implications of the *multitask agency problem* for incentive contracts, a theory first suggested by Holmstrom and Milgrom (1991). The core principal of *multitasking* is that, when a job consists of many tasks, workers will direct their attention to tasks that are easy to measure and rewarded, and away from those that are not - i.e. the higher effort in one task increases the marginal cost of other tasks (see Prendergast (1999) for a review of the theory and its relations to contract theory). In health care, economists have also constructed models to examine the implications of the problem of *multitasking* for purchasing agency contracts with hospitals (Chalkley and Malcomson, 1998) and doctors (Eggleston, 2005), whose output is multidimensional.

There are a few empirical examples of the *multitasking problem* in the health economics literature. Campbell et al. (2009) studies the effects of a pay-for-performance — the Quality and Outcomes Framework (QOF) — on the quality of primary care in England and find a decrease in quality for non-incentivised activities, such as continuity of care and certain aspects of care in patients with asthma and heart disease. Feng Lu (2012) shows that the introduction of mandatory public reporting of quality measures for nursing homes in the US Nursing Home Quality Incentive (NHQI) programme improved scores of quality measures for incentivised dimensions of quality but led to deteriorations for those unreported. That is, nurses allocated resources and efforts based on whether the programme publicly disclosed or not those measures. These studies emphasise the importance of understanding potential spillover effects when designing and implementing incentive schemes.

3 Data

3.1 Data Sources and Sample Definition

The data used for this study pertain to consultants practicing in the English NHS between April 2009 and March 2015. These data are derived from three sources: routinely collected hospital discharge data for all publicly-funded hospital care in England taken from the Hospital

Episode Statistics (HES); consultant data on demographics and medical education from the List of Registered Medical Practitioners (LRMP)[19] maintained by the General Medical Council (GMC); and lists of successful candidates to the national Clinical Excellence Awards published annually by the Advisory Committee on Clinical Excellence Awards (ACCEA). All three data sets were linked on the basis of consultants' GMC registration numbers, which are uniquely attributed to each medical doctor practicing in the UK.

After merging all data, my sample includes $32,619$ consultants and $160,202$ consultant-quarters. I focus on consultants practicing between April 2009 and March 2015 for two reasons. First, the lists for successful candidates to the national awards are only publicly available from 2009 onwards. Second, the UK government introduced a major change in the tax relief system in April 2016 which affected high-income earners and, ultimately, reduced their effective take-home pay[20]. Many NHS consultants fall into the policy target group and I, therefore, anticipate that the reform may have had implications for the activity rates of consultants from 2016 onwards that are unrelated to the effect of awards. To avoid contamination, I limit my analysis to activity data up to March 2015, i.e. a year before the tax relief reform was enacted.

I apply four restriction rules to my primary sample. The impact of selection on sample size is laid out in Table B.2. A first selection rule is that I only include consultants who work (predominantly) in one of 18 medical or surgical specialties - as defined by the 'mainspef' field in the HES data. I therefore exclude consultants working in, for example, psychiatry, pathology and radiology because their practice seldom involves responsibility over inpatient hospital stays. Second, I exclude consultants with observed activity in a single year only. Third, I only include consultants with continuously observed activity during the time period of analysis due to the effect of temporary leave of absence (e.g. sabbatical, secondments and maternity) on both activity rates and the probability of getting an award. Finally, I select consultants who are granted new awards and those who lost an award, with pre- and post-award data - a requirement of the within-consultant (fixed effects) approach I adopt.

[19] The GMC is the national body that determines doctors' qualification to practice. All doctors licensed to practice in the UK must be registered with the GMC and, hence, the LRMP constitutes a full list of all eligible NHS consultants.

[20] Assessing the impact of the 2016 reform on the clinical activity rates of consultant is the focus of Chapter 3 of this thesis.

After imposing the sample restrictions described above, the analysis sample includes $23,518$ consultant-quarters corresponding to 963 award winners: 476 Bronze awards, 213 Silver, 46 Gold, 26 Platinum, and 202 awards lost.

3.2 Measures of Consultant Activity

My main outcome variable of interest is the number of care episodes provided by consultant $j = 1, \ldots, J$ in calendar quarter $t = 1, \ldots, T$, which serves as a measure of consultants' clinical activity as previously employed by Bloor et al. (2004, 2012) and Lafond et al. (2017). I use administrative, pseudonymised records from HES for all NHS-funded inpatient care provided in England. Each observation in HES reflects a finish consultant episode (FCE)[21], which represents the time spent under the care of a single consultant. To compute activity, I extract patient episodes from HES covering the period April 2009 and March 2015. I include all FCEs in public hospitals and all NHS-funded episodes delivered in private providers, including both elective and emergency care.

In secondary analysis, I use a more refined measure of clinical activity that accounts for differences in case complexity across consultants and specialties. For example, a consultant may handle fewer cases relative to her peers, not because she exerts less effort or works fewer hours, but because her patients have more severe and complex health problems that require a greater time input. Previous studies of consultant activity have addressed this issue by weighing each FCE by the national average reference cost for the Healthcare Resource Group (HRG)[22] to which the care episode is assigned (Bloor et al., 2004, 2012; Lafond et al., 2017). I argue that in the context of my study this procedure might not be suitable because it assumes that the labour input required to deliver different types of activities is directly related to their cost. This may not be the case. For example, hip replacement surgery is a fairly quick surgical procedure (1 to 2 hours in the operating room) and requires only a brief hospitalisation period

[21] An FCE is thus generated for every new admission and for every time responsibility for the care of a patient is transferred from one consultant to another. For example, a patient having a stroke may be admitted to an accident and emergency department, then transferred to a neurology department, and finally be transferred to a rehabilitation ward under the care of a geriatrician before being discharged. Three FCEs are generated in one single hospital stay.

[22] HRG is a patient classification system which groups conditions and procedures that use similar levels of resources.

(< 5 days) under the supervision of a consultant. However, the main cost of a hip replacement is for the joint implant. Cost-based weighting may therefore bias the results towards clinical care that requires substantial non-labour inputs. Instead, I weigh consultant activity by the national average length of stay (LOS) for the HRG to which the care episode is assigned. I use the national average to avoid issues due to differences across consultants in the propensity to discharge patients at a given time point in the recover pathway. This approach reflects that consultants have time budgets rather than cost budgets because they work under salaried contracts. Furthermore, since consultants remain responsible for patients under their care until they are discharged or transferred, LOS-weighted activity is arguably a more accurate measure of individual workload.

To formalize my procedure, let l_{ijht} be the length of stay for episode of care i undertaken by consultant j and belonging to HRG group h and quarter of the year t. I compute the national average length of stay for the HRG group h over the entire sample period as $\bar{l}_h = \frac{\sum_{i \in P_h} l_{ijht}}{n_h}$, where P_h and n_h describe the set and the number of episodes of care belonging to HRG h, respectively. I then compute LOS-weighted activity rates for consultant j at time t:

$$A_{jt} = \sum_h x_{jht} \bar{l}_h \tag{2.1}$$

where x_{jht} is the volume of output from consultant j in HRG group h in time t.

3.3 The Clinical Excellence Awards

I derive information on the year and level of awards from the annual ACCEA publications. Because applying to a new award requires holding the lower award level (e.g. only Bronze awards can apply to Silver awards, and so on), I construct dummy variables for each award that take the value of one in the period of the new award, and zero for the period before when a consultant holds the previous CEA (or no award in the case of Bronze). I also construct a dummy variable indicating whether the award is lost, which takes the value of zero in the period before.

3.4 Controls

I account for the level of seniority of consultants in the models to capture trends in working patterns. This is computed as the time, in years, since medical qualification.

3.5 Summary Statistics and Descriptive Patterns

Table 2.1 shows summary statistics for consultants who won Bronze (column (1)), Silver (column (2)), Gold (column (3)), and Platinum (column (4)), and for those who lost an award (column (5)). In each level of the awards, the majority of the consultants are male, hold a UK medical qualification, and work in medical specialties. Consultants who are given Platinum awards are, on average, older than those who win Bronze awards, by more than nine years. The average number of FCEs undertaken by consultants per quarter over the entire sample varies between approximately 77 (Platinum) and 153 (Bronze), and between 261 (Platinum) and 516 (Bronze) days of care for my measure of LOS-adjusted activity. This shows that the activity rates of Platinum award winners are almost half of those of consultants who are granted a Bronze award. The averages are significantly higher than the respective medians, reflecting that activity rates follow positively skewed distributions. Furthermore, the reported standard deviations show large variations between and within consultants. Variation across consultants is larger than that observed for each consultant over time - the latter being the relevant source of variation for the fixed-effects estimation.

Figure 2.1 provides time series of both activity measures in quarters relative to the time the award is given (or withdrew). For Bronze awards (Panel A), the data suggest that there is no clear pre-post trend. For Silver award winners (Panel B), the figure displays reductions in both outcomes in the post-award period, and possibly no pre-award trend. For Gold (Panel C)) and Platinum (Panel D) awards, activity rates decline across the observed period. Finally, those who lose an award (Panel E) experience a decrease in activity before the award is lost, and no clear pattern afterwards (unfortunately, my data do not include information beyond the first year after a consultant looses an award).

Table 2.1: Summary Statistics

	Bronze	Silver	Gold	Platinum	Lost
	(1)	(2)	(3)	(4)	(5)
Male	0.84	0.87	0.91	0.92	0.93
UK medical qualification	0.84	0.84	0.79	0.81	0.89
Seniority, first observed					
Mean (years)	22.65	26.50	28.82	31.77	33.19
SD	4.75	4.25	3.81	3.47	3.85
Specialty					
General Surgery	0.12	0.15	0.21	0.15	0.18
Urology	0.04	0.05	0.06	0.04	0.03
Trauma and Orthopaedics	0.06	0.05	0.09	0.00	0.06
Otorhinolaryngology	0.04	0.02	0.06	0.04	0.06
Ophthalmology	0.06	0.08	0.06	0.12	0.09
Neurosurgery	0.02	0.02	0.00	0.00	0.03
Plastic Surgery	0.02	0.01	0.00	0.00	0.01
Cardiothoracic Surgery	0.04	0.03	0.00	0.00	0.02
General Medicine	0.11	0.11	0.18	0.12	0.14
Gastroenterology	0.05	0.05	0.06	0.12	0.07
Clinical Haematology	0.05	0.03	0.00	0.04	0.04
Cardiology	0.09	0.07	0.06	0.08	0.06
Dermatology	0.02	0.03	0.03	0.04	0.02
Respiratory Medicine	0.04	0.06	0.00	0.08	0.04
Neurology	0.06	0.04	0.06	0.04	0.02
Rheumatology	0.05	0.06	0.03	0.00	0.05
Paediatrics	0.09	0.11	0.09	0.15	0.07
Geriatric Medicine	0.02	0.02	0.03	0.00	0.01
Share in surgical specialties	0.41	0.41	0.47	0.35	0.48
Activity					
Mean	153.29	144.08	144.28	77.65	147.13
Median	104.00	98.00	80.00	68.00	108.00
Overall-SD	184.39	148.28	255.96	57.05	148.09
Between-SD	165.87	130.74	195.35	43.11	123.22
Within-SD	75.31	64.81	151.93	38.41	78.88
LOS-adjusted Activity					
Mean	515.58	509.46	452.05	261.27	472.37
Median	353.10	322.60	288.68	226.53	303.96
Overall-SD	563.04	624.21	484.90	195.53	571.15
Between-SD	503.35	553.93	381.38	146.64	458.76
Within-SD	245.71	274.65	288.53	133.07	330.29
Consultants	476	213	46	26	202
Consultant-quarters	12,513	5,375	1,104	663	5,117

Notes: In all rows, except for “Seniority, first observed”, “Activity”, and “LOS-adjusted activity”, the table reports the share of consultants within the given population with the indicated characteristic. Specifically, for “Seniority, first observed” it reports the mean (row 3) and standard deviation (row 4), and for “Activity”, and “LOS-adjusted activity” it shows the mean (rows 24 and 29, respectively), and the overall (rows 25 and 30), between (rows 26 and 31), and within (rows 27 and 32) standard deviation.

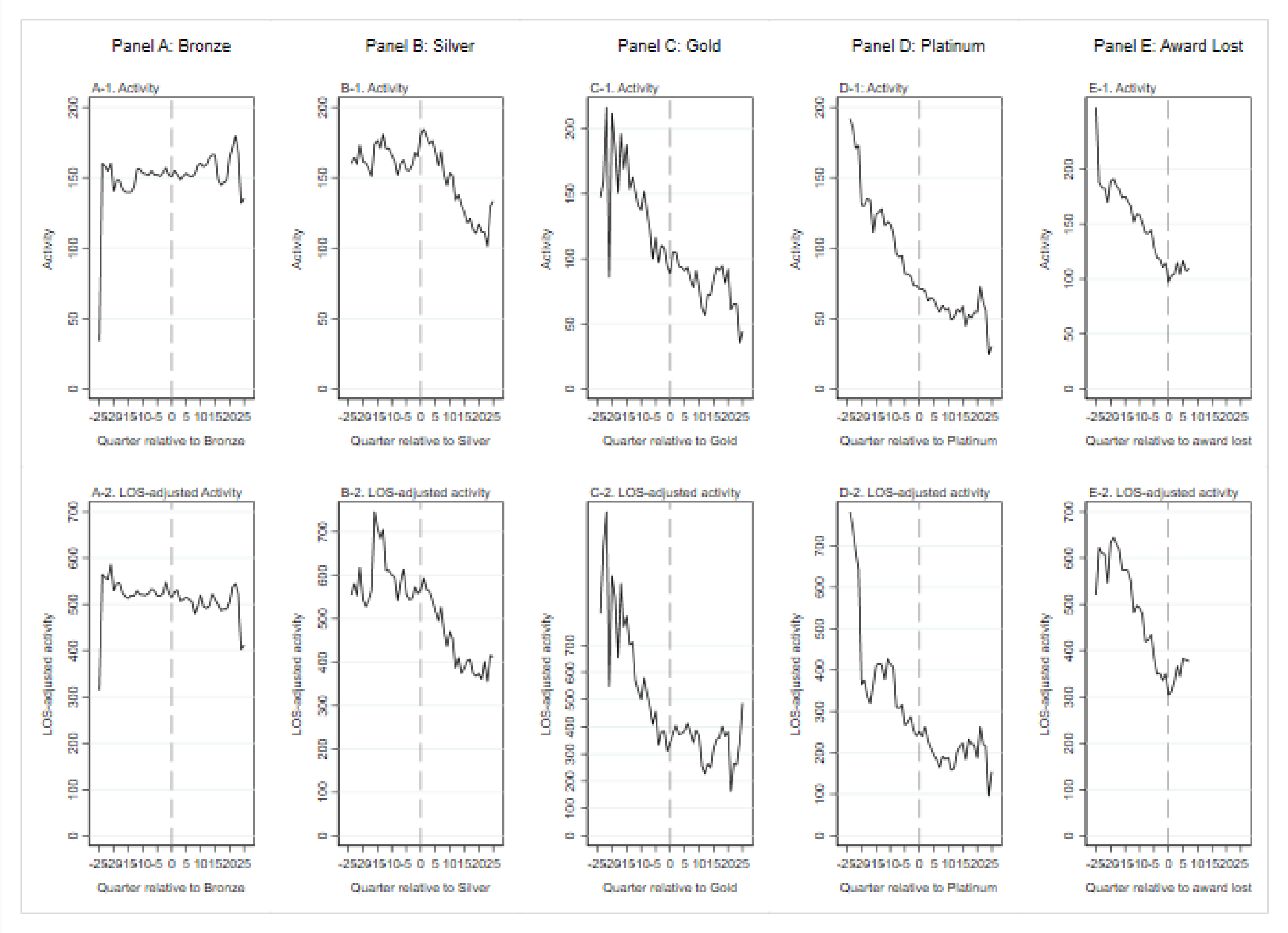

Figure 2.1: Levels of Activity and LOS-adjusted Activity Around the Time of Award

4 Empirical Strategy

4.1 Base Model

I seek to identify the effect of a change in award status (winning or losing an award) on the clinical activity of NHS consultants. I split my analysis by award level because i) the marginal increase in CEA payments provided by each new award, relative to the previous level, differs across them and ii) the descriptive analysis in Figure 2.1 shows that the trends in clinical activity differ for each award level around the time that the award is given. I, therefore, explore the panel nature of the data and start by estimating separately for each new award level regression models of consultant j's clinical activity in quarter t, denoted Y_{jt}, on her change in award status k in quarter t, W^k_{jt}. In each of the five models, W^k_{jt} is constructed to take the value of one in quarters t when a consultant j holds award k, and zero when she holds the previous award level, $k-1$. Recall that for the award lost dummy the previous award level is any of the awards a consultant has held before losing it[23]. The general form of my regressions is:

$$Y_{jt} = \beta^k W^k_{jt} + X'_{jt}\theta + \phi_j + \delta_t + \epsilon_{jt} \tag{2.2}$$

where X_{jt} is consultant's seniority measured in 5-year bands with separate categories for < 20 and ≥ 31 years; ϕ_j is a vector consultant fixed effects to account for time-invariant consultant heterogeneity; δ_t is a vector of quarter dummies to capture technological change and other shocks that are common to all consultants; and ϵ_{jt} is an idiosyncratic error term with zero mean and finite variance. The coefficient β^k in each model is the parameter of interest. It denotes deviations between pre- and post-award activity outcomes.

The fixed-effects strategy that I adopt differences out any time-invariant heterogeneity in consultant characteristics that may simultaneously affect the likelihood of winning (or losing) an award and their clinical activity. Since the awards are not randomly assigned to

[23] Due to small numbers of observations in this category, I am unable to split the analysis by each award level that is lost.

consultants, but instead result potentially from consultants' unobserved individual differences (e.g. individual ability, innate motivation, and background), eliminating by means of consultant fixed effects such relevant time-constant factors reduces the sources of potential selection bias. Naturally, this strategy also cancels out any of the time-invariant differences reported in Table 2.1, as well as institutional factors specific to the hospitals where they practice (e.g. teaching status) and the clinical specialties they belong to (e.g. job characteristics).

As a result of employing a fixed-effects approach, the effect of the awards on clinical activity is estimated only using within-consultant variation. This has two limitations. First, it follows that the effects of the awards can only be identified through changes in award levels, and thus any consultant who does not experience a change in award status during the sample period does not contribute to the estimation, which substantially reduces the sample size. Second, by means of consultant fixed effects any between-individual variation is discarded, and thus this approach effectively trades off efficiency for a reduction in bias.

One other limitation of my approach is that the before-after estimator treats each consultant as their own control, and thus any change in the observed values of Y_{jt} in the post-award period is assumed to result exclusively from the change in award status. However, this assumption may be violated if other concurring events in the post-award period affect consultants' clinical activity. To relax this assumption a difference-in-differences approach could be employed if a suitable control could be found. This would require identifying a control group that would respond similarly to the awards and which is subject to the same external influences as the group of award winners. However, the pool of potential controls in my primary sample is ill-suited for this task because it includes two distinct groups of consultants: those who do not pursue the awards at all, and those who did but failed in their application. And because the latter group might react differently due to failure, this approach would bias the estimation of the effect of interest in my work.

The sample descriptive evidence in Table 2.1 suggests that the distributions of the measures of clinical activity are strongly right-skewed with heavy tails due to a small group of consultants with exceptionally high levels of activity. Under such distributional qualities of the dependent variables, traditional OLS regression has been shown to perform poorly (Jones, 2009). I also detour from log-linear forms frequently found in the literature to avoid known re-transformation issues when heteroskedasticity is present in the data on the transformed scale (Manning and Mullahy, 2001; Buntin and Zaslavsky, 2004; Mihaylova et al., 2011).

Alternatively, Poisson regression has been proven useful to obtain unbiased and precise estimates of the parameters of interest in the context of panel data with substantial skewness in outcomes (Buntin and Zaslavsky, 2004; Santos Silva and Tenreyro, 2006). The econometrics literature has shown that the distributional assumption of the Poisson estimator, namely that the mean is equivalent to the variance, can be relaxed through use of robust standard errors (see Santos Silva and Tenreyro (2006) for a detailed discussion). The only assumption required for consistent estimation is that, as in ordinary least squares, the conditional mean is correctly specified (Santos Silva and Tenreyro, 2006; Wooldridge, 2010). This result extended the application of the Poisson regression beyond the realm of count data to any dependent variable that follows a process with non-negative integer values. Therefore, equation 2.2 is estimated as a Poisson model with heteroscedasticity-robust standard errors (Wooldridge, 2010).

4.2 Model Specification Allowing for Anticipation or Adaptation

The regression model in specification 2.2 estimates the effect of the awards on the measures of consultant activity by comparing the activity levels of consultants who experience a change in award status before and after the change occurred. The effect is consistently estimated under the assumption that neither anticipation nor adaptation is present. In other words, the clinical activity of consultants is unaffected until the award is given, and thereafter the effect is permanent (or constant).

However, it is possible that consultants change their clinical activity in the five years of the assessment period that precedes the award application. In this case, a simple before-after comparison of outcomes would lead to a biased estimation of the effect of winning (or losing) CEAs on clinical activity, which in fact becomes a mixture of both the effect of the award at t

and an anticipation effect[24]. Furthermore, neglecting any adaptation in the post-award period may lead to short-run effects being covered by longer average effects.

I address these issues by studying the dynamic pattern of consultants' clinical activity relative to the awards by replacing the award dummy W_{jt}^k in equation 2.2 with a series of lead and lag year dummies using the following specification:

$$Y_{jt} = \sum_{s=-5}^{4} \beta_s^k W_{jts}^k + X_{jt}'\theta + \delta_t + \phi_j + \epsilon_{jt} \tag{2.3}$$

where $s = s' - s_0'$ is the year-specific index relative to the year of the change in award status, s_0'. The award dummies for $s < 0$ allow for year-on-year changes in clinical activity resulting from consultants altering their practice behaviour before the award is given (anticipation effects). The award dummies for $s \geq 0$ allow for any post-award effects to change across by year (adaptation effects). Similar to Equation (2.2), Equation (2.3) is estimated as a Poisson model with heteroscedasticity-robust standard errors (Wooldridge, 2010).

5 Results

5.1 Baseline Models

Table 2.2 presents the results from estimating Equation (2.2) independently for each award level, showing the estimated coefficients and standard errors. Each panel in the table refers to a model estimating the effect of Bronze (panel A), Silver (panel B), Gold (panel C), and Platinum awards (panel D), and for when an award is lost (panel E). The outcomes are consultant activity in columns (1) and (2), and LOS-adjusted activity in columns (3) and (4). In each column, I present five models, one for each of the award levels. Models in columns (1) and (2) - and (3) and (4) - only differ in the inclusion of the seniority in the latter. Note that in each model, the award dummies are constructed using the previous award level as the reference category so that, for example, the coefficient for the Gold award dummy denotes

[24] Anticipation effects are well-documented in the labour economics literature. Ashenfelter (1978) showed that neglecting a pre-program dip in earnings in participants of government-funded job training programs in the US overestimates the training effect.

deviations in outcome measures from the Silver award, and so on. Because the effect can only be identified through changes in award levels in a fixed effects approach, the results in Table 2.2 are estimated using only consultants for whom a change in award status is observed.

For Bronze awards (Panel A), the estimate in column (1) does not change when I control for seniority in column (2). The estimated coefficient is negative and statistically significant at the 1% level. Consultants who are granted Bronze awards experience an average reduction in quarterly activity of almost 8 episodes of care, a 5% reduction relative to the mean. When using LOS-adjusted activity as outcome in column (3) and (4), the coefficients halve in magnitude and become statistically indistinguishable from zero. For Silver awards (Panel B), controlling for the seniority level of consultants in column (2) produces an estimate marginally larger than that in column (1) that reaches statistically significance at the 10% level. However, the positive effect observed is small: winning a Silver award is associated with an average increase of 5 episodes of care, an increase of nearly 3.7% relative to the mean. This effect becomes negative when using LOS-adjusted activity as the outcome in column (3) and (4) and is statistically insignificant. The coefficients for Gold in Panel C are negative and statistically significant at the 5% and 1% level in columns (1) and (3), respectively, but the observed effect is substantially reduced when I control for consultant seniority in columns (2) and (4) and is no longer statistically significant. For Platinum (Panel D), the results are negative, but show no statistically detectable differences between pre- and post-award activity. Although the coefficients associated with this award suggest that there might be a modest decrease in both activity and LOS-adjusted activity, very much as for Gold awards, my sample may be too small to detect a statistically significant effect. Finally, losing an award (Panel E) does not seem to be associated with changes in clinical activity.

5.2 The Dynamic Effect of the CEAs

Estimation of the leads and lags model (Equation (2.3)) results in a large number of coefficient estimates. For ease of presentation, I plot the estimated coefficients for each award level in Figures 2.2 and 2.3 and present full regression outputs in Tables B.3 and B.4 in the Appendix.

Table 2.2: Consultant Activity and the Awards

Dependent Variable:	Activity		LOS-adjusted activity	
	Quarter FE	+ Seniority	Quarter FE	+ Seniority
	(1)	(2)	(3)	(4)
Panel A: Bronze				
Award	-0.048^{***}	-0.048^{***}	-0.023	-0.023
	(0.015)	(0.015)	(0.014)	(0.014)
Consultant-quarters	12,513	12,513	12,513	12,513
Consultants	476	476	476	476
Panel B: Silver				
Award	0.033	0.037^{*}	-0.014	-0.011
	(0.022)	(0.021)	(0.027)	(0.026)
Consultant-quarters	5,375	5,375	5,375	5,375
Consultants	213	213	213	213
Panel C: Gold				
Award	-0.128^{**}	-0.035	-0.145^{***}	-0.078
	(0.064)	(0.061)	(0.070)	(0.063)
Consultant-quarters	1,104	1,104	1,104	1,104
Consultants	46	46	46	46
Panel D: Platinum				
Award	-0.081	-0.096	-0.090	-0.079
	(0.059)	(0.062)	(0.066)	(0.068)
Consultant-quarters	663	663	663	663
Consultants	26	26	26	26
Panel E: Lost				
Award	-0.056	-0.027	0.025	0.064
	(0.046)	(0.046)	(0.057)	(0.057)
Consultant-quarters	5,117	5,117	5,117	5,117
Consultants	202	202	202	202
Seniority	NO	YES	NO	YES
Quarter FE	YES	YES	YES	YES
Consultant FE	YES	YES	YES	YES

Notes: */**/*** indicate statistical significance at the 10%/5%/1% level. Standard errors in parentheses under coefficients are robust to arbitrary heteroskedasticity and autocorrelation.

Each point estimate is shown with associated 95% confidence intervals, which are calculated from standard errors that are robust to heteroscedasticity. Year zero is the first year of the award and is identified in the plots by the vertical dashed lines. I also normalize to zero the coefficient in the year before the award (reference base).

The dynamic patterns of each award level are similar for activity (in Figure 2.2) and LOS-adjusted activity (in Figure 2.3). The coefficients for Bronze award (Panel A) show no pre- and post-award relationship with activity measures (except for a positive statistically significant increase in year four after the award). For Silver (Panel B), there is no evidence of changes in the years before and one year after the award. However, I observe suggestive evidence of post-award reduction in activity and LOS-adjusted activity, which is statistically significant in both measures in the second and third year of the post-award period and - for unweighted clinical activity - also in year four. Gold (Panel C) award winners experience a positive and statistically significant (at the 1% level) increase in activity rates two years prior to the award (in relation to the year before the award), which is significantly reduced when using LOS-adjusted activity as the outcome. For Platinum (Panel D), there are no observed statistically significant dynamic patterns in leads and lags. For both Gold and Platinum awards, the estimates are imprecise, displaying with large confidence intervals as we move further away into future and past years relative to the award due to small counts in cells (see counts of leads and lags in Table B.5). Moreover, I do not observe statistically significant pre-event trends for consultants who loose an award (Panel E), nor any effects in the year that follows (unfortunately, my data do not include information beyond the first year after a consultant looses an award). Overall, there is no clear evidence of lead and lagged effects of the awards, except for a modest post-award reduction for Silver awards.

6 Conclusions

This chapter addresses the question of whether and to what degree the post-award clinical activity of NHS consultants is affected by the receipt and withdrawal of CEAs using seven years of data information on NHS doctors. The data I use follows NHS consultants over time,

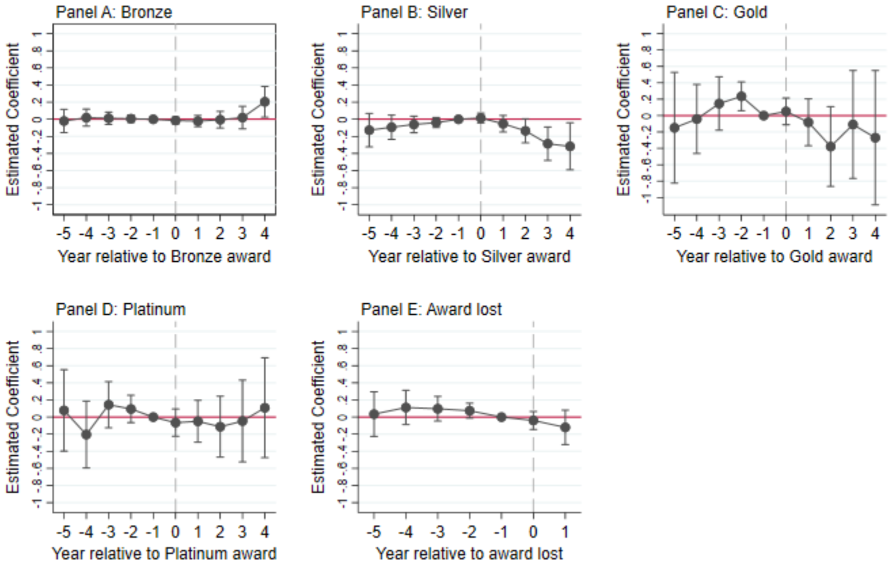

Figure 2.2: The Dynamic Effect of a Change in Award Status on Clinical Activity

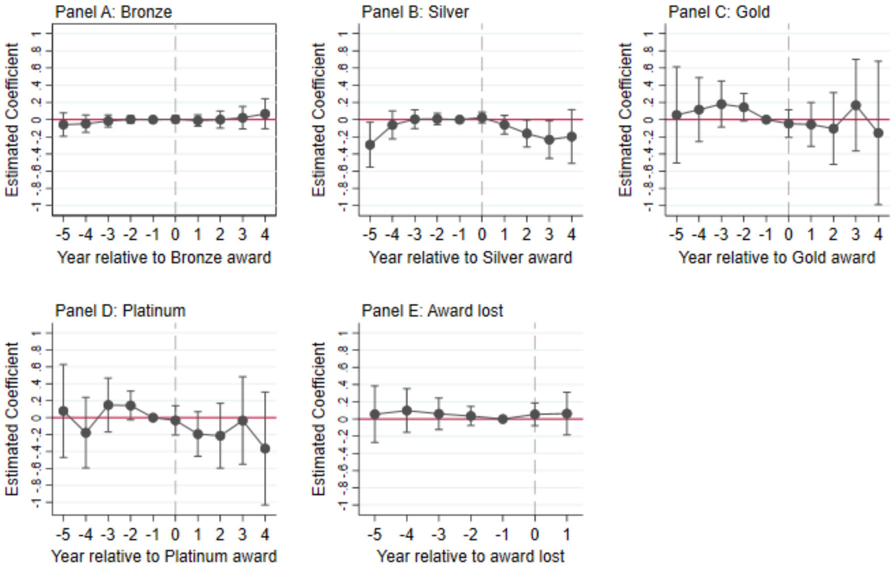

Figure 2.3: The Dynamic Effect of a Change in Award Status on LOS-adjusted Activity

and thus provide a unique opportunity to identify the effect of the awards on two different measures of activity while accounting for unobserved heterogeneity across consultants (e.g., if innate motivation or skill drive both the probability of receiving an award and activity rates). To account for dynamic effects of the awards on clinical activity, I also extend the models to include leads and lags of award status changes. I argue that failing to account for anticipation related to the pre-award incentive effects may bias the results due to consultants adjusting their workload while pursuing higher award levels.

My analysis shows weak evidence of a negative effect of Bronze and Silver awards on the clinical activity of consultants, but these findings are sensitive to the econometric modelling approach. Winning Gold and Platinum awards does not seem to have an impact on post-award consultant activity. However, this nil effect must be interpreted with caution as it could be due to the limited power of the analysis. Moreover, losing an award does not have an impact on the activity rates of consultants. Finally, I do not find evidence of anticipation effects for any of the award levels.

This chapter has a number of limitations. First, and most importantly, CEAs are not randomly assigned to consultants but the observed allocation is potentially the result of unobserved factors that can lead to selection bias in the estimated relationships of interest. Second, my analysis is likely low powered due to small sample sizes, particularly at the highest award levels. Third, the retirement plan of consultants is likely to affect both the probability of applying for an award and the labour supply decisions of consultants. If consultants approaching retirement simultaneously change their activity rates and stop pursuing awards, then my findings would be biased by unobserved retirement plans. Fourth, a residual endogeneity bias may remain due to the confounding effects of unobserved time-varying consultant characteristics such as their life arrangements (e.g., family background) or health status. Finally, the small sample size hampers meaningful stratified analysis to identify heterogeneous effects of CEAs (e.g. differences by gender or place of training).

Future research should seek to overcome some of the limitations of this chapter by using larger, longer and more detailed panel data that are better able to capture the application decision process (i.e. which doctors applied but failed) and contain more information on possible time-varying determinants of clinical activity that are potentially correlated with awards.

Furthermore, the econometric literature provides a few satisfying solutions to address some of these threats if more detailed data on the awards become available to researchers. For example, the before-after comparison of activity rates of award winners I adopt would be improved in a difference-in-differences framework if a suitable comparison group could be identified. This approach would reduce or eliminate the importance of trends in the activity rates of consultants. The pool of potential controls in my sample is, however, ill-suited for this task without information on the candidates to the awards because it includes both consultants who do not pursue the awards at all, and those who did but failed in their application.

In conclusion, this study finds weak evidence of unintended consequences of CEAs in the form of a reduction in clinical activity rates following the successful application for a CEA. Some critics advocate that the scheme should be abolished altogether (Bloor et al., 2012), but the awards may not "hurt" the clinical activity of consultants in a meaningful way and are likely to contribute to hard-to-measure activities (e.g. innovation) which are relevant for the NHS. Furthermore, the CEA are an attractive feature of consultant contracts, likely to offer high-powered incentives for the retention of high-skilled doctors (and turnover of low achievers over time).

B Appendix

Table B.1: Domains in the Assessment of Applications to the Clinical Excellence Awards

Domain	Description
1 - Delivering a high-quality service	Evidence of achievements in delivering a service that is safe, has measurable effective clinical outcomes, good patient experience, and where opportunities for improvement are consistently sought and implemented. This should include quantified measures if these exist that reflect the whole service they provide, using Indicators for Quality Improvement or Quality Standards and other reference data sources in England. The evidence on patient safety should refer where possible to the new quality indicators and the evidence on the patient experience should indicate how they have addressed the issues of dignity, compassion and integrity with patients.
2 - Developing a high-quality service	Evidence should demonstrate how applicants have significantly enhanced clinical effectiveness (the quality, safety and cost effectiveness) of their local service or related clinical service widely within the NHS.
3 - Leadership and managing a high-quality service	Evidence should show how applicants have made a substantial personal contribution to leading and managing a local service, or to national or international health policy development.
4 - Research and innovation	Applicants should outline contributions to research or to innovation including developing evidence base for the measurement of quality improvement
5 - Teaching and Training	Evidence should show how teaching and training forms a major part of the contribution applicants make to the NHS, over and above contractual obligations.

Source: Advisory Committee on Clinical Excellence Awards (2014).

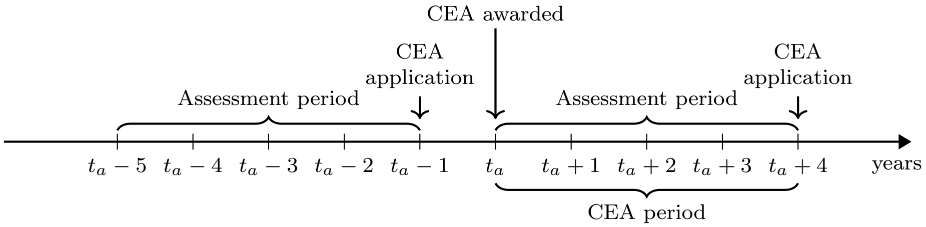

Figure B.1: Timeline Illustrating Different Periods of the Application Process to the Clinical Excellence Awards

Notes: The figure shows a timeline of the application process to the Clinical Excellence Awards. Consultants apply to a new CEA in year $t_a - 1$ by providing evidence of work performance from the five years prior to the date of the application (i.e. assessment period, which runs from $t_a - 5$ to $t_a - 1$). If successful, CEA payments are made for five years from year t_a to $t_a + 4$, after which a new application is required to renew the award (for another five years) or to pursue a higher level CEA. Evidence for the new application refers to the period between t_a and $t_a + 4$.

Table B.2: Sample Selection

Exclusion Rules	Sample Size (% of primary sample)
0. Primary sample (after merging all three data sets)	32,619 (100)
1. Only consultants in 18 specialties	31,661 (97.1)
2. Excluding consultants with single year observation	27,499 (84.3)
3. Excluding consultants with gaps in panel	25,389 (77.8)
4. Study sample	963 (0.03)

Notes: The table reports the number of consultants in the sample under different restrictions on the set of all consultants in my primary study sample (in Row 1). Each row puts a further restriction on the sample compared to the row above it (e.g. Row (2) is a strict sub-sample of Row (1), and so on).

Table B.3: Leads and Lags of the Awards, Activity - Full regression output

	Bronze	Silver	Gold	Platinum	Lost
Years	(1)	(2)	(3)	(4)	(5)
Leads and Lags					
-5	-0.020	-0.127	-0.149	0.078	0.034
	(0.069)	(0.100)	(0.345)	(0.243)	(0.133)
-4	0.019	-0.093	-0.042	-0.204	0.112
	(0.051)	(0.072)	(0.215)	(0.199)	(0.102)
-3	0.012	-0.060	0.147	0.144	0.097
	(0.037)	(0.049)	(0.166)	(0.138)	(0.073)
-2	0.007	-0.039	0.235^{***}	0.095	0.074^{*}
	(0.022)	(0.030)	(0.090)	(0.082)	(0.044)
-1	(ref)	(ref)	(ref)	(ref)	(ref)
0	-0.014	0.015	0.051	-0.065	-0.040
	(0.020)	(0.029)	(0.083)	(0.081)	(0.053)
1	-0.022	-0.052	-0.082	-0.049	-0.120
	(0.034)	(0.049)	(0.146)	(0.125)	(0.103)
2	-0.006	-0.135^{*}	-0.378	-0.113	
	(0.050)	(0.072)	(0.248)	(0.182)	
3	0.020	-0.286^{***}	-0.108	-0.046	
	(0.067)	(0.099)	(0.336)	(0.244)	
4	0.205^{**}	-0.316^{**}	-0.270	0.109	
	(0.091)	(0.140)	(0.418)	(0.298)	
Seniority	YES	YES	YES	YES	YES
Quarter FE	YES	YES	YES	YES	YES
Consultant FE	YES	YES	YES	YES	YES
Consultants	476	213	46	26	202
Consultant-years	12,513	5,375	1,104	663	5,117

Notes: */**/*** indicate statistical significance at the 10%/5%/1% level. Standard errors in parentheses under coefficients are robust to arbitrary heteroskedasticity and autocorrelation.

Table B.4: Leads and Lags of the Awards, LOS-adjusted activity - Full regression output

	Bronze	Silver	Gold	Platinum	Lost
Years	(1)	(2)	(3)	(4)	(5)
Leads and Lags					
-5	-0.058	-0.292^{**}	0.054	0.079	0.055
	(0.069)	(0.133)	(0.285)	(0.280)	(0.168)
-4	-0.048	-0.062	0.117	-0.178	0.098
	(0.052)	(0.083)	(0.189)	(0.213)	(0.130)
-3	-0.016	0.005	0.181	0.149	0.060
	(0.036)	(0.056)	(0.137)	(0.163)	(0.093)
-2	0.002	0.008	0.144^{*}	0.143	0.035
	(0.021)	(0.034)	(0.082)	(0.087)	(0.057)
-1	(ref)	(ref)	(ref)	(ref)	(ref)
0	0.005	0.024	-0.047	-0.033	0.054
	(0.021)	(0.032)	(0.082)	(0.089)	(0.067)
1	-0.010	-0.060	-0.056	-0.194	0.063
	(0.034)	(0.055)	(0.130)	(0.134)	(0.126)
2	-0.000	-0.161^{**}	-0.104	-0.214	
	(0.051)	(0.080)	(0.213)	(0.195)	
3	0.022	-0.233^{**}	0.168	-0.034	
	(0.067)	(0.111)	(0.272)	(0.264)	
4	0.067	-0.197	-0.155	-0.365	
	(0.089)	(0.159)	(0.425)	(0.340)	
Seniority	YES	YES	YES	YES	YES
Quarter FE	YES	YES	YES	YES	YES
Consultant FE	YES	YES	YES	YES	YES
Consultants	476	213	46	26	202
Consultant-years	12,513	5,375	1,104	663	5,117

Notes: */**/*** indicate statistical significance at the 10%/5%/1% level. Standard errors in parentheses under coefficients are robust to arbitrary heteroskedasticity and autocorrelation.

Table B.5: Number of observations of lags and leads

	Bronze	Silver	Gold	Platinum	Lost
Years	(1)	(2)	(3)	(4)	(5)
-5	59	28	5	1	91
-4	189	106	24	6	202
-3	265	144	36	14	202
-2	347	171	40	20	202
-1	476	213	46	26	202
0	476	213	46	26	202
1	412	152	31	25	76
2	336	118	22	22	.
3	258	86	13	16	.
4	74	19	3	2	.

Chapter 3

Doctors' Take-Home Pay and NHS Activity: Evidence from a UK Pension Reform

1 Introduction

Throughout the world, healthcare policymakers are confronted with the need to ensure a sufficient supply of doctor services to meet the growing demand for health care. Attempts to expand the medical workforce through increases in medical school enrollments (McPake et al., 2014) may be suitable long-term solutions but do little to ease short-term shortages.[25] The pay structure of doctors is a potential short-term policy tool through which payers can promote increased doctor productivity (Sloan, 1975), and it remains a recurring topic of debate between doctor associations and payers in several high-income countries including the UK (Quentin et al., 2018).[26] However, there is limited empirical evidence on how changes in remuneration of salaried doctors, such as hospital doctors working in the English NHS, influence their labour supply decisions to guide appropriate and 'fair' remuneration policies that secure doctor productivity.

In this paper, we exploit a 2016 UK pension reform to study how the short-run labour

[25] The market for medical labour is characterised by long periods of training and high levels of specialisation (Lee et al., 2019; Nicholson and Propper, 2011) so that supply cannot easily be expanded in the short term.

[26] The NHS consultant workforce has experienced a series of pay 'freezes' over the last decade (Lee et al., 2019) and the system has relied heavily on the goodwill of consultants, many of whom work in excess of their contracted hours without adequate compensation (British Medical Association, 2020). The British Medical Association has called for a pay rise of 5% for consultants based on the assertion that average consultant pay in 2019/20 remains below 2008/09 levels in real terms, i.e. after adjusting for inflation, following a period of public sector austerity cuts (British Medical Association, 2020). The UK government has promised to deliver a 1% pay increase this year (Rimmer, 2021).

supply of salaried senior hospital doctors working in the NHS, known as consultants[27], responds to changes in take-home pay. UK tax payers are able to make tax-free annual contributions to their pension scheme up to a maximum limit - known as the *annual allowance* (AA) - that is set by the UK government. Any pension payments above the AA are subject to a tax charge at an individual's marginal income tax. In 2016, the government introduced a taper to reduce the annual allowance of individuals with higher taxable income. The tapering applied to those whose 'threshold' income[28] exceeds £110,000 and whose 'adjusted' income[29] exceeds £150,000 per annum. For individuals with a threshold income of less than £110,000 the tapering did not apply, regardless of the level of adjusted income. For those exceeding both limits, every £2 of adjusted income over £150,000 reduced the annual amount they can contribute to their pension free of income tax (their AA) by £1, from a standard value of £40,000 down to a minimum of £10,000. Once the threshold is reached, the impact of the reform can be severe creating what has been described as a 'cliff edge' effect. Furthermore, the complexity of the tapered annual allowance calculation prevents workers from making informed decisions of their tax position *a priori* (Thurley, 2020). The upshot of the policy change was an unexpected tax surcharge to be paid upfront, which in effect impacted individuals' take-home pay.

Standard economic models of labour supply suggest that the effect of the pension reform on after-tax income may shift consultants into a different point of the labour-leisure trade-off, yet the expected response is ambiguous (Borjas, G. J., 2008). Because the income effect[30] and the substitution effect[31] work in opposite directions, consultants' responses to the pension reform are determined by their individual preferences over consumption and leisure. Put differently, some consultants may respond by working more to maintain pre-reform income levels, whereas others may work less because the leisure that would be forgone is valued higher than the consumption enabled through income. This ambiguity generally motivates empirical investigations to guide policy on the labour supply responses to changes in pay.

[27] Consultants are the most senior medical and surgical hospital staff in the NHS, who have expert knowledge in their specialties and lead the delivery of publicly-funded care.

[28] 'Threshold' income is the total taxable income, but net of the value of any employee pension contributions.

[29] 'Adjusted' income is the total taxable income plus the real growth in value of pension rights over the year.

[30] For example, workers choose to work less and enjoy more leisure time as take-home pay rises — provided leisure is a normal good.

[31] That is, the opportunity cost of leisure, which incentivises more work as income rises.

There is some evidence that consultants may have reduced their individual labour supply to reduce tax exposure following the new pension arrangements. In 2019, the Royal College of Physicians conducted a survey of 2,800 medical consultants aged 50 to 65, which revealed that 38% of clinicians reported to have received a tax charge due to exceeding their annual pension allowance threshold (British Medical Association, 2020). As a consequence, half of the consultants surveyed said that they would retire at a younger age than previously planned; two in three said that they have avoided taking on additional paid work beyond their core contract (e.g. to cover for colleagues who are on sick leave); one in four reduced their contracted hours; and one in five reported having stepped down from a leadership or other role with extra remuneration (British Medical Association, 2020). These findings suggest that the pension reform may have had an impact both on the number of hours worked (intensive margin) and doctors' decisions to work or not (extensive margin).

Neither work hours nor effort are directly observable in routinely collected healthcare data. We construct doctors' pay using its two largest components, the Basic Pay Scale, and publicly available information on the national Clinical Excellence Awards, a form of bonus payment available to NHS hospital doctors. We measure doctors' activity as the annual count of finished episodes of care per doctor using English hospital discharge data, where a single episode is defined as a period of health care under the responsibility of one consultant in one hospital. To account for case complexity and doctors' effort, we employ the adjustment procedure developed in Chapter 2, which weights each episode of care using the national average of length of stay for comparable cases.

At least two econometric issues have hindered rigorous empirical investigations of the effect of doctor earnings on labour supply: the endogeneity of earnings, which may themselves be determined by hours worked, and omitted variable bias due to unobserved effects (e.g. motivation and ability) driving both earnings and hours worked (Lee et al., 2019). The empirical strategy we adopt to address these endogeneity issues builds on the exogenous decrease in AA due to the UK pension reform to estimate the causal relationship between doctors' take-home pay and their level of clinical activity. We conduct a difference-in-difference (DID) analysis of clinical activity rates between high-earning doctors affected by the pension reform, those earning in excess of £110,000, and lower-earning doctors, those with earnings below £95,000, that are unlikely to be affected directly. We exploit a large pool of potential controls in our study sample to pre-process the data using entropy balancing (EB) (Hainmueller, 2012; Hainmueller and Xu, 2013) prior to parametric modelling, to balance pre-reform activity trends and baseline observable characteristics of doctors (Cefalu et al., 2020). By combining EB with DID (Marcus, 2013; Freier et al., 2015; Everding and Marcus, 2020), our policy estimates are robust against selection on observable and unobserved time-invariant omitted variables. We further test for pre-existing trends and lagged effects of the reform using a complementary event study approach. Finally, we test whether any observed effect is due to a cohort effect (i.e. the cohort of doctors in our sample ages over time and, thus, becomes more likely to reduce activity towards retirement) through a placebo test.

This chapter provides new evidence on the effect of remuneration on the clinical activity rates of consultants working in the English NHS. It contributes to a scant literature on the labour supply responses of salaried and employed doctors to changes in remuneration (see literature review below). We find that exposure to the 2016 UK pension reform led to a 8.4 percent post-reform drop in episodes of care provided, and a 8.7 percent reduction in work effort equivalent to days of patient care. Both results are robust to a range of sensitivity analysis. The placebo test suggests that our results are unlikely to be driven by cohort effects.

In what follows, we provide a review of the literature on the doctor labour supply responses to changes in remuneration in Section 2, and background on the 2016 pension reform in Section 3. Sections 4 and 5 present our empirical strategy and describe the data and variable construction. Section 6 presents our results and report a set of robustness tests. Finally, Section 7 summarises and concludes.

2 Literature on Doctor Labour Supply

Despite a large economic literature on the effects of taxes and wages on labour supply (e.g. Hausman (1985); Blundell, R. and MaCurdy (1999); see Keane (2011) for a survey of this literature), empirical evidence on the labour supply of doctors and the relative importance of income and substitution effects remains limited (Nicholson and Propper, 2011; Lee et al., 2019).

Early estimates of doctor labour supply elasticities were derived from cross-sectional and aggregated data (Feldstein, 1970; Sloan, 1975; Vahovich, 1977; Brown and Lapan, 1979; Yang, 1987; Brown, 1989). These studies typically find that doctors are not very responsive to changes in remuneration, and may even decrease effort with increases at the top of the distribution (giving rise to what is termed a backward-bending supply curve). However, they fail to address one of the central issues in the labour supply literature: the endogeneity of wages. That is, wages are likely to be correlated with unobserved preferences for work, which also determine the amount of hours worked.

More recent studies attempt to correct for endogeneity issues using doctor experience (Rizzo and Blumenthal, 1994), variation across US states in the maximum marginal tax (Showalter and Thurston, 1997), and market-level demand variables such as per capita income and degree of urbanization (Thornton and Eakin, 1997) as instrumental variables for doctors' wages. Rizzo and Blumenthal (1994) and Showalter and Thurston (1997) estimate small and positive short-run wage elasticities for self-employed doctors of 0.23 and 0.33, respectively. In contrast, Thornton and Eakin (1997) report small negative uncompensated wage and income elasticities, suggesting that the labour supply curve may indeed be backward bending.

The labour supply studies above are from the US, where doctors are self-employed, rather than employed (and salaried) as in the English NHS. The few studies inferring labour supply responsiveness in settings where doctors are employed are mostly from Norway. Baltagi et al. (2005), Sæther (2005), and Andreassen et al. (2013) use comprehensive micro data on employed Norwegian doctors to estimate labour supply models that identify the wage effect by exploiting exogenous changes in national wage settlements. Their estimates of short-run uncompensated wage elasticities vary between 0.18 and 0.55, which are broadly in line with those found in the US.

Our paper is most closely related in setting to the work of Ikenwilo and Scott (2007), which analyses data from a Scottish survey of consultants working in the NHS to estimate a modified labour supply model that includes job satisfaction. They report small short-run uncompensated earnings elasticities equal to 0.09, without controls for job satisfaction, and 0.12, when adjusting for job quality. Furthermore, elasticities are lower for consultants in full-time employment and male doctors. Although Ikenwilo and Scott (2007) use a Generalised Method of Moments estimator to account for the endogeneity of earnings and job quality, their estimates are limited by the cross-sectional nature of the data and measurement error in self-reported measures of earnings and hours worked.

Taken together, the studies above suggest that doctors' labour supply is relatively inelastic to changes in income and earnings, much like workers in other labour markets (Keane, 2011).

3 The UK Pension Reform and its Impact on Consultants

The UK pension system consists of a modest tax-funded state pension and a large private pension sector, which grew substantially in the last century and exists primarily in the context of employer-employee arrangements. Most workplace pension plans are either organised in defined-benefit plans, which pay pension benefits based on years of service and salary or on career average revalued earnings, or in defined-contribution schemes that pay according to employer and employee contributions made over an employee's entire career plus any investment returns on their pension pot. NHS consultants make contributions to the NHS Pension scheme, which, like many other public service pension schemes, is a defined-benefits plan with the size of the pension benefits linked to career average revalued earnings (Danzer et al., 2016).

The UK government incentivises workers to save into private pension schemes by allowing them to make contributions from their pre-tax income, thus saving on income tax. The amount of tax-free contributions that individuals can make is limited: The AA limits the amount by which a worker's pension pot can grow tax-free in a single year, whilst the lifetime allowance (LTA) limits the size of tax-free contributions made by taxpayers during their entire career.[32] Where pension contributions exceed the AA and LTA, individuals are allowed to

[32] Both allowances were one of the main features of a radical reform of the tax treatment of private pension

use any unused allowance from the previous three years to absorb any accruals in excess of these limits. If these are used up, a tax charge is applied at the marginal income tax on the remaining excess, which must be paid upfront.

The generosity of the AA and LTA upon introduction - the AA was set at £215,000 and the LTA at £1.5 million pounds - was successively reduced in stages as part of a set of policies designed to increase tax revenue and reduced the disproportionately high benefits going to high earners. In 2016, the AA was set at £40,000 and the LTA at £1,055,000. The new pension reform, which we explore in this paper, further reduced the benefits of private pension savings for high-income earners by introducing a taper to the AA.

The interaction between the new pension tax regime, the NHS Pension Scheme, and the nature of consultants roles has posed significant challenges to this relevant group of NHS staff for several reasons. First, the English NHS faces excess demand for hospital care - as is evidenced by substantial waiting lists for publicly-funded treatments - and consultants are regularly asked to carry out additional clinical sessions over and above their contracted working hours to reduce waiting lists and fill rota gaps. These additional sessions attract further remuneration and are generally lucrative for consultants. Pay derived from this service is non-pensionable and does not contribute to growth of the pension pot. However, it counts towards the tapering calculation, thereby potentially increasing the size of the tax charge. Consequently, declining or reducing extra clinical work is one mechanism through which consultants may adjust their tax exposure. Second, members of the NHS Pension Scheme pay a fixed proportion of their income into their pension pot and cannot adjust these personal pension contributions to fit within the AA. This inflexibility means that increases in pensionable pay due to, for example, a new clinical or medical director role or a Clinical Excellence Award generates a spike in pension growth that lead to additional tax charges. Third, the tapering calculation is complex and extremely difficult to conduct, which reduces consultants ability to predict or accurately adjust their behaviour *before* they receive their first tax charges. Finally, consultants may opt out of the NHS Pension scheme but they would lose valuable ancillary benefits such as death-in-service entitlements and initial survivor pensions for their family

savings in the UK on April 2006 - which became known as the “A-day” reform (Thurley, 2020).

members, or enhanced ill-health retirement options. As a consequence, very few consultants opt out of the NHS Pension scheme.

Motivated by the issues expressed above, the UK government announced in their 2020 Spring Budget an increase in the tapered AA thresholds by £90,000 to restore the incentive for consultants to take on additional NHS commitments. We therefore focus on the period between April 2013 and April 2018 where the lower thresholds were in place.

4 Data

We construct a panel of NHS consultants with associated NHS inpatient activity, annual income and consultant characteristics spanning the financial years[33] 2013/14 to 2018/19. This database is compiled from three routinely collected data sources: patient-level hospital discharge data for all publicly funded hospital care in England from the Hospital Episode Statistics (HES) maintained by NHS Digital; consultant-level demographic and medical education data from the List of Registered Medical Practitioners maintained by the General Medical Council[34] (GMC); and lists of successful candidates to the national Clinical Excellence Awards published by the Advisory Committee on Clinical Excellence Awards (ACCEA). We link the three data sets at the consultant level using consultants' GMC registration numbers, which uniquely identify them across all data sources.

We apply four selection rules to our primary sample of consultants ($J = 47,563$). First, we exclude consultants for whom we observe clinical activity in a single financial year only (Number of consultants left in the sample: $J = 41,126$). Second, we restrict our sample to consultants for whom most of their clinical activity falls within one of 18 medical and surgical specialties ($J = 26,705$) (see Table 3.1 for a list of these). The choice of specialties was made to cover those in which most of the clinical work is captured by inpatient activity recorded in HES data, so that our measures of consultant activity are sensitive to changes in consultants' working patterns. Therefore, we exclude consultants working, for example,

[33] Financial years run between 1 April and 31 March.

[34] The GMC is the national body that determines doctors' qualification to practice.

in anaesthesia, pathology and radiology because their practice seldom involves taking lead responsibility for admitted patients. Third, we exclude consultants who win (or loose) CEA because such increases (or decreases) in pay may affect their activity rates (Chapter 2 shows that the awards may have a negative effect on the activity rates of consultants) ($J = 26,072$). Fourth, we define as inclusion rule that consultants must have continuously observed activity from the beginning of the sample period ($J = 18,156$). Consultants may take temporary leave of absence (e.g. due to sickness, maternity or study leave) that generates gaps in our panel and potentially biases our findings. For consultants who are continuously observed but drop out of the sample before the end of the period in analysis, we impute zero activity for the missing quarters of the year. By imputing a volume of zero for those consultants, we account for the effect that the policy may have had in pushing consultants to early retirement or to leaving the NHS altogether. We explore the effect of this modelling decision as part of sensitivity analyses.

4.1 Measure of Consultant Activity

Our main outcome variable of interest is the number of care episodes provided by consultant $j = 1, \dots, J$ in calendar quarter $t = 1, \dots, T$, which serves as a measure of consultants' clinical activity. We use administrative, pseudonymised records from HES for all NHS-funded inpatient care provided in England. HES contains detailed information on patient demographics (e.g. age, sex, and place of residence), clinical data (e.g. diagnoses codes (ICD-10) and procedure codes (OPCS4)), information about the admission pathway including dates and mode of admission and discharge, as well as consultant and hospital identifiers. Each observation in HES reflects a finish consultant episode (FCE), which represents the time spent under the care of a single consultant.

Similar to previous studies (Bloor et al., 2004, 2012; Lafond et al., 2017), our primary measure of consultant activity is the aggregated counts of FCEs per consultant and quarter of the year. To compute this, we extract patient episodes from HES covering the period from April 2013 to March 2019. We include all FCEs in public hospitals and all NHS-financed episodes delivered in private providers, including both elective and emergency care.

In secondary analysis, we refine our measure of clinical activity using the weighing procedure developed in the previous chapter (see Subsection 3.2 in Chapter 2 for a detailed discussion on the motivation for this approach), which accounts for differences in case complexity across consultants and specialties. Briefly, we weight consultant activity by the average length of stay (LOS) for the Healthcare Resource Group (HRG)[35] to which the care episode is assigned, rather than by the national average reference cost for the HRG, commonly used in previous studies (Bloor et al., 2004, 2012; Lafond et al., 2017). The advantage of this approach over a cost-based weighting procedure is that it reflects time budgets rather than cost budgets. Because consultants work under salaried contracts, their labour input is likely independent from the cost of the different activities they perform, and thus a cost-based weighting may bias the results towards clinical care that requires substantial non-labour inputs. Given this, LOS-weighted activity is likely a more accurate measure of individual workload.

To recall the procedure, let l_{ijht} be the length of stay for episode of care i undertaken by consultant j and belonging to HRG group h and quarter of the year t. We compute the national average length of stay for the HRG group h over the entire sample period as $\bar{l}_h = \frac{\sum_{i \in P_h} l_{ijht}}{n_h}$, where P_h and n_h describe the set and the number of episodes of care belonging to HRG h, respectively. We then compute LOS-weighted activity rates for consultant j at time t:

$$A_{jt} = \sum_h x_{jht} \bar{l}_h \tag{3.1}$$

where x_{jht} is the volume of output from consultant j in HRG group h in time t.

4.2 Measure of Consultant Income and Exposure to the Pension Reform

Total consultant annual income is not directly observable.[36] Instead, we approximate it using the two largest elements of consultant pay - their basic pay, and additional payments they

[35] HRG is a patient classification system which groups conditions and procedures that use similar levels of resources.

[36] Consultants may have other sources of taxable income, e.g. from property or other investments, and from private clinical work outside the NHS.

receive from any national Clinical Excellence Awards (CEAs) they may hold.

Consultants' basic pay is the wage they receive under their contract with their employing hospital. A full-time position entails 10 programmed activities per week, each having a timetabled value of four hours. Basic pay is determined by a pay scale that is dependent on years completed as a consultant (that is, years from anniversary of appointment), rather than performance. Consultants progress through a 19-year scale that comprises seven pay thresholds: the first four pay thresholds are awarded at one-year intervals and the next three thresholds are awarded at five-year intervals. In 2015, the pay scale spanned from £75,249 for first-appointed consultants to £101,451 for the most senior doctors (NHS Employers, 2015). Since doctors working in the NHS must apply for specialist registration if they want to practice as consultants, the year of entry in the GMC's List of Registered Medical Practitioners provides a near accurate date for the year a doctor became a consultant. We use this date from the GMC database and pay thresholds for 2015 (the year before the reform was introduced) to construct consultant basic pay.

CEAs are a financial incentive scheme rewarding consultants for their commitment to the NHS. There are four different levels of awards at the national level - bronze (£35,484, in 2015), silver (£46,644), gold (£58,305), and platinum (£75,796). CEAs are lifetime bonuses reviewed every five years that, in relative terms, can almost double a consultant's salary at the highest level of the awards. We derive information on the level and year of the award from the lists of successful candidates to the national CEAs that are published by the ACCEA on an annual basis. We only include consultants who retained the same level of award through the sample period (including those that never held an award) to avoid other income shocks that could potentially drive our findings.

We use the combined income from basic pay and CEAs to define a binary measure of exposure to the pension reform (Appendix Table C.1 shows values of the consultant pay scale and the clinical excellence awards in 2015/16). Consultants earning in excess of £110,000 in 2015/16 are in the treated group, and those earning less are in the control group. We recognise that our measure of consultant pay is likely to be a lower bound estimate of true earnings because we do not observe other elements of consultant income. To accommodate the measurement error due to unobserved components of pay, we restrict our controls to consultants earning less than £95,000 in an attempt to avoid contaminating our comparison group with consultants that might have been affect by the reform. We vary this pay threshold as part of sensitivity analyses.

4.3 Control Variables

We select a set of control variables that potentially affect consultant earnings and activity. The choice of variables is driven by data availability, previous studies, and our economic reasoning. In our regression models we include years since medical qualification (coarsened into 5 age groups, with thresholds set at 20, 25, 30 and 35) derived from the GMC database, which is a proxy for age. We also obtain information on consultants' sex, country of medical qualification, main clinical specialty, and the Strategical Health Authority (SHA) where consultants practice. The first two variables are taken from the GMC database, whereas the latter two are derived from HES data. We use SHA areas to capture factors which may affect geographic differences in consultant activity rates and earnings. SHAs were formally abolished in 2013 but the geographic coding continues to be recorded in HES data. This choice of geography captures not only place-specific variations in consultant activity but also differences in the allocation of national CEAs: regional committees (defined geographically using the divisions of SHA) assess and short-list candidates that are then considered for submission to the national committee. Hence, place of work may reasonably affect annual income via CEAs. The consultants in the estimation sample are located in 17 SHA.

5 Empirical Framework

At the core, our empirical analysis compares the activity rates of consultants with annual income in excess of £110,000 p.a., who were affected by the new UK pension legislation, to those earning less than £95,000 p.a., who were unlikely to be affected. Using this binary split in income as a measure of exposure, we exploit the timing of the reform to implement a difference-in-differences (DID) approach combined with pre-processing of data using Entropy Balancing (EB) (Hainmueller, 2012; Hainmueller and Xu, 2013). Our empirical approach thus follows past work in the economics literature that uses matching estimators to create balanced samples in observed characteristics for subsequent estimation of treatment effects with binary treatment variables in DID analysis (Heckman et al., 1997, 1998; Blundell et al., 2004; Abadie, 2005; Todd, 2007; Marcus, 2013; Everding and Marcus, 2020; Freier et al., 2015).

The estimator is described using potential outcomes in Equation (3.2), as previously defined by Todd (2007). Let N_1 and n_1 be the set and number of treated consultants, respectively, and N_0 and n_0 the equivalents for controls. The states associated with receiving treatment and not receiving treatment are denoted "1" and "0", respectively. Thus, Y_{0i}^{before} and Y_{1i}^{after} represent the outcomes of interest measured before and after treatment for treated consultant $i \in N_1$, and in the same fashion Y_{0j}^{before} and Y_{0j}^{after} the outcomes for control $j \in N_0$. Finally, $\omega(i,j)$ are the weights from the EB algorithm (Hainmueller, 2012).

$$\hat{\alpha}_{DID} = \frac{1}{n_1} \sum_{i \in N_1} \left[(Y_{1i}^{after} - Y_{0i}^{before}) - \sum_{j \in N_0} \omega(i,j)(Y_{0j}^{after} - Y_{0j}^{before}) \right] \tag{3.2}$$

The key identifying assumption in this research design is that in the absence of the new UK pension legislation consultants with income in excess of £110,000 would have experienced changes in activity rate (e.g. due to technological advancements that affect the healthcare production function) similar to a re-weighted group of consultants earning less than £95,000. This is equivalent to the common trend assumption in the traditional DID approach, but after using EB to reweigh the control group to equalise the moments of the treatment group, as formally presented in Equation (3.3).

$$\mathbf{E}\left[Y_0^{after} - Y_0^{before} | \omega(i,j), D = 1\right] = \mathbf{E}\left[Y_0^{after} - Y_0^{before} | \omega(i,j), D = 0\right] \tag{3.3}$$

The estimator in Equation (3.2) is implemented in two steps. In the first step, we use EB to pre-process the data and balance consultant characteristics and pre-policy outcome trends between treated and control consultants. Unlike other matching algorithms such as propensity score methods, EB is more effective in reducing covariate imbalance as it searches weights that achieve exact balancing on prespecified moments of the distribution by imposing a series of balancing constraints[37]. This produces a sample in which the moments specified in the procedure are the same in treatment and control groups. In turn, EB circumvents several iterations of model estimation, matching, and "manual" balance checking to achieve satisfactory balancing solutions. We follow Cefalu et al. (2020) and match treatment and control groups on individual consultant characteristics (i.e. all time-invariant characteristics shown in Table 3.1) and pre-policy outcome trends. We use linear regression to parametrically model pre-policy trends for each consultant in our sample as shown below:

$$Y_{jt} = \beta_{0j} + \beta_{1j} s + \epsilon_{jt} \tag{3.4}$$

where Y_{jt} is our measure of consultant activity for consultant j in quarter t; s represents time in the pre-policy period measured in quarters of the year, and ϵ_{jt} is an error term. The coefficient β_{1j} captures the consultant j's linear trend in pre-policy outcomes, which we include in the balancing constraints.

In a second step, we estimate DID regression models, with each observation being weighted using EB weights $\omega(i, j)$, as follows:

$$Y_{jt} = \alpha Post_t + \theta \left[Post_t \times \mathbf{1}[> \pounds 110,000_j]\right] + \beta X_{jt} + \phi_j + \gamma_t + \epsilon_{jt} \tag{3.5}$$

The term $Post_t$ is an indicator variable equal to one in post-policy years 2016-2018, whilst $\mathbf{1}[> \pounds 110,000_j]$ is equal to one for consultants earning in excess of £110,000, and zero otherwise; X_{jt} is the seniority variable measured in years since medical qualification in 5-year bins; ϕ_j captures unobserved time-invariant consultant effects such as taste for work and practice style; γ_t are time (in calendar quarters) fixed effects that capture changes in workload and clinical staffing common to all consultants in our sample; and ϵ_{jt} is an error term. Finally, θ is the parameter of interest and describes the differential effect of the UK pension reform for

[37] See Hainmueller (2012) for a discussion on the numerical implementation and theoretical properties of EB, and Hainmueller and Xu (2013) for details on implementation in Stata 16 using the programme "ebalance"

consultants who are likely to be affected by it compared to those assumed to be unaffected. Standard errors are robust to arbitrary serial correlation and heteroskedasticity and clustered at the consultant level.

We also examine whether the reform has immediate or lagged effects on consultant activity in an event study analysis. This is implemented by augmenting equation (3.5) with a set of quarter indicators relative to the reform interacted with the policy variable of interest:

$$Y_{jt} = \sum_{k=-T'}^{T'} 1[k = t'](\alpha_k + \theta_k 1[> \pounds 110,000]_j) + \beta X_{jt} + \phi_j + \gamma_t + \epsilon_{jt} \tag{3.6}$$

where $t' = t - t^0$ with $t' \in [-T', T']$, is a period-specific index in quarters relative to the time of the reform, t^0.

6 Results

Our panel consists of 11,957 consultants, which we split into two groups based on their 2015/16 earnings (Figure C.1 in the Appendix displays the distribution of income for all consultants before construction of our measure of policy exposure ($J = 18,156$)): (1) treated consultants with earnings above £110,000 p.a. ($J = 605$) and (2) controls earning less than £95,000 p.a. ($J = 11,352$). Table 3.1 presents summary statistics for both groups in column (1) and (2), and two-tailed p-values for a t-test of equality of means in column (3). There are clear differences in pre-policy trends, as well as sex, country of medical qualification, seniority and specialty membership. High earners are more likely male, older, and in a downward trend of clinical activity. A larger share of consultants in the group of high earners drops out of the sample over the study period relative to the group of low earners. Average seniority at drop out is almost 14 years higher in the group of high-earners. There are a variety of reasons for which consultants exit the sample, including death, retirement, or leaving the NHS altogether. Average clinical activity and average LOS-adjusted activity are higher in consultants earning less than £95.000 p.a., and all measures are right-skewed: averages are significantly larger than median values.

Table C.2 in the Appendix reports standardised percent differences in means between treated (£110,000 p.a.) and control (£95,000 p.a.) groups, both unbalanced and after applying preprocessing based on EB. As expected, EB exactly adjusts imbalance with respect to the first moment of the covariate distributions.

Panel A and B in Figure 3.1 show time-series of average quarterly clinical activity for the treatment and control groups (unweighted and EB weighted). The treated and unweighted control group seem to follow different trends in activity rates in pre-policy years, with the treatment group displaying a faster decline in average activity. This suggests that the common trend assumption underpinning DID analysis (Angrist and Pischke, 2008) is unlikely to be satisfied, which may result in biased estimates of the policy effect. The EB-weighted control group offers a better approximation of the observed behaviour of the treatment group.

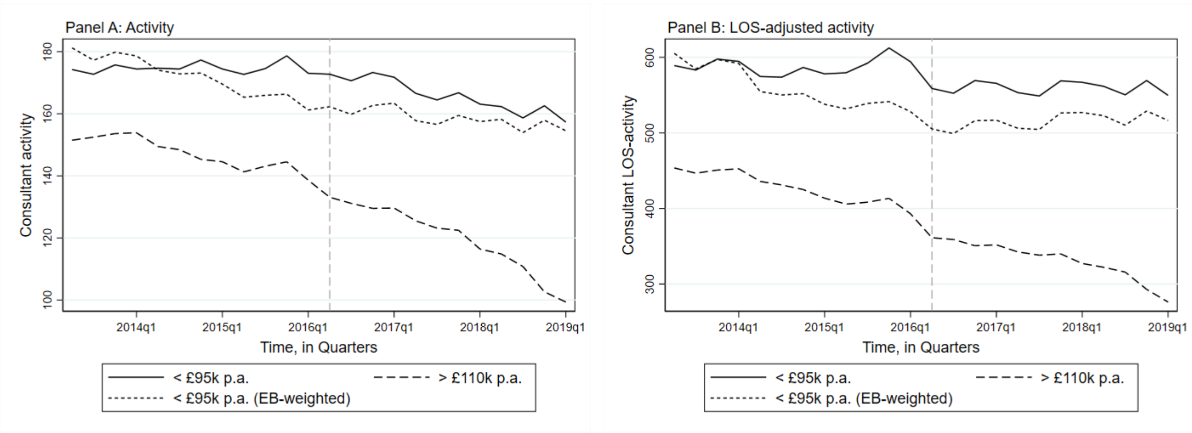

Figure 3.1: Trends in Clinical Activity

Notes: Figure shows time series over the sample period for activity in panel A, and LOS-adjusted activity in panel B, by treated consultants (> £110k p.a.), and unweighted and EB weighted controls (< £95k p.a.).The vertical dashed grey line represents the first post-policy quarter (Q2/2016).

6.1 Policy Estimates

Table 3.2 provides the difference-in-differences estimates for consultant activity in panel A, and for LOS-adjusted activity in Panel B. As a reference, we first report difference-in-differences OLS estimates in column 1 without controls for seniority, and in column 2 controlling for

Table 3.1: Summary Statistics

	< £95k p.a.	> £110k p.a.	P-value
	(1)	(2)	(3)
Male	0.73	0.88	<0.001
UK medical qualification	0.63	0.86	<0.001
Pre-policy trend in activity (Q2/2013 to Q1/2016)	0.60	-1.24	<0.001
Seniority, first observed			
$\leq$ 20 years	0.64	0.01	<0.001
21-25 years	0.26	0.13	<0.001
26-30 years	0.07	0.32	<0.001
31-35 years	0.02	0.36	<0.001
$\geq$ 36 years	0.01	0.18	<0.001
Average Seniority (in years)	19.20	31.02	<0.001
Specialty			
General Surgery	0.12	0.12	0.132
Urology	0.04	0.04	0.065
Trauma & Orthopaedics	0.13	0.05	<0.001
Otorhinolaryngology	0.03	0.04	<0.001
Ophthalmology	0.05	0.06	<0.001
Neurosurgery	0.02	0.02	<0.001
Plastic Surgery	0.03	0.01	<0.001
Cardiothoracic Surgery	0.01	0.04	<0.001
General Medicine	0.12	0.10	<0.001
Gastroenterology	0.05	0.05	0.146
Haematology	0.03	0.06	<0.001
Cardiology	0.06	0.08	<0.001
Dermatology	0.02	0.02	0.003
Respiratory Medicine	0.04	0.06	<0.001
Neurology	0.03	0.05	<0.001
Rheumatology	0.02	0.05	<0.001
Paediatrics	0.13	0.09	<0.001
Geriatric Medicine	0.05	0.02	<0.001
Share consultants dropped out of sample	0.13	0.23	<0.001
Average seniority at drop out (in years)	23.92	37.80	<0.001
Activity in Q1/2016			
Mean	173.06	138.60	
Median	128.00	76.00	
Std. dev.	234.89	230.14	
LOS-adjusted activity in Q1/2016			
Mean	594.21	393.02	
Median	407.38	258.01	
Std. dev.	875.35	472.37	
Consultants	11,352	605	
Consultant-years	266,969	14,250	

Notes: In column (1) and (2) all rows, except for "Average seniority", "Average seniority at drop out", "Activity in Q1/2016", and "LOS-adjusted activity in Q1/2016", report the share of consultants within the given population with the indicated characteristic. "Average seniority" and "Average seniority at drop out" are the averages for each group. "Seniority, first observed" is defined as years since medical qualification in the first year a consultant is observed in the sample. Column (3) shows the two-tailed p-value for a t-test of equality of means between both groups. The pre-policy trend is the coefficient β_{1j} estimated from Equation (3.4). ($J = 11,957$).

seniority in 5 age groups. In column 1 panel A, we estimate that consultants affected by the reform experienced a relative average decrease in activity rates of almost 20 episodes of care per quarter, or an 11.8% reduction relative to the mean, compared to the unweighted group of consultants with earnings below £95.000. In panel B, we estimate a reduction of work effort equivalent to 70 days of patient care (-12.6%). Both coefficient estimates are statistically significant at the 1 percent level. When we add consultant seniority in column 4, our estimates decrease in magnitude to approximately 14 episodes of care (Panel A) and 49 days of patient care (Panel B), but remain statistically significant at the 5 and 1 percent level, respectively.

Table 3.2: Policy Estimates

	OLS		OLS-EB	
Outcome	(1)	(2)	(3)	(4)
Panel A. Activity (mean= 168.43)				
Post-Policy x (> £110k p.a.)	-19.868^{***}	-13.703^{**}	-15.426^{**}	-14.259^{**}
	(5.280)	(5.589)	(5.509)	(6.077)
Panel B. LOS-adjusted activity (mean =563.92)				
Post-Policy x (> £110k p.a.)	-70.855^{***}	-48.686^{***}	-55.725^{***}	-49.167^{***}
	(11.971)	(12.648)	(12.470)	(12.609)
Consultant seniority (5-year bins)	NO	YES	NO	YES
Time FE	YES	YES	YES	YES
Consultant FE	YES	YES	YES	YES
Consultant-quarters	281,219	281,219	281,219	281,219

Notes: Table reports estimates of */**/*** indicate statistical significance at the 10%/5%/1% level. Standard errors (in parentheses) are clustered at the consultant level to allow for arbitrary serial correlation and heteroskedascity. The sample is consultants with earnings above > £110k p.a. and below < £95k p.a ($J = 11,957$)

The last two columns report estimates for the difference-in-differences weighting the controls by their weights produced by EB, which we consider to be our most robust and reliable estimates. In column 3, we estimate a 15.4 reduction in post-reform activity in panel A (9.2 percent relative to the mean) and 55.7 days of patient care in panel B (9.8 percent) relative to the weighted comparison group. Once again, adding consultant seniority (column 4) leads to a reduction in the estimated effect and we now find that consultants with earnings in excess of £110.000 experienced a decrease of 14.3 in activity rates (8.4 percent) in panel A, and 49.1 days of patient care in panel B (8.7 percent). Both estimates are statistically significant at the 5 and 1 percent level, respectively.

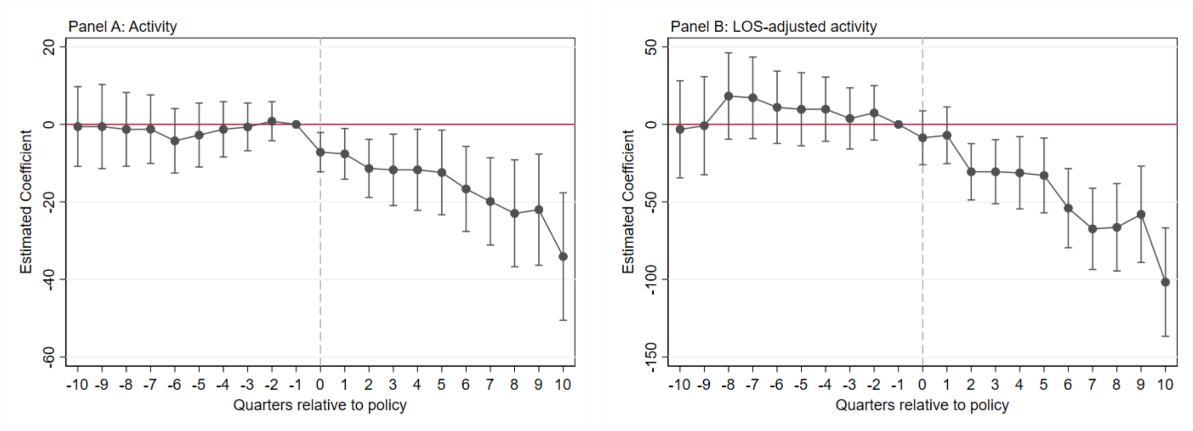

Figure 3.2: Event Study Estimates

Notes: Each panel displays the estimated coefficients θ_k from the event study specification in equation 3.6, using weights for the control group generated by the EB algorithm. The dependent variable Y_{jt} is consultant activity in panel A and LOS-adjusted activity in panel B. Both specifications include consultant fixed effects and seniority measured categorically in 5 age groups. The vertical dashed grey line at quarter zero marks the first post-policy quarter. The coefficient for relative quarter -1 is normalized to zero. Each plotted estimate includes 95% confidence intervals adjusted for clustering at the consultant level. The sample is all consultant-quarters ten quarters before and after the reform (N=247,107).

We present an event study for consultant activity in panel A and for LOS-adjusted activity in panel B of Figure 3.2, using ten quarters before and after the policy was introduced. The plotted estimates are the event study coefficients θ_k from estimating equation 3.6, which includes weights for the control group generated by the EB algorithm, and conditions on consultant seniority and consultant fixed effects. Since the event study coefficients are only identified up to a constant, we normalize the relationship between our measure of exposure to the policy and activity rates to zero in quarter -1 relative to the reform (i.e. Q1/2016). Upper and lower bounds of the 95% confidence interval are reported for each estimate, adjusted for clustering at the consultant level. We observe statistically significant reductions in consultant activity and LOS-adjusted activity rates beginning as the reform take effect and continuing gradually in a downward trend in the first ten post-policy quarters.

6.2 Robustness Checks

In this section, we explore several further remaining threats to our identification strategy. Table 3.3 shows the results of a range of robustness checks to explore the sensitivity of our findings to alternative model specifications and analysis samples. We discuss each of these in turn.

Table 3.3: Robustness Checks

	Baseline	Placebo Test	< £90k p.a.	Private Practice	Dropouts	Poisson FE
Outcome	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Activity						
Post-Policy x (> £110k p.a.)	−14.259**	-1.439	−21.365***	−15.284**	-12.122	−0.125***
	(6.077)	(3.073)	(5.694)	(6.576)	(7.383)	(0.047)
Panel B. LOS-adjusted activity						
Post-Policy x (> £110k p.a.)	−49.167***	4.290	−64.763***	−49.326***	−43.845***	−0.170***
	(12.609)	(12.191)	(13.607)	(13.574)	(14.137)	(0.033)
Consultant seniority	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Consultant FE	YES	YES	YES	YES	YES	YES
Consultant-quarters	281,219	257,405	186,561	233,396	244,259	281,219

Notes: */**/*** indicate statistical significance at the 10%/5%/1% level. Standard errors (in parentheses) are clustered at the consultant level to allow for arbitrary serial correlation and heteroskedascity.

Placebo Test

Our empirical strategy is to select a group of consultants who were clinically active between April 2013 and March 2015 and follow them until March 2019. Over time, this cohort ages and some consultants may decide to retire or reduce their clinical activity independent of the tax rules. Because older consultants earn more — NHS pay is related to tenure — they are also more likely to be affected by the policy change. Our analysis may therefore be confounded by unrelated retirement decisions.[38]

As a placebo test, we estimate the same specification in equation 3.5, and apply the same sample selection rules defined in Section 4 but use a cohort of consultants clinically active in years 2009-2014, i.e. prior to the pension reform. We define the second quarter of 2011 to be the “policy-on” quarter. If we observe similar reduction in clinical activity following this fictitious pension reform, this would indicate that the estimated effect in our main analysis reflects the ‘natural’ decline in clinical activity as consultants age. However, we only estimate small and not statistically significant effect of our fictitious pension reform (Column 2 in

[38] Although note that in the baseline models we find that adding seniority to the analysis has no substantial effect on the findings.

Table 3.3). This finding provides reassurance that the observed reduction in clinical activity is plausibly driven by the 2016 pension reform; not a general cohort effect.

Varying Income Threshold for Controls

As noted in Subsection 4.2, our measure of income is prone to measurement error and is likely to be an underestimation of actual consultant income. For example, we do not observe other elements of income, such as those from private clinical practice or returns from private investments. In our main analysis, we defined a headroom of £15,000 in potential income from other sources when selecting controls. In doing so, we aim to exclude consultants with known earnings below the AA limit but who may be over when such other elements of their income are taken into account. In column 3 of Table 3.3, we increase the headroom to £20,000 and define a limit of £90,000 in our measure of income to select a control group which is less likely to experience annual allowance charges. As expected, the estimates are larger in magnitude than the main estimates of the paper, suggesting that when a stricter comparison group is chosen the policy effects become more evident.

Private Practice

Income derived from private practice is possibly the largest source of consultant income outside of the NHS for most consultants. In column 4 of Table 3.3, we re-estimate our main specification excluding consultants working in three specialties that are known to often derive sizeable earnings from private practice (Morris et al., 2008). These are plastic surgery, trauma and orthopaedics, and neurosurgery. The estimated policy effects on clinical activity are similar in magnitude to the baseline estimates.

Excluding Dropouts

Our main result suggests that the new pension tax reduced consultant activity, but it is unclear whether this reduction occurs due to consultants being pushed into early retirement (i.e. leaving the NHS altogether)(extensive margin) or due to consultants reducing their additional clinical sessions (intensive margin). To more directly examine the impact of the reform at the intensive margin, we re-estimate our model excluding consultants who drop out of the sample at any point during the post-policy period. The results are given in column 5 of Table 3.3. The estimated coefficients are slightly smaller in magnitude than in the baseline model; although estimate in panel B remains statistically significant at the 1% level, the estimate in panel A becomes statistically indistinguishable from zero.

Choice of Estimator

As a final robustness check, we estimate Equation (3.5) using a Poisson fixed effects model. We show in Table 3.1 that our measures of consultant activity are non-negative and right-skewed due to a small number of consultants with very high activity rates. It has been shown that the Poisson estimator is better suited under such distributional qualities of the dependent variable compared to its log-linear counterpart and the negative binomial model (Buntin and Zaslavsky, 2004; Santos Silva and Tenreyro, 2006) as long as the conditional mean is correctly specified (Manning and Mullahy, 2001; Wooldridge, 2010). Because of limitations of the statistical software we are unable to estimate Equation (3.5) using the EB weights, and therefore present results for the unweighted model in column 6 of Table 3.3. The Poisson estimates are larger than the OLS estimates - a 12.5% post-policy decrease in activity rates (vs. -11.8% in the OLS model) and a 17.0% drop in days of patient care (vs. -12.6%). This suggests that our main findings may be at best an underestimation of the effect of the reform.

7 Conclusions

This chapter set out to identify the short-term effect of decreased take-home pay on the clinical activity of senior hospital doctors working in the NHS, using an exogenous shock in income generated through the 2016 UK pension reform. We compare two measures of clinical activity across consultants earning in excess of £110,000 p.a. relative to those earning less than £95,000 p.a., where the latter group is unlikely to be affected by the pension reform. We combine a difference-in-differences estimation strategy with pre-processing based on entropy balancing. This empirical strategy, we argue, makes it less likely that our findings are neither driven by unobserved consultant fixed effects such as ability and motivation, nor by differential pre-policy trends between treated and control groups.

Our estimates suggest that the policy had significant effects on the activity rates of consultants directly affected. We find that consultants who were subject to increased taxation on their income reduced their clinical activity rates by approximately 8.4 percent compared to their colleagues that were unaffected by the pension reform. Similar results are obtained when we use a bed-day based measure of activity (8.7 percent). Our results are not directly comparable with previous studies of the elasticity of labour supply in self-employed and salaried doctors but the estimates appear relatively large. One possible reason for this may be that the pension reform was highly punitive and produced substantial reductions in consultants take-home pay, possibly higher than the changes in pay examined in other studies.

We find that the estimated effect of the pension reform appear to increase over time. The UK pension tax regime allows individuals to "carry forward" unused annual allowances from up to three previous years to absorb or reduce any annual allowance excess in that tax year. As consultants use up this facility over time, we expect them to become progressively more exposed to the new pension arrangements. Furthermore, because of the complexity of the tax system, many doctors may have become aware of the implications of the pension reform only when they began to receive higher than expected tax bills, i.e. after the end of the first post-policy year.

Finally, we show that our findings are robust to alternative specifications and samples. Most importantly, we validate our empirical strategy by examining whether similar activity decreases occurred in a previous year when the policy was not in place, i.e. a placebo test. We do not find evidence of any significant reductions in clinical activity between high and low earning consultants over that time period, suggesting that the findings of the main study are plausibly causal estimates of the policy effect.

This study has two main limitations. First, our derived measured of consultant pay is likely to be underestimated because some components of consultant pay are not observed by us. As a result, some of the consultants included in the control group may have been exposed to the policy, which would contaminate the control group design and lead to downward bias in our policy estimates. Second, the results may suffer from confounding effects related to time-varying unobserved consultant characteristics such as their health status and life arrangements (e.g., family background).

In conclusion, our findings suggest that in the short run NHS doctors respond to changes in take-home pay.

C Appendix

Table C.1: Consultant Pay Scale and Financial Value of the Awards, 1st April 2015

Years as consultant	£p.a.	Level	£p.a.
0	£75,249	Bronze	£35,484
1	£77,605	Silver	£46,644
2	£79,961	Gold	£58,305
3	£82,318	Platinum	£75,796
4-8	£84,667		
9-13	£90,263		
14-18	£95,860		
≥19	£101,451		

Notes: The pay scale presented above is for consultants appointed on or after 31st October 2003. Source: NHS Employers (2015)

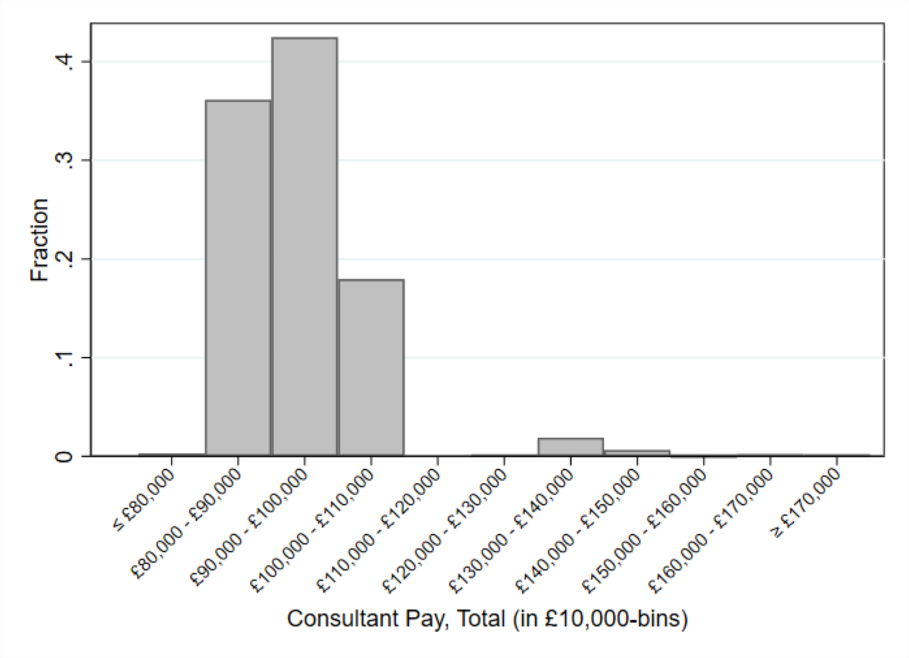

Figure C.1: Distribution of Consultant Pay

Notes: Figure shows the distribution of consultant income for all consultants in our sample before construction of measure of exposure to the reform $(J = 18,156)$.

Table C.2: Balancing of Means

	Bias (Std % diff. in means)	
	Unbalanced	EB
	(1)	(2)
Male	47.7	0.0
UK medical qualification	66.1	0.0
Pre-policy trend in activity (Q2/2013 to Q1/2016)	-15.3	0.0
Specialty		
General Surgery	-1.5	0.0
Urology	2.0	0.0
Trauma & Orthopaedics	-36.6	0.0
Otorhinolaryngology	3.8	0.0
Ophthalmology	4.8	0.0
Neurosurgery	5.0	0.0
Plastic Surgery	-12.9	0.0
Cardiothoracic Surgery	14.4	0.0
General Medicine	-6.6	0.0
Gastroenterology	1.3	0.0
Haematology	9.4	0.0
Cardiology	8.0	0.0
Dermatology	2.0	0.0
Respiratory Medicine	9.0	0.0
Neurology	6.4	0.0
Rheumatology	12.9	0.0
Paediatrics	-14.0	0.0
Geriatric Medicine	-16.8	0.0

Notes: EB, Entropy balancing.

Conclusions

The common theme of the chapters in this thesis is the role of both financial and non-financial incentives on NHS hospital doctors' clinical behaviour and labour supply.

In Chapter 1, we examine whether hospital effects drive NHS surgeons' choice between cemented and cementless methods of fixation in hip replacement surgery to shed new light on drivers of supply-side variation in medical care. We separate the role of physician-specific factors and clinical environment-specific factors by investigating the behaviour of doctors who move their practice between hospitals in the English NHS. There are two main findings in this chapter. First, we find robust evidence suggesting that the hospital environment in which surgeons practice determines to a large extent their treatment decisions, as previously found by Molitor (2018) and Avdic et al. (2021) in different settings. This result highlights the importance of non-financial incentives in determining the treatment decisions of consultants, or more broadly, their practice style.

A second important result of this chapter is that movers adapt immediately rather than gradually to the new hospital, with very limited further adjustments over time. This suggests that policies directed towards changing the organizations in which doctors work are likely to have significant short-run effects. Concurrently, the limited adjustment of consultants' practice style after the move indicates that policies attempting to change physician-specific factors, such as the beliefs and preferences of consultants, may only have an effect in the long run. Lastly, our analysis suggests that the sudden change in practice behaviour had no detrimental effects for patient care, potentially due to the lack of superiority in the effectiveness of one method of fixation over the other found in the medical literature (Abdulkarim et al., 2013).

Chapter 1 contributes to a scant literature attempting to open the "black box" of supply-side drivers of variation in medical care (Chandra et al., 2011). Our finding that the clinical environment significantly impacts surgeons' practice style suggests that future research attempting to separate specific factors of the clinical-environment component should be encouraged. One potential fruitful area for further research that has not been explored in great detail yet is the relative importance of the social vs. physical context, in particular the role of peers in shaping doctors' behaviours (Huesch, 2011; Epstein and Nicholson, 2009; Barrenho et al., 2021). Avdic et al. (2021) provides a first attempt at separating hospital-specific and a peer group-specific factor by examining cardiologists working in the same hospital on a given day. The context of our analysis (hip replacement surgery) is also likely to provide the opportunity to better understand the role of peers. Because hip replacement surgery is typically performed by a team of medical professionals, information on how movers are matched to teams in the new hospital would allow for the identification of peer effects. However, this information is not available, as English hospital data only record the consultant that is responsible for the episode of care.

Chapter 2 examines the effect of the CEAs on the post-award clinical activity of NHS consultants, which is not directly incentivised by the scheme. The key findings are as follows. First, I find weak evidence of a negative effect for the two lowest award categories (Bronze and Silver awards), but these findings are sensitive to the econometric modelling approach. Second, the two highest award categories, Platinum and Gold awards, have no effect on post-award clinical activity. However, this null effect must be interpreted with caution as it could be due to the limited power of my analysis. Third, loosing an award does not impact the activity rates of consultants. Fourth, I do not find clear evidence of anticipation effects for any of the award levels, which could bias the results.

Future research should seek to address the limitations of this chapter by using longer and larger panels with information on all candidates to the awards and time-varying factors that are correlated with both the awards and clinical activity.

Despite these methodological challenges, I argue that assessing the impact of the awards on both incentivised and non-incentivised dimensions of consultant work is an endeavour worth pursuing. First, in a healthcare system that is concerned with efficiency gains and increasing overall productivity (Lafond et al., 2017; NHS, 2019), assessing the added value of a costly and contentious incentive scheme, including its consequences for patient care, is paramount.

Second, although some critics advocate that the scheme should be abolished (Bloor et al., 2012), the awards may contribute to hard-to-measure activities (e.g. innovation) and the retention of high-skilled doctors.

In Chapter 3, we exploit the 2016 UK pension reform, which disproportionately reduced take-home pay for high-earners in the UK, to examine the short-run responsiveness of NHS consultants' doctor supply to changes in remuneration. We find evidence that consultants who were subject to increased taxation on their income reduced their clinical activity. This result is robust to different specifications and sample definitions, including a placebo test. Our estimates appear large compared to short-run wage elasticities found in previous studies of self-employed and salaried doctors. One possible explanation may be that the reductions in consultants' take home-pay were larger than the changes in pay of previous work, thus eliciting a stronger response.

A second result is that the estimated effect of the pension reform appears to increase over time. This is likely to be explained by one of the features of the UK pension arrangements. The pension tax regime allows workers to "carry forward" unused annual allowances from up to three previous years to absorb or reduce annual allowance excess in that tax year. Thus, consultants become progressively more exposed to the new tax pension arrangements as this facility is used up.

The work in this chapter could be refined and extended in future research in two ways. First, our measure of consultant pay is likely to be underestimated as some elements of consultant pay are unobserved to us. One possible solution to address the measurement error introduced in our regression models by this variable would be to use accurate measures of consultant pay derived from tax returns held by Her Majesty's Revenue and Customs. However, access to tax returns information is highly restricted, and unlikely to be linked to HES data. Second, when new data becomes available, future research can evaluate whether the increase in the annual allowance threshold announced at the 2020 Spring Budget restored pre-2016 levels of clinical activity, or whether the 2016 Pension reform produced permanent changes in the labour supply of consultants.

Bibliography

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72(1):1–19.

Abdulkarim, A., Ellanti, P., Motterlini, N., Fahey, T., and O'Byrne, J. M. (2013). Cemented versus uncemented fixation in total hip replacement: a systematic review and meta-analysis of randomized controlled trials. *Orthopedic reviews*, 5(1):e8–e8.

Advisory Committee on Clinical Excellence Awards (2014). NHS Consultants' Clinical Excellence Awards Scheme. Guide for assessors 2014. Technical report.

Advisory Committee on Clinical Excellence Awards (2019). 2019 ACCEA Annual Report (covering the 2018 Awards Round). Technical report.

Andreassen, L., Di Tommaso, M. L., and Strøm, S. (2013). Do medical doctors respond to economic incentives? *Journal of Health Economics*, 32(2):392–409.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Anthony, D. L., Herndon, M. B., Gallagher, P. M., Barnato, A. E., Bynum, J. P. W., Gottlieb, D. J., Fisher, E. S., and Skinner, J. S. (2009). How much do patients' preferences contribute to resource use? *Health Affairs*, 28(3):864–873.

Appleby, J., Raleigh, V., Frosini, F., Bevan, G., Gao, H., and Lyscom, T. (2011). Variations in health care: the good, the bad, and the inexplicable. Technical report, The King's Fund, London.

Armour, B. S., Pitts, M. M., Maclean, R., Cangialose, C., Kishel, M., Imai, H., and Etchason, J. (2001). The effect of explicit financial incentives on physician behavior. *Archives of Internal Medicine*, 161(10):1261–1266.

Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review*, 53(5):941–973.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics*, 60(1):47–57.

Avdic, D., Ivets, M., Lagerqvist, B., and Sriubaite, I. (2021). Providers, Peers and Patients: How do Physicians' Practice Environments Affect Patient Outcomes? *CINCH working paper series*, 2021(01).

Baker, G. P. (1992). Incentive Contracts and Performance Measurement. *Journal of Political Economy*, 100(3):598–614.

Baltagi, B. H., Bratberg, E., and Holmås, T. H. (2005). A panel data study of physicians' labor supply: the case of Norway. *Health Economics*, 14(10):1035–1045.

Barnato, A. E., Herndon, M. B., Anthony, D. L., Gallagher, P. M., Skinner, J. S., Bynum, J. P. W., and Fisher, E. S. (2007). Are Regional Variations in End-of-Life Care Intensity Explained by Patient Preferences? *Medical care*, 45(5):386–393.

Barrenho, E., Miraldo, M., Propper, C., and Walsh, B. (2021). The importance of surgeons and their peers in adoption and diffusion of innovation: An observational study of laparoscopic colectomy adoption and diffusion in England. *Social Science & Medicine (1982)*, 272:113715.

Bloor, K., Freemantle, N., and Maynard, A. (2008). Gender and variation in activity rates of hospital consultants. *Journal of the Royal Society of Medicine*, 101(1):27–33.

Bloor, K., Freemantle, N., and Maynard, A. (2012). Trends in consultant clinical activity and the effect of the 2003 contract change: retrospective analysis of secondary data. *Journal of the Royal Society of Medicine*, 105(11):472–479.

Bloor, K., Maynard, A., and Freemantle, N. (2004). Variation in activity rates of consultant surgeons and the influence of reward structures in the English NHS. *Journal of Health Services Research & Policy*, 9(2):76–84.

Blundell, R., Meghir, C., Dias, M. C., and Reenen, J. V. (2004). Evaluating the Employment Impact of a Mandatory Job Search Program. *Journal of the European Economic Association*, 2(4):569–606.

Blundell, R. and MaCurdy, T. (1999). Chapter 27 - Labor supply: A review of alternative approaches. In *Handbook of Labor Economics*, volume Volume 3 of Handbook of Labor Economics, pages 1559–1695.

Borjas, G. J. (2008). *Labor economics.* McGraw-Hill/Irwin, Boston.

Briggs, T. (2015). A national review of adult elective orthopaedic services in England - Getting It Right the First Time. Technical report, British Orthopaedic Association.

British Medical Association (2020). Consultant workforce shortages and solutions: Now and in the future. Technical report.

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., and Wiesen, D. (2017). The Effects of Introducing Mixed Payment Systems for Physicians: Experimental Evidence. *Health Economics*, 26(2):243–262.

Brown, D. M. and Lapan, H. E. (1979). The Supply of Physicians' Services. *Economic Inquiry*, 17(2):269–279.

Brown, M. C. (1989). Empirical determinants of physician incomes — Evidence from Canadian data. *Empirical Economics*, 14(4):273–289.

Buntin, M. B. and Zaslavsky, A. M. (2004). Too much ado about two-part models and transformation?: Comparing methods of modeling Medicare expenditures. *Journal of Health Economics*, 23(3):525–542.

Campbell, S. M., Reeves, D., Kontopantelis, E., Sibbald, B., and Roland, M. (2009). Effects of Pay for Performance on the Quality of Primary Care in England. *New England Journal of Medicine*, 361(4):368–378.

Cefalu, M., Vegetabile, B. G., Dworsky, M., Eibner, C., and Girosi, F. (2020). Reducing bias in difference-in-differences models using entropy balancing. *Working Paper.*

Chalkley, M. and Malcomson, J. M. (1998). Contracting for Health Services with Unmonitored Quality. *The Economic Journal*, 108(449):1093–1110.

Chandra, A., Cutler, D., and Song, Z. (2011). *Chapter Six - Who Ordered That? The Economics of Treatment Choices in Medical Care*, volume 2 of *Handbook of Health Economics*. Elsevier.

Christianson, J. B., Leatherman, S., and Sutherland, K. (2008). Lessons from evaluations of purchaser pay-for-performance programs: a review of the evidence. *Medical care research and review: MCRR*, 65(6 Suppl):5S–35S.

Clark, A. E., Diener, E., Georgellis, Y., and Lucas, R. E. (2008). Lags And Leads in Life Satisfaction: a Test of the Baseline Hypothesis. *The Economic Journal*, 118(529):F222–F243.

Corallo, A. N., Croxford, R., Goodman, D. C., Bryan, E. L., Srivastava, D., and Stukel, T. A. (2014). A systematic review of medical practice variation in OECD countries. *Health Policy*, 114(1):5–14.

Currie, J., MacLeod, W. B., and Van Parys, J. (2016). Provider practice style and patient health outcomes: The case of heart attacks. *Journal of Health Economics*, 47:64–80.

Danzer, A. M., Dolton, P., and Bondibene, C. R. (2016). Who wins? Evaluating the impact of UK public sector pension scheme reforms. *National Institute Economic Review*, 237:R38–R46.

Davies, C. and Lorgelly, P. (2013). Hospital Procurement with Concentrated Sellers: A Case Study of Hip Prostheses. Technical Report 2013-13, Centre for Competition Policy, University of East Anglia, Norwich, UK.

Dawson, J., Fitzpatrick, R., Carr, A., and Murray, D. (1996). Questionnaire on the perceptions of patients about total hip replacement. *The Journal of Bone and Joint Surgery. British Volume*, 78(2):185–190.

Department of Health (2008). Guidance on the routine collection of Patient Reported Outcome Measures (PROMs) For the NHS in England 2009/10. Technical report, Department of Health.

Diriwaechter, P. and Shvartsman, E. (2018). The anticipation and adaptation effects of intra- and interpersonal wage changes on job satisfaction. *Journal of Economic Behavior & Organization*, 146:116–140.

Doran, T., Fullwood, C., Kontopantelis, E., and Reeves, D. (2008). Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. 372(9640):728–736.

Doran, T., Maurer, K. A., and Ryan, A. M. (2017). Impact of Provider Incentives on Quality and Value of Health Care. *Annual Review of Public Health*, 38:449–465.

Eggleston, K. (2005). Multitasking and mixed systems for provider payment. *Journal of Health Economics*, 24(1):211–223.

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical Care*, 36(1):8–27.

Epstein, A. J. and Nicholson, S. (2009). The formation and evolution of physician treatment styles: An application to cesarean sections. *Journal of Health Economics*, 28(6):1126–1140.

Epstein, A. M. (2006). Paying for performance in the United States and abroad. *The New England Journal of Medicine*, 355(4):406–408.

Essex, R., Talagala, I., Dada, O., and Rao, M. (2021). Clinical Excellence Awards—time for a fairer NHS rewards scheme. *BMJ*.

Evans, R. G. (1974). Supplier-Induced Demand: Some Empirical Evidence and Implications. In Perlman, M., editor, *The Economics of Health and Medical Care: Proceedings of a Conference held by the International Economic Association at Tokyo*, International Economic Association Series, pages 162–173. Palgrave Macmillan UK, London.

Everding, J. and Marcus, J. (2020). The effect of unemployment on the smoking behavior of couples. *Health Economics*, 29(2):154–170.

Feldstein, M. S. (1970). The Rising Price of Physician's Services. *The Review of Economics and Statistics*, 52(2):121–133.

Feng Lu, S. (2012). Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes. *Journal of Economics & Management Strategy*, 21(3):673–705.

Ferguson, B. and Keen, J. (1996). Transaction costs, externalities and information technology in health care. *Health Economics*, 5(1):25–36.

Finkelstein, A., Gentzkow, M., and Williams, H. (2016). Sources of geographic variation in health care: evidence from patient migration. *The Quarterly Journal of Economics*, 131(4):1681–1726.

Fisher, E. S. (2006). Paying for performance–risks and recommendations. *The New England Journal of Medicine*, 355(18):1845–1847.

Freier, R., Schumann, M., and Siedler, T. (2015). The earnings returns to graduating with honors — Evidence from law graduates. *Labour Economics*, 34:39–50.

Frijters, P., Johnston, D. W., and Shields, M. A. (2008). Happiness Dynamics with Quarterly Life Event Data. Technical Report 3604, Institute of Labor Economics (IZA).

Gaynor, M., Moreno-Serra, R., and Propper, C. (2013). Death by Market Power: Reform, Competition, and Patient Outcomes in the National Health Service. *American Economic Journal: Economic Policy*, 5(4):134–166.

Glover, J. A. (1938). The Incidence of Tonsillectomy in School Children. *Proceedings of the Royal Society of Medicine*, 31(10):1219–1236.

Gosden, T., Forland, F., Kristiansen, I., Sutton, M., Leese, B., Giuffrida, A., Sergison, M., and Pedersen, L. (2000). Capitation, salary, fee-for-service and mixed systems of payment: effects on the behaviour of primary care physicians. *Cochrane Database of Systematic Reviews*, (3).

Gravelle, H., Sutton, M., and Ma, A. (2010). Doctor Behaviour under a Pay for Performance Contract: Treating, Cheating and Case Finding? *The Economic Journal*, 120(542):F129–F156.

Greb, S., Focke, A., Hessel, F., and Wasem, J. (2006). Financial incentives for disease management programmes and integrated care in German social health insurance. *Health Policy*, 78(2-3):295–305.

Grytten, J. and Sørensen, R. (2003). Practice variation and physician-specific effects. *Journal of Health Economics*, 22(3):403–418.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46.

Hainmueller, J. and Xu, Y. (2013). ebalance: A Stata Package for Entropy Balancing. *Journal of Statistical Software*, 54(1):1–18.

Hanglberger, D. and Merz, J. (2015). Does self-employment really raise job satisfaction? Adaptation and anticipation effects on self-employment and general job changes. *Journal for Labour Market Research*, 48(4):287–303.

Hausman, J. (2001). Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left. *Journal of Economic Perspectives*, 15(4):57–67.

Hausman, J. A. (1985). Chapter 4 - Taxes and labor supply. In Auerbach, A. J. and Feldstein, M., editors, *Handbook of Public Economics*, volume Volume 3, pages 213–263. Elsevier, New York.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66(5):1017–1098.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4):605–654.

Holmstrom, B. and Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization*, 7:24–52.

Horrace, W. C. and Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3):321–327.

Huesch, M. D. (2011). Is blood thicker than water? Peer effects in stent utilization among Floridian cardiologists. *Social Science & Medicine*, 73(12):1756–1765.

Hurley, J. (2000). Chapter 2 - An Overview of the Normative Economics of the Health Sector. In Culyer, A. J. and Newhouse, J. P., editors, *Handbook of Health Economics*, volume 1 of *Handbook of Health Economics*, pages 55–118. Elsevier.

Hutchison, B. (2008). Pay for Performance in Primary Care: Proceed with Caution, Pitfalls Ahead. *Healthcare Policy*, 4(1):10–15.

Ikenwilo, D. and Scott, A. (2007). The effects of pay and job satisfaction on the labour supply of hospital consultants. *Health Economics*, 16(12):1303–1318.

Institute of Medicine (2007). *Rewarding Provider Performance: Aligning Incentives in Medicare*. The National Academies Press, Washington, DC.

John Appleby and Devlin, N. (2005). *Measuring NHS Success: Can patients' views on health outcomes help to manage performance?* King's Fund.

Jones, A. M. (2009). Panel data methods and applications to health economics. *Palgrave handbook of econometrics*, pages 557–631.

Kantarevic, J. and Kralj, B. (2013). Link Between Pay for Performance Incentives and Physician Payment Mechanisms: Evidence from the Diabetes Management Incentive in Ontario. *Health Economics*, 22(12):1417–1439.

Keane, M. P. (2011). Labor Supply and Taxes: A Survey. *Journal of Economic Literature*, 49(4):961–1075.

Laffont, J.-J. and Tirole, J. (1993). *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA, USA.

Lafond, S., Charlesworth, A., and Roberts, A. (2017). A year of plenty? An analysis of NHS finances and consultant productivity. Technical report, The Health Foundation.

Lagarde, M. and Blaauw, D. (2017). Physicians' responses to financial and social incentives: A medically framed real effort experiment. *Social Science & Medicine (1982)*, 179:147–159.

Lee, T., Propper, C., and Stoye, G. (2019). Medical Labour Supply and the Production of Healthcare. *Fiscal Studies*, 40(4):621–661.

Li, J., Hurley, J., DeCicca, P., and Buckley, G. (2014). Physician Response to Pay-for-Performance: Evidence from a Natural Experiment. *Health Economics*, 23(8):962–978.

Manning, W. G. and Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20(4):461–494.

Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, 60(3):531–542.

Maradit Kremers, H., Larson, D. R., Crowson, C. S., Kremers, W. K., Washington, R. E., Steiner, C. A., Jiranek, W. A., and Berry, D. J. (2015). Prevalence of Total Hip and Knee Replacement in the United States. *The Journal of Bone and Joint Surgery. American Volume*, 97(17):1386–1397.

Marcus, J. (2013). The effect of unemployment on the mental health of spouses – Evidence from plant closures in Germany. *Journal of Health Economics*, 32(3):546–558.

Maynard, A. (2012). The powers and pitfalls of payment for performance. *Health Economics*, 21(1):3–12.

McGuire, T. G. (2000). Chapter 9 - Physician Agency. In Culyer, A. J. and Newhouse, J. P., editors, *Handbook of Health Economics*, volume 1, pages 461–536. Elsevier.

McGuire, T. G. and Pauly, M. V. (1991). Physician response to fee changes with multiple payers. *Journal of Health Economics*, 10(4):385–410.

McLennan, D., Barnes, H., Noble, M., Davies, J., Garratt, E., and Dibben, C. (2011). *The English indices of deprivation 2010*. Department for Communities and Local Government, London.

McPake, B., Scott, A., and Edoka, I. (2014). *Analyzing Markets for Health Workers: Insights from Labor and Health Economics*. Directions in Development - Human Development. The World Bank.

Michael Green, Peter Howard, Martyn Porter, Mark Wilkinson, and Nick Wishart (2017). 14th Annual Report National Joint Registry for England, Wales, Northern Ireland and the Isle of Man, Surgical data to 31 December 2016.

Mihaylova, B., Briggs, A., O'Hagan, A., and Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 20(8):897–916.

Mitchell, A. J., Crowfoot, D., Leaver, J., and Hughes, S. (2011). Does the academic performance of psychiatrists influence success in the NHS Clinical Excellence Award Scheme? *JRSM Short Reports*, 2(3):1–9.

Molitor, D. (2018). The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration. *American Economic Journal: Economic Policy*, 10(1):326–356.

Morris, S., Elliott, B., Ma, A., McConnachie, A., Rice, N., Skåtun, D., and Sutton, M. (2008). Analysis of consultants' NHS and private incomes in England in 2003/4. *Journal of the Royal Society of Medicine*, 101(7):372–380.

Moura, A., Salm, M., Douven, R., and Remmerswaal, M. (2019). Causes of regional variation in Dutch healthcare expenditures: Evidence from movers. *Health Economics*, 28(9):1088–1098.

National Institute for Health and Care Excellence (2011). Hip fracture: management. Technical Report [NICE Guideline No. 124].

National Institute for Health and Care Excellence (2014). Total hip replacement and resurfacing arthroplasty for end-stage arthritis of the hip. Technical Report [Technology appraisal guidance No. 304].

NHS (2019). The NHS long term plan. Technical report.

NHS Employers (2013). Pay and Conditions Circular (M&D) 1/2013. Technical report.

NHS Employers (2015). Pay and Conditions Circular (M&D) 1/2015. Technical report.

Nicholson, S. and Propper, C. (2011). Chapter 14 - Medical Workforce. In *Handbook of Health Economics*, volume 2, pages 873–925. Elsevier.

OECD (2014). *Geographic Variations in Health Care.* OECD Health Policy Studies. OECD Publishing, Paris.

O'Hare, A. M., Rodriguez, R. A., Hailpern, S. M., Larson, E. B., and Kurella Tamura, M. (2010). Regional variation in health care intensity and treatment practices for end-stage renal disease in older adults. *JAMA*, 304(2):180–186.

Ostendorf, M., van Stel, H. F., Buskens, E., Schrijvers, A. J. P., Marting, L. N., Verbout, A. J., and Dhert, W. J. A. (2004). Patient-reported outcome in total hip replacement. A comparison of five instruments of health status. *The Journal of Bone and Joint Surgery. British Volume*, 86(6):801–808.

Papanicolas, I. and McGuire, A. (2015). Do financial incentives trump clinical guidance? Hip Replacement in England and Scotland. *Journal of Health Economics*, 44:25–36.

Pauly, M. (1980). Physicians as Agents. In *Doctors and Their Workshops: Economic Models of Physician Behavior*, pages 1–16. University of Chicago Press.

Petersen, L. A., Woodard, L. D., Urech, T., Daw, C., and Sookanan, S. (2006). Does pay-for-performance improve the quality of health care? *Annals of Internal Medicine*, 145(4):265–272.

Phelps, C. and Mooney, C. (1993). Chapter 7 - Variations in Medical Practice Use: Causes and Consequences. In Arnould, R. J., Rich, R. F., and White, W. D., editors, *Competitive Approaches to Health Care Reform*. Urban Institute Press, Washington, D.C.

Phelps, C. E. (2000). Chapter 5 - Information diffusion and best practice adoption. In *Handbook of Health Economics*, volume 1, pages 223–264. Elsevier.

Powdthavee, N. (2011). Anticipation, Free-rider problems, and adaptation to trade unions: re-examining the curious case of dissatisfied union members. *Industrial and Labor Relations Review*, 64(5):1000–1019. Publisher: Sage Publications, Inc.

Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature*, 37(1):7–63. Publisher: American Economic Association.

Prieto, D. C. and Lago-Peñas, S. (2012). Decomposing the determinants of health care expenditure: the case of Spain. *The European Journal of Health Economics*, 13(1):19–27.

Quentin, W., Geissler, A., Wittenbecher, F., Ballinger, G., Berenson, R., Bloor, K., Forgione, D. A., Köpf, P., Kroneman, M., Serden, L., Suarez, R., van Manen, J. W., and Busse, R. (2018). Paying hospital specialists: Experiences and lessons from eight high-income countries. *Health Policy*, 122(5):473–484.

Rimmer, A. (2021). Doctors unhappy at proposed 1% pay rise for consultants and salaried GPs in England. *BMJ*, 372.

Rizzo, J. A. and Blumenthal, D. (1994). Physician labor supply: do income effects matter? *Journal of Health Economics*, 13(4):433–453.

Roland, M. (2004). Linking physicians' pay to the quality of care–a major experiment in the United kingdom. *The New England Journal of Medicine*, 351(14):1448–1454.

Rosenthal, M. B. and Frank, R. G. (2006). What is the empirical basis for paying for quality in health care? *Medical care research and review: MCRR*, 63(2):135–157.

Salm, M. and Wübker, A. (2020). Sources of regional variation in healthcare utilization in Germany. *Journal of Health Economics*, 69:102271.

Santos Silva, J. and Tenreyro, S. (2006). The Log of Gravity. *The Review of Economics and Statistics*, 88(4):641–658. Publisher: MIT Press.

Sappington, D. E. M. (1991). Incentives in Principal-Agent Relationships. *Journal of Economic Perspectives*, 5(2):45–66.

Scott, A., Sivey, P., Ait Ouakrim, D., Willenberg, L., Naccarella, L., Furler, J., and Young, D. (2011). The effect of financial incentives on the quality of health care provided by primary care physicians. *The Cochrane Database of Systematic Reviews*, (9).

Scott, I. A. (2007). Pay for performance in health care: strategic issues for Australian experiments. *Medical Journal of Australia*, 187(1):31–35.

Showalter, M. H. and Thurston, N. K. (1997). Taxes and labor supply of high-income physicians. *Journal of Public Economics*, 66(1):73–97.

Skinner, J. (2011). *Chapter 2 - Causes and Consequences of Regional Variations in Health Care*, volume 2 of *Handbook of Health Economics*. Elsevier.

Sloan, F. A. (1975). Physician Supply Behavior in the Short Run. *Industrial and Labor Relations Review*, 28(4):549.

Song, Y., Skinner, J., Bynum, J., Sutherland, J., Wennberg, J. E., and Fisher, E. S. (2010). Regional Variations in Diagnostic Practices. *New England Journal of Medicine*, 363(1):45–53.

Sæther, E. M. (2005). Physicians' Labour Supply: The Wage Impact on Hours and Practice Combinations. *Labour*, 19(4):673–703.

Thornton, J. and Eakin, B. K. (1997). The Utility-Maximizing Self-Employed Physician. *The Journal of Human Resources*, 32(1):98–128. Publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System].

Thurley, D. (2020). Pension tax rules - impact on NHS consultants and GPs. Briefing Paper CBP-8626, House of Commons Library, London.

Todd, P. E. (2007). Chapter 60 - Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated. In Schultz, T. P. and Strauss, J. A., editors, *Handbook of Development Economics*, volume 4, pages 3847–3894. Elsevier.

Town, R., Kane, R., Johnson, P., and Butler, M. (2005). Economic incentives and physicians' delivery of preventive care: A systematic review. *American Journal of Preventive Medicine*, 28(2):234–240.

Vahovich, S. G. (1977). Physicians' Supply Decisions by Specialty: 2SLS Model. *Industrial Relations: A Journal of Economy and Society*, 16(1):51–60.

Wennberg, J. and Gittelsohn, n. (1973). Small area variations in health care delivery. *Science (New York, N.Y.)*, 182(4117):1102–1108.

Wennberg, J. E. (1984). Dealing With Medical Practice Variations: A Proposal for Action. *Health Affairs*, 3(2):6–33. Publisher: Health Affairs.

Wennberg, J. E. (2002). Unwarranted variations in healthcare delivery: implications for academic medical centres. *BMJ*, 325(7370):961–964. Publisher: British Medical Journal Publishing Group Section: Education and debate.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data, second edition*. MIT Press.

Yang, B. M. (1987). Supply and demand elasticities of physician services: disequilibrium analysis. *Asia-Pacific Journal of Public Health*, 1(2):26–31.

Zuckerman, S., Waidmann, T., Berenson, R., and Hadley, J. (2010). Clarifying Sources of Geographic Differences in Medicare Spending. *New England Journal of Medicine*, 363(1):54–62.