University College London

Submission for the degree of Ph.D. in Economics

Essays on the Economics of Health Care Provision

Thomas P. Hoe

May 14, 2018

Declaration

I, Thomas Peter Hoe, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed

Date

This thesis represents a major chapter of my life to date. It would not have been possible without the patience of my partner and the love of my parents. I am especially grateful to Richard Blundell for the opportunity to start and for his thoughtful advice through to the finish.

Abstract

Health care provision is a major sector of the economy in all developed economies. Productivity in these settings impacts levels of taxation and insurance costs, and for patients can often mean the difference between life and death. This thesis studies the economics of health care production in a hospital setting. I use uniquely rich administrative data from England over the period 2006 to 2013. I present three new findings. First, the number of patients admitted to hospital ('crowding') has an adverse impact on the quality of care delivered in hospitals. This features in Chapter 2, where I show that more crowding, despite its adverse effects, can benefit consumers because it allows for shorter waiting times for hospital appointments. Second, the number of days a patient spends in a hospital inpatient department has a material impact on the likelihood that a patient subsequently returns to hospital for further treatment ('readmission'). I quantify this relationship in Chapter 3 and argue that it partially explains the increases in readmissions that has accompanied the adoption of price regulation through prospective payment systems. Third, policies that constrain the amount of time patients can spend in a hospital emergency department can induce cost-effective reductions in patient mortality. This finding stems from Chapter 4, which is joint work with Jonathan Gruber and George Stoye, where we use an innovative application of 'bunching' techniques to study a landmark policy in emergency departments.

Impact statement

This thesis has immediate policy relevance for the National Health Service in England. Each chapter presents new empirical evidence that can help shape future health care policy, potentially benefiting the huge volumes of patients that are treated by the NHS each year. In particular, I propose specific changes to policies that regulate: (i) waiting times for elective patients; (ii) hospital reimbursements; and (iii) emergency department waiting times.

Many of these policy issues are not specific to England and are of wider international interest: many countries face similar challenges with long waiting times for elective care, especially Canada; the English model of hospital reimbursement mirrors the U.S. system, which has also faced increases in readmissions; and, pressures on emergency departments have recently been described as an international crisis. This thesis provides a framework for analysing these issues in a general context.

The thesis also makes novel academic contributions. The central contribution is to show the importance of time in the health production process. This theme emerges in each chapter, first in the context of the trade-off between crowding and waiting times for appointments (Chapter 2), and then through the estimated health impacts of inpatient length of stay (Chapter 3) and emergency department waiting times (Chapter 4).

I have disseminated this research to policymakers including the U.K. Department of Health and the U.K. healthcare regulator, NHS Improvement, and to academic audiences in the U.K., U.S., Canada, and Europe. I will further disseminate this research through scholarly publications.

Contents

1 Introduction 13

2 Are Public Hospitals Overcrowded? 21

2.1 Background . . . 29

2.2 Data . . . 33

2.3 The impact of hospital crowding on patients . . . 39

2.4 Marginal welfare analysis . . . 64

2.5 Conclusion . . . 80

3 Efficiency Gains or Quality Cuts? 83

3.1 Background . . . 88

3.2 Economic framework . . . 94

3.3 Data . . . 97

3.4 Empirical specification . . . 98

3.5 Results . . . 102

3.6 Policy implications . . . 106

3.7 Conclusion . . . 110

4 Saving Lives by Tying Hands 113

4.1 Background . . . 119

4.2 Data . . . 124

4.3 Empirical methodology . . . 129

4.4 Results . . . 151

4.5 Identifying Mechanisms . . . 157

4.6 Conclusion . . . 166

5 Concluding remarks 169

A Average crowding effects 173

B Additional charts and tables 175

Bibliography 191

List of Tables

2.1 Mean characteristics of patients in trauma and orthopaedics and other specialties . . . 36
2.2 OLS regression estimates of hospital scheduling decisions on expected emergency admissions . . . 43
2.3 Estimated effects of emergency admissions on health outcomes . . . 49
2.4 Estimated effects of emergency admissions on inflows of emergency patients . . . 52
2.5 Estimated effects of emergency admissions on inflows of elective patients . . . 53
2.6 Estimated effects of emergency admissions on inpatient care 56
2.7 Estimated effects of emergency admissions by patient type . 60
2.8 OLS regression estimates of hospital-year average crowding effects for length of stay on hospital characteristics . . . 63
2.9 OLS and IV estimates of the effect of elective admissions on elective waiting times . . . 74

3.1 Summary statistics . . . 97
3.2 Estimated effects of emergency admissions on health outcomes . . . 104

4.1 Summary statistics . . . 127
4.2 Demographic tests of the no-selection assumption . . . 153
4.3 Estimated distortion effects of the target on treatment decisions and costs . . . 154
4.4 Estimated distortion effects of the target on mortality . . . 156
4.5 OLS regressions of the estimated 30-day mortality reductions on other effects of the target . . . 165

B.1 Summary statistics for the panel dataset . . . 180
B.2 Summary statistics for the inpatient dataset . . . 181
B.3 Summary statistics for the ED dataset . . . 182
B.4 Estimated effects of emergency status on length of stay . . . 183
B.5 Top 10 causes of unplanned readmission . . . 184

B.6 Regression estimates of emergency shocks on admissions at other hospital departments 185
B.7 Estimated effects of emergency admissions on inpatient care and health outcomes using different horizon outcomes 186
B.8 Estimated effects of emergency admissions by inpatient care and health outcomes with different sets of control variables . 187
B.9 Estimated effects of emergency admissions on inflows of emergency patients 188
B.10 Estimated effects of emergency admissions on the characteristics of admitted elective patients 189
B.11 Estimated effects of emergency admissions on inpatient care and health outcomes for emergency patients by expected mortality risk 190

List of Figures

1.1 Schematic diagram of a traditional hospital 15

2.1 Heterogeneity between trauma and orthopaedic patients . . . 37
2.2 Example of the decomposition of daily emergency admissions 41
2.3 Poisson property of daily emergency admissions 44
2.4 Non-parametric estimates of the effect of emergency admissions on 7-day unplanned readmission for discharge cohorts 50
2.5 Non-parametric estimates of the effect of emergency admissions on inflows of elective patients 54
2.6 Non-parametric estimates of the effect of emergency admissions on inpatient care 57
2.7 Binned scatter plot of hospital-level average crowding effects for 7-day unplanned readmission and length of stay for emergency patients 61
2.8 Revealed preference estimates of C across the patient population 79

3.1 First-stage estimates by patient severity 106
3.2 Diagnosis types by patient severity and age 107

4.1 Distribution of wait times 129
4.2 Inpatient admission probability conditional on wait time . . . 131
4.3 Estimated counterfactual wait time distribution 135
4.4 Mean wait times by patient severity and expected volumes of ED arrivals 137
4.5 Impact of queues on wait times for arriving patients by early- and late-exit groups 140
4.6 Estimated counterfactual admission probability conditional on wait times 146
4.7 Constructing the composition-adjusted counterfactual for admission probability 148
4.8 Demographic test of the no-selection assumption using age . . 150

4.9 Proportion wait beyond the threshold vs. predicted mortality by diagnosis groups . . . 159
4.10 Estimated effects of the target vs. predicted mortality by diagnosis groups . . . 160
4.11 Estimated effects of the target vs. inpatient crowding by crowding-severity groups . . . 163

B.1 Hospital-level tests of first-order serial correlation . . . 176
B.2 Distribution of estimated hospital-level effects of emergency admissions on length of stay by patient type . . . 177
B.3 Non-parametric estimates of the effect of elective admissions on elective waiting times . . . 178
B.4 Distribution of wait times at a large hospital in California . . . 179

Chapter 1

Introduction

Health care provision is a major sector of the economy in all developed economies, representing around 9% of GDP in OECD countries. The level of spending and associated health outcomes across countries, however, varies dramatically. Annual spending is around 3,000 USD per capita in the U.K. and several other European nations, while in the U.S. it is approximately three times higher. In comparison, the likelihood of death following a heart attack, to take just one example, is similar in Italy and the U.S. but over 40% higher in the U.K. (OECD, 2015). It is clear then, that health care productivity can have a substantive impact on levels of taxation and insurance costs, and for patients can often mean the difference between life and death.

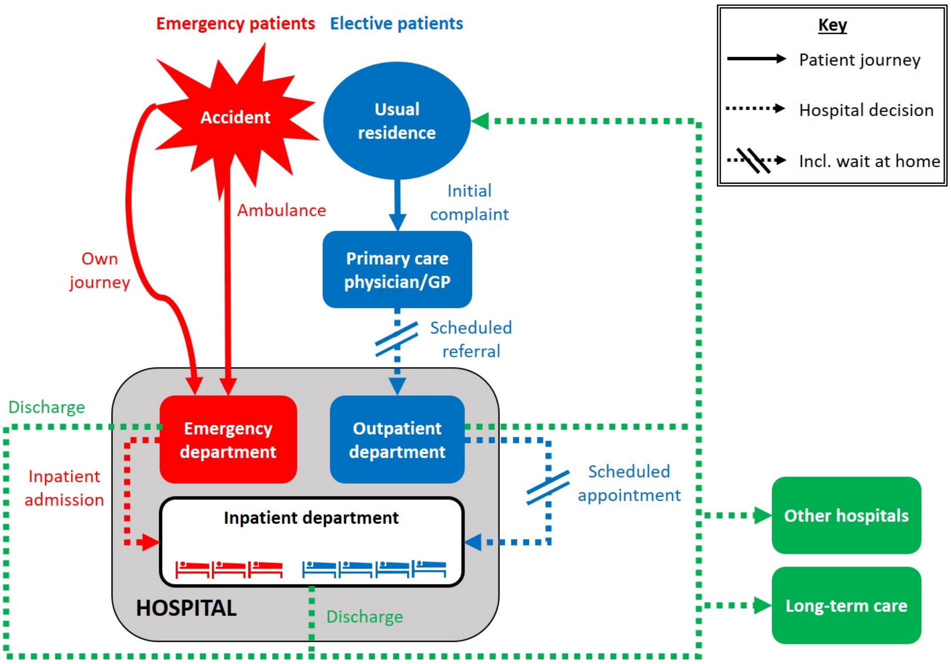
Economics plays a central role in the delivery of health care services. Providers are subject to a myriad of market and regulatory pressures, many of which are based on economic considerations. Consider some of the most prominent examples: providers compete for patients, which has long been the case in the U.S. and pro-competitive reforms have recently been implemented in several European countries; regulators determine which products can be provided to patients, sometimes according to explicit cost-effectiveness benchmarks; provider payments are often subject to explicit price regulations; and regulators can impose operational targets on providers, such as maximum waiting times. These policies all shape the incentives of health care providers.

The issues of provider productivity and economic policy are at the heart of this thesis. I focus my attention on public hospitals in the the English National Health Service (NHS). This is one of the most famous health services globally, being the largest single-payer system and offering full insurance to all U.K. residents. To set the scene for the three major chapters that follow, I begin by briefly outlining the merits of studying the English setting to learn about health care provision more generally, and provide a description of how hospitals are typically organised.

A major strength of the English NHS for research purposes is that it collects data centrally, meaning that data on the universe of patient visits to public hospitals are available to researchers. The Hospital Episodes Statistics (HES) is the database containing this information, and I rely on extracts for the years 2006 through 2013. Unlike many other hospital datasets, HES contains the entirety of a single health care system. HES therefore allows researchers to track how each user of the service fully interacts with the secondary health care system, linking their behaviour across time and hospitals, and also allows for a complete view of how busy a hospital is at each point in time. I exploit these features of the data several times in this thesis.

As well as the rich availability of data, England also offers an excellent setting to study economic policy. Since the early 2000s, there has been a wave of health care reforms in England, as command-and-control style policies, such as waiting time targets, have been implemented alongside

Figure 1.1: Schematic diagram of a traditional hospital



decentralised and market-based policies, such as competition for patients. These changes provide an ideal opportunity to study and contrast policies that are being used across health care systems globally.

Turning now to the organisation of a hospital, Figure 1.1 presents a schematic diagram of a traditional hospital. There are two classes of patient, emergencies and electives. Emergency patients often arrive at hospital after an accident (e.g. fall) or health shock (e.g. heart attack), and travel to hospital either by their own means or by ambulance. In contrast, elective patients reach the hospital through a longer, staged process, after first visiting their primary care physician and seeking a hospital referral. Elective patients will usually have medical conditions that are not time sensitive (e.g. hip replacement).

The hospital itself can be broadly divided into three types of department. The emergency department (ED) is where emergency patients are first seen, and here patients are triaged, assessed, and initial treatment is provided. Most hospitals have a single ED, and I study this setting in Chapter 4. An outpatient department is another area of the hospital, and this is where elective patients first arrive following a referral from primary care. As an outpatient, patients are evaluated in relation to a specific illness or injury; for example, they may receive a CT scan, an assessment for a hip replacement, or have a number of blood tests taken. There can be several outpatient departments in a hospital, each assigned to a specific medical specialty (e.g. cardiology, oncology, trauma and orthopaedics).

Any patient that requires further treatment after an ED or outpatient visit is admitted to an inpatient department. For emergency patients admissions happens on the same day as the ED visit, while elective patients will be given an appointment for admission at a future date. Inpatient departments house patients during their receipt of surgical or medical care, and for a number of days during any required recovery period. As with outpatient departments, inpatient departments are organised by medical specialty. In Chapters 2 and 3, I focus on trauma and orthopaedic inpatient departments, which deal with injuries of the musculoskeletal system (e.g. hip replacements, broken arms and legs).

Once hospital treatment is complete, whether that be in the ED, outpatient or inpartient department, patients are discharged to their home, another hospital, or an alternative care facility (e.g. long term care).

I now turn to a description of the three core chapters of the thesis and outline their respective contributions. Chapter 2, entitled 'Are Public Hospitals Overcrowded? Evidence from Trauma and Orthopaedics', studies the trade-off that hospitals face between how many non-emergency patients to admit each period ('crowding') and how long these patients must wait for an appointment ('waiting'). I first exploit pseudo-random variation in emergency admissions to estimate the short-run effects of crowding on patient health outcomes. I find that crowding has adverse effects, causing the rate of unplanned readmission to vary by at least 22%. I show that variation in length of stay caused by binding bed constraints is a plausible mechanism for these effects. I then evaluate policies which reduce crowding by rationing elective admissions and thereby increase waiting times. I estimate the impact of elective admissions on equilibrium waiting times by exploiting technological change and compare this impact to the crowding effects using a model of consumer welfare. The optimal crowding condition derived from the model is strongly rejected by the data and the results indicate that hospitals' incentives undervalue patients' preferences for waiting times. Policies which increase elective admissions, reducing waiting times but increasing readmissions, are therefore predicted to improve consumer welfare.

Chapter 3, entitled 'Efficiency Gains or Quality Cuts? How Prospective Payment Can Reduce Health Care Quality', retains the same setting as the preceding chapter and focuses attention on the relationship between length of stay and readmission outcomes. I begin by noting that the dominant form of price regulation for health care providers, Prospective Payment Systems (PPS), has proven highly effective at reducing costly time spent in hospital. Estimates range between 20 and 25%. I investigate whether these shorter stays represent efficiency gains, such that health outcomes were unaffected, or quality cuts such that health outcomes were impaired. For patients on the margin of being discharged, I estimate that changes in length of stay have a large impact on the likelihood of experiencing an unplanned readmission. Increasing length of stay for marginal patients by around 11% is predicted to reduce readmissions by up to 7%. This mechanism suggests that PPS regulation, which reduced length of stay, also led to reductions in health care quality. Moreover, I show why recent policies that penalize hospitals for readmissions are ineffective at reversing these effects.

Chapter 4, entitled 'Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers', changes tack and turns to the emergency department (ED). This complex node of healthcare delivery is facing market and regulatory pressure across developed economies to reduce wait times. In this chapter, together with Jonathan Gruber (MIT) and George Stoye (UCL), we study how ED doctors respond to such incentives, by focussing on a landmark policy in England that imposed strong incentives to treat ED patients within four hours. Using bunching techniques, we estimate that the policy reduced affected patients' wait times by 19 minutes, yet distorted a number of medical decisions. In response to the policy, doctors increased the intensity of ED treatment and admitted more patients for costly inpatient care. We also find a striking 14% reduction in mortality. To determine the mechanism behind these health improvements, we exploit heterogeneity in patient severity and hospital crowding, and find strongly suggestive evidence that it is the reduced wait times, rather than the additional admits, that saves lives. Overall we conclude that, despite distorting medical decisions, constraining ED doctors can induce cost-effective reductions in mortality.

Two common themes emerge from the thesis: first, the importance of time–whether that be waiting for hospital access, or spent in different parts of the hospital–as an input to the health production process; and second, the way in which economic research can be utilised in a medical setting to learn about health production and inform economic policy. I discuss these themes more fully in the concluding remarks, and hope that the insights from this thesis can help other researchers and policy makers. I now present the three, largely self-contained, chapters of research.

Chapter 2

Are Public Hospitals Overcrowded? Evidence from Trauma and Orthopaedics[1]

In recent years there have been dramatic changes in hospital capacity. Across OECD countries hospital beds per capita fell by 13% between 2000 and 2013, with pronounced drops of over 30% in some countries (OECD, 2015). While partially driven by technological improvements, the downsiz-

[1] I am extremely grateful to Richard Blundell for advice and suggestions, and to Orazio Attanasio, David Chan, David Cutler, Alejandro Estefan, Jonathan Gruber, Amanda Kowalski, Neale Mahoney, Costas Meghir, Magne Mogstad, Lars Nesheim, David Oliver, Aureo de Paula, Imran Rasul, John Van Reenen, Pasquale Schiraldi, Marcos Vera-Hernández, and Heidi Williams for helpful discussions. I thank seminar participants at Arizona, Cornell PAM, The Department of Health, EMCON 2016, The Institute for Fiscal Studies, Manchester CHE, Mannheim, McGill, NHS Improvement, Oxford HERC, RES 2016, TADC 2017, Toulouse, and University College London. I also thank many healthcare professionals from the English NHS for providing me with insights on the industry. This work was supported by the Economic and Social Research Council grant number ES/J500185/1. The data was made available by NHS Digital under data sharing agreement NIC-369748-C1Y0V.

ing and closures of hospitals have led to widespread concerns about hospital crowding and access to hospital care. High profile examples include emergency department crowding (Hoot and Aronsky, 2008), inpatient bed shortages (British Medical Association, 2017), and the Veterans Health Administration scandal (Kizer and Jha, 2014).

The notion that crowding may have adverse effects on patient health outcomes implies a trade-off: a hospital can moderate crowding, potentially improving the quality of care delivered, by admitting fewer patients and thereby making some patients wait longer for admission. This type of rationing with queues is routinely used by healthcare providers although the incentives that influence the trade-off are complex. Market pressures shape incentives to deliver quality (Cooper et al. 2011; Gaynor et al. 2013; Chandra et al. 2016), as can intrinsic incentives (Kolstad, 2013), and both quality and access are highly regulated. Examples of relevant policies include quality-based financial penalties (Gupta, 2017), malpractice liabilities (Kessler and McClellan, 1996), mandatory quality reporting (Dranove et al., 2003), and waiting time targets (Propper et al., 2008). An important question for policymakers in this context is whether this combination of incentives delivers an appropriate trade off between quality of care (crowding) and access to care (waiting).

This chapter examines this trade-off in the context of the one of the largest public healthcare systems, the English National Health Service. The first question I address is whether the volume of hospital admissions affects patient health outcomes (an effect I refer to as ‘crowding’). This presents an endogeneity problem because admissions - or any other measure of crowding - will be correlated with unobservable patient characteristics and hospital production inputs. I deal with this by exploiting shocks to emergency admissions which I show are pseudo-random and plausibly exogenous. I use this variation to estimate the short-run effects of crowding and explore the mechanisms that generate these effects using rich administrative data.

I then ask whether reductions in hospital crowding would be desirable. I focus on consumer welfare and policies that would reduce crowding by rationing non-emergency ('elective') admissions. I derive an optimal crowding condition that characterises the welfare maximising trade-off between crowding and waiting times and test whether this condition holds empirically. This provides an assessment of whether marginal changes in elective admissions would improve consumer welfare and the distributional impacts of such changes.

Throughout the chapter I use linked administrative data on the universe of medical records for publicly funded hospital visits in England during the period 2006 to 2013. The data links hospital emergency and inpatient departments, and allows me to track patients across hospitals and over time. For each hospital visit, I observe extensive information about the patient's health conditions and the treatments received. This data allows for a highly granular study of admissions, health outcomes, waiting times, and many other dimensions of hospital care. I focus on trauma and orthopaedic departments, which are one of the largest hospital inpatient departments and treat diseases and injuries of the musculoskeletal system such as arthritis and broken bones. The setting is well suited to the study: emergency trauma admissions occur frequently (which I use as a source of variation in the crowding analysis), unplanned readmissions are observed in the data (which are a relevant health outcome for these patients), and rationing is commonly adopted in this setting.

I begin by studying daily emergency admissions at each hospital. I show that after conditioning on seasonal and within-week variation, emergency admissions closely approximate a Poisson process. This indicates that the shocks to emergency admissions are pseudo-random and the result of a series of low probability and independent events, consistent with the types of accidents that lead to emergency trauma admissions (e.g. road traffic accidents, falls, and sports injuries). The Poisson property implies that hospitals are unable to forecast and plan for admission shocks and that, under certain conditions, the shocks will be unrelated to the characteristics of admitted patients. I use this rich source of variation in emergency admissions to assess the short-run effects of crowding.

I find that emergency admissions have substantial adverse impacts on patient health outcomes. A one standard deviation increase in admissions (2.8 admissions relative to an average department size of 60 beds) increases the likelihood of unplanned readmission by 0.163 percentage points (5.8% per cent relative to the baseline). Using a flexible non-parametric specification I show that these effects occur across the admission distribution and cause the readmission rate to vary by at least 22%.

I explore several mechanisms that could drive these health impacts. I find that bed constraints and physicians varying how soon they discharge patients is the most plausible explanation. This is evident in correlations between the effect of emergency admissions on length of stay and on readmission: bigger reductions in length of stay are consistently associated with bigger increases in readmissions. This correlation holds across subgroups of patients and hospitals. The implication is that hospitals do not appear to reserve sufficient bed capacity to absorb even small variations in emergency admissions.

The data also allows me to rule out a number of mechanisms. I analyse the inflows of emergency and elective patients and find that selection is not a plausible explanation for the health impacts. I find no evidence that emergency admissions impact ambulance diversion, admission decisions in the emergency department, transfers to other hospitals, or the discharge destination. I do find evidence of delays (in the emergency department and prior to inpatient surgery) and cancellations of elective appointments but the magnitude and timing of these effects is not sufficient to explain the health impacts.

By investigating the heterogeneity in the crowding effects, I find that size is important: smaller hospitals and those with smaller trauma and orthopaedic departments exhibit larger crowding impacts. This is consistent with these smaller units being less able to find physical space to accommodate volatility in the emergency admissions. I also find that, conditional on size, hospitals which admit higher volumes of elective patients exhibit larger crowding impacts. By admitting these patients, departments effectively leave less capacity available for emergency patients and this highlights the importance of hospitals' incentives with respect to elective admissions.

I then turn to a marginal welfare analysis. This asks whether consumer welfare would be improved by marginal changes that ration the number of elective admissions. Fewer elective admissions will reduce hospital crowding (and thus readmissions) but increase the length of time elective patients wait for hospital appointments. To analyse this question I set up a model of patients and hospitals with a regulator that sets elective admissions to maximise consumer welfare. The model implies an optimal crowding condition that states the marginal rate of technical substitution between waiting times and readmission outcomes should be proportional to consumers' relative preferences for these outcomes. This condition provides an empirical test of whether hospitals' incentives with respect to rationing maximise consumer welfare. A particular advantage of the test is that it can be implemented with reduced-form estimates.

To implement the test I require estimates of how equilibrium waiting times respond to changes in elective admissions. I exploit within-region variation in technological change for this purpose. I use the introduction of 'fast-track surgery' - an innovation in post-operative recovery procedures for elective patients - which led to major reductions in length of stay without impairing health outcomes (Kehlet, 2013). Shorter hospital stays in turn allowed hospitals to increase elective admissions while maintaining the same capacity. I use this plausibly exogenous shift in admissions to estimate the response of equilibrium waiting times. I find that a one patient decrease in hospital occupancy across a region is estimated to increase mean waiting times for elective patients by 5.6 days (6.6% relative to baseline). This compares to the crowding estimates which, under a similar change in occupancy, imply a decrease in the likelihood of readmission by 0.058 percentage points (2.1% relative to the baseline).

Combining these estimates with benchmarks for preferences, I can strongly reject the optimal crowding condition. This implies that hospitals' incentives and current levels of elective admissions do not maximise consumer welfare. The results also show that hospitals' incentives undervalue preferences for waiting times, and that marginal increases in elective admissions would improve consumer welfare. As a result those policies which reduce waiting times but increase readmissions are predicted to improve consumer welfare. Moreover, I find that such policies, while benefiting consumer welfare overall, will generate benefits disproportionately for younger males and older females. These groups are particularly exposed to elective waiting times but shielded from the impacts of crowding.

This chapter contributes to two literatures. The first concerns the effects of crowding on medical care and patient health outcomes. This has a long history in medical research where studies typically focus on the association between measures of hospital occupancy and patient health outcomes (see the reviews by Hoot and Aronsky (2008) and Eriksson et al. (2017)). One channel through which these effects can arise is if the available medical resources per patient declines as hospitals become busier. In work related to this same notion, economists have instead focussed directly on variation in medical resources per patient - induced by exogenous factors other than crowding - and examined how this affects patient health outcomes (examples include Almond et al. (2010), Almond and Doyle (2011), Bartel et al. (2014), Card et al. (2009), Doyle (2005), Doyle (2011), Gruber and Kleiner (2012), and Friedrich and Hackmann (2017)). From this perspective, crowding could be viewed purely as an instrument that creates variation in medical care.

The second literature concerns non-price rationing of hospital care. This includes studies that have examined waiting times (Lindsay and Feigenbaum 1984, Windmeijer et al. 2005, Propper et al. 2008) and admission decisions (Joskow 1980, Fiedler 2016, Freedman 2016).[2] Of these rationing studies, the recent papers by Fiedler (2016) and Freedman (2016) are closest to my work. Both examine how daily variation in the occupancy of intensive care units (ITUs) affects the likelihood of ITU admission, finding that higher

[2] There are three related literatures. The first concerns hospital capacity, which is the natural alternative to rationing policies, and a series of papers have studied the implications of demand variation for capacity and hospital costs (Friedman and Pauly 1981, Friedman and Pauly 1983, Gaynor and Anderson 1995, Keeler and Ying 1996, Hughes and McGuire 2003). The second is from operational research, which has studied hospital waiting times extensively and dates back to Young (1962), Shonick (1970) and Shonick and Jackson (1973). The third is an early literature in economics that incorporated these operational research techniques into models of service industries (De Vany 1975, De Vany 1976, De Vany and Saving 1977, De Vany and Frey 1982).

levels of occupancy reduce the likelihood of admission. Relative to this chapter, this type of response can be viewed as rationing when there is no prospect of waiting (either ITU care is provided or not) or as a crowding effect where quality of care deteriorates because of ITU capacity constraints (non-ITU care is provided when there is no ITU care available).

I make the following two principle contributions. First, I use a novel source of variation - shocks to emergency admissions - to estimate the causal impact of crowding on patient health outcomes. I combine this identification strategy with linked administrative data to provide a detailed study of the mechanisms that generate these impacts on health outcomes and I am able to rule out a number of mechanisms. Second, I develop a framework for evaluating hospitals' incentives to moderate these crowding effects by rationing elective admissions. This connects the literature on crowding with the literature on rationing and illustrates an important trade-off between two dimensions of hospital production. I show that despite the adverse effects of crowding on patients, it is not optimal to reduce crowding because the increase in waiting times this would imply more than offset the gains in consumer welfare. More generally these results illustrate that policies targeted at quality of care (e.g. malpractice liability) may have unintended consequences for access to care and vice versa.

The chapter proceeds as follows. Section 2 provides information about hospital inpatient departments and the institutional setting. Section 3 describes the data. Section 4 sets out the empirical analysis of crowding. Section 5 sets out the marginal welfare analysis including the analysis of waiting times. Section 6 concludes.

2.1 Background

2.1.1 Hospital inpatient departments

Inpatient departments are where the majority of care for serious injuries and illnesses is provided. These departments are organised by medical specialty, which group together related diagnoses and medical procedures. Examples include cardiology (diagnoses relating to the heart), neurology (brain), and trauma and orthopaedics (musculoskeletal system). Inpatient departments account for a large part of physical hospital capacity as many patients require accommodation for overnight stays.

Patients in inpatient departments are classified as either elective or emergency cases. Elective patients are those that require treatment but it is not urgent. A common example is a hip replacement. Elective patients obtain an inpatient appointment after first seeking a referral from a primary care physician and then having an initial assessment at an outpatient consultation with a secondary care physician. If treatment is required, the patient will join a waiting list and be given an inpatient appointment at a prespecified time in the future, which may be several weeks or months later. Emergency patients in contrast often have severe conditions that require immediate treatment. Common examples include broken bones. These patients first attend the emergency department (ED), arriving by their own means or via an ambulance. The ED provides triage and initial treatment and then a decision is made about whether further treatment is required. The majority of ED cases are discharged without additional treatment, but those that do require treatment are admitted to an appropriate inpatient department.

Upon admission, both elective and emergency patients experience a similar overall pathway: a surgical or medical procedure is provided on or shortly after admission, after which they are monitored and nursed through the recovery process until they are considered fit for discharge. The specifics of a pathway will vary according to the diagnosis. In the case of high volume elective surgeries, such as a total hip replacement procedure, the pathway can be very standardised. These patients will often have set goals for each day of their stay and will be discharged as soon as they can navigate a flight of stairs unaided. In contrast, emergency patients with more complex and varied health conditions require a more flexible pathway. Examples include patients with multiple or very severe injuries, who will be assessed on a day-by-day basis according to their needs.

Hospitals have a degree of control over the flow of patients in and out of inpatient departments. The inflow of elective patients is primarily controlled through appointments. These are set in advance but can be cancelled or rescheduled at short notice. There is far less control over the inflow of emergency patients, since when dealing with urgent and severe cases there is often no option but to accept patients. For less urgent or severe cases, there is potentially more control, as hospitals can divert ambulances to alternative hospitals or adjust the threshold for inpatient admissions from the ED, although these responses have potential for adverse effects on patients. The outflow of elective and emergency patients is controlled by discharge decisions. Patients are evaluated daily and discharged once they are medically fit and able to leave the hospital. Upon discharge they may either return to their home residence or be transferred to another hospital or an alternative care facility.

Decisions over patient flow are made by a combination of physicians and managers. Physicians are responsible for all decisions about individual patients. This includes whether to admit (following an outpatient consultation or an ED visit), all treatment decisions, and when a patient is fit for discharge. Managers are responsible for operational decisions such as when to cancel elective appointments or divert ambulances. Patient flow is monitored closely throughout each day and managers will communicate information to physicians through meetings and via electronic means.

2.1.2 Institutional setting

The empirical application focuses on public hospitals in the English National Health Service. This is a single-payer healthcare system funded through the proceeds of general taxation. All approved hospital treatments are provided to residents for free.[3] Public hospitals provide the large majority of elective inpatient care and all emergency care in England. These hospitals are centrally managed and regulated by a number of government departments. Policies are set that specify targets and incentives for hospitals to operate by and these can apply to financial, operational, and clinical performance. The majority of policies are set at the national level and apply equally to all public hospitals.

The sample covers the period 2006 to 2013. During this period two policies had a major influence on the incentives of hospitals to admit elective admissions. The first policy is the ‘Referral to Treatment’ waiting time target which specified the maximum time between a referral and admission for all elective patients. The target was introduced in 2006, setting a maximum of six months, and then tightened to 18 weeks from 2008. The target was strongly enforced through senior management incentives (Propper et al., 2008) and financial penalties (£300 per patient that waits above the thresh-

[3] The exception to this is prescription drugs which are subject to a small co-payment.

old). This policy proved very effective: average waiting times fell by over 50% between 2000 and 2010 for trauma and orthopaedic elective patients.

The second policy is the 'Payment by Results' tariff that specifies hospital reimbursements. Hospitals were paid on the basis of a prospective payment system which, similar to the DRG system in the U.S., specified fixed payments per admission of each diagnosis type (Department of Health, 2012). Hip replacements, for example, were reimbursed at a rate of approximately £6,000 per patient. The tariff, which was implemented for most hospitals in 2006, created a financial incentive for hospitals to treat higher volumes of elective patients.[4] This incentive was material: hospitals were typically failing to meet their financial obligation to break even during the sample period and increasing admissions was the primary way to raise revenue. For example, by the end of 2005 hospitals were on average running a financial deficit of 2.5% of total revenue (with 10-90th percentile range of -8.5% to 0.5%) and this only improved marginally by 2013 when the average deficit was still 1.6% (with 10-90th percentile range of -7.0% to 1.0%).[5]

There were also a number of other policy changes that were implemented around the beginning of 2006. This includes: the removal of restrictions on hospital choice (Cooper et al. 2011, Gaynor et al. 2013); increased monitoring of clinical performance, especially for elective patients (NHS Digital, 2017); and, capacity expansions achieved by enabling private hospitals to conduct publicly-funded elective care (Kelly and Stoye, 2015).

[4] The tariff was implemented on a limited scale between 2003 and 2005, covering a small number of hospitals and a limited range of activities. It was rolled out to all hospitals and all activity over the period 2006 to 2008.

[5] Financial data obtained from the annual accounts of NHS Trusts. Data in 2005 was not available for all Trusts. Financial figures exclude any financial support.

2.2 Data

I use linked administrative data on medical records for inpatient and ED visits from the Hospital Episodes Statistics (HES). This data provides a complete picture of secondary care use at public hospitals in England. It allows me to observe each patient's care history and track each episode of care from initiation through to discharge via any transfers. Rich information is available for each episode, including the hospital site, admission and discharge dates, a complete listing of diagnoses (5-digit ICD-10 codes) and procedures (OPCS codes), and a standard set of demographic information. I have inpatient records available for the period 2006 to 2013 and the ED records for the period 2010 to 2013.[6]

The empirical application focuses on trauma and orthopaedic departments at general acute hospitals with an active ED in England. These departments treat musculoskeletal conditions such as broken bones and arthritis and are the third largest department measured by admissions (6.6% in 2013). The trauma and orthopaedic setting is well suited to the analysis: they are strongly influenced by the policy pressures discussed above, relevant outcome measures can be constructed from the data, and emergency admissions, which I use as a source of identification for the analysis of crowding, are common in these departments.

2.2.1 Sample construction

I construct three data samples for the analysis. In each sample I identify general acute hospitals using the Estates Return Information Collection data and define hospitals by their postcode, which references a specific geographic

[6] The dates refer to financial years beginning in April and ending in March the following year. This convention is used throughout the chapter.

location. I define an ED as active in a year if the trauma and orthopaedic department received on average five emergency admissions per week in each quarter of the year. Trauma and orthopaedic patients are identified by the medical specialty that they are treated under.

The first sample, referred to as the panel dataset, contains hospital-day level information on the number of elective and emergency admissions to trauma and orthopaedic departments. I exclude hospital-years with incomplete information, which can occur if the specific hospital site is not recorded accurately.[7] After making these exclusions I further exclude departments with fewer than three years of data. These exclusions, which together account for 30% of hospital-days, ensure that each department has a reliable and reasonably long time-series of data which is important when I decompose emergency admissions. The qualitative results are robust to changes in these exclusion rules.[8]

The second sample, referred to as the inpatient dataset, contains medical records for patients admitted to trauma and orthopaedic departments. I limit this dataset to patients admitted and discharged on days contained in the panel dataset. I construct the following variables: an indicator for whether the primary operation received involves no overnight stay for the median patient ('daycase operation'); an indicator for whether surgery occurred after the day of admission ('delayed operation'); a count of the number of medical procedures received; length of stay; an indicator for discharge

[7] I define a hospital-year as incomplete using two rules: (i) the data contains fewer than 51 weeks in a year; (ii) the data contains a week where emergency and elective admissions are both at least 80% below the annual average. After these exclusions I also remove four weeks either side of any data break to ensure that the data is not missing information from adjacent but excluded periods.

[8] The majority of results are also robust to removing all of the exclusion rules. One exception to this is the analysis of elective admissions. Periods of incomplete data, which cause elective and emergency admissions to move together simultaneously, unduly influences this analysis by forcing a positive correlation between the two types of admission.

to another hospital (‘transfers out’); an indicator for discharge to the patient’s home residence (‘home discharge’); a count of the number of diagnoses recorded; the Charlson co-morbidity index (a proxy for the severity of underlying health conditions); the count of ED admissions in the past year (another proxy for underlying health conditions); 7-day unplanned readmission; and, 30-day in-hospital mortality. Unplanned readmissions are defined as any emergency inpatient admission to any hospital within a specified time horizon from the previous discharge. I use a 7-day horizon in the baseline analysis and conduct robustness tests using other horizons.

The third sample, referred to as the ED dataset, contains medical records for all visits to emergency departments. I match this data to the inpatient dataset which later allows me to evaluate how inpatient crowding affects ED outcomes. This matching process is incomplete because the ED data does not always contain information on the specific hospital site. I am able to match 65% of hospitals in the ED dataset to the inpatient dataset. I also exclude patients that visit the same ED multiple times on the same day as these patients cannot be matched uniquely to the inpatient data (2.5% of visits) and limit the data for matched hospitals to the days present in the panel dataset. I compute three variables using the ED data: an indicator for whether a patient attended their nearest hospital; the (straight-line) distance travelled to hospital; and, an indicator for whether admission was to the trauma and orthopaedic department (rather than another inpatient department).

Together these three samples provide information on 149 trauma and orthopaedic departments, 3.9 million inpatient visits (2006-2013), 97 emergency departments, and 22.5 million ED visits (2010-2013). Tables B.1 to B.3 present basic summary statistics for each sample.

Table 2.1: Mean characteristics of patients in trauma and orthopaedics and other specialties

	T&O	Other patients	% diff.
Age	52.9	56.3	−6
Male, %	48.4	45.4	7
White, %	85.4	89.4	−5
Emergency, %	39.4	36.9	7
Elective waiting time, days	84.8	58.9	44
Diagnosis count	3.4	3.5	−3
Charleson index	1.7	2.8	−40
ED admissions in past year	0.8	1.1	−30
7-day unplanned readmission, %	2.8	4.1	−31
30-day in-hospital death, %	1.1	2.4	−55

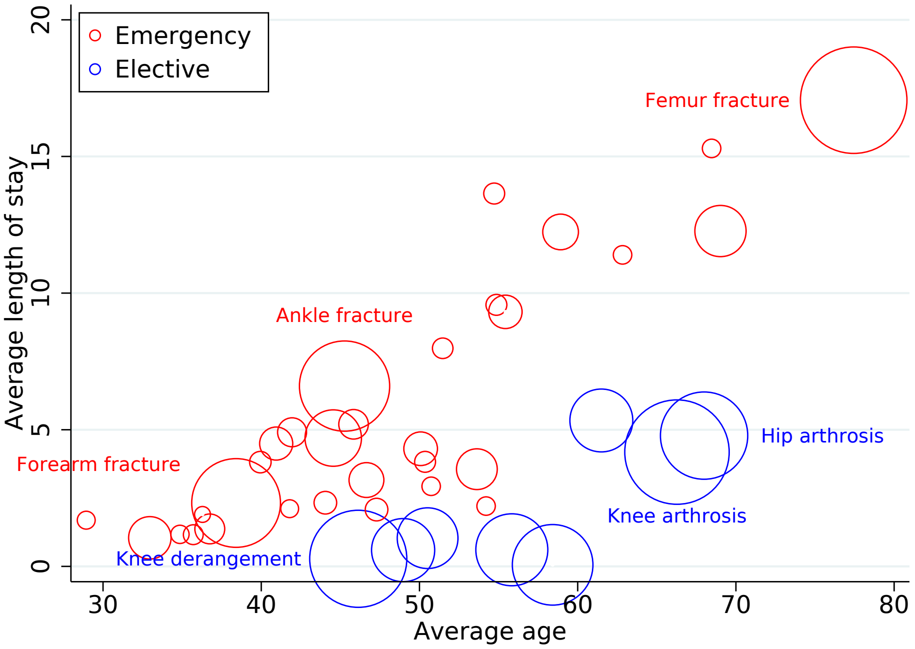
Notes: (1) 'Other specialties' excludes paediatrics and maternity care and is based on a 1% sample of the full inpatient HES data; (2) Charleson index is a measure of co-morbidities.

2.2.2 Descriptive statistics

Characteristics of trauma and orthopaedic patients

Table 2.1 presents the mean characteristics for trauma and orthopaedic patients in comparison to patients from other specialties at general acute hospitals (excluding maternity and paediatric care). Trauma and orthopaedic patients are similar along several dimensions to patients admitted to other specialties. The demographic mix is comparable in terms of age, gender and ethnicity, but trauma and orthopaedic patients wait significantly longer for elective care, and are healthier in terms of pre-existing conditions (diagnoses, co-morbidities, past ED admissions) and health outcomes (likelihood of unplanned readmission, in-hospital mortality). Trauma and orthopaedic patients are on average 53 years old, with an even gender balance, and are predominantly white.

Figure 2.1: Heterogeneity between trauma and orthopaedic patients



Notes: (1) Each marker is a three-digit ICD-10 category; (2) Market size indicates the relative number of admissions; (3) Labels shown for the three largest diagnosis groups for elective and emergency patients; (4) Data extracted from the inpatient dataset in 2010 for the top tertile of elective and emergency patients when sorted by diagnosis frequency.

Differences between elective and emergency patients

There is substantial heterogeneity between different trauma and orthopaedic patients. This is particularly evident when comparing between elective and emergency cases and across the many different diagnosis types.

Figure 2.1 illustrates the heterogeneity between patients in the data for 2010. It shows the most common diagnosis groups for elective and emergency patients, plotted by average length of stay and age with the size of the marker indicating the number of patients. There are two notable difference between elective and emergency patients. First, they are located in different regions of the length of stay and age space. Elective patients typically have shorter stays and are on average between 45 and 70 years old. The most common elective diagnosis is arthrosis (commonly known as osteoarthritis) and the majority of these patients will require a hip or knee replacement. In contrast, emergency patients stay longer and the age distribution is bimodal: there is a group with an average age of around 40 presenting with broken arms and legs, and a group with an average age of around 80 presenting with broken hips. These emergency patients will often receive an 'open reduction and internal fixation' which involves open surgery and the use of metal places or screws to realign and secure a broken bone. The second difference between elective and emergency patients is the degree of heterogeneity among diagnoses: there is significantly more heterogeneity and dispersion for emergency patients. The figure contains equal volumes of elective and emergency patients, and the elective patients are concentrated within 8 diagnosis groups while the emergency patients are spread over more than 30 groups.

In some cases elective and emergency patients have a similar or the same diagnosis. Even here there are strong differences between the two patient types. In Table B.4 I show that hospital stays for emergency patients are on average 94% longer than elective patients and after controlling for the observable characteristics of patients, including the specific diagnosis, this difference is still as large as 46%.

Health outcomes

I use unplanned readmission as my primary measure of health outcomes. This is widely used in academic studies and by healthcare regulators (e.g. NHS Improvement in England, Centers for Medicare and Medicaid Services in the U.S.). Unplanned readmission is also used specifically in relation to trauma and orthopaedic patients by regulators and in medical research to evaluate orthopaedic surgery (Kehlet, 2013). Common diagnoses among readmitted trauma and orthopaedic patients include complications with internal devices (e.g. mechanical components of a hip replacement), infections, inflammation, and bleeding (see Table B.5). The average length of stay for a readmission is 7.3 days, which is approximately equal to the length of stay in the index admission of readmitted patients.

Alongside readmission, I report results using 30-day in-hospital mortality. Relative to readmission this outcome has two drawbacks: first, it is a very extreme outcome that does not occur very often in the sample; and second, I only observe mortality that occurs within the hospital, which makes this outcome conditional on other events such as admission and length of stay. Across trauma and orthopaedic patients, the 7-day unplanned readmission rate is 2.8% and the 30-day in-hospital mortality rate is 1.1%.

2.3 The impact of hospital crowding on patients

I now turn to the question of how crowding affects patients. I focus on the relationship between the number of hospital admissions each day and the health outcomes of patients. This relationship can in general be written as

$$y_{iht} = \alpha_h + \beta_{iht} q_{hs} + \varepsilon_{iht} \tag{2.1}$$

where y_{iht} is an outcome for patient i at hospital h in cohort t, α_h is a hospital fixed effect, q_{hs} is the number of admissions at hospital h on day s, and ε_{iht} is an error term.

The empirical challenge when trying to estimate the effect of q_{hs} on y_{iht} is that admissions are endogenous in the sense that $\mathbb{E}[q_{hs}\varepsilon_{iht}] \neq 0$. This is because admissions are correlated with factors contained in ε_{iht} such as patient composition and inputs to hospital production. These correlations may arise because of seasonality, where the type and volume of patients presenting varies during the year, and hospital scheduling decisions, where resources and workload are organised to match anticipated peaks and troughs in admissions.

I address this endogeneity problem by focusing on the variation in emergency admissions. I show that these admissions can be decomposed into 'expected admissions' and 'emergency shocks', where the latter is pseudo-random. The variation in admissions caused by the pseudo-random shocks changes the number of patients in the hospital (the extent of 'crowding') and, under conditions that I make explicit, is plausibly exogenous.

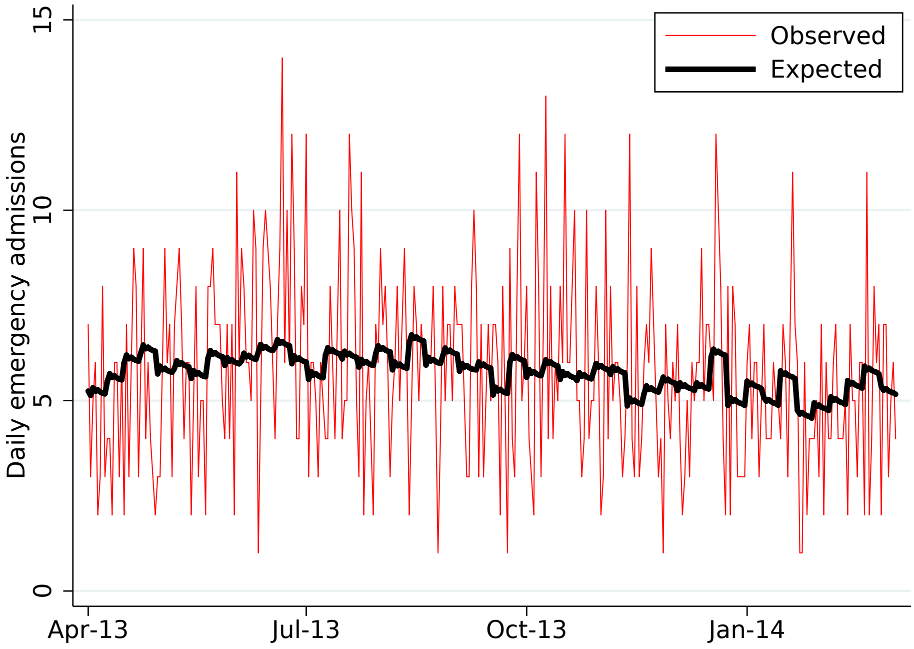
2.3.1 Pseudo-random variation in emergency admissions

I begin by decomposing emergency admissions for each hospital into seasonal and within-week components and a random shock component. The idea is that while the seasonal and within-week variation may be correlated with patient composition and hospital scheduling decisions, the shock component is exogenous to these factors. I decompose emergency admissions using the panel dataset and the following additive specification

$$q_{1,hs} = \lambda_{hy} + \phi_{hw} + \pi_{hd} + z_{hs}, \tag{2.2}$$

where $q_{1,hs}$ is the number of emergency admissions at hospital h on day s, λ_{hy}, ϕ_{hw} and π_{hd} are hospital-specific year, weekly-seasonal, and day-of-week fixed effects (which together comprise the 'expected emergency admissions'), and z_{hs} is the 'emergency shock'.

Figure 2.2: Example of the decomposition of daily emergency admissions



Notes: (1) Data shown for one hospital in one year; (2) Expected emergency admissions defined by a regression of emergency admissions on hospital-specific year, week-seasonal, and day-of-the-week fixed effects.

Figure 2.2 provides an example of the decomposition for one hospital in one year. Observed emergency admissions (red line) have a mean of around five and exhibit significant variation with low admission days and high admission days often in close succession. The expected admissions (black line) show that the seasonal pattern is slightly higher in summer and lower in winter, with minor variations across days of the week. This pattern is consistent with the causes of many trauma admissions, which involve outside activities such as road traffic accidents, slips and falls, and sports injuries. Emergency shocks are defined as the difference between the observed data and the expected admissions.

Hospital responses to expected emergency admissions

I first examine the properties of the expected emergency admissions. If expected admissions are known to the hospital then it should be apparent that hospitals respond to these predictions. I conduct two tests of these responses. I compare expected admissions with the number of senior physicians present each day (as an example of resource scheduling) and the number of elective admissions each day (workload scheduling). If hospitals are aware of the pattern in expected admissions, then this should show up as a positive correlation with the number of physicians working (more physicians are scheduled when it is expected to be busier) and a negative correlation with the number of elective appointments (fewer elective patients are admitted to moderate overall admissions). In both cases I control for hospital fixed effects and the alternative outcome (physicians or elective admits); the latter is important because hospitals may schedule fewer elective appointments in periods when there is less staff availability and this may correlate with emergency admissions (e.g. during holiday periods).

Table 2.2 presents the results of these tests. Column (1) shows that more physicians are indeed present on days with higher expected admissions. Each additional admission is associated with 0.34 additional senior physicians. Column (2) also confirms that elective admissions are negatively associated with expected emergency admissions. Each additional expected emergency admission is associated with 0.26 fewer elective admissions. These tests indicate that hospitals are at least partially aware of the seasonal pattern in emergency admissions and plan the scheduling of physicians and elective admissions around these expectations. These results highlight the importance of controlling for seasonal variations in emergency admissions.

Table 2.2: OLS regression estimates of hospital scheduling decisions on expected emergency admissions

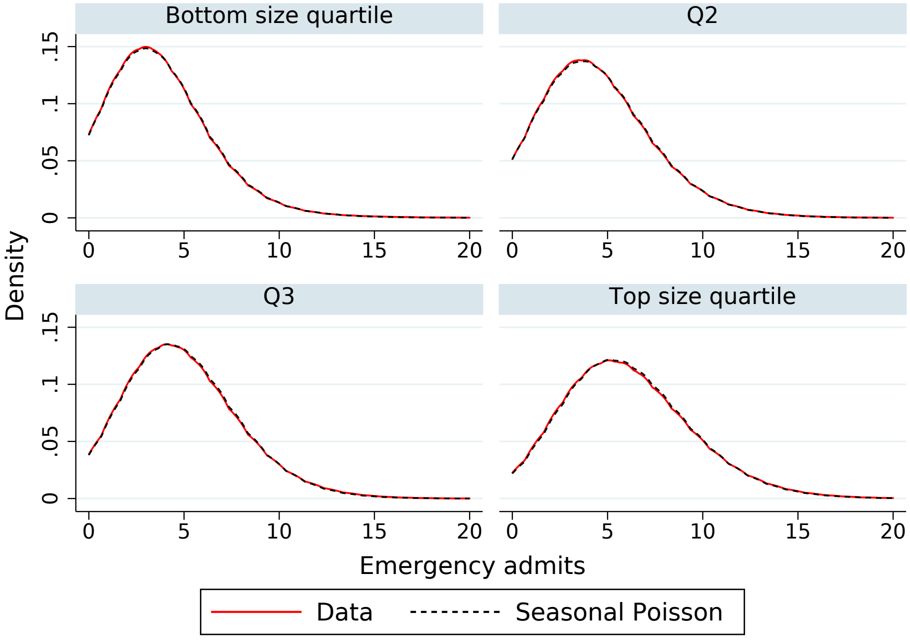
	Physician count (1)	Elective admits (2)
Expected emergency admits	0.339^{***} (0.023)	-0.261^{***} (0.081)
Hospital fixed effects	✓	✓
Elective admits	✓	
Physician count		✓
N	335,508	335,508

Notes: (1) Expected emergency admissions are defined using a regression of daily emergency admissions on hospital-specific year, weekly-seasonal, and day-of-the-week fixed effects; (2) Standard errors clustered at the hospital-level (149 clusters); (3) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

Poisson property of emergency admissions

I now consider the properties of the emergency shocks. If admissions approximate a Poisson arrivals process, it implies that the shocks are the result of independent draws from a large series of low probability Bernoulli trials. This interpretation fits intuitively with the types of accident that often cause emergency trauma admissions and if this property holds then it has a number of useful statistical implications for the analysis. I therefore examine how the observed data compares with simulated data from a Poisson distribution with a time-varying mean based on the seasonal and within-week variation. To do this I simulate a Poisson process for each hospital with a mean equal to the expected admissions from Equation (2.2). Figure 2.3 presents this simulated data alongside the observed data for hospitals grouped into quartiles based their mean daily emergency admissions. I split hospitals into these groups to check a key feature of the Poisson distribution: its mean is equal to its variance. These charts confirm that the data very

Figure 2.3: Poisson property of daily emergency admissions



Notes: (1) Size quartiles are defined by the mean daily emergency admissions of each hospital; (2) Poisson data simulated for each hospital with a mean equal to expected emergency admissions, which is defined by a regression of emergency admissions on hospital-specific year, week-seasonal, and day-of-the-week fixed effects.

closely approximates the Poisson distribution in each case.

The Poisson property has three implications for how the variation in emergency admissions caused by the shocks should be interpreted. First, the emergency shock each day is the result of a series of independent events. This rules out that the variation is being caused by large-scale events such as major road traffic accidents, terrorist attacks, and epidemics. This in turn mitigates the concern that the patients arriving on high-shock days are different to those on low-shock days. Second, the emergency shocks are independent across time, which suggests that these admissions are not being restricted or moderated in any way (i.e. admissions today are unaffected by previous levels of admissions). This mitigates selection concerns that might arise from ambulance diversion or admission decisions in the ED. Third, independence across time also means that hospitals cannot forecast the shocks. This renders sophisticated forecasting techniques redundant, since there is no short-term information contained in the shocks, and implies that hospital scheduling decisions are uncorrelated with shocks.[9]

These properties indicates that, after conditioning on expected admissions, the variation in emergency admissions is highly suitable for the empirical analysis: the realisations are pseudo-random and uncorrelated with several key factors.

2.3.2 Empirical specifications

I use the following baseline specification

$$y_{iht} = \delta_d + x_{ht}\gamma + \beta_{dxq} q_{1,hs} + u_{iht} \tag{2.3}$$

where y_{iht} is an outcome for patient i at hospital h in cohort t (described further below), δ_d is a series of fully interacted diagnosis, age category, and emergency status fixed effects (over 45,000 patient types), x_{ht} is a vector containing hospital-specific year, weekly-seasonal and day-of-the-week fixed effects, $q_{1,hs}$ is the number of emergency admissions at hospital h on day s, and $\beta_{dxq} \equiv \mathbb{E}[\beta_{iht} \mid d_i, x_{ht}, q_{1,hs}]$. This specification can be derived explicitly by substituting Equation (2.2) into Equation (2.1) and decomposing the error term into δ_d and u_{iht}.

I also use a non-parametric specification, which replaces the linear $q_{1,hs}$ term in Equation (2.3) with a series of indicators for each discrete value

[9] In Appendix B I illustrate directly that there is little or no serial correlation between hospital-specific shocks (Figure B.1) and shocks in trauma and orthopaedic departments are uncorrelated with admissions at other inpatient departments (Table B.6).

of $q_{1,hs}$. This specification allows me to examine any non-linearities in the impact of emergency admissions on outcomes.

I use these specifications to examine four groups of outcomes: health outcomes; inflows of emergency patients; inflows of elective patients; and treatment and discharge decisions in the inpatient department. I am primarily interested in the impact on health outcomes and the remaining outcomes help to understand the mechanisms behind the health impacts. Depending on the particular outcome I specify s and t accordingly, examining either admission cohorts ($t = s + 1$, where s is the admission date) or discharge cohorts ($t = s$, where s is the discharge date).

Identification

Under the assumption that $\mathbb{E}[q_{1,hs}u_{iht}] = 0$, applying OLS to Equation (2.3) will identify a weighted-average of the β_{dxq} terms (Angrist and Krueger, 1999). I refer to this as the 'average crowding effect' and it provides a summary measure of the relationship between emergency admissions and outcomes. To explore this relationship further, I later disaggregate these estimates across the distribution of patients, hospitals, and emergency admissions.

To illustrate the features of the average crowding effect, I simplify the notation and consider a case with a single hospital and no seasonality. This allows me to drop x_{ht} from the conditioning set and h from the notation. In this case, first define

$$\Delta\beta_{dp} \equiv \mathbb{E}[y_{it} \mid d_i, q_{1,s} = p] - \mathbb{E}[y_{it} \mid d_i, q_{1,s} = p - 1] \tag{2.4}$$

where $p = 0, ..., P$ are discrete value of emergency admissions. Equation

(2.4) is the difference in expected outcome for patient type d between cohorts that experience admissions p relative to $p-1$, such that $\Delta\beta_{dp}$ can be interpreted as a 'treatment effect on the treated' parameter. OLS then produces the following sample estimate,

$$\hat{\beta}^{OLS} = \frac{\sum_d \sum_{p=0}^{P}(q_{1,p} - \bar{q}_{1,d})(n_{dp}/n) \sum_{r=1}^{p} \Delta\hat{\beta}_{dr}}{\sum_d \sum_{p=0}^{P}(q_{1,p} - \bar{q}_{1,d})^2 (n_{dp}/n)} \tag{2.5}$$

where r is a summation index, n_{dp} is the sample size of patient type d and admissions p, n_d is the sample size of patient type d, n is the total sample size, and $\bar{q}_{1,d} \equiv \sum_{p=0}^{P} q_{1,p}(n_{dp}/n_d)$. I derive this result in Appendix B, and the derivation in the more general setting with β_{dxq} is similar but with a larger conditioning set. Equation (2.5) shows that OLS produces a weighted average of the treatment effect parameters across the distribution of patient types and the distribution of emergency admissions.

The identification assumption, $\mathbb{E}[q_{1,hs}u_{iht}] = 0$, states that the emergency admissions, conditional on d_i and x_{ht}, are uncorrelated with other period-specific shocks. Potential threats to this assumption include changes in patient composition, labour and capital inputs, and admissions at other inpatient departments. After conditioning on x_{ht}, many of these concerns are ruled out by the Poisson property of the emergency admissions (see Section 2.3.1).

One remaining concern is that patients arriving on high shock days may still differ to those arriving on low shock days. For example, weather shocks may mean that emergency admissions maintain the Poisson property but shift both the number and type of admissions that occur in some periods (e.g. a cold weather shock may mean fewer admissions but more slips than sports injuries). The role of the age-diagnosis-emergency fixed effects δ_d is to mitigate against this source of correlation. These fixed effects allow for over 45,000 patient types and include five digit ICD-10 diagnosis codes, which record the exact location of the injury (e.g. fracture of lower end of tibia) and severity (e.g. whether the wound is open or closed). The identification assumption therefore implies that any differences in patient composition not accounted for by the Poisson property are uncorrelated with emergency admissions within an age-diagnosis-emergency type.[10]

Estimation

I estimate the average crowding effect by regressing y_{iht} on $q_{1,hs}$ and the relevant fixed effects. Since the set of fixed effects is large (over 55,000), I implement the estimation using an algorithm by Correia (2016).[11] Standard errors are clustered at the hospital-level (149 clusters).

2.3.3 Results

I now present the results for each group of outcomes. I begin with health outcomes, followed by hospital responses, and conclude with a subgroup analysis that further explores the heterogeneity in the average crowding effects.

[10] This issue is limited to discharge cohorts. It arises here because some patients arrive (affecting q_{ht}) and are discharged on the same day (affecting y_{iht}). A change in the composition of these 'daycase' patients may therefore directly affect patient outcomes. The same issue does not arise with admission cohorts where there is a clear separation in the timing between the shock (patients arriving at $t-1$) and the outcomes (patients arriving at t). An alternative approach to this issue is to exclude the daycase patients from the analysis. However, this introduces selection concerns because hospitals may respond to emergency shocks by discharging patients as a daycase when they would otherwise have stayed overnight. Despite these concerns, the qualitative features of the results are robust to taking this approach.

[11] This is implemented in Stata using the -reghdfe- command. The same analysis can be implemented with OLS, although with extensive estimation times, and this approach produces near-identical estimates.

Table 2.3: Estimated effects of emergency admissions on health outcomes

Dependent variable	Coeff	Std error	N
Panel A: Admission cohorts			
7-day unplanned readmission, %	0.011^{***}	(0.003)	$3,940,878$
30-day in-hospital mortality, %	0.003	(0.002)	$3,940,878$
Panel B: Discharge cohorts			
7-day unplanned readmission, %	0.047^{***}	(0.007)	$3,940,878$
30-day in-hospital mortality, %	-0.001	(0.002)	$3,940,878$

Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable; (2) All specifications include a fully interacted set of diagnosis, age category, and emergency status fixed effects, and hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) Standard errors clustered at the hospital-level (149 clusters); (4) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

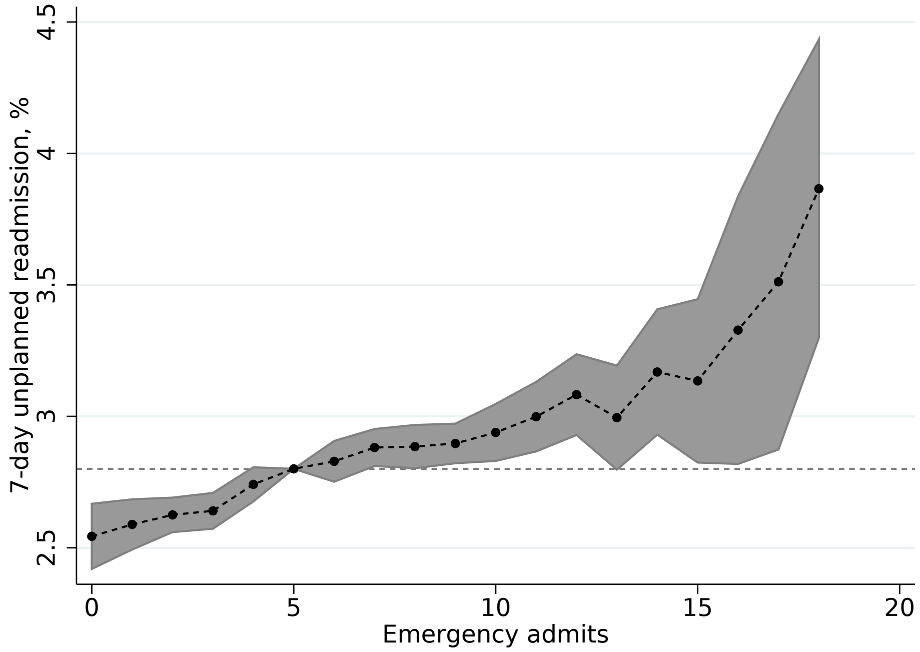
Health outcomes

Table 2.3 presents estimates of the baseline specification using the inpatient dataset for admission cohorts (Panel A) and discharge cohorts (Panel B). For both cohorts I find that emergency admissions have a positive and statistically significant impact on readmission but I find no statistically significant impacts on mortality. The readmission impacts for discharge cohorts are substantially larger than those for admission cohorts: a one standard deviation increase in emergency admissions (2.8 patients) is estimated to increase the readmission rate by 0.031 percentage points (1.1% relative to the baseline) for admission cohorts compared to 0.132 percentage points (4.7%) for discharge cohorts. The estimated impacts on mortality are very small in magnitude and precisely estimated.[12]

I further examine these readmission effects using the non-parametric specification. Figure 2.4 presents the estimates for discharge cohorts. This shows that the impacts on readmissions are near linear across the distri-

[12] Table B.7 and B.8 shows that these estimates are robust to changes in the health outcome horizon and the inclusion of additional control variables.

Figure 2.4: Non-parametric estimates of the effect of emergency admissions on 7-day unplanned readmission for discharge cohorts



Notes: (1) Base category of 5 emergency admissions normalised to the unconditional mean of emergency admissions; (2) Estimates for values of emergency admissions above 15 omitted from the figure and are mostly statistically insignificant; (3) N = 3,940,878; (4) Standard errors clustered at the hospital-level (149 clusters) with 95% confidence intervals shown in the shaded region.

bution of emergency admissions, showing that readmissions decrease when there are low realisations of emergency admissions and increase when there are high realisations. Across the distribution of emergency admissions the readmission rate varies between 2.54 and 3.17 (with a mean of 2.80). The implication is that the likelihood of readmission varies by 22% depending on the day of discharge. The equivalent non-parametric estimates for admission cohorts are small in magnitude and generally statistically insignificant.

Three features of these results stand out. First, the measurable impact of emergency admissions on health outcomes is on readmission and not mortality. Second, the impact of emergency admissions is felt primarily by discharge cohorts, where the impact is almost five times larger than for admission cohorts. Third, even small variations in the number of daily admissions have material effects on readmission, suggesting that hospitals are making daily adjustments that affect health outcomes. I now examine several hospital responses to emergency admissions that might be causing these variations in readmission.

Inflows of emergency patients

Table 2.4 presents estimates of the baseline specification using the ED dataset, where I focus on visits to the ED that occur the day after the emergency admissions and examine three outcomes: the likelihood a patient attends their nearest ED; time spent in the ED; and the likelihood of inpatient admission. For ED attendance, I use the number of emergency admissions at the nearest ED which allows me to test whether hospital choice is affected by the how busy trauma and orthopaedic departments are. The estimates show that emergency admissions have no statistically significant impact on hospital choice or inpatient admission but do have a small and statistically significant impact on time spent in the ED. A one standard deviation increase in emergency admissions (2.8 patients) is estimated to increase time spent in the ED by 47 seconds on average.[13]

These results show that emergency admissions have only very limited impacts on patients in the ED. This is consistent with the Poisson property of emergency admissions, which suggested that emergency admissions are independent across time. The implication is that the inflow of emergency

[13] I explore these results further in Table B.9 and show that the estimates are similar for various subgroups of patients (e.g. those more likely to be admitted to a trauma and orthopaedic department, ambulance arrivals).

Table 2.4: Estimated effects of emergency admissions on inflows of emergency patients

Dependent variable	Coeff	Std error	N
Attended nearest ED, %	0.008	(0.010)	$22,519,392$
Time spent in the ED, mins	0.281***	(0.064)	$22,519,392$
Inpatient admission, %	0.014	(0.013)	$22,519,392$

Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable (time spent in ED, inpatient admission) or the daily emergency admissions at the nearest hospital (attended nearest ED) on the day prior to the inflows of emergency patients; (2) All specifications include a fully interacted set of diagnosis, age category, and ambulance arrival fixed effects, and hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) The nearest ED is defined according to straight-line distances from the patient's home to the set of general acute hospitals in the panel dataset; (4) Standard errors clustered at the hospital-level (149 clusters); (5) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

patients to inpatient departments is not moderated in the ED, and emergency patients arriving at the inpatient department are not subject to any pre-selection by the hospital.[14]

Inflows of elective patients

Table 2.5 presents estimates using the panel dataset with daily elective admissions as the dependent variable.[15] In column (1) I report estimates of a specification containing contemporaneous emergency admissions. The estimates show that emergency admissions have a negative and statistically significant effect on the number of elective admissions. In column (2), I use the same specification but include additional lags of emergency admissions. These estimates show that, as well as the contemporaneous effect, emergency admissions lead to fewer elective admissions over the next four days. In

[14] One surprising finding here is that ambulance diversion is not used to insure hospitals against the volatility in emergency admissions. This may be possible in some cases (e.g. patients without time sensitive injuries and when the spatial correlation between shocks is low).

[15] As this variable is measured at the hospital-day level I omit the patient-level fixed effects from the baseline specification.

Table 2.5: Estimated effects of emergency admissions on inflows of elective patients

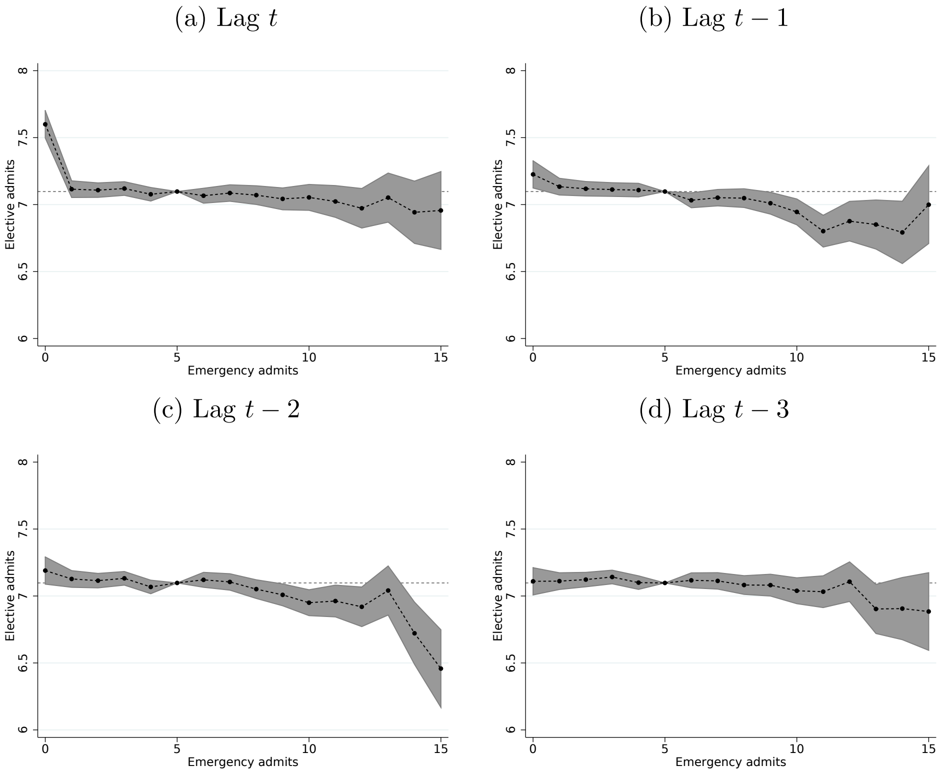
	(1)		(2)	
	Coeff	Std error	Coeff	Std error
Emergency admits, t	-0.013^{***}	(0.003)	-0.014^{***}	(0.003)
Emergency admits, $t-1$			-0.022^{***}	(0.003)
Emergency admits, $t-2$			-0.017^{***}	(0.003)
Emergency admits, $t-3$			-0.009^{***}	(0.003)
Emergency admits, $t-4$			-0.009^{***}	(0.003)
Emergency admits, $t-5$			-0.004	(0.003)
Emergency admits, $t-6$			-0.002	(0.003)
N	338,746		321,481	

Notes: (1) Dependent variable is the number of daily elective admissions at a hospital on day t; (2) All specifications include hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

subsequent days the effects become smaller and statistically insignificant.

Figure 2.5 presents estimates of the non-parametric specification for elective admissions where I include separate terms for each lag of emergency admissions (up to three lags). The estimates show that the impacts on elective admissions are modest, occur in response to extreme realisations of emergency admissions, and vary with the time horizon. For example, a high realisation of emergency admissions today has no impact on elective admissions today but decreases elective admissions for the next two days. This compares to a low realisation today which increases elective admissions today but not thereafter. With respect to magnitude, an increase of 5 emergency admissions today (approximately 2 standard deviations) relative to the mean number of admissions is estimated to cumulatively decrease elective admissions by 0.3 (0.15 fewer tomorrow and the same reduction the day after), and an equivalent decrease of 5 emergency admissions today is

Figure 2.5: Non-parametric estimates of the effect of emergency admissions on inflows of elective patients


(a) Lag t

(b) Lag $t-1$

(c) Lag $t-2$

(d) Lag $t-3$

Notes: (1) Base category of 5 emergency admissions normalised to the unconditional mean of emergency admissions; (2) Estimates for values of emergency admissions above 15 omitted from the figure and are mostly statistically insignificant; (3) N = 321,481; (4) 95% confidence intervals shown in the shaded region.

estimated to increase elective admissions by 0.5 today.

In Table B.10 I explore whether these effects on elective admissions create selection among admitted patients. I find no evidence of selection or effects of a negligible magnitude across a range of observable patient characteristics. For example, a one standard deviation increase in emergency admissions increases the observed average waiting time of admitted elective patients by less than 1 day (0.2% of the baseline).

These results are consistent with high realisations of emergency admissions causing cancellations of elective appointments and cancelled patients being rescheduled at short notice on days when low realisations occur. This cancellation mechanism moderates the total number of admissions and operates by shuffling patients between the extreme tails of the emergency admissions distribution. However, the effects are very small and do not appear to introduce any substantial selection to the pool of admitted elective patients.[16]

Inpatient care

The analysis of inflows of emergency and elective patients indicates that the readmission effects are not being caused by selection. This leaves treatment decisions in the inpatient department as a potential explanation. Table 2.6 presents estimates of the baseline specification for several aspects of inpatient care. For admission cohorts I examine: the likelihood of receiving a daycase operation; the likelihood of having a delayed operation; and the number of procedures. For discharge cohorts I examine: length of stay (logged); the likelihood of being transferred to another hospital; and the likelihood of being discharged to home.[17] The results show that increases in emergency admissions have three statistically significant effects on inpatient care. First, there are more delays, with a higher proportion of patients waiting at least a day before receiving their primary operation. Second, patients receive fewer procedures. Third, patients have shorter hospital stays, which implies that physicians are discharging patients earlier. I find no statistically

[16] A comparison between daily elective admissions and a simulated Poisson process confirms that, unlike emergency admissions, the elective admissions are not pseudo-random.

[17] To incorporate patients with a length of stay of zero, I use log(length of stay+1). The results are similar if instead the specification uses length of stay as a linear variable.

Table 2.6: Estimated effects of emergency admissions on inpatient care

Dependent variable	Coeff	Std error	N
Panel A: Admission cohorts			
Daycase operation, %	0.018^{*}	(0.011)	$3,940,878$
Delayed operation, %	0.185^{***}	(0.013)	$3,940,878$
Number of procedures	-0.001^{***}	(0.000)	$3,940,878$
Panel B: Discharge cohorts			
Length of stay (log)	-0.009^{***}	(0.000)	$3,940,878$
Transfers to other hospitals, %	0.001	(0.005)	$3,940,878$
Discharges to home, %	0.001	(0.009)	$3,940,878$

Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable; (2) All specifications include a fully interacted set of diagnosis, age category, and emergency status fixed effects, and hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) Standard errors clustered at the hospital-level (149 clusters); (4) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

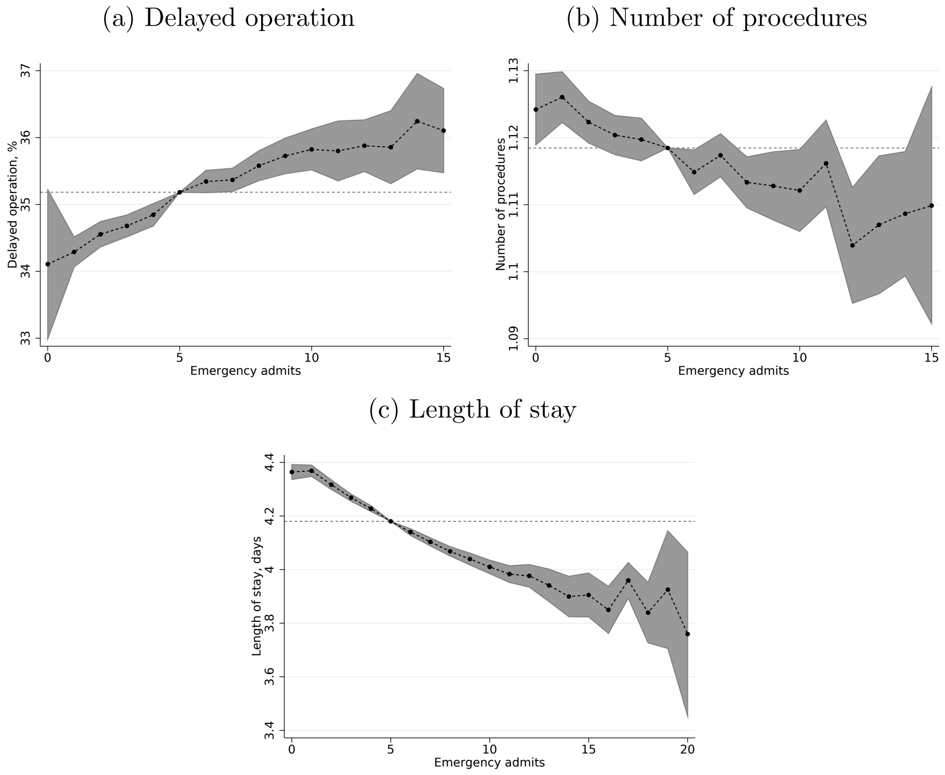
significant evidence for other outcomes.[18]

Figure 2.6 presents estimates of the non-parametric specification for delays, procedures and length of stay. Across the distribution of emergency admissions, delays and procedures vary by up to 5.7% and 1.3% (both affecting admission cohorts) and length of stay varies by 11.0% (affecting discharge cohorts). The comparatively large impact on length of stay is especially notable and interesting to understand further since this affects discharge cohorts which is where the majority of the readmission impacts were identified earlier.

There are two potential explanations for the discharge effect. First, as physicians become busier and deal with more patients, this may result in fewer checks or tests and, in turn, more mistakes when discharging patients (a 'staff constraints' effect). The second possibility is that as the hospital becomes busier physicians are required to lower the threshold at which they

[18] Table B.8 shows that these results are robust to the inclusion of additional control variables.

Figure 2.6: Non-parametric estimates of the effect of emergency admissions on inpatient care



(a) Delayed operation

(b) Number of procedures

(c) Length of stay

Notes: (1) Admission cohorts shown in Panel (a) and (b) and discharge cohorts shown in Panel (c); (2) Base category of 5 emergency admissions normalised to the unconditional mean of emergency admissions; (3) Estimates for values of emergency admissions above 15 omitted from the figure and are mostly statistically insignificant; (4) N = 3,940,878; (5) Standard errors clustered at the hospital-level (149 clusters) with 95% confidence intervals shown in the shaded region.

discharge patients to free up space for newly arriving patients (a 'bed constraints' effect). The staff constraints explanation suggests the effects would be more prominent for positive realisations of emergency admissions while the bed constraints explanation suggests that the effects would feed through linearly to length of stay as new arrivals imply increases or decreases in bed availability. The non-parametric estimates, which are approximately linear, indicate that bed constraints are the most likely explanation.[19]

Another test of the bed constraints hypothesis is to compare the estimated length of stay effect with the implied magnitude of effect that would occur if each new emergency admissions led to one existing patient being discharged. To do this I first rescale the estimated length of stay effect to give the impact of 1 emergency admission that requires a bed, which I do by excluding the proportion of emergency admissions that did not require a bed. This gives a length of stay impact of 0.75%. I then compute the number of patients present in hospital on average across hospital-days, and calculate the length of stay impact that would be observed if one of these patients was discharged a day early relative to the mean. This gives an implied length of stay impact of 0.63%. This is only very marginally outside the lower 95% confidence interval of the estimated impact (0.65%) suggesting that the bed constraints explanation is plausible.

Together these results indicate that the first-order impact of emergency admissions on hospital treatment is to cause physicians to discharge patients earlier from the inpatient department, and bed constraints are a plausible reason for this happening. The reductions in length of stay that this causes are felt by discharge cohorts, which were the same cohorts shown earlier to suffer the larger readmission impacts.

[19] The bed constraints explanation implies that new arrivals effectively 'push out' recovering patients. This would be self-reinforcing if it leads to emergency readmissions that then crowd the hospital in future periods. To assess this I estimate the impact of emergency admissions on total length of stay over a 90-day period (logged). A one standard deviation increase in emergency admissions is estimated to decrease total bed-days by 0.007% (0.45 days). This indicates that discharging patients early is not self-reinforcing: it may create some readmissions but the net effect is to decrease total bed-days.

Subgroup analysis

I now explore the heterogeneity in the data and probe two issues: first, the link between the length of stay and readmission effects (mechanisms); and second, the hospital characteristics and behaviour that may help explain the crowding effects (heterogeneity). I focus throughout on discharge cohorts, which is where emergency admissions have the largest impacts.

Mechanisms. To explore the readmission mechanism further I evaluate how the average crowding effects for length of stay and readmission correlate across groups of patients and hospitals. Table 2.7 presents the results split by elective and emergency patients. These estimates show that, across all of the outcomes, the impacts on elective patients are more muted than for emergency patients. Focussing specifically on the length of stay and readmission impacts, there is no impact on length of stay for elective patients and only a very small impact on readmission, while for emergency patients there are substantial impacts on both outcomes. These estimates support the notion that the changes in length of stay are a causal factor in readmission. The difference in the length of stay impacts also suggests that physicians do not routinely substitute beds between the two patient types.[20]

In Table B.11 I present an equivalent analysis that splits patients by expected mortality risk. This also supports length of stay being a causal factor in readmission: the effects on length of stay and readmission are larger in magnitude for low risk patients compared to high risk patients. This is consistent with physicians rationing care according to clinical need or simply being unable to discharge high risk patients at short notice (e.g. because they may require more discharge planning).[21]

[20] I show in Figure B.2 that the distinction in the length of stay effect between elective and emergency patients is consistent across hospitals.

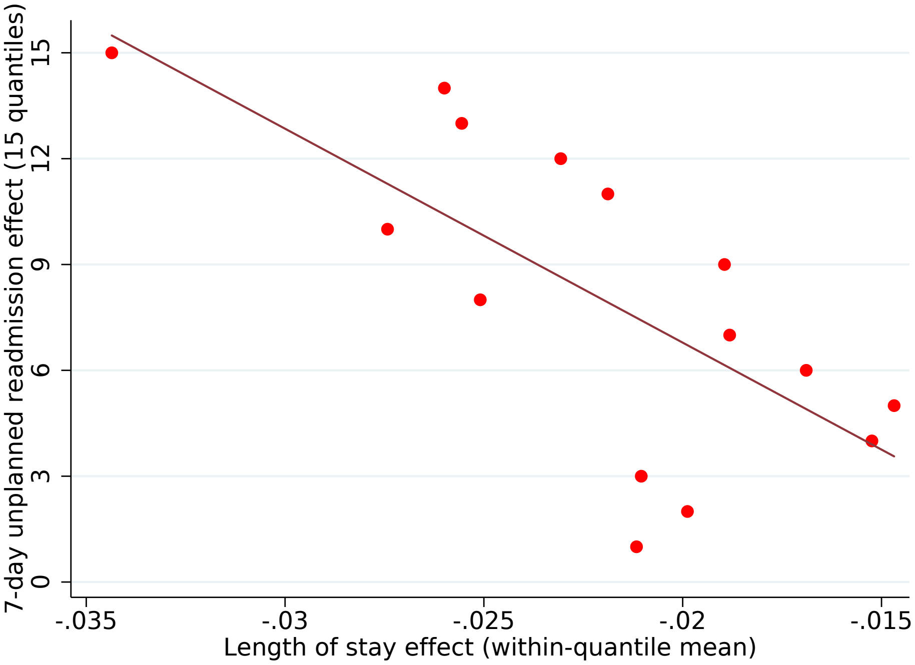
[21] I do not analyse the elective patients split by expected mortality risk since almost all

Table 2.7: Estimated effects of emergency admissions by patient type

	Electives		Emergencies	
Dependent variable	Coeff	Std err	Coeff	Std err
Panel A: Admission cohorts				
Daycase operation, %	0.004	(0.011)	−0.001	(0.006)
Delayed operation, %	0.056***	(0.014)	0.369***	(0.020)
Number of procedures	−0.001***	(0.000)	−0.002***	(0.000)
7-day unplanned readmission, %	0.004	(0.003)	0.020***	(0.007)
30-day in-hospital mortality, %	0.001	(0.001)	0.004	(0.005)
Panel B: Discharge cohorts				
Length of stay (log)	0.000	(0.000)	−0.020***	(0.001)
Transfers to other hospitals, %	−0.002	(0.003)	0.006	(0.010)
Discharges to home, %	−0.006	(0.007)	0.008	(0.014)
7-day unplanned readmission, %	0.008***	(0.003)	0.100***	(0.014)
30-day in-hospital mortality, %	0.001	(0.001)	−0.004	(0.005)

Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable; (2) All specifications include a fully interacted set of diagnosis, age category, and emergency status fixed effects, and hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) N = 2,387,641 and 1,553,237 for elective and emergency patients; (4) Standard errors clustered at the hospital-level (149 clusters); (5) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

Figure 2.7: Binned scatter plot of hospital-level average crowding effects for 7-day unplanned readmission and length of stay for emergency patients



Notes: (1) Estimates of the baseline specification for length of stay and 7-day unplanned readmission conducted separately for each hospital in the sample; (2) Each point on the plot represents approximately 10 hospitals, where hospitals are grouped into 15 quantiles according to the magnitude of the hospital-level readmission effects.

Finally, I show that the correlation between the length of stay and readmission effects is even more pronounced if the analysis is segmented by hospital. Figure 2.7 presents hospital-level estimates of the length of stay and readmission effects for emergency patients, grouping hospitals together by the magnitude of readmission effects. This shows that hospitals which make greater reductions in length of stay in response to emergency admissions are the same hospitals that exhibit greater increases in readmissions.

These results indicate that bed constraints and changes in length of stay are a plausible mechanism behind the impacts on readmission. An

elective patients have zero mortality risk.

implication of this mechanism is that additional admissions, whether they are elective or emergency patients, will lead to similar impacts on outcomes because what matters is the number of admissions rather than the type. This implication is useful for the welfare analysis that follows below.

Heterogeneity. I now consider what explains the variation in the average crowding effects. The sample size for analysing this issue in the cross-section is limited, not only by the number of hospitals but also by the data on hospital characteristics which is often available only for groups of hospitals. I therefore analyse heterogeneity by examining how the average crowding effects have changed across time and within hospital. To do this I estimate average crowding effects at the hospital-year level and focus on the length of stay impacts since these are estimated more precisely than the readmission impacts.

Table 2.8 presents regression estimates of the average crowding effects for length of stay at the hospital-year level on hospital characteristics, where all variables are standardised. Column (1) shows that the average crowding effects are larger in magnitude (more negative) for hospital-years with fewer emergency admissions and more elective admissions. Column (2) includes hospital fixed effects and shows that the same relationship holds when using only the within-hospital variation. Column (3) includes the total number of hospital beds in the regression and shows that larger hospitals have smaller average crowding effects. Column (4) includes the relative size of the trauma and orthopaedic department, measured by the proportion of total bed-days taken up by trauma and orthopaedic patients. This shows that hospitals with larger departments have smaller average crowding effects, and that after controlling for department size, the number of emergency admissions is no longer statistically significant.

Table 2.8: OLS regression estimates of hospital-year average crowding effects for length of stay on hospital characteristics

	(1)	(2)	(3)	(4)
Mean emergency admits	0.365^{***}	0.363^{***}	0.317^{**}	0.058
	(0.052)	(0.128)	(0.131)	(0.125)
Mean elective admits	-0.079^{**}	-0.365^{***}	-0.291^{**}	-0.361^{***}
	(0.038)	(0.129)	(0.137)	(0.107)
Total acute hospital beds			0.287^{**}	0.446^{***}
			(0.116)	(0.12)
Size of T&O department				0.652^{***}
				(0.121)
Hospital fixed effects		✓	✓	✓
N	960	960	937	937

Notes: (1) Dependent variable is the estimated effect of emergency admissions on the length of stay of emergency patients based on the baseline specification for each hospital-year; (2) All variables are standardised by their standard deviation; (3) Total acute hospital beds is counted at the hospital-level; (4) Size of T&O department estimated using the proportion of total bed-days at the hospital taken up by trauma and orthopaedic patients; (5) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

This analysis shows that the average crowding effects are larger at smaller hospitals, smaller trauma and orthopaedic departments, and departments that admit more elective patients. The correlation with hospital and department size is intuitive, since smaller hospitals with smaller departments may face tighter bed constraints and have less ability to substitute resources with other departments. The correlation with the volume of elective admissions is also intuitive: admitting greater volumes of elective patients leaves fewer beds available to accommodate the uncertain volumes of emergency admissions. The role of elective admissions in causing crowding is especially important and is taken up in the next section.

2.4 Marginal welfare analysis

The results to now show that hospital crowding has adverse effects on patients, mostly notably through increases in readmissions. I now turn to the question of whether it would be desirable to reduce hospital crowding. I take the perspective of consumer welfare and consider policies that ration elective admissions. This will reduce hospital crowding and create benefits for patients as there will be fewer readmissions and other adverse events (delays, cancellations). But it also creates costs for patients. By decreasing the rate at which elective patients are admitted, the waiting list and time spent waiting for elective appointments will increase. Rationing policies therefore trade off the effects of crowding with waiting times for elective appointments.

I analyse this trade-off using a model of consumer welfare where a regulator sets the incentives for hospitals to admit elective admissions. I use this model to derive an optimal crowding condition that I then test empirically. To implement the test I use the crowding estimates from earlier in the chapter, combined with estimates how of equilibrium waiting times respond to changes in elective admissions, and benchmarks from the literature of preferences for waiting times and readmissions. The test allows me to evaluate whether the current rationing incentives are optimal from the perspective of consumer welfare, and whether marginal changes in elective admissions would be welfare improving.

2.4.1 A model of consumer welfare

I begin by setting out a model of consumer welfare with the following features. I consider consumers that have exogenous demand for elective and emergency care. The utility they receive from attending hospital includes a health outcome (a benefit) and the waiting time before receiving care (a cost). I focus on health outcomes purely in terms of the likelihood of readmission. Hospital technology is such that the volume of elective admissions affect patients' health outcomes (a crowding effect) and the waiting time (a queuing effect). These two hospital characteristics are in direct conflict because increasing the flow of admissions will increase crowding but reduce queues. I do not explicitly specify hospital incentives but I assume that hospitals do not fully internalise the costs for patients of waiting for care. This provides an economic rationale for policymakers to regulate elective admissions. A regulator is assumed to set policies that determine elective admissions with an objective of maximising consumer welfare. The solution to the regulator's problem gives an optimal crowding condition that characterises the trade-off between crowding and waiting times.

Each of these features of the model is now described in more detail.

Consumer preferences

There are N consumers and consumer i demands inpatient care at a general acute public hospital with probability ρ_{ei}, where $e = 0$ for elective care and $e = 1$ for emergency care. The probabilities are independent across patients.[22] Consumer utility from receiving inpatient care is a function of the hospital characteristics which include the likelihood of readmission r_e and a waiting time of w days. I assume that $w = 0$ for all emergency patients. Utility in each state of inpatient care can be written as

$$u_{0i} = -\theta r_0 - \psi_i w, \tag{2.6}$$

$$u_{1i} = -\theta r_1, \tag{2.7}$$

[22] In a dynamic model these assumptions would imply Poisson demand as $N \to \infty$.

where θ is the utility cost of a readmission (e.g. a second visit to hospital, any additional days in hospital, any impacts of readmission on well-being or health outcomes) and ψ_i is the utility cost of waiting a day for elective care (e.g. impaired mobility, any impacts of waiting on well-being or health outcomes). Utility when no care is demanded is normalised to zero.

Hospital technology

I model hospital technology using reduced-form functions for the likelihood of readmission and the equilibrium waiting times that reflect the crowding and queuing mechanisms. These functions can be thought of as representing a single hospital or a group of hospitals.

Likelihood of readmission. This outcome is assumed to be affected by elective admission through a crowding mechanism. I assume that the likelihood of readmission is increasing and convex in admissions. Let q_0 be the number of elective admissions and denote the readmission odds for elective and emergency patients as $r_e(q_0)$ with the properties $r_e'(q_0) > 0$ and $r_e''(q_0) > 0$. This function does not specify the cause of the crowding effects but, as shown earlier, binding bed constraints are one explanation. It is on this basis that I implicitly assume elective and emergency admissions have the same impact on readmission outcomes since what matters for the crowding effect is the volume of admissions rather than the type of admissions. The earlier empirical results also showed that the impact of admissions on the likelihood of readmission was approximately linear; the assumption here is that this relationship would be convex over a wider range of values for q_0.[23]

[23] The crowding functions could also incorporate the effect of emergency admissions, which will also have crowding effects, but I omit this from the notation as it is not central to the model.

Waiting times. Equilibrium waiting times are assumed to be affected by the number of elective admissions through a queuing mechanism. To specify the properties of the equilibrium waiting time function I draw on intuition from queuing theory.

Consider the following standard 'M/M/c' queuing model. This model specifies that patient arrivals follow a Poisson distribution with mean λ, length of stay follows an exponential distribution with mean $1/\mu$, and c beds are devoted to these patients. Waiting times arise in this model because of short-term mismatches between supply (available beds) and demand (patient arrivals). If $c > \lambda/\mu$ then in equilibrium the expected waiting time in this model is finite and weakly decreasing and convex in c. This is intuitive: increasing the number of beds means more patients can be admitted at once, and this increase in flow will cause the equilibrium queue length and expected waiting time to decrease, but the benefits to additional beds will diminish as the queue tends to zero.

Another feature of this queuing model is that queues will become infinite if $c \leq \lambda/\mu$, at which point demand exceeds supply in the long-term. This property limits the range of beds over which this particular model predicts variation in equilibrium waiting times; too few beds results in infinite queues, and too many beds result in no queues. An extension to the model that avoids the infinite queue outcome incorporates 'baulking': patients may decide not to join the queue if the expected waiting time is above some threshold. This demand property naturally moderates arrivals as the queue grows longer and mean that the range of beds is less restricted.

Based on the intuition from this queuing model, I specify equilibrium waiting times as a function $w(q_0)$ with the properties $w'(q_0) \leq 0$ and $w''(q_0) \geq 0$. This approach is more tractable than explicitly specifying a queuing model, especially when it comes to empirically estimating responses in equilibrium waiting times.[24] The function $w(q_0)$ can be interpreted as an aggregation over time periods in a dynamic queuing model, where the number of admissions q_0 is the outcome of the number of beds made available and the length of stay. Baulking in my context can be thought of as elective patients opting for care in the private healthcare sector rather than at a public hospital. This can be accommodated in the model by allowing for a negative correlation between p_{0i} (likelihood of attending a public hospital for elective care) and ψ_i (preferences for waiting times).

Regulator behaviour

A regulator is assumed to maximise consumer welfare by setting a lower bound on elective admissions for hospitals. The economic rationale for the regulator is that the hospital does not fully internalise the effect of waiting times, and because of this the lower bound imposed by the regulator will always bind. The regulator problem is therefore equivalent to setting elective admissions and can be derived as follows.

After including the hospital technology functions the expected utility of consumer i conditional on elective admissions can be written

$$\mathbb{E}[u_i \mid q_0] = \mathbb{E}\Big[-\rho_{0i}\big(\theta r_0(q_0) + \psi_i w(q_0)\big) - \rho_{1i}\theta r_1(q_0) \mid q_0 \Big]. \qquad (2.8)$$

[24] In particular because the full queuing process at hospitals is more complex than an M/M/c queue and several features of the queue are not observable in the data (e.g. arrivals, prioritisation, multiple queues, intermediate queues).

Weighting consumers equally, consumer welfare can then be written as

$$U = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[u_i \mid q_0] = -\rho_0\big(\theta r_0(q_0) + \psi w(q_0)\big) - \rho_1\theta r_1(q_0) \\ - \operatorname{cov}(\rho_{0i}, \psi_i)w(q_0), \tag{2.9}$$

where ρ_e and ψ are population averages. The covariance term arises when the demand probabilities and preferences for waiting are not independent and will be negative if patients that baulk at the queues (low p_{0i}) are those with high preferences for waiting times (high ψ_i). The regulator problem is therefore

$$\max_{q_0} \quad - \rho_0\big(\theta r_0(q_0) + \psi w(q_0)\big) - \rho_1\theta r_1(q_0) - \operatorname{cov}(\rho_{0i}, \psi_i)w(q_0), \tag{2.10}$$

which has the following first order condition

$$\underbrace{-\frac{\rho_0}{\rho_1}\frac{w'(q_0)}{r_1'(q_0)}}_{\text{Crowding ratio} \equiv C} = \underbrace{\frac{\theta}{\psi}}_{\text{Preference ratio} \equiv P} + \frac{\operatorname{cov}(\rho_{0i}, \psi_i)w'(q_0) + \rho_0\theta r_0'(q_0)}{\psi\rho_1 r_1'(q_0)}. \tag{2.11}$$

This is the optimal crowding condition. On the left-hand side is a term that reflects the marginal impact of elective admissions on the likelihood of waiting a day for elective surgery relative to the likelihood of experiencing a readmission as an emergency patient. Ignoring readmission outcomes for elective patients, this can be interpreted as the marginal rate of technical substitution between expected readmissions and expected waiting times. I refer to this term as the 'crowding ratio', C. The first term on the right-hand side is the relative preferences for readmission and waiting time and I refer to this as the 'preference ratio', P. Optimal crowding sets incentives such that the crowding ratio equals the preference ratio plus an additional term, which is a function of the covariance in preferences and the three marginal effects.

2.4.2 Testing the optimal crowding condition

I now test whether the optimal crowding condition holds empirically. There are several reasons why this may not be the case. For example, the regulator may have imperfect knowledge of preferences or hospital technology or it may be that the policy environment is determined through an imperfect political process. The test I conduct both establishes whether the optimal crowding condition holds and indicates whether there would be gains in consumer welfare from marginal increases or decreases in the level of elective admissions.

Implementing the test

To implement the test I estimate both sides of Equation (2.11) and test the hypothesis that these are equal. Obtaining estimates for some inputs to Equation (2.11) is straightforward. In particular, on the assumption that the impact of an elective and emergency admission is similar, the earlier empirical results provide approximations to the crowding functions (i.e. $r'_e(q_0) \cong r'_e(q_1)$). This assumption follows directly if bed constraints and length of stay are the primary mechanism behind the readmission impacts since it is only the volume of patients that matters and not their type. The crowding estimates therefore show that $r'_0(q_0) \cong 0$ and $r'_1(q_0) > 0$ (see Table 2.7). Estimates of ρ_0 and ρ_1 are also easy to obtain from national population statistics and the HES data. I estimate that $\rho_0/\rho_1 = 1.97$ which indicates that the average patient is twice as likely to demand elective care than emergency care at a general acute public hospital.

Obtaining estimates for the other inputs to Equation (2.11) requires additional work. I directly estimate the marginal impact of elective admissions on waiting times, $w'(q_0)$, using the HES data and I describe this analysis in the next section. With this I can then construct estimates of the crowding ratio C. To construct estimates of the preference ratio P, I use benchmarks for θ and ψ from external sources and the existing literature.

This leaves the covariance term. Estimating this object directly is difficult but it is possible to gain a sense of its sign and magnitude. As noted earlier it is is likely to be negative due to patients that baulk at the queue length and use private sector hospitals. It is also plausible that the magnitude of any covariance is small: around 6% of the population have a private medical insurance policy (Laing & Buisson, 2013) and over three-quarters of these policies are not purchased by individuals making an active choice but instead offered as a fringe benefit by their employer (Olivella and Vera-Hernández, 2013). On this basis I assume that $\text{cov}(\rho_{0i}, \psi_i) = 0$.

The optimal crowding condition can therefore be written as

$$C = P \tag{2.12}$$

which states that the crowding ratio should approximately equal the preference ratio. I now describe my estimates of $w'(q_0)$ which is used to construct estimates of C and the benchmarks I use to estimate P, after which I discuss the results of test.

Estimating the impact of elective admissions on waiting times

To estimate $w'(q_0)$ I require exogenous variation in elective admissions that shifts equilibrium waiting times. Revisiting the intuition from queuing theory, a particular concern here is that changes in elective admissions may be correlated with changes in the rate of arrivals which will also affect equilibrium waiting times. For example if the rate of arrivals (demand for hospital care) increases then hospitals may respond by increasing elective admissions. This type of response would bias OLS estimates of a regression of waiting times on elective admissions. Examples of these trends in the sample period include the decline of the private healthcare sector (Competition and Markets Authority, 2014) and the increasing use of private hospitals for publicly funded care (Kelly and Stoye, 2015). To mitigate these concerns I use an IV strategy and exploit a technological change that shifted elective admissions and is plausibly unrelated to demand for hospital care.

To set up the empirical specification, I first aggregate the HES data to the regional-year level and include all hospitals that treat trauma and orthopaedic patients. This includes some hospitals that were previously excluded from the analysis (e.g. private hospitals conducting publicly-funded work, specialist hospitals with no ED) and I use a regional definition corresponding to the 28 local government healthcare authorities at the start of the sample period. The estimating equation is

$$w_{rt} = \kappa_r + \tau q_{rt}^{G} + \eta q_{rt}^{NG} + v_{rt} \tag{2.13}$$

where w_{rt} is the mean waiting time in days for elective surgery at public hospitals in region r during year t, κ_r are regional fixed effects, q_{rt}^{G} is the number of elective admissions at general acute public hospitals, and q_{rt}^{NG} is the number of elective admissions at other publicly-funded hospitals (e.g. private hospitals conducting publicly-funded work, specialist public hospitals without an ED). The parameter of interest is τ, which is the (average) impact of elective admissions at general acute public hospitals on the waiting times for elective surgery at public hospitals. I control for q_{rt}^{NG} since these admissions will also impact equilibrium waiting times and may be correlated with admissions at general acute public hospitals.[25]

I use an IV strategy that exploits a technological innovation called 'fast track surgery' (FTS). FTS revolutionised post-surgical care for elective procedures and led to substantial reductions in length of stay without impairing health outcomes (Kehlet, 2013). As FTS was rolled out across general acute public hospitals throughout the sample period it led to increases in elective admissions. The validity of the FTS instrument rests on the assumption that its roll-out was uncorrelated with changes in demand for elective care at public hospitals; this will hold if the roll-out was idiosyncratic across time and FTS did not impact purchase decisions for private medical insurance. I do not observe FTS directly in the data but do observe elective length of stay which I use as a proxy for FTS. I therefore instrument for q_{rt}^{G} using mean length of stay of elective patients at general acute hospitals, l_{rt}^{G}. The identification assumption is that l_{rt}^{G} is uncorrelated with other factors that affect waiting times contained in v_{rt}.

Table 2.9 presents estimates of Equation (2.13). In the first column I present OLS estimates. These indicate that elective admissions have a negative and statistically significant impact on elective waiting times: 1,000 additional admissions in a region over a year is estimated to reduce average waiting times by around 2 days. In the second and third columns I present the first-stage and reduced-form regressions. These show that the FTS instrument is relatively strong, with an F-statistic of 8.3 in the first-stage

[25] As with the variable of interest, this admissions variable is also potentially endogenous. Using an equivalent instrument to the one I describe below, and instrumenting for both admissions variables, produces similar results.

Table 2.9: OLS and IV estimates of the effect of elective admissions on elective waiting times

	Waiting time	Elective admits	Waiting time	
	OLS	First-stage	Reduced-form	IV
GAH elective admits, 000s	-2.152^{***} (0.690)			-6.764^{***} (2.140)
Length of stay, days		-3.164^{***} (1.096)	21.404^{***} (3.892)	
Regional fixed effects	✓	✓	✓	✓
Non-GAH elective admits	✓	✓	✓	✓
N	220	220	220	220

Notes: (1) GAH = general acute hospital (i.e. those included in the panel dataset); (2) Standard errors clustered at the regional level (28 clusters); (6) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

regression, and it has a statistically significant positive impact on waiting times in the reduced-form regression. In the final column I present the IV estimates. Similar to the OLS estimates, these estimates indicate a negative and statistically significant impact of admissions on waiting times but the effect is larger in magnitude: 1,000 additional admissions is estimated to reduce elective waiting times by around 7 days. The difference between the IV and OLS estimates suggests that the OLS estimates are upward biased, which is consistent with usage of the private market being an omitted variable.[26] I use these IV estimates to help construct estimates of the crowding ratio C for use in the optimal crowding test.

Benchmarks of the preference ratio

The preference ratio is defined as $P = \theta/\psi$, where θ is the utility cost of a readmission event and ψ is the utility cost of waiting one day for elective surgery. Both of these are difficult objects to measure accurately and I therefore adopt two benchmarks, each based on a different source and set of assumptions.

The first benchmark is from Beckert and Kelly (2017). This study estimates mixed logit demand models to study the hospital choices of hip replacement patients in England during 2012. This sample overlaps substantially with my data, where hip replacement patients are a large category of orthopaedic patient. From their reported results I obtain estimates of the mean preference for waiting times and unplanned readmission. I divide these two parameters to obtain an approximate estimate of the preference ratio.

[26] In Figure B.3 I show non-parametric estimates of Equation (2.13). These show that the estimated relationship does exhibit some non-linearities, most notably indicating that waiting times are decreasing and convex in elective admissions as queuing theory would predict, but that marginal effects are negative through the sample variation and do not reach zero. As an alternative specification I also estimate Equation (2.13) using a log dependent variable and this produces similar results.

One potential source of bias in these estimates is that the readmission rate at the hospital-level, which features in the demand model, may be correlated with other dimensions of hospital quality and this could overstate the magnitude of preferences for readmissions. This gives $P_{BK} = 115$.[27]

The second benchmark is based on opportunity cost. For this I assume that readmissions causes patients to spend additional days in hospital relative to their hospital stay in the absence of a readmission. This assumption implies that readmissions have an opportunity cost, although it does not incorporate any welfare impact outside of opportunity cost (e.g. health impacts). I use the mean length of stay for readmissions as an upper bound of this potential opportunity cost and quantify it using estimates of post-tax median earnings from the national statistics authority. This gives an estimate of θ. I then take estimates of ψ from a study by Propper (1990), which used contingent valuation survey methods to estimate the willingness to pay for reductions in waiting times. Together these estimates give a second benchmark of $P_{OC} = 235$.

This gives two benchmarks of the preference ratio, $P_{BK} < P_{OC}$, to compare with my estimates of the crowding ratio. These benchmarks are approximate, both potentially overvaluing readmissions relative to waiting times.

Results

To test the optimal crowding condition all that remains is to construct estimates of the crowding ratio C using the reduced-form crowding and waiting time estimates. Doing this requires rescaling the estimates since the crowding estimates are at the hospital-day level and the waiting time

[27] This is taken from the 'Distance choice set' results in Table 5 of Beckert and Kelly (2017).

estimates are at the region-year level.[28] I rescale to a scenario where hospitals on average accommodate one additional patient per day throughout the year. The crowding estimates already satisfy this, so it only involves rescaling the waiting time estimates.[29] After rescaling, a one patient increase in hospital occupancy is estimated to decrease mean waiting times for elective patients by 5.6 days (6.6% relative to baseline) and increase the likelihood of readmission for emergency patients by 0.120 percentage points (2.3% relative to baseline).

Assembling estimates of the crowding ratio, I find that $\hat{C} = 9,267$. This clearly exceeds the two benchmarks of the preference ratio ($P_{BK} = 115, P_{OC} = 235$). The data strongly reject the hypothesis that $C = P$ which indicates that the readmission and waiting time allocations do not satisfy the optimal crowding condition.[30,31] Since $C > P$ it implies that hospitals' incentives undervalue preferences for waiting times relative to readmission. In turn this implies that policies which increase elective admissions, thereby reducing waiting times but increasing readmissions, are predicted to improve consumer welfare.

[28] In principle there are two additional scaling factors to consider. The first is the weighting in each regression. The different levels of aggregation imply differences in the weighting and it is possible to re-weight both regressions on a comparable basis (e.g. by regional population). In practice I find this makes little difference. The second factor is the potential impact on occupancy of the readmission events themselves. Since the majority of patients are not readmitted I do not incorporate this second-order effect.

[29] The waiting time estimates give the impact of 1,000 admissions in a region-year. I scale this by the average number of hospitals per region and the number of elective patients required to occupy one bed throughout the year (i.e. 365 divided by the average length of stay). For the crowding estimates I use the sum of the impacts on admission and discharge cohorts from the baseline specification.

[30] I test the hypothesis that $C = P$ by taking a log transformation and applying the Delta method, treating ρ_0/ρ_1 as fixed and assuming that $\text{cov}\big(\ln(w'(q_0)), \ln(r_1'(q_0))\big) = 0$. This gives a p-value of 0.000.

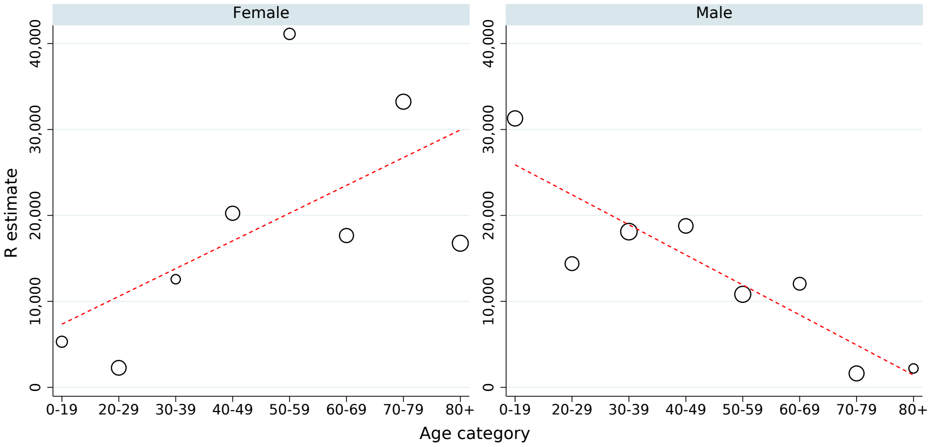
[31] The same conclusions hold if the OLS estimates for the waiting time effect are used as a lower bound. This gives $\hat{C} = 2,948$. An alternative way to do the calculation is to consider what waiting time effect would be optimal relative to the benchmarks for P. Taking P_{BK} as the benchmark, the marginal impact of elective admissions on waiting times that would be optimal is approximately 1.5% of the magnitude of my IV estimate.

It is possible to dissect the estimate of C to gain further intuition for this result. Two factors contribute to the welfare conclusion. First, while the crowding effect is a sizeable magnitude when considered in isolation, it is small relative to the reductions in waiting times that can be achieved by increases in crowding (i.e. $w'(q_0)/r_1'(q_0)$ is large). Second, the proportion of the population that affected by the adverse effects of crowding (emergency patients) is small relative to the proportion that benefit from the reductions in waiting times (elective patients) (i.e. ρ_0/ρ_1 is also large). For a randomly selected consumer prior to any hospital admission, this means that the marginal benefit from the reduction in expected waiting times exceeds the marginal cost from the increase in the likelihood of experiencing a readmission.

The results can also be further understood by disaggregating across demographic groups. Figure 2.8 presents estimates of C across the population split by gender and age categories.[32] Looking first at the left panel for females, it shows that the policies implicitly assume older females value readmissions relative to waiting times more highly than younger females, by a factor of around three. This is plausible if readmissions have worse health impacts for older patients. In comparison, the results in the right panel for males, shows the opposite result: the policies implicitly assume younger males value readmissions relative to waiting times more than older males. Even without knowledge of the true preferences of these sub-populations, the opposing gradients for males and females in the estimates of C is hard to rationalise. This suggests another potential inefficiency in the policies that regulate this trade-off: the distributional allocation of readmissions and waiting times may also be misallocated across the population.

[32] These estimates assume that $w'(q_0)$ is constant across the population.

Figure 2.8: Revealed preference estimates of C across the patient population



Notes: (1) Estimates assume $w'(q_0)$ are constant across the population; (2) Size of the market indicates the proportion of the population in each category.

There are two potential sources of bias in the optimal crowding test. The first is from the assumption that the covariance term in Equation (2.11) is negligible. If this covariance is negative and substantial in magnitude, for example because of baulking, then since $w'(q_0) < 0$ the optimal crowding condition would state that $C > P$. In this situation, finding that $C > P$ does not necessarily indicate that the optimal crowding condition is violated. The second potential bias is from crowding effects that are not observed in the data. I focus on readmission outcomes but there may be other crowding effects in the short-run (e.g. patient satisfaction) and long-run (e.g. hospital-transmitted infections). Depending on how these other crowding effects impact elective relative to emergency patients then it may bias my estimates of C upwards or downwards. Notwithstanding these issues, the magnitude of the estimated crowding ratio is such that it seems unlikely that these biases would undermine the marginal welfare conclusion.

To summarise, the estimates of C and P indicate that the optimal crowding condition does not hold empirically. This implies that hospitals are not overcrowded in the sense that further increases in elective admissions are predicted to improve consumer welfare despite the adverse effects that crowding would have on patients. These policies changes, however, will have strong distributional impacts and are predicted to disproportionately benefit older females and younger males.

2.5 Conclusion

This chapter shows that hospital crowding has adverse effects on patient health outcomes but that reducing crowding by rationing elective admissions would have negative impacts on consumer welfare. I identify the crowding effects from pseudo-random variation in emergency admissions and explore these effects using rich, linked administrative data. This shows that increases in admissions ('crowding') causes patients to be discharged sooner and readmitted more often. I evaluate hospitals' incentives to moderate crowding with rationing elective admissions in a model of consumer welfare. The data strongly rejects an optimal crowding condition I derive from the model, implying that the benefits of reducing crowding (fewer readmissions) do not outweigh the costs of increased waiting times.

These findings highlight an important trade-off that healthcare providers face when allocating capacity: by admitting more patients the hospital becomes more crowded, which has adverse effects on quality of care, but this reduces the equilibrium queue length, which improves waiting times and access to care. In the present setting, hospitals' incentives with respect to quality of care and access to care do not maximise consumer welfare. More generally this highlights that policies which target quality or access may have unintended consequences for hospitals managing this trade-off. A malpractice policy, for example, designed to safeguard quality of care may encourage hospitals to limit access to care and thereby increase waiting times; while a policy designed to reduce waiting times, in contrast, may have negative effects on quality of care because of crowding. These types of policies can therefore act as substitutes when regulating certain aspects of hospital quality and access.

Chapter 3

Efficiency Gains or Quality Cuts? How Prospective Payment Can Reduce Health Care Quality[1]

Over the past several decades, the reimbursement of healthcare providers has seen a sharpening of financial incentives. This trend, beginning with the adoption of Prospective Payment System (PPS) reforms in the U.S. in the 1980s, has quickened pace in recent years as European health care systems have adopted similar policies. PPS is a reimbursement scheme that provides zero compensation on the margin for providing care, and marked a sharp change from the previous Fee-For-Service (FFS) scheme which fully compensated providers. Alongside the widespread adoption of PPS, the U.S.

[1] I thank many healthcare professionals from the English NHS for providing me with background about the medical procedures I discuss in this chapter. This work was supported by the Economic and Social Research Council grant number ES/J500185/1. The data was made available by NHS Digital under data sharing agreement NIC-369748-C1Y0V.

and other nations have also began to impose policies that penalise providers based on performance metrics such as readmission.

Stronger financial incentives have been associated with significant improvements along several margins. It has been robustly demonstrated that PPS led to major reductions in the length of patients' hospital stays (Cutler and Zeckhauser, 2000; Dranove, 2012), and the Readmission Reduction Program in the U.S. has at least partially achieved its aims (Gupta, 2017). Yet there have been long standing concerns with PPS policies (Ellis and McGuire, 1986) and penalizing outcomes in a health care setting naturally raises multitasking concerns (Holmstrom and Milgrom, 1991).

Disentangling how these policies have impacted quality is difficult. Causal evidence is somewhat limited (Acemoglu and Finkelstein, 2008), and hospitals often respond to these policies through a variety of margins. For example, in response to readmission penalties, hospitals responded by denying future admissions (Gupta, 2017) and extending out-of-hospital care (Kocher and Adashi, 2011), although there is less evidence about their responses through in-hospital care. Even where there is clear evidence that certain margins have responded, such as with length of stay and PPS, the paucity of knowledge on hospitals' production functions makes it difficult to evaluate the impact on quality. Understanding these impacts is important for the design of future health care policy.

In this chapter I investigate a key aspect of hospital production: the effects of length of stay in inpatient departments on the likelihood of a readmission event. I illustrate first how the time spent in hospital can itself be an important input to the production process, as patients naturally heal from surgery and likelihood of a future health shock disrupting the recovery reduces. I then address an identification problem that arises because length of stay is set by doctors and is therefore correlated with unobservable patient characteristics. I deal with this by exploiting the empirical variation in emergency admissions discussed in Chapter 2. I conclude by discussing the relevance of these estimates for our understanding of PPS and readmission penalty policies. In particular, I show that the reductions in length of stay achieved by PPS may explain a sizeable fraction of the observed increases in readmissions, and that the incentives created by readmission penalties are ineffective at mitigating these increases.

Estimating the impact of length of stay on readmission requires addressing a standard identification problem in the study of production functions (Ackerberg et al., 2007). Consider a patient-level production function, where the patient's health status is an output and the inputs include length of stay and hospital treatments (e.g. surgery, nursing care). In a regression of outputs on inputs, the latter will be endogenous as health care personnel allocate inputs to patients based on a number of unobservable patient characteristics. Even for two identical admitted patients, their health status may evolve differently during a hospital stay, and this will lead doctors to allocate more resources to the patient with greater marginal benefits. This is comparable to the behaviour of a multiproduct firm that allocates inputs to products with the highest marginal profits.

To resolve the identification problem, I retain the setting of Chapter 2 and exploit variation in the number of emergency admissions each day. As shown earlier, the shocks to emergency admissions in England induce quasi-experimental variation in the length of stay of trauma patients. Here I argue that this variation is suitable for identifying the Local Average Treatment Effect (LATE) of length of stay on readmission odds. The compliers in this setting are those patients that are discharged early from hospital due to the variation in emergency admissions. These patients are the policy relevant population and the LATE estimates thus provide the appropriate input to evaluate PPS and readmission penalty policies.

I first present OLS estimates, which suggest length of stay has a statistically significant, positive, but negligible impact on readmission. This is consistent with the medical literature that has generally found only a weak correlation between readmission and length of stay (Kaboli et al., 2012; Vorhies et al., 2011). As I demonstrate in a simple economic model, the unobservable patient information that doctors use when making discharge decisions will attenuate the OLS estimates towards zero. Moving to the IV approach, the results confirm the attenuation bias, and the LATE estimates show that length of stay has a substantial negative impact on readmission odds. A one-day increase in length of stay for marginal patients is estimated to reduce the likelihood of readmission by 1.3 percentage points which, given some assumptions on the population of marginal patients, amounts to a 6.6% reduction in the number of readmissions.

To characterise the treatment group implicit in these LATE estimates, I compare the first-stage estimates across the distribution of patient severity. I measure severity using predictions from a patient-level regression of mortality on an exceptionally rich set of individual-level information in the data. I show a clear pattern in the first-stage estimates: in response to the emergency shocks, hospitals discharge low-severity patients early while the length of stay of high-severity patients is generally unchanged. By exploring the patient-level regression that generates my severity measure, I show that the compliers are younger than average patients (between 40 and 50 years old) with injuries such as fractured lower legs and forearms.

Finally, I consider the implications of these estimates for PPS and readmission penalty policies. PPS policies have been shown to reduce length of stay by up to 25% (Cutler and Zeckhauser, 2000). The estimates here suggest that reversing these increases may reduce readmissions by around 0.64 percentage points. While this estimate is for a specific patient group, and part of the reduction in length of stay following PPS is driven by technological change, the estimate is of a comparable magnitude to the change in readmissions that followed the introduction of PPS in the U.S. and the U.K. The implication is that length of stay is one channel through which PPS may have partially driven the increases in readmission, consistent with early predictions (Ellis and McGuire, 1986) and evidence (Cutler, 1995).

The implications for readmission penalties are also notable. Using the estimates, I compute the financial costs and benefits to a hospital of reducing its readmissions by increasing length of stay. Costs accrue because extending length of stay leaves less capacity available for discretionary, elective admissions. This opportunity cost is compared to the benefits that accrue through fewer readmissions penalties. I show that readmission penalties create no incentive to adjust length of stay, and in order do so penalties would need to be at least an order of magnitude higher. This highlights a drawback of the readmission penalty policy: it may induce low-cost changes that reduce observable readmissions (e.g. out-of-hospital care) but fail to induce high-cost changes that improve the quality of care during the initial hospital visit (e.g. extending length of stay).

These findings underline the importance of the hospital production function for designing and improving hospital incentive schemes. It makes two principle contributions to the existing literature. First, by characterising the length of stay mechanism, I show an important channel relevant to PPS and readmission penalty policies. These policies have been widely studied in past work, with notable examples including Cutler (1995), Acemoglu and Finkelstein (2008), Gupta (2017), and Kristensen and Sutton (2016). Cutler (1995) studies the introduction of PPS to U.S. hospitals by Medicare, finding that it led to a trend-increase in readmissions. Gupta (2017) and Kristensen and Sutton (2016) study the impact of readmission penalties in the U.S. and U.K. respectively, finding that it reduced readmissions in the former but had no impact in the latter. Neither readmission penalty study finds an impact on length of stay. The present chapter offers a convincing explanation of how PPS may have led to increases in readmissions, and why readmission penalties may have lacked impact through this causal channel.

The second literature is the study of production functions. From its origins in firm production (Ackerberg et al., 2007), this framework has proved exceptionally useful for studying a range of other settings including child development (Attanasio et al., 2012; Heckman et al., 2010) and education (Macartney et al., 2015). It is perhaps surprising that there have been fewer examples in a health care setting, and this chapter provides an example of how within-firm variation can be used to study these production functions.

The chapter proceeds as follows. Section 3.1 provides background information on the policy context and institution setting. Section 3.2 sets out a simple economic framework of discharge decisions. Section 3.4 presents the empirical specifications I use. Section 3.5 presents the empirical results. Section 3.6 discusses the policy implications of the results for PPS and readmission penalties. Section 3.7 concludes.

3.1 Background

The link between length of stay and readmission is especially relevant to two major healthcare policies present in the U.S. and a number of other countries. In this section I first describe those two policies and the related evidence, and then turn to the setting in England that I focus on for the empirical application.

3.1.1 Prospective payment systems

A Prospective Payment System (PPS) is a form of price regulation for hospitals. It specifies that hospitals receive fixed prices for hospital admissions, where the price depends on the diagnosis of the patient. PPS was first introduced for Medicare payments in U.S. hospitals in 1983 and has since been adopted in other areas of U.S. healthcare provision (e.g. psychiatric care, long-term care) as well as in other healthcare systems. In England, for example, a PPS tariff was implemented for the majority of hospital payments in 2006.

The prevalent reimbursement system in the U.S. and England prior to PPS was Fee-For-Service (FFS). FFS tariffs specify that hospitals costs are fully reimbursed. The adoption of PPS therefore marked a major change in incentives: under FFS providers were compensated on the margin for care, whereas under PPS there is zero marginal compensation. As one might expect, PPS was introduced to control growing healthcare costs owing to the cost-control incentives it places on providers.

Numerous empirical studies have assessed the impact of PPS adoption, a literature comprehensively reviewed by Cutler and Zeckhauser (2000) and Dranove (2012). One of the most consistent findings from this literature is that PPS is associated with major reductions in length of stay. As Cutler and Zeckhauser (2000) write, ‘The effect of prospective payment on hospital stays is uniformly strong and impressive; many studies find reductions of 20 to 25 percent over a period of 5 years or less. These studies provide among the clearest evidence that supply-side reimbursement changes do affect medical treatments.'

A common concern with PPS reimbursement, however, is that it may lead hospitals to discharge patients 'quicker and sicker' (Morrisey et al., 1988). This concept was formalised by Ellis and McGuire (1986), who showed unless physicians act as perfect agents for their patients, then they will have an incentive to underprovide levels of care. These concerns have only been partially borne out in empirical studies. Cutler and Zeckhauser (2000) and Dranove (2012) both conclude that, under certain conditions, PPS has been shown to impact quality negatively. Yet as Acemoglu and Finkelstein (2008) point out, much of this evidence comes from studies that rely on aggregate data and time-series comparisons. Cutler (1995) is a notable exception, and studies the roll-out of PPS to U.S. hospitals by exploiting the differential impact of PPS across hospitals. He finds that PPS led to a change in the time profile of patient mortality but did not affect the overall level of mortality, and PPS led to a trend-increase in the number of readmissions.

Alongside these length of stay and readmission impacts, Cutler and Zeckhauser (2000) and Dranove (2012) report a multitude of evidence on other responses to PPS adoption. This includes labour inputs and admissions, and Acemoglu and Finkelstein (2008) also find evidence of increased technology adoption. This combination of responses to PPS adoption make it difficult to uncover the mechanisms through which PPS impacts readmissions.

3.1.2 Readmission penalties

The impact of PPS on readmissions is related to recent policy changes that have seen 'unplanned readmission' increasingly used as a measure of quality. Unplanned readmission events ('readmissions' hereafter) are defined as occurring when a patient returns to hospital and is admitted as an emergency inpatient within a narrow window (often 30 days) after being discharged from an initial hospital admission. Under PPS, readmissions generate additional revenue for the hospital. To the extent that readmissions are a marker of poor quality of care in the initial admission, PPS therefore rewards these episodes of poor quality.

Healthcare regulators have recently endorsed readmission as a quality measure by imposing penalties based on these measures. The Centres for Medicare and Medicaid Services (CMS) in the U.S. and NHS Improvement (NHSI) in England have both implemented policies in recent years. These policies aim to withdraw PPS payments associated with readmission events for hospitals with 'high' levels of readmissions. While the CMS and NHSI policies share this common principle, they differ in how the penalty is implemented particularly with regard to how 'high' is defined.[2] The penalties can be large: the CMS policy, for example, can be up to 3 per cent of hospital revenues, and in 2015 this resulted in fines totalling $420 million.[3,4]

These policies have been controversial. Proponents argue that it acts to correct incentive problems with readmissions under PPS, and that readmissions are a proxy for quality of care in the initial visit. Critics argue that readmission is a poor quality measure and is simply a measure of resource

[2] The CMS penalty calculates the 'excess readmissions' at a given hospital relative to a national benchmark after adjusting for observable patient characteristics. For further details, see: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html. The NHSI penalty in contrast, calculates the excess of 'avoidable readmissions' at a given hospital on the basis of a clinical review of past cases at the same hospital. There are also differences in scope: the CMS penalty is imposed on a narrow list of conditions (e.g. heart failure, hip replacement), while the NHSI penalty is imposed on most admissions with only a small number of carve-outs (e.g. child birth, cancer care).

[3] https://khn.org/news/a-guide-to-medicare-readmissions-penalties-and-data/

[4] https://khn.org/news/half-of-nations-hospitals-fail-again-to-escape-medicares-readmission-penalties/

utilisation that varies with socio-demographic factors (Kangovi and Grande, 2011). Consistent with this second interpretation, studies have shown that a substantial proportion of readmits are unavoidable (Axon and Williams, 2011).

Despite numerous studies of readmissions and readmission penalties, it remains unclear which factors under the control of physicians and hospitals affect readmission. Gupta (2017), for example, studies the CMS penalty and finds that the policy reduced the readmission rate for Medicate patients by 1 percentage point (5% of the baseline). Half of the reduction is caused by hospitals denying admission to returning patients, but the mechanisms responsible for the remaining half are unclear. In contrast, Kristensen and Sutton (2016) study the NHSI policy and find that the policy had no impact on readmissions or other outcomes.

Intuitively, it seems very plausible that length of stay could drive some of the readmissions since patients that are discharged very early in their recovery process would inevitably return to hospital. Yet this does not preclude there being very little returns to length of stay on the margin, and this may explain the low correlation between length of stay and readmissions in existing studies (Kaboli et al., 2012; Vorhies et al., 2011). In this chapter, I evaluate the relationship between length of stay and readmission using quasi-experimental variation and test for a causal relationship. To the extent that there is a causal impact of length of stay on readmission, this has important implications for the design of readmission penalties, and for how we interpret the impact of PPS on quality of care.

3.1.3 Institutional and medical setting

I study the length of stay and readmission relationship in the context of the English National Health Service (NHS). A description of the English system is given in Chapter 2, and it is notable that the payment incentives include PPS from 2006 and readmission penalties from 2009.

The medical setting that I focus on is the experience of emergency patients in trauma and orthopaedic departments ('trauma patients'). These patients are typically suffering from broken bones. I choose these patients because the recovery process of the patients is simple and, in combination with empirical variation I describe below, this setting allows me to examine how doctor's discharge decisions influences readmissions.

The most common type of medical intervention for emergency trauma patients is an 'open reduction and internal fixation' (ORIF). This is highly intrusive surgery that requires exposing the broken bone, reassembling the bone pieces, and then fixing these pieces together with metal implants (e.g. screws, plates, pins or rods). The recovery process is then very gradual. It involves an extended period of not using the injured body part (often a limb), while it heals over a period of time. For leg injuries this involves bed rest. The healing process is aided by a course of pain management (e.g. opioid-based painkillers), regular wound care, and (towards the end of the process) a course of physiotherapy.

From the perspective of a doctor managing ORIF patients, the post-surgical recovery process involves monitoring and deciding when to discharge the patient. The discharge considerations include how well the wound has healed, the patient's general health, and the support available outside of the hospital (e.g. informal care at home, or any transfer to formal care that is required). Each discharge decision involves a risk assessment that the patient can cope without further hospital care.

3.2 Economic framework

I set out a simple model of discharge decisions below and use it to motivate the empirical analysis. Let the health status of an individual patient i in hospital after surgery evolve according to

$$h^*_{ir} = g_i(r), \tag{3.1}$$

where r is the number of periods since surgery and $g_i(r)$ is the patient's health production function while in hospital. The function $g_i(r)$ is assumed to have the properties $g_i'(r) > 0$ and $g_i''(r) < 0$, which reflects both the medical treatments given each period and the natural healing process over time.

For a patient i discharged in period s, let health status evolve according to

$$h^*_{i,s+t} = g_i(s) + h_i(t) + \varepsilon_{it}, \tag{3.2}$$

where t is the number of periods since discharge, $g_i(s)$ is the total health production prior to discharge, $h(t)$ is the patient's health production function outside of hospital, and ε_{it} are *iid* health shocks that occur each period after discharge. The function $h_i(t)$ is assumed to have the properties $h_i'(t) > 0$, $h_i''(t) < 0$ and $h_i(0) = 0$.

A readmission will occur after discharge at any point t if

$$g_i(s) + h_i(t) + \varepsilon_{it} \leq \kappa_i, \tag{3.3}$$

where κ_i is a threshold at which the patient feels sufficiently unwell to return to hospital for further treatment. Readmissions can thus be caused by a patient being patently discharged too soon, when $g_i(s) - \kappa_i \leq 0$, or by a substantially large adverse health shock after discharge, when $g_i(s) - \kappa_i > 0$ but $g_i(s) + h_i(t) - \kappa_i \leq -\varepsilon_{it}$.

Over a given time period T, the likelihood of a readmission is

$$\begin{aligned} \Pr_i(r_i^T = 1 \mid s_i, \kappa_i) &= \Pr_i(h^*_{i,s+0} \leq \kappa_i, h^*_{i,s+1} \leq \kappa_i, \\ &\qquad \ldots, h^*_{i,s+T} < \kappa_i \mid s_i, \kappa_i) \qquad (3.4) \\ &= \prod_{t=0}^{T} \Pr_i(h^*_{i,s+t} \leq \kappa_i \mid s_i, \kappa_i) \qquad (3.5) \\ &= \prod_{t=0}^{T} F_i(\kappa_i - g_i(s_i) - h_i(t)), \qquad (3.6) \end{aligned}$$

where the second line follows from the *iid* health shocks and $F_i(.)$ is the *cdf* of the health shocks.

I assume doctors maximise an unspecified utility function by deciding when to discharge patients. The utility function contains the costs and benefits associated with discharge decisions. Benefits could include the revenues associated with discharging patients (the option value of empty beds) while costs could include any lost health production that doctors internalise (such as through readmission penalties). The solution to the doctor problem will be to discharge patients once the probability of readmission is below a certain threshold. Specifically doctors set s for each patient such that

$$\Pr_i(r_i^T = 1 \mid s_i, \kappa_i) \leq \tau(\phi), \qquad (3.7)$$

where τ is a function of a vector of parameters ϕ that characterise doctors' preferences and the relevant incentive schemes. For example, ϕ could contain the marginal revenue associated with treatment (zero under PPS) or the marginal cost of readmissions (positive above the threshold under a readmission penalty).

It is clear from Equation (3.7) that doctors' preferences and incentive schemes, through the function $\tau(\phi)$, play a key role in shaping discharge decisions. Take the change to PPS from FFS as an increase in ϕ. This reduces the marginal benefit from retaining patients in hospital and will increase the discharge threshold, $\tau'(\phi) > 0$. By increasing the discharge threshold, doctors will take more risk by discharging patients earlier and this will increase readmissions. The link between length of stay and readmission can be seen directly by differentiating Equation (3.6) to give

$$\begin{aligned}\frac{\partial \mathrm{Pr}_i(r_i^T = 1 \mid s_i, \kappa_i)}{\partial s_i} &= -g_i'(s_i) \sum_{t=0}^{T} f_i(\kappa_i - g_i(s_i) - h_i(t)) \\ &\qquad \prod_{r \neq t} F_i(\kappa_i - g_i(s_i) - h_i(r)) \qquad (3.8) \\ &< 0, \qquad (3.9)\end{aligned}$$

where $f_i(.)$ is the *pdf* of the health shocks, and the second line follows by the assumptions on $g_i(s)$. The impact of length of stay on readmission can arise directly through $g_i(s)$, as well as through the properties of the distribution of ε_{it} if this were to depend on s.[5] While Equation (3.8) is predicted to be negative, depending on the shape of the $g_i(s)$, the magnitude of the effect at current levels of s may be close to zero.

The goal of this chapter is to test and quantify this anticipated effect of length of stay on unplanned readmission. The above model predicts that this effect will be negative. If this is the case, then it is a mechanism by which the adoption of PPS tariffs would lead to increases in readmissions,

[5] For example, the variance of ε_{it} may be decreasing in s.

and a margin by which hospitals could avoid readmissions and readmission penalties.

3.3 Data

I use administrative data on medical records for inpatient visits from the Hospital Episodes Statistics (HES). The data sample is taken from Chapter 2 and restricted to emergency patients in trauma and orthopaedic departments ('trauma patients').

Table 3.1 presents summary statistics for the sample of trauma patients. The average patient is 52 years old, 50% of patients are male, and 85% are white. Relative to the general population of hospital patients, trauma patients are generally sicker, with more diagnoses, more previous ED admissions, and a higher likelihood of readmission and death. As discussed earlier, trauma patients spend substantial time in hospital. The average length of stay around 8 days, which is around three times longer than the average hospital patient.

Table 3.1: Summary statistics

	Trauma	Other	Difference, %
Age	52.2	56.3	−7
Male, %	50.3	45.4	11
White, %	85.4	89.5	−5
Diagnosis count	4.2	3.5	21
Co-morbitidities	2.6	2.8	−8
Past ED admits	1.6	1.1	40
Length of stay	7.7	2.6	193
30-day unplanned readmission, %	9.6	8.1	19
30-day in-hospital death, %	2.6	2.4	9

Notes: (1) Trauma patients are defined as those admitted via the emergency department into trauma and orthopaedic inpatient departments; (2) Co-morbidities defined by the Charleson index; (3) Past ED admits measured over a 12 month period.

3.4 Empirical specification

Identifying the impact of length of stay on readmission is complicated by two issues. The first is that the available data includes only discharged patients and, because of the discharging process, this results in an endogeneity problem. To see this, consider a simple regression of readmission (r_i^T) on length of stay (s_i) that is identified from cross-patient heterogeneity. Here s_i will be endogenous because the discharge decision, shown in Equation (3.7), is based on individual-level factors that will be unobservable in the regression equation. In general this bias will attenuate the estimated effects towards zero. The intuition behind this result is that doctors will allow sicker patients to stay longer, inducing variation in s_i, but will discharge all patients at a similar likelihood of readmission. Observable variations in s_i are therefore not associated with variations in readmission odds.[6]

The second issue is one of interpretation since length of stay is a measure that encapsulates both time and resource use. The potential benefits of resource inputs are clear since medical treatments are specifically designed to improve health. But, holding resources constant, time itself may also be a valuable input to the health production process. Patients naturally heal while in hospital and as this happens a patient's health status increases (through $g_i(s)$ in the model) and may also become more stable upon discharge (through $F(.)$ in the model). These returns to time can reduce the

[6] Consider a simple example where $g_i(s) = \alpha_i \ln(s)$ and there are two otherwise-identical groups of patients with $\alpha_a < \alpha_b$. According to the discharge rule, type b will be discharged sooner than type a. In the data, there will be variation in s_i, with $s_a > s_b$, but no variation in readmission odds since both groups are discharged at the same readmission threshold. In this example a regression of r_i^T on s_i will return a coefficient of zero. In more realistic settings, the same problem will attenuate the estimated effect towards zero. An alternative approach to dealing with this problem might try to control for the heterogeneity between patients through observables. Yet even if this were to successfully control for patient heterogeneity, there would be no variation left in (conditional) length of stay and the effect of interest would not be identified.

risk of a health shock outside of hospital causing a readmission. In settings where the time spent in hospital coincides with intensive medical care, it can therefore be difficult to interpret how length of stay affects readmission odds since both time and resources are changing. This interpretation issue matters because the policy recommendations vary depending on whether time or resources are important.

The setting adopted here offers an ideal opportunity to address these issues. With respect to the first issue, as Chapter 2 shows, admission shocks from new arrivals in trauma and orthopaedic departments create exogenous variation in trauma patients' length of stay. This mitigates the first problem by providing a quasi-experiment: the arrival shocks cause the option value of empty beds (a component of ϕ) to vary, and this means that comparable patients are discharged at different points in their recovery process. The identification strategy essentially compares the readmission likelihood across these patients with different hospital stays.

The setting also mitigates the second factor, since emergency trauma patients have a remarkably simple production function. As explained in section 3.1.3, once surgery has been completed, the recovery process of trauma patients is largely focussed around recovery time with minimal resource usage. This provides a simple interpretation of the impact of length of stay: it is the benefit of further time spent in the hospital healing prior to discharge while holding resources (approximately) constant.

3.4.1 Baseline specification

I use the following linear specification

$$r_{iht}^{T} = \alpha_d + \beta_d s_{iht} + u_{iht} \tag{3.10}$$

where r_{iht}^T is an indicator equal to one if patient i at hospital h that is discharged on day t is readmitted within T days, α_d is a fully interacted set of age-category and diagnosis codes, s_{iht} is the length of stay of patient i at hospital h discharged on day t, and u_{iht} is an error term.

The parameter of interest is β_d which is the impact of length of stay on readmission odds, averaged across patient type d. This parameter is an approximation to the partial effect shown in Equation (3.8). I use a linear probability model rather than a non-linear model (e.g. Probit) on the basis that both offer approximations to Equation (3.8) and the linear probability model requires fewer assumptions on $g_i(s)$ and the error term.

3.4.2 Identification

In general length of stay and the error term will be correlated because of the discharge decisions of doctors. This implies $E[s_{iht}u_{iht}] \neq 0$ and precludes identification of β_d from cross-patient heterogeneity. For identification, I instead exploit variation in the emergency admissions each day. As in Chapter 2, I define emergency shocks using the following regression

$$q_{ht} = \lambda_{hy} + \phi_{hw} + \pi_{hd} + z_{ht}, \tag{3.11}$$

where $q_{1,hs}$ is the number of emergency admissions at hospital h on day t, λ_{hy}, ϕ_{hw} and π_{hd} are hospital-specific year, weekly-seasonal, and day-of-week fixed effects (which together comprise the 'expected emergency admissions'), and z_{hs} is the 'emergency shock'.

I use these emergency shocks as an instrument for length of stay in Equation (3.10). The IV estimator with heterogeneity in β_d, and under appropriate assumptions, will identify the Local Average Treatment Effect

(LATE). The LATE estimate in this setting can be interpreted as the average causal effect of length of stay on readmission for compliers, where compliers are the marginal patients that are discharged as a result of the emergency shocks. This is the specific group of patients that has policy relevance in this setting.

The identification assumptions required to obtain the LATE are that: (i) emergency shocks are random; (ii) emergency shocks are excluded from the readmission equation; and (iii) emergency shocks have a monotonic effect on length of stay.

The first of these assumptions is addressed in Chapter 2, where I show that the shocks are serially uncorrelated and approximate a Poisson process.

The second assumption is also partially addressed in Chapter 2, where I show that emergency shocks do not affect a number of margins that could affect readmission outcomes. For example, it shows that hospitals do not adjust the composition of admitted patients in response to shocks, either through through ambulance diversion or selective admission decisions in the ED. It also shows that patients are discharged to the same locations irrespective of the emergency shock.

There are, however, several margins through which hospitals could respond to emergency shocks that are not observable in the data. My identification assumption rules out responses along these margins. Examples include changes to the discharge process, such as the amount of information, drugs, or equipment provided to the patient at discharge. The responses I am ruling out include deliberate responses, such as doctors and nurses reallocating their time from discharge activities, and other responses such as mistakes being made during the discharge process.

The final assumption of monotonicity implies that increases in emergency shocks only cause patients to be discharged early and not late. This is highly plausible, since hospitals discharge patients to free up capacity in response to emergency shocks. Delaying discharge would offer no benefits to doctors trying to alleviate capacity constraints.

Under these assumptions, IV will deliver LATE estimates that reflect the average causal effect of length of stay on readmission for the marginal patients discharged by doctors.

3.4.3 Estimation

I estimate the model using standard IV techniques. I regress r_{iht}^T on α_d, s_{iht} and the fixed effects in Equation (3.11), and instrument for s_{iht} with q_{ht}. This is equivalent to computing z_{ht} from Equation (3.11) and then using this as an instrument for s_{iht} in Equation (3.10). Owing to the large number of fixed effects, I implement the IV estimator using the *reghdfe* package in Stata (Correia, 2016). Standard errors are clustered at the hospital level (149 clusters).

3.5 Results

I begin by report estimates from the baseline specification of the LATE. I then explore variation in patient severity to characterise the compliers that the treatment effects I estimate relate to.

3.5.1 Baseline estimates

Table 3.2 presents the baseline regression estimates. Column 1 and 2 report OLS estimates, and Columns 3 through 5 report the first-stage, reduced-form and IV estimates. Column 1, which is the baseline specification excluding all fixed effects and estimated by OLS, reports a positive and statistically significant impact of length of stay on readmission, but the magnitude of this effect is negligible. A one-day increase in length of stay is estimated to increase readmission odds by 0.12 percentage points (1.3% of baseline). A similar result is shown in Column 2, which is the OLS estimates of the full baseline specification. The OLS estimates therefore suggest that either there is no return to length of stay on the margin, either suggesting that hospitals are inefficient (such that they could achieve similar readmission outcomes with shorter stays) or that the OLS estimates being biased towards zero (as hypothesised in Section 3.4).

The IV estimates explore these two possibilities. Column 3 reports the first-stage regression of length of stay on emergency admissions. The estimates show that the emergency shock instrument is strong: there is a negative and statistically significant impact of emergency shocks on length of stay, and with a large F-statistic of over 200. Column 4 reports the reduced-form regression, which mirrors the key result in Chapter 2, showing that emergency shocks are associated with increases in readmissions. Finally, Column 5 presents the IV estimates which show that the impact of length of stay on readmission is negative and statistically significant. The magnitude of the estimate is large: the LATE estimate indicates that a one-day increase in length of stay across marginal patients is estimated to reduce readmission odds by 1.3 percentage points.

These IV estimates show that length of stay does have a causal impact on readmission at least for the marginal patients. The OLS estimates in contrast, which indicated no causal impact, are biased towards zero in a manner consistent with the model of discharge decisions in Section 3.2. This provides one explanation for why estimates in the medical literature have found no association with readmissions and length of stay.

Table 3.2: Estimated effects of emergency admissions on health outcomes

Dependent variable	Readmission		Length of stay	Readmission	
Specification	OLS		First stage	Reduced form	IV
	(1)	(2)	(3)	(4)	(5)
Length of stay, days	0.120***	0.042***			-1.274^{***}
	(0.006)	(0.004)			(0.234)
Emergency admissions per day			-0.064^{***}	0.081***	
			(0.004)	(0.015)	
Age-diagnosis fixed effects		✓	✓	✓	✓
Hospital-season fixed effects		✓	✓	✓	✓
N	1,553,278	1,540,716	1,540,716	1,540,716	1,540,716

Notes: (1) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level; (2) Standard errors clustered at the hospital-level (149 clusters).

3.5.2 Compliers

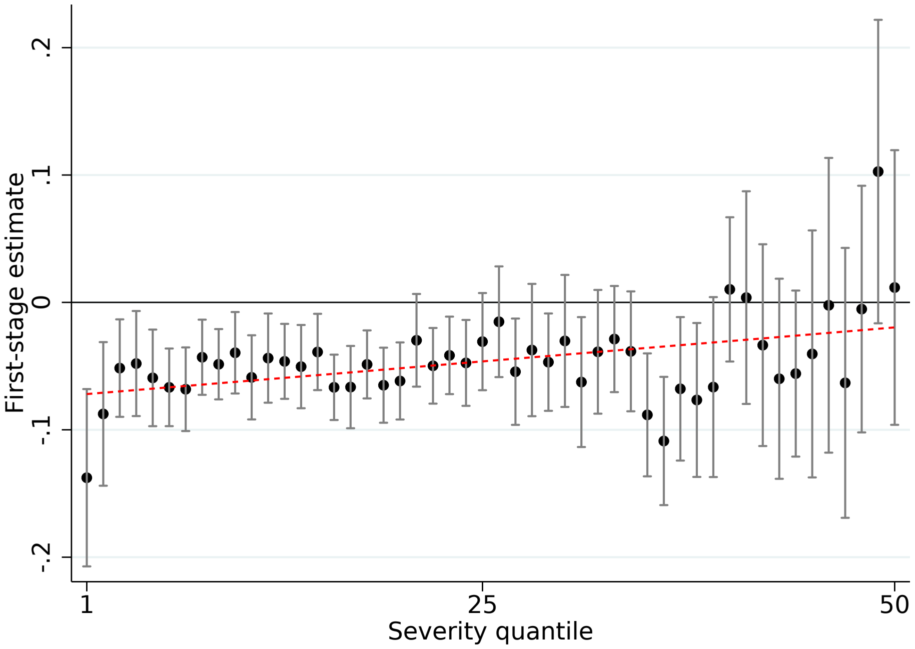
I now turn attention to the marginal patients that are implicit in the IV estimate. These patients are the compliers in the LATE delivered by the IV estimator. To help characterise these patients, I estimate the first-stage regression across quantiles of patient severity. To proxy for severity, I calculate predicted mortality using a regression of 30-day in-hospital death on a fully interacted set of age and diagnosis fixed effects, as well as controls for ethnicity, co-morbidities, past ED admits, and the number of diagnoses. I then group patients into 50 quantiles.

Figure 3.1 presents the results of this exercise. The plot shows that the magnitude of the first-stage coefficient declines with predicted mortality. Those with very low risk of death have an estimated first-stage coefficient of around -0.1, while those at the high end of risk have an estimated coefficient near to zero. This is consistent with hospitals discharging patients according to the expected mortality risk.

The implication for the interpretation of the IV estimates is that the compliers are generally patients at low risk of death. To further characterise these patients, Figure 3.2 plots the average age of each diagnosis group across the severity distribution, where I aggregate the previous severity quantiles into 10 categories. The size of the markers in the plot shows the number of admissions for each diagnosis. The largest four diagnosis groups–injuries to the hip/thigh, knee/lower leg, elbow/forearm, and wrist/hand–are highlighted in colour.

At the lower end of the severity distribution, patients are typically aged 40 to 60, and most often suffering from knee/lower leg and elbow/forearm injuries. Most of these patients will have fractured bones and have received ORIF procedures. These are the compliers for whom the LATE estimates

Figure 3.1: First-stage estimates by patient severity



Notes: (1) Severity defined as predicted 30-day in-hospital mortality, calculated from a regression of mortality on a series of fully interacted age and diagnosis categories, as well covariates for age, ethnicity, diagnosis count, past ED admits, and co-morbidities.

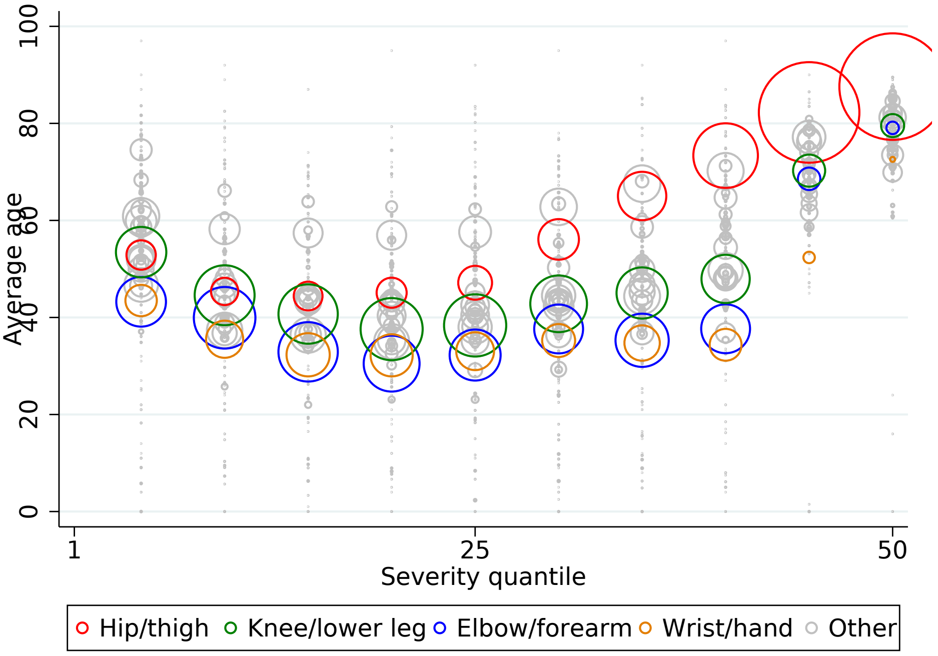
relate to.

At the high end of the severity distribution, patients are older, on average above 70, and most are suffering from hip and thigh injuries. The most common injury is a broken hip. These elderly and high risk patients are rarely being discharged in response to the emergency shocks and thus contribute very little to the LATE estimates.

3.6 Policy implications

I now assess the quantitative importance of the estimated relationship for PPS and readmission penalty policies. While I do not characterise the

Figure 3.2: Diagnosis types by patient severity and age



Notes: (1) Each circle corresponds to a two-digit ICD-10 diagnosis category, where the size of the circle represents the number of admissions; (2) Highlighted circles correspond to the most common categories of injury, and these commonly involve fractures and ORIF procedures; (3) Severity defined as predicted 30-day in-hospital mortality, calculated from a regression of mortality on a series of fully interacted age and diagnosis categories, as well covariates for age, ethnicity, diagnosis count, past ED admits, and co-morbidities; (4) Severity aggregated into 10 quantiles for illustration purposes.

weighting function implicit in the average causal responses I estimate (Angrist and Krueger, 1999), for the purposes of these calculations I assume that compliers can be roughly characterised as the patients with below-median predicted mortality (i.e. those patients with the largest first-stage estimates in Figure 3.1). These patients have a readmission rate of 8.2% and a length of stay of 4.2 days.

3.6.1 Prospective payment systems

The LATE estimates highlight a mechanism by which PPS can drive increases in readmissions. To get a sense of the potential magnitude of the PPS impact, consider the benchmark from Cutler and Zeckhauser (2000) that PPS is estimated to have reduced length of stay by 20 to 25%. I assume that this reduction in hospital stays came from efficiencies and not technological change. This approach will overstate the impact of PPS, since some technological change reduced length of stay without impairing readmission outcomes (Kehlet, 2013), and so the estimates I discuss provide an upper bound estimate on the potential impact of PPS on readmissions.

Reducing length of stay by 25% for marginal patients is roughly equivalent to a one-day decrease. This is predicted to increase the likelihood of readmission for these patients by around 1.3 percentage points, or 16% relative to their baseline readmission rate. At the aggregate, given the assumptions on marginal patients and if the readmission rate of other patients is unchanged, the readmission rate would drop by 0.64 percentage points (6.6% relative to baseline).

Comparing this estimate to trends in readmissions around the introduction of PPS in the U.S. and the U.K. is intriguing. The increase in 30-day readmission in England for trauma patients was 1.6 percentage points over 2006 through 2013, while the increase in 6-month readmission for Medicare patients in the U.S. was approximately 1.4 percentage points over 1981 to 1988 (Cutler, 1995). These changes are of a similar magnitude to the estimates presented in this chapter. It is therefore at least plausible that the length of stay mechanism may explain a sizeable fraction of the growth in readmissions since PPS was imposed.

3.6.2 Readmission penalties

The estimates also have important implications for the design of readmission penalties. With length of stay being a salient mechanism that determines readmissions, it is possible to evaluate whether the penalties create a sufficient incentive to cause hospitals to respond through the length of stay margin.

To evaluate this I estimate the financial consequences of extending length of stay for marginal trauma patients. A revenue loss occurs because increasing hospital stays for emergency patients means less capacity is available for elective admissions and fewer admissions will mean a reduction in the associated PPS payments. A revenue gain occurs because longer hospital stays will reduce readmissions and the hospital will therefore incur fewer penalties.

I adopt a scenario in which hospitals increase the average length of stay of marginal patients by 0.43 days. This is equivalent to the change in length of stay caused by a one standard deviation emergency shock, a benchmark that captures the variation in length of stay that hospitals routinely make. A change of this magnitude is predicted to reduce readmissions for marginal patients by 0.55 percentage points or, given the volume of marginal patients, 534 readmission events per annum. At an average readmission PPS payment of £2,014, and assuming full incidence of the penalty, hospitals will avoid penalties of around £1 million per annum. Fewer readmissions also provides some spare capacity in the future, since less patients are returning to hospital, and this is estimated to create an additional 3,912 bed-days per annum.

To obtain the net impact on the required bed-days, the increase in spare capacity needs to be offset by the longer stays of trauma patients. These longer stays are estimated to increase the required by bed-days by 41,744 per annum, meaning there is a net increase in bed-days of 37,832 per annum. Given an average elective patients length of stay (1.9 days) and the average PPS payment for these patients (£2,507), by extending the length of stay of trauma patients hospitals would forgo approximately £50 million in revenue (0.07% of total revenue across all NHS hospitals).

Readmission penalties, as currently designed, therefore create no incentive to increase length of stay to avoid readmissions: the costs of doing this are approximately 50 times as high as the benefits. This calculation also overstates the benefits, since readmission penalties in practice operate at less than full incidence.

This exercise highlights two policy implications for the design of readmission penalties. First, to influence the length of stay margin readmission penalties would need to be at least an order of magnitude higher. Second, the penalty for readmissions should be linked to the value of PPS payments for patients that are substitutable for readmissions (elective admissions) rather than the PPS payments of the readmission events themselves.

3.7 Conclusion

Health care providers are increasingly being subject to stronger financial incentives. This has led to major successes, often interpreted as efficiency gains, but has also raised concerns about the potential impacts on quality of care. Underlying many of these concerns is the multitasking idea: the difficulties measuring health outcomes could mean health care providers respond to strong incentive schemes by switching effort away from tasks that benefit unobservable dimensions of health towards tasks that benefit the observable dimensions of health specified in the incentive scheme (Holmstrom and Milgrom, 1991).

I study one part of the production process: how length of stay affects the likelihood of readmission. Understanding this relationship is important for how we think about two major reforms: PPS-style reimbursement tariffs and readmission penalties. Identifying the impact of length of stay requires quasi-experimental variation in this input, analogous to the study of firm production functions. I exploit variation in the emergency admissions which, under capacity constraints in the English setting, generate the required variation in length of stay. I argue that this variation meets the requirement needed to identify local treatment effects of early discharge.

I find that length of stay is a critical input to readmission outcomes. Increasing length of stay can substantially reduce the frequency of readmission events, and thus represents a straightforward adjustment to the production process that improves quality of care. Increasing length of stay by one-day for marginal patients could reduce readmissions by as much as 6.6%. The marginal patients that these estimates relate to in my setting are typically patients aged between 40 and 60 suffering from injuries such as fractured forearms and lower legs.

These estimates provide a convincing explanation for how PPS and readmission penalty policies have affected quality of care. The reductions in length of stay spurred by the adoption of PPS tariffs may have caused at least part of the subsequent increase in readmissions observed in the U.S. and U.K.. In contrast, length of stay has proven to be unaffected by the introduction of readmission penalties, a surprising fact once it is acknowledged that this is one margin for reducing readmissions. I show that these penalties are too weak to incentivise responses through the length of stay margin, primarily because the opportunity cost of scarce bed-capacity is high.

Overall this chapter underlines the importance of understanding the health care production process for designing policies and understanding the effects on welfare. The increasing availability of administrative data has begun to shed light on these processes although much remains to be done.

Chapter 4

Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers[1]

Perhaps the most complicated node of health delivery in any modern health care system is the emergency department (ED). Patients arrive at the ED with a wide array of different problems. ED nurses and physicians must quickly assess where patients should slot in what can be a very large queue, deciding almost instantly who needs to be treated right away and who can wait. And ultimately these providers need to decide whether those going

[1] This chapter is joint work with Jonathan Gruber (MIT) and George Stoye (UCL). Each author contributed equally. We thank Richard Blundell, Aureo de Paula, Eric French, Peter Hull, and Henrik Kleven for useful comments, as well as seminar participants at the Institute for Fiscal Studies, MIT, and UCL. The authors thank NHS Digital and the Office for National Statistics for access to the Hospital Episode Statistics and official mortality statistics under data sharing agreement CON-205762-B8S7B. Hoe and Stoye gratefully acknowledge financial support from the UK Economic and Social Research Council through the Centre for the Microeconomic Analysis of Public Policy (CPP) at IFS (ES/M010147/1).

to the ED are to be admitted to the hospital or sent back to their homes – a decision that can, in many instances, have life or death consequences.

Despite its critical role, EDs often face budgetary pressures and a shortfall in resources. These pressures have been especially acute in recent years, with ED performance having been described as an international crisis in several developed economies (Hoot and Aronsky, 2008). Practising doctors are especially vocal, referring to 'battlefield medicine' and 'third world conditions' caused by ED overcrowding in England.[2] Alongside these tensions, EDs are increasingly facing public pressure to advertise and reduce their wait times. U.S. cities are replete with digital billboards highlighting wait times at local EDs. And other nations use regulatory and financial tools to reward reductions, or penalize increases, in wait times.

Many are concerned that external pressures on wait times could reduce the ability of EDs to maximize the quality of the care that they provide. At the same time, however, it is not clear that ED personnel would maximize patient quality in the absence of such pressures. Emergency rooms are not compensated on the margin based on wait times. Moreover, while health-maximizing ED personnel will internalize the costs of waiting to the extent that they impact patient outcomes, this will be imperfect in the presence of uncertainty and may not account for patient well-being beyond health outcomes. Theoretical ambiguities such as this have motivated a growing number of empirical studies of hospital production in the ED setting (Chan, 2016, 2017; Gowrisankaran et al., 2017; Silver, 2016).

The 'four-hour wait' policy in the England provides a natural environment in which to address this critical question. This policy was first announced in 2000 as part of a wide ranging set of government pledges to

[2] https://www.nytimes.com/2018/01/03/world/europe/uk-national-health-service.html?smid=tw-share&_r=0

decrease wait times for different types of care, and came into force in all English public hospitals in 2004.[3] The policy set arbitrary targets for wait times, initially requiring that 98% of all patients be treated within four hours of arrival. The ability of hospitals to meet this target became an important part of overall hospital evaluation in England, with managers in some cases losing their jobs because of poor wait time performance. In addition, there were strong financial penalties associated with breaching the target – hospitals were penalized by an amount that was more than twice the average revenue of an ED patient, and total fines for missing ED and elective wait time targets were equivalent to a third of hospital deficits.

This policy has been controversial. Some stakeholders have argued that the focus on patient wait times has improved patient care. As one ED nurse quoted in Mortimore and Cooper (2007) said, "it was worse [before the targets were introduced], definitely it just seemed to be more hectic, there were people on trolleys for 12 hours and you'd leave here at 8pm and come back in the morning and there would still be some patients here". Others have argued that care quality has been sacrificed. One medical student stated, "patients are no longer known by their names or by their conditions, they're not even known by a number, patients are referred to by their time. By this I mean how long they've been in the department, as soon as a patient ticks past 3 hours their name lights up like a Christmas tree. If their stay approaches 3 hours 30, the managers start to appear, they don't actually care about Mr Jones who is having a heart attack. He's got to go, wherever it may be, as long as it's not ED".

Despite the controversy, there is little consistent evidence from either the UK or other nations that have introduced wait time targets on the impact

[3] Other targets included maximum limits on wait times for elective surgery.

of those targets on patient costs and health outcomes. This is because the policies are generally introduced nation-wide, with no 'hold-out' or control populations, making it impossible to apply quasi-experimental methods such as difference-in-difference estimation. An additional challenge in the case of the English wait time policy is that no systematic data on wait times are available before the policy was introduced in 2004.

In this chapter we take a different approach. We apply the bunching techniques that have been used widely in other contexts (see Kleven, 2016) to analyze wait times and outcomes. This approach allows us to model how the four-hour target impacts wait times, costs and outcomes, conditional on the underlying hospital technology in place to monitor patient wait times. That is, we estimate here the short term impact of changing wait times, but hold constant the underlying technological changes that might be associated with the introduction or removal of a wait time target. This counterfactual focuses attention on the impact of incentives rather than technology adoption.

We initially examine the distribution of wait times around the four-hour target, finding a very large spike right at four hours. We then turn to estimating counterfactual distributions of wait times in order to measure the effect of the four-hour policy. We estimate that, relative to the counterfactual, the four-hour target led wait times to be 19 minutes (8%) lower for patients affected by the policy.

We then use these data to study the impact of the policy on patient treatment and outcomes. Without pre-period data and exogenous variation in policy effects across hospitals, we cannot directly use data on treatments and outcomes to identify policy effects. But we argue that under a set of minimal and testable assumptions we can directly identify policy effects from bunching at the four-hour target.

In particular, to assess the impact of the target on outcomes such as hospital admissions and mortality, we need to separate a ‘composition effect’ (because some patients are moved from after to before four hours of wait time due to the target, and they may not be randomly chosen) and a ‘distortion effect’ (the target itself may have a direct distortion on the treatment of randomly chosen patients). To separately identify the distortion effect, we estimate a ‘composition-adjusted counterfactual outcome’ by imposing a ‘no-selection’ assumption on the distribution of patients that obtain shorter wait times because of the policy. We can test this assumption directly using patient observables, showing that along multiple dimensions there is no difference between these and other patients.

We estimate that there is a significant distortion effect of the English policy. We find that there is more intensive testing of patients in the ED, leading to a modest rise in ED costs. We also find that there is a significant increase in hospital admissions as a means of meeting the target, with corresponding reductions in those discharged to home. Among those marginal admits, inpatient resource use is insignificant, suggesting that such admissions were just placeholders to meet the four-hour target. These admissions were not costless, however, and we estimate that inpatient payments from the government to hospitals rose by roughly 5% due to the target.

Most interestingly, we find significant improvements in patient outcomes associated with the four hour policy. We estimate that 30-day patient mortality falls by 14% among patients who are impacted by the wait time change, a very sizeable positive effect. This effect falls slightly over time while baseline mortality rises, so that by one year after ED admission this amounts to a 3% mortality reduction, which is still quite large.

We then turn to understanding the mechanism behind the outcome improvement that we observe. To do so we exploit heterogeneity across patient groups that are affected along different margins. The first is patients of different severity: across severity groups, the four-hour policy is associated with differential impacts on wait times, but not admission probabilities. The second is patients facing different levels of crowding of the inpatient department when they arrive at the ED: across different levels of crowding, the four-hour policy is associated with differential impacts on admission probabilities but little variation in the wait times impacts. We then show that the mortality effect we estimate varies strongly across patient severity, but not across inpatient crowding. Taken together, this evidence suggests that it is the wait time mechanism, and not the admissions mechanism, that is driving our mortality effect.

We contribute to two literatures. First, there is a growing literature that has begun documenting features of hospital production relevant for incentive setting (Chan, 2016, 2017; Gowrisankaran et al., 2017; Silver, 2016). Chan (2016) and Chan (2017), for example, study how ED physicians respond to team environments and work schedules, while Silver (2016) studies peer effects in the ED. Gowrisankaran et al. (2017) also study the ED and estimate different measures of physician skill. Adjacent to these studies, a medical literature has documented robust correlations between mortality rates and measures of ED crowding and wait times (Hoot and Aronsky, 2008). Our contribution is to show how ED production is affected when doctors are put under pressure to make decisions quicker. We find that the wait time policy generated cost-effective mortality improvements through reduced wait times but at the expense of distorting medical decisions. These findings are consistent with the medical literature and highlight that ED wait times are an important input to the health production process. The findings also illustrate how constraining healthcare providers through regulatory interventions can improve health outcomes even in the presence of significant distortions.

The second contribution we make is to the literature using bunching estimators. From its origins in the tax setting (Saez, 2010; Chetty et al., 2013; Kleven and Waseem, 2013), these estimators have now been deployed in other settings such as health insurance (Einav et al., 2015, 2017, 2018), mortgage markets (Best et al., 2017; Best and Kleven, 2018) and education (Diamond and Persson, 2016). We apply these estimators in a healthcare provision setting, adapting them to study outcomes indirectly affected by a discontinuity in the incentives associated with the running variable, and devise new empirical tests to evaluate the credibility of the bunching assumptions required in our context.

The chapter proceeds as follows. Section 4.1 provides background information on emergency care in England and on the four-hour target policy. Section 4.2 describes the data. Section 4.3 sets out our methodology, beginning with an overview and followed by the details of our analysis of wait times, treatment decisions, and health outcomes. Section 4.4 describes our results for wait times, treatment decisions and health outcomes. Section 4.5 explores heterogeneity and mechanisms. Section 4.6 concludes.

4.1 Background

4.1.1 Emergency care in England

Emergency care in England is publicly funded and is available free at the point of use for all residents. There is no private market for emergency care.

The majority of care is provided at emergency departments (EDs) attached to large, publicly owned hospitals. These major emergency departments are physician-led providers of 24-hour services, based in specifically built facilities to treat emergency patients that contain full resuscitation facilities. In 2011/12, 9.2 million patients made 13.6 million visits to 174 emergency departments. In addition, 2.1 million patients made an additional 2.7 million visits to specialist emergency clinics and 'walk in' or minor injury centres where simple treatment is provided for less serious diagnoses; as discussed below, we exclude patients from these centres due to the minor nature of their injuries and our results are unaffected if they are included.

EDs provide immediate care to patients. Hospitals are reimbursed by the government for the care they provide, receiving a nationally fixed payment for providing certain types of treatment.[4] In 2015/16, there were 11 separate tariffs for ED treatment depending on the severity of the patient and the type of treatments administered.[5] These tariffs ranged from $77 to $272 per visit.[6] Revenue from the ED accounted for 5.3% of total hospital income in 2015/16.[7]

Treatment in the ED follows one of two pathways depending upon the method of arrival. Non-ambulance patients register at reception upon arrival, where they must identify themselves and provide basic details of their condition. Patients then undergo an initial assessment to establish the se-

[4] Treatments are assigned to a Healthcare Resource Group (HRG), similar to DRGs in the US, with a set of national tariffs for each HRG announced each year by the Department of Health.

[5] https://www.gov.uk/government/publications/confirmation-of-payment-by-results-pbr-arrangements-for-2012-13

[6] All cost figures in 2017/18 US Dollars. Figures are deflated using the UK GDP deflator, and then converted from sterling to dollars using an exchange rate of 1GBP:1.35USD (US Treasury, 31st Dec 2017, $https://www.fiscal.treasury.gov/fsreports/rpt/treasRptRateExch/currentRates.htm$).

[7] Figures calculated from the 2015/16 UK Department of Health Reference Costs. See: https://www.gov.uk/government/publications/nhs-reference-costs-2015-to-2016

riousness of their condition. This triage process is carried our either by a specialist triage nurse or doctor, and includes taking a medical history, and, where appropriate, conducting a basic physical examination of the patient. Patients are then prioritized according to severity.

Alternatively, patients can arrive at the ED by ambulance following an emergency call out. In 2011/12, 29.4% of ED patients arrived by ambulance. For these patients, ambulance staff collect medical details en route, and report these details to hospital staff upon arrival.[8] This information feeds into a separate triage process, where patients will be categorized by their severity.

These triage processes sort patients into 'minor' and 'major' cases. Minor cases require relatively simple treatment, and can often be treated in a short space of time. Major cases are often those who arrive by ambulance, although there are some exceptions to this (for example, a patient with chest pain may arrive independently at the hospital). Major cases will receive treatment more quickly, as they often present with more severe symptoms, but will usually require more treatment and investigations within the ED, and are therefore likely to spend longer in the ED. Treatment of the two types often requires the use of different resources (including staff and machines), and in most large hospitals, treatment for minor conditions will take place in a separate part of the emergency department (for example, in the hospital's 'urgent care centre').

Following triage, patients are placed into a queue on the basis of their severity and time of arrival. Patients are not aware of their position in the queue. Patients are assigned to individual doctors as they become available. These doctors will carry out a series of further examinations and tests. The

[8] Ambulance staff also provide emergency treatment in the ambulance to patients where required.

nature of these investigations depend on the symptoms presented by the patients, and range from physical examinations to tests such as x-rays or MRI scans. Patients can also receive treatment in the ED, ranging from sutures to resuscitation, before being admitted for further treatment in an inpatient ward, or discharged from the hospital.

4.1.2 The 4-hour target

All public hospitals with EDs in England are subject to a wait time target. This target specifies that 95% of ED patients must be admitted for further inpatient treatment, discharged or transferred to another hospital within four hours of their arrival. The target level was initially set at 98% when it was first introduced in December 2004, before being relaxed to its current level in November 2010.[9]

This target is important to hospitals in two ways. First, the target is widely used by policy makers and the media as a measure for the wider performance of the public health service in England.[10] Hospital managers who consistently fail to meet this target are likely to be fired, and therefore have a strong incentive to organise emergency care in a way that minimises the number of patients who take more than four hours to treat.

Second, hospitals face significant financial incentives to meet the target. As the target came into force between March 2004 and March 2005, hospitals were offered payments (to be used only for hospital investment)

[9] Interviews with hospital managers, doctors and regulators suggest that it is the 'four-hour' component of the target that matter to hospitals rather than the absolute level of the target. Hospitals attempt to meet the target on a daily basis, and aim to achieve the highest proportion possible. This suggests that certain behaviours, such as relaxing or improving performance in later parts of the reporting period, are unlikely. Consistent with this, we do not find any systematic evidence of differences in our results by time period or at different points of the reporting period.

[10] For example, see http://www.mirror.co.uk/news/uk-news/ae-crisis-exposed-only-three-9801509.

if they met the target level early (National Audit Office, 2004). In recent years, significant financial penalties have been imposed for missing the target. In 2011/12, hospitals were fined \$300 for every patient who failed to be treated within 4 hours if the hospital missed the overall 95% target during that week.[11] This compares to an average payment of \$140 per patient in the same year. In 2015, a report commissioned by a number of hospitals indicated that public hospitals paid \$325 million in fines due to missed performance targets (including the 4 hour target), with total penalties equal to around a third of the average deficit of public hospitals in that year.[12]

Hospital staff therefore face pressure from hospital management to meet the target. As a result, the organisation of EDs has changed significantly since the target was introduced.[13] Changes include the use of new IT systems, which track patient wait times in real time. The exact systems vary by hospital, but will indicate when patients reach particular waiting thresholds (e.g. 3 hours) and alert physicians (for example through changing the colour of the computer screen).[14] Most departments also now employ specific members of staff to monitor the progress of all patients against the clock, and to alert physicians that an admission decision is required soon.

[11] This penalty was decreased to \$170 in 2015.

[12] https://www.theguardian.com/society/2016/mar/29/nhs-bosses-slam-600m-hospital-fines-over-patient-targets

[13] Interviews with senior member of the Emergency Care Improvement Programme (ECIP), a clinically led programme intended to improve the performance of EDs, clearly describe significant changes to the technology used in EDs since the target was introduced. One manager in the programme claimed that "This [the target] is the most monitored part of the entire healthcare system with software specifically designed for it".

[14] One medical student in an ED describes the IT system in the following way: "Displayed prominently on an electronic whiteboard is a list of all the patients currently in A&E and waiting to be seen, and the second a patient ticks past a 3 hour wait, their name lights up like a Christmas tree in bright red". See: https://imamedicalstudentgetmeoutofhere.blogspot.co.uk/2008/03/there-is-338-in-bay-5.html

4.2 Data

4.2.1 Hospital Episodes Statistics

Our primary source of data are the Hospital Episode Statistics (HES). These contain the administrative records of all visits to public hospitals between April 2011 and March 2013, and include information on both ED visits and inpatient admissions.[15]

The ED data record treatment at the visit level, and include information on the precise time of arrival, initial treatment and the admission decision. We define ED 'wait times' as total time spent in the ED, consistent with the definition of the policy. This includes time being examined and treated. We calculate ED wait times as the time elapsed between arrival and the admission decision.

The data also include a hospital identifier, whether the patient is admitted or discharged, details of basic diagnoses, the number and types of ED investigations and treatments, whether the patient arrived by ambulance, and some basic patient characteristics such as age, sex and local area of residence.

Patients are identified by a psuedo-anonymized identifier that allows patients to be followed over time and across hospitals, and enable linkage between ED and inpatient records. Inpatient records contain detailed information on treatment undergone in the hospital. The data contain the dates of admission and discharge, and information on up to twenty diagnoses and procedures undertaken. Treatment is recorded at the episode level, defined as a period of treatment under the care of a single senior doctor.[16] We

[15] Data on EDs is available prior to 2011, covering 2008 and 2010, although data from the earlier period is less complete than in the years we study.

[16] Senior doctors in England are known as 'consultants', and are equivalent to attending physicians in the US.

combine information across all episodes within the same admission to create visit-level variables for total length of stay (in days) and number of inpatient procedures. Each episode also contains a Healthcare Resource Group (HRG) code, similar to Diagnosis Related Groups (DRGs) in the US. English hospitals are compensated by the government through a system of national tariffs for each HRG.[17] We calculate 'costs' for each episode by matching tariffs to the appropriate HRG, which gives us a measure of the cost to the government, and revenue received by the hospital, associated with each visit. We then sum all treatment costs over a 30 day period to estimate the cost associated with each ED visit and any follow-up treatment.

Mortality outcomes are recorded in administrative records made available by the UK Office for National Statistics (ONS). These records are linked to HES through anonymized identifiers based on patient National Insurance (Social Security) numbers. The data include the date of death for all individuals who died in the UK, or UK citizens who died abroad, between April 2010 and March 2014. We create indicators of whether a patient dies within 30, 90 and 365 days of an ED visit. An indicator of in-hospital mortality is also calculated using HES.

Sample construction

Our analysis focuses on a sample of emergency patients treated in 'major' emergency departments.[18] We exclude patients treated at specialist clinics that treat only particular diagnoses (e.g. dental) and minor injury ('walk

[17] National tariffs are calculated for each HRG on the basis of annual cost reports submitted by hospitals to the UK Department of Health. These tariffs are meant to reflect the average cost of providing the procedure. Payments are then adjusted for unavoidable regional differences in providing care, and unusually long hospital stays.

[18] Major emergency departments are defined as consultant-led providers of 24-hour services, based in specifically built facilities to treat emergency patients that contain full resuscitation facilities.

in') centres. Patients treated by these units typically have simple diagnoses and short wait times, and are therefore unlikely to be affected by the target. This excludes 18% of emergency visits.

We keep all patients with full information relating to the timing of treatment and their exit route from the ED, in addition to their age, gender and whether they arrived by ambulance. Dropping patients with some missing information reduces the number of visits in the sample by 14.5%.[19] This yields an analysis sample of 14.7 million patients, who made 24.7 million visits to 184 EDs between April 2011 and March 2013.

Summary statistics

Table 4.1 reports summary statistics. The first two columns present the mean and standard deviation for a range of patient characteristics, treatments and outcomes for all ED patients in the sample. Mean ED patient age was 39 years, and 51% of patients were male. 29% of patients arrived by ambulance. 5.8 million visits, or 24% of all ED episodes, resulted in an inpatient admission at the same hospital. 58% of visits did not require further hospital treatment and led to a patient being discharged. The remaining visits resulted in a transfer to an outpatient clinic or another hospital for further treatment. Mean 30-day treatment costs were $1,676, of which 89% was accounted for by subsequent inpatient treatment. In the short term, mortality among ED patients is relatively rare. 2% of patients died within 30 days of visiting the ED. This increases to 3% over a 90 day period, and 5% during the following year.

Table 4.1 also shows summary statistics separately for visits that result in an inpatient admission. As expected, these case are typically more severe,

[19] Results are unaffected by the inclusion of patients with full information relating to treatment times and decisions, but who are missing demographic information.

Table 4.1: Summary statistics

	All patients		Admitted inpatients	
	Mean	Std. dev.	Mean	Std. dev.
Patient characteristics				
Age	38.99	26.22	54.64	27.84
Male	0.51	0.50	0.48	0.50
Ambulance arrival	0.29	0.45	0.60	0.49
Treatment decisions				
Inpatient admission	0.24	0.42	1.00	0.00
ED discharge	0.58	0.49	0.00	0.00
ED referral	0.19	0.39	0.00	0.00
Wait time (mins)	154.56	100.20	222.50	120.46
ED treatments	1.81	1.38	2.22	1.68
ED investigations	1.54	2.03	3.18	2.50
Inpatient stay (days)	1.28	5.63	5.41	10.58
Inpatient procedures	0.16	0.64	0.69	1.18
Costs				
30-day ED cost	172.35	117.21	203.98	114.98
30-day inpatient cost	1,503.58	5,321.99	4,558.00	8,524.53
30-day total cost	1,675.93	5,358.37	4,761.98	8,559.73
Mortality outcomes				
30-day mortality	0.02	0.13	0.05	0.23
60-day mortality	0.03	0.16	0.09	0.29
365-day mortality	0.05	0.22	0.16	0.37

Notes: (1) Costs reported in 2018 USD and refer to payments from the government to hospitals based on the prospective payment system; (2) All inpatient variables (e.g. length of stay, costs) take on the value zero for patients that are not admitted.

with an older average age (55 years) and twice the likelihood of arriving in an ambulance (60%). Mortality rates (5% over 30 days, 16% over a year) are substantially higher than in the main sample. ED treatment is more intense for this sample, with a higher mean number of treatments and investigations than in the main sample. Their treatment is also more expensive, with an average total cost over a 30-day period of $4,762.

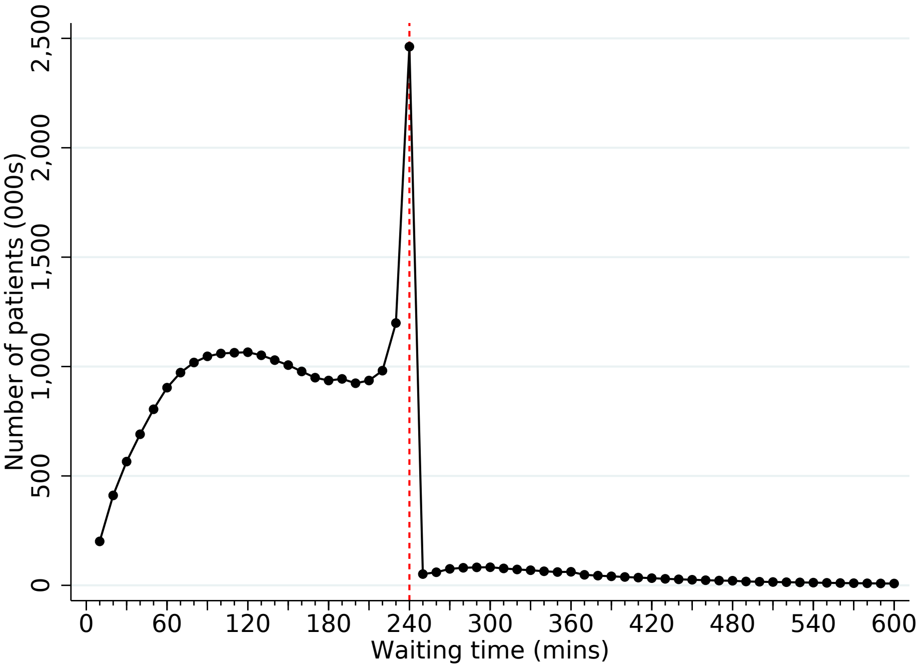
Inpatients also experienced longer mean wait times in the ED than those who are not admitted. Mean wait times were 223 minutes for patients who were eventually admitted as inpatients, compared to a mean of 155 minutes for all ED patients. This demonstrates that the level of patient complexity, and the intensity of treatment for these patients, is likely to vary by wait time. This variation is important to account for when analysing the impact of the target.

Figure 4.1 shows the distribution of ED wait times. There is a noticeable discontinuity in the proportion of patients who exit the ED in the period immeditely prior to 4 hours. This spike is unlikely to naturally occur, and is instead induced by the target. We cannot illustrate the absence of this spike prior to the wait times target, since we do not have systematic data available from that period. But it is worth noting, as we do in Figure B.4, that such a spike is not present in data on ED wait times from a major U.S. hospital.[20]

One possibility is that this spike in wait times simply reflects recoding and is not a real change in patient wait times. Two features suggest that this is not the case. First of all, a sizeable share of hospitals pay large penalties and are publicly criticized as a result. Indeed, a substantial number of

[20] Of course, different ED objectives and technologies across countries means that the U.S. data does not provide a natural comparison group, but the lack of any spike confirms our conclusion that the large spike here is particular to the wait time policy.

Figure 4.1: Distribution of wait times



Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy.

hospitals only just miss the target, with 23% of hospitals missing the target by less than two percentage points in 2011/12. If recoding explained the spike then those hospitals should do more recoding to avoid the penalty altogether. Second, we show below that there are comparable spikes in a number of real outcomes, such as hospital admissions, costs, and mortality, which are inconsistent with this simply being a coding response.

4.3 Empirical methodology

We now set out our empirical methodology. We begin with an outline of our approach and then describe our analysis of wait times followed by our analysis of treatment decisions and health outcomes.

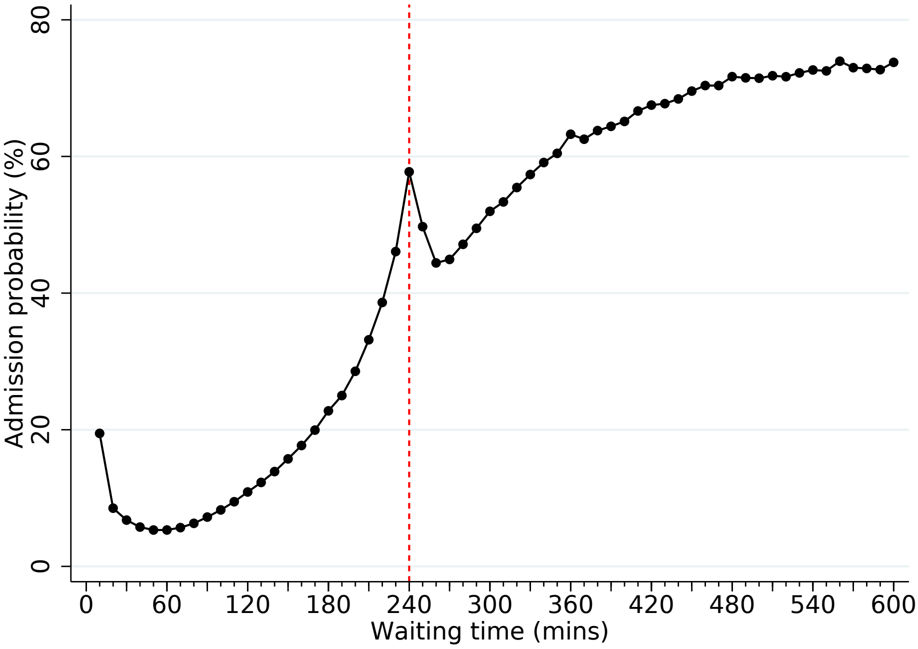
4.3.1 Overview

A key challenge when analysing the four-hour target is that without pre-policy data or a control sample, quasi-experimental methods cannot be used to construct the counterfactual outcome. To address this issue we use and extend bunching estimators that were developed in the tax literature (Saez 2010, Chetty et al. 2013). We argue that these methods can be used in our setting to estimate the counterfactual outcomes that would occur if the target were removed but other aspects of hospital production were held constant. This allows us to quantify the short-run impact of the policy.

We first apply a bunching estimator to the distribution of wait time outcomes. This involves interpolating how the wait time distribution would look in the absence of the target. As is typical in other bunching settings, we make a 'local effects' assumption; namely, that the target only affects the wait time distribution within a certain segment of the distribution. We argue that this assumption holds if hospitals do not substitute resources between patients located in different segments of the wait time distribution, and present empirical evidence that supports this assumption. The estimated counterfactual distribution from the bunching estimator allows us to quantify the impact of the target on wait times.

We then turn to an analysis of treatment decisions and health outcomes. Plotting these outcomes conditional on the wait time shows that they also exhibit 'bunching' at the four-hour discontinuity point. Figure 4.2 gives an example for the likelihood of inpatient admission. The plot shows that admission odds are generally increasing with wait times, and there is a clear spike in admission odds at 240 minutes. Our analysis decomposes this spike

Figure 4.2: Inpatient admission probability conditional on wait time



Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy.

into two channels.

The first channel is the ‘composition effect’. As Figure 4.1 suggests, the target causes a substantial number of patients to be moved from later to earlier in the distribution of wait times (a group we refer to as ‘post-threshold movers’). Since admission probabilities are increasing with wait time, this movement of patients would increase the observed pre-threshold admission probability even if the target led to no additional admissions. This effect arises purely because the target changes the composition of patients observed at each wait time.

There is also potential for a ‘distortion effect’ if the target has a direct effect on treatment decisions and health outcomes. The distortion effect implies identical patients receive different treatment depending on whether or not the target is in place. In the case of admissions, for example, it would imply that part of the spike in observed outcomes is because the target causes additional admissions, in addition to the composition effect shifting some admissions from after to before the target.

To decompose these two effects we construct a 'composition-adjusted counterfactual' (CAC). This is the outcome that would occur in the presence of composition effects but the absence of distortion effects. Since the observed data contains both effects, the difference between the observed data and the CAC identifies the distortion effect. Estimates of the distortion effects and tests of whether these are significantly different from zero are the central results of this chapter.

We construct estimates of the CAC by first showing it can be written as a weighted average of counterfactual outcomes for patients situated in different parts of the wait time distribution. We then argue that the required counterfactual outcomes can be constructed by applying bunching techniques to the *expected outcomes conditional on the wait time.*[21] This relies on a 'no-selection' assumption about the distribution of post-threshold movers: that those patients moved forward in time are representative of all post-threshold patients.

To evaluate the validity of the no-selection assumption, we devise a test based on observable patient characteristics such as age. These variables, conditional on the wait time, also exhibit bunching at the four-hour point but in these cases the spike can only be explained by a composition effect since there is no distortion effect by definition. If the no-selection assumption

[21] This is in contrast to a typical bunching application that would work with the distribution of a variable that is subject to a discontinuity in incentives. Here we work with outcomes conditional on a variable that is subject to a discontinuity in incentives.

is valid then for these variables the observed data and the CAC should be equal. Tests of this hypothesis therefore act as a placebo test, where rejection of the null hypothesis would suggest that the no-selection assumption has been violated. We pass these placebo tests for a range of demographic variables.

We proceed by outlining each of these steps and assumptions more formally.

4.3.2 Wait times

Let w be the wait time in minutes, where $w^* = 240$ (the target threshold). Denote the density function of w in the targeted regime as $f_t(w)$ where $t = \{0, 1\}$ signifies whether the function relates to the targeted or non-targeted regime. We observe data on $f_1(w)$ and use a bunching estimator to obtain $f_0(w)$.

To implement the bunching estimator we aggregate the data to 10-minute wait time bins and then interpolate parts of the distribution using a polynomial regression. Following Kleven (2016) we define $\hat{f}_0(w) \equiv \sum_{i=0}^{p} \hat{\beta}_i w^i$ and obtain the estimates $\hat{\beta}_i$ from the following regression

$$c_j = \sum_{i=0}^{p} \beta_i (w_j)^i + \sum_{i=w^-}^{w^+} \gamma_i \mathbf{1}[w_j = i] + u_j, \tag{4.1}$$

where c_j is the number of individuals in wait time bin j, w_j is the maximum wait time in bin j (e.g. $w_j = 10$ for the 1-10 minute wait time bin, $w_j = 20$ for the 11-20 minute wait time bin, etc), p is the order of the polynomial, and $[w^-, w^+]$ is an 'exclusion window' that contains w^* and is the period during which we assume that the target may have had local effects on the wait time.

Equation (4.1) makes the following assumption in relation to the exclusion window.

Assumption 1 (Local wait time effects)**.** *Wait times of patients outside of an 'exclusion window', defined locally around the threshold w^*, are unaffected by the target:*

$$f_0(w) = f_1(w) \qquad \forall w \notin [w^-, w^+]. \tag{4.2}$$

As we explain shortly, this assumption will hold if hospitals do not respond to the target by substituting resources between patients that are inside and outside of the exclusion window.[22]

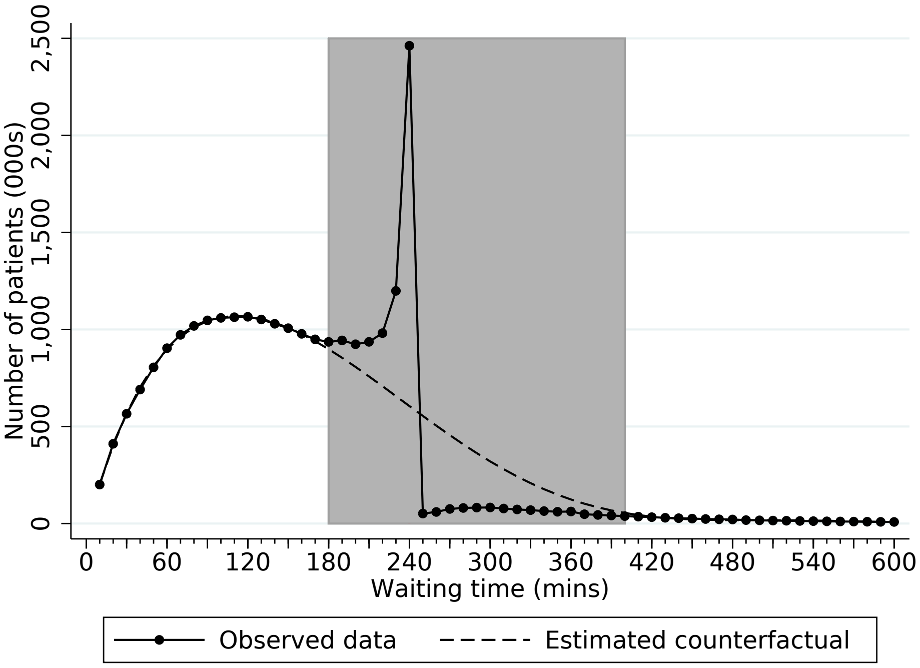
To establish the bounds of the exclusion window, we follow Kleven and Waseem (2013) and set w^- visually by examining when the distribution changes sharply and determine w^+ using an iterative procedure that equates the excess mass in the period $[w^-, w^*]$ with the missing mass in the period $(w^*, w^+]$.[23] In the baseline analysis we use a polynomial of order 10 and set $w^- = 180$. After applying the iterative procedure this produces an upper cut-off of $w^+ = 400$.

The observed data and our estimated counterfactual distribution are shown in Figure 4.3, which indicates that the target moves a number of patients from the post-threshold period to the pre-threshold period ('post-threshold movers'). We later use these distributions to estimate the impact of the target on wait times.

[22] A comparable assumption is required when using bunching techniques to study taxable income responses. In that setting the local effects assumption is often innocuous because the income distribution is the result of optimization decisions of many unrelated individuals, with those situated far from the tax scheme discontinuity having no incentive to adjust their behaviour. In our setting, the distribution of patient wait times is not determined by patients' decisions but by the decisions of doctors and nurses, and this raises the concern that there may be an incentive to substitute wait times between patients across different parts of the wait time distribution.

[23] This implicitly assumes that the target does not affect patient demand for ED care in the short-term.

Figure 4.3: Estimated counterfactual wait time distribution



Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy; (4) The estimated counterfactual is obtained from a polynomial regression that omits the exclusion window shown in grey.

Testing for substitution effects

If hospitals respond to the target by substituting time or resources between patients inside the exclusion window and those outside the window then Assumption 1 will not hold. A particular concern is that the target may induce doctors to substitute time or resources from patients typically discharged early in the wait time distribution (often high severity patients with unambiguous symptoms, e.g. knife attack victims) to patients that might be at risk of breaching the four-hour target (often high severity patients with uncertain diagnoses, e.g. headaches). By making this type of substitution, doctors would extend some shorter wait times in order to treat a greater proportion of patients within four-hours and thus perform better relative to the target. The wait time distributions, with and without the target, would then differ outside of the exclusion window and violate Assumption 1.

We test for two types of substitution effects. The first is 'planned substitution', where the target causes a permanent change in the priority given to certain types of patients. The second is 'temporary substitution', where the target causes short-term deviations from planned priorities when the ED is momentarily overrun with patients.

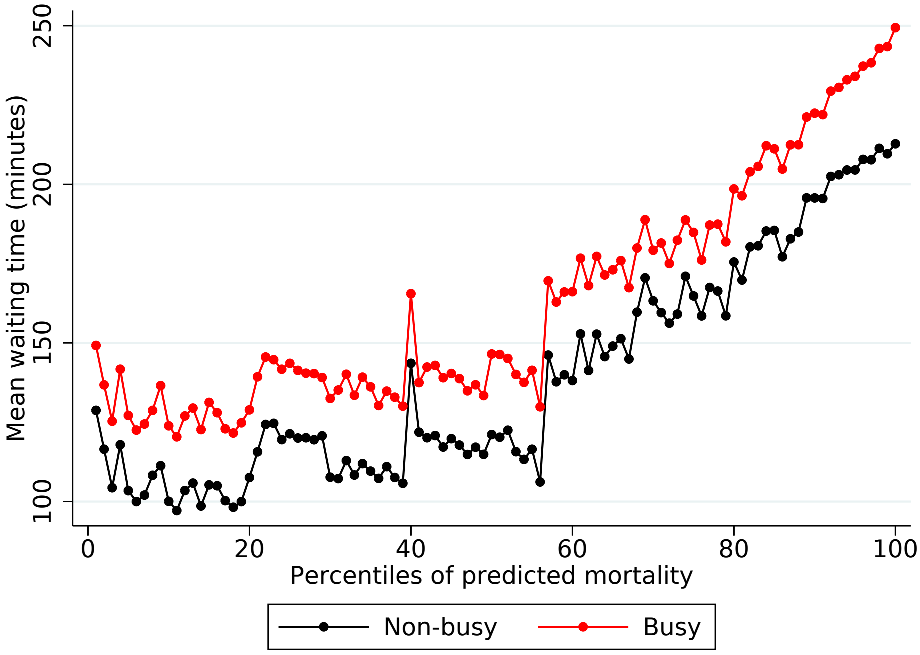
To test for planned substitution effects we exploit variation in expected volumes of ED arrivals. This variation changes how tightly the target binds since when there are higher volumes of arrivals the target is more challenging to meet in relative terms. With planned substitution effects we anticipate that hospitals would change patient prioritisation across periods that are expected to be more or less busy. While the volume of arrivals may be correlated with other factors, such as the number of doctors scheduled to be on shift, this would not necessarily impact the patient prioritisation that we compare in this test.[24]

Figure 4.4 plots average wait times for each percentile of predicted mortality (as a measure of severity) for patients that arrive during 'busy' and 'non-busy' periods. We define busy periods by first predicting the number of patients present in the ED during each hour in our data, using a regression with hospital-specific week-of-year, day-of-week, and hour-of-day fixed effects. We then divide periods into the top-third of predicted volumes (busy) and bottom-third of predicted volumes (non-busy).

The plot shows that higher severity patients typically have longer wait

[24] One potential concern could be that any increase in scheduled doctors may offset any increase in expected arrivals. If we repeat the same test but use shocks to ED arrivals then we find similar results.

Figure 4.4: Mean wait times by patient severity and expected volumes of ED arrivals



Notes: (1) Wait times defined as the time from arrival in the ED to leaving the ED; (2) Predicted mortality defined using a regression of 30-day in-hospital mortality on a fully interacted set of age, gender, ambulance arrival fixed effects and diagnosis fixed effects; (3) Busy and non-busy periods defined by predicting the volume of ED arrivals during each hour in our data, using a regression with hospital-specific week-of-year, day-of-week, and hour-of-day fixed effects, and then dividing periods into the top-third of predicted volumes (busy) and bottom-third of predicted volumes (non-busy).

times, and that busier periods have longer wait times for patients of all severity. Most importantly for our purposes, the relative wait times of high and low severity patients are very similar in both types of period. This suggests that as the target binds more or less tightly, hospitals maintain the same prioritisation of patients and there are no planned substitution effects.

To test for temporary substitution effects, we examine whether there is any evidence that hospitals substitute resources away from patients that we would expect to exit in the early part of the distribution in order to ensure that patients approaching the target do not wait over 4 hours. Intuitively, we compare the wait times of newly arrived patients on the basis of how many patients in the ED have waited almost four hours. If there are temporary substitution effects between these individuals, we would anticipate large effects of the presence of existing patients near the four-hour threshold on the wait times of new patients.

We examine two groups of newly arriving patients. We first focus on 'early exit' patients (those predicted to have wait times below 180 minutes, such that they exit prior to the exclusion window) and regress their wait times on the volume of existing patients wait ahead of them at each 10-minute interval of the queue. We then compare these results to an equivalent analysis of 'late exit' patients (those predicted to have wait times above 180 minutes). The late exit group act as a control group in the sense that Assumption 1 allows for temporary substitution effects to occur for this group (inside the exclusion window) but not for the early exit group (outside the exclusion window). We predict early or late exit using a regression of wait times on age, gender, diagnosis fixed effects and an ambulance indicator.

To implement the test we aggregate the data to the hospital-period level, where periods are defined at 10-minute intervals, and estimate the following equation

$$w_{ht}^{g} = \sum_{k} \beta_k q_{h,t-k} + \mu_{hw} + \delta_{hd} + \gamma_{hp} + e_{ht} \tag{4.3}$$

where w_{ht}^{g} is the mean wait time for newly arriving patients of type g (early or late exit) at hospital h in period t (e.g. between 12:01 and 12:10), $q_{h,t-k}$ is the number of existing patients waiting ahead in the queue at horizon $t-k$ (e.g. the number of patients that have been waiting 1-10 minutes, 11-20 minutes, and so on), and μ_{hw}, δ_{hd} and γ_{hp} are hospital-specific week-of-year, day-of-week, and period-of-day fixed effects.

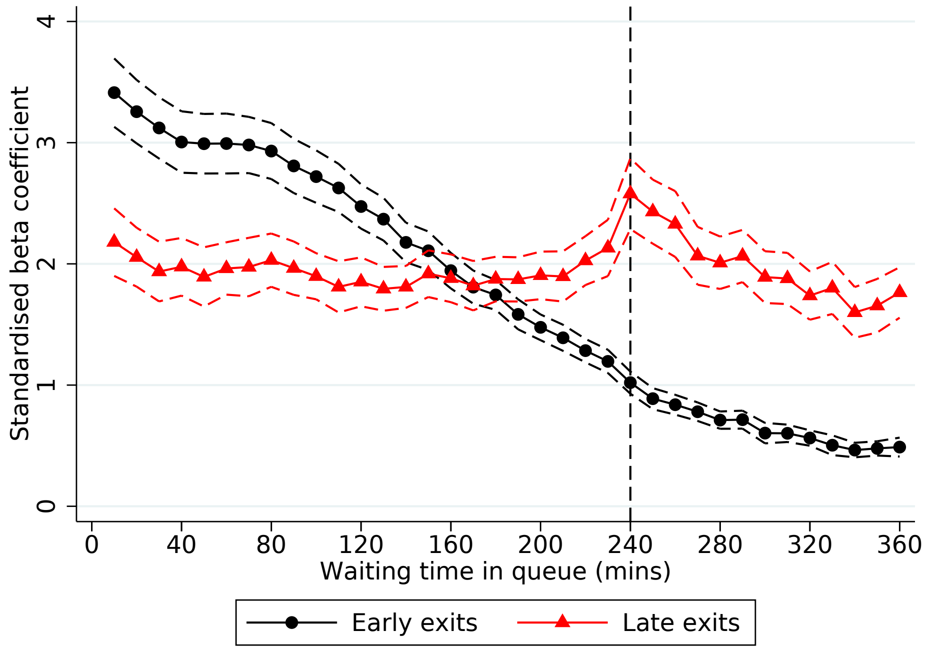
Figure 4.5 presents the estimated β_k coefficients from Equation (4.3). We normalise coefficients so they can be interpreted as the impact of a one standard deviation increase in the queue length at each horizon on newly arriving patients' wait times. Looking first at the early exit group, the plot shows that longer queues increase wait times and the impacts decline with the time horizon. There is no evidence of disproportionate impacts around the four-hour threshold. Looking now at the late exit group, there is again evidence of longer queues increasing wait times but for this group there is clear evidence of a discontinuity at the four-hour threshold. This indicates that, for the late exit group, doctors actively substitute resources away from newly arriving patients towards those patients that are at risk of breaching the target. These results suggest that there are temporary substitution responses for patients predicted to be within the exclusion window (late exits) but not for those predicted to be in the earlier part of the distribution (early exits).

Taken together, Figures 4.4 and 4.5 suggest that there are no planned substitution responses, and temporary substitution responses do not occur outside of the exclusion window. This is consistent with Assumption 1.

Interpreting the counterfactual

The counterfactual that the bunching estimator delivers in our context is the short-run outcome that would occur if the four-hour discontinuity in incentives were removed. The counterfactual holds constant other aspects of hospital production, such as patient prioritisation, capital and labour inputs, and government funding. As a benchmark, the counterfactual focuses attention on the role of incentives in determining outcomes rather than the

Figure 4.5: Impact of queues on wait times for arriving patients by early- and late-exit groups



Notes: (1) Wait times defined as the time from arrival in the ED to leaving the ED; (2) We normalise coefficients so they can be interpreted as the impact of a one standard deviation increase in the queue length at each horizon on newly arriving patients' wait times; (3) Early or late exit is defined using a regression of wait times on age, gender, diagnosis fixed effects and an ambulance indicator, from which we predict wait times and group individuals into early (below 180 minutes) and late (above 180 minutes) exit groups.

specifics of the production function in our setting. We see it as a logical benchmark for understanding how wait time incentives affect outcomes.

Our counterfactual differs from the pre-policy or long-run outcomes. To give an example of the difference, we know from anecdotal evidence that the pre-policy outcome had different production inputs (particularly the volume of staff) and different production technology (e.g. IT systems). The full policy impact relative to the pre-policy situation would include the impact of these changes as well as the discontinuity in incentives introduced by the target.

We refer to our results as the 'impact of the target' for brevity but with the above understanding in mind. This interpretation applies to the results for wait times and other outcomes.

4.3.3 Treatment decisions and mortality outcomes

We now extend the analysis to consider outcomes other than the wait time, such as treatment decisions (e.g. inpatient admission) and mortality outcomes. We first introduce some notation to define the different channels through which the target can affect outcomes and then show how we identify and estimate the 'distortion effects' of the target.

Composition and distortion effects

Letting y_t be an outcome (treatment decision or mortality outcome) and w_t be the wait time in regime $t \in \{0, 1\}$, we define two conditional expectation functions. The first is $E[y_t \mid w_t]$, which is the expected outcome conditional on the wait time. This allows us to express average outcomes (either in the targeted or non-targeted regime) for groups of patients located in different parts of the wait time distribution (either in the targeted or non-targeted regime). For example, the observed data can be written as $E[y_1 \mid w_1]$. It is also possible to think about $E[y_0 \mid w_0]$, outcomes in the absence of the target, and combinations such as $E[y_0 \mid w_1]$ which are the outcomes in the non-targeted regime for patients at certain points of the wait time distribution in the targeted regime.

We also define $E[y_t \mid w_1, w_0]$, which is the expected outcome for patients with wait time w_1 in the targeted regime and wait time w_0 in the non-targeted regime. This notation allows us to denote outcomes for groups of individuals that have had a change in wait time due to the target. For example, $E[y_t \mid w^- < w_1 \leq w^*, w^* < w_0 < w^+]$ is the expected outcome for post-threshold movers. Since we will repeatedly refer to this and other related groups, we abbreviate these conditioning inequalities in the following way: $E[y_t \mid \underline{w}_1^-, \underline{w}_0^+]$.

Using this notation we can decompose the observed outcomes in the pre-threshold period. Note that, from the wait time analysis, we know that the target causes a number of patients to shift from the post-threshold to the pre-threshold period ('post-threshold movers'). So with the target, outcomes in the pre-threshold period are a weighted-average of pre-threshold non-movers and post-threshold movers. Abbreviating the pre-threshold period as $\underline{w}_1^-$, outcomes can be written as

$$E[y_1 \mid \underline{w}_1^-] = \rho E[y_1 \mid \underline{w}_1^-, \underline{w}_0^-] + (1-\rho) E[y_1 \mid \underline{w}_1^-, \underline{w}_0^+], \tag{4.4}$$

where $\rho \equiv \big[F_0(w^*) - F_0(w^-)\big] / \big[F_1(w^*) - F_1(w^-)\big]$ and F_t is the *cdf* of wait times. The parameter ρ is defined by the observed and counterfactual wait time distributions, where ρ is the proportion of pre-threshold non-movers and $1 - \rho$ is the proportion of post-threshold movers.

The composition and distortion effects are then defined as follows.

Definition 1 (Composition effect). *The composition effect is the change in expected outcomes conditional on the wait time that occurs in the pre-threshold period because the target shifts some patients into this period from* *the post-threshold period:*

$$\begin{aligned}\Delta_C \equiv\ & \rho\big(E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]\big) \\ & + (1-\rho)\big(E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]\big) \qquad (4.5) \\ =\ & (1-\rho)\big(E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]\big). \qquad (4.6)\end{aligned}$$

Definition 2 (Distortion effect)**.** *The distortion effect is the change in expected outcomes conditional on the wait time that occurs in the pre-threshold period because the target has a direct effect on the outcomes in each regime:*

$$\begin{aligned}\Delta_D \equiv & \rho\big(E[y_1 \mid \underline{w}_1^-, \underline{w}_0^-] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]\big) \\ & + (1-\rho)\big(E[y_1 \mid \underline{w}_1^-, \underline{w}_0^+] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+]\big). \qquad (4.7)\end{aligned}$$

With these definitions the observed outcomes in the pre-threshold period can be written as

$$\underbrace{E[y_1 \mid \underline{w}_1^-]}_{\text{Targeted regime (observed)}} = \underbrace{E[y_0 \mid \underline{w}_0^-]}_{\text{Non-targeted regime}} + \underbrace{\Delta_C}_{\text{Composition effect}} + \underbrace{\Delta_D}_{\text{Distortion effect}} \tag{4.8}$$

which can be verified by substituting in Equations (4.4), (4.6) and (4.7) and rewriting the non-targeted regime outcome as $E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]$.

Identification of the distortion effect

To identify the distortion effect we make use of the following definition.

Definition 3 (Composition-adjusted counterfactual)**.** *The composition-adjusted counterfactual (CAC) is the outcomes from the non-targeted regime in the pre-threshold period that would occur in the presence of the composition ef-*

fect only:

$$\begin{aligned} E[y_0 \mid \underline{w}_1^-] &\equiv E[y_0 \mid \underline{w}_0^-] + \Delta_C \qquad (4.9) \\ &= \rho E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-] + (1-\rho) E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+]. \qquad (4.10) \end{aligned}$$

where the second line follows from the definition of Δ_C*.*

With this definition it is straightforward to show that the distortion effect is identified as the difference between the observed data and the CAC: $\Delta_D = E[y_1 \mid \underline{w}_1^-] - E[y_0 \mid \underline{w}_1^-]$. Moreover, Equation (4.10) shows the CAC can be constructed as a weighted average of the counterfactual outcomes for two groups, the pre-threshold non-movers and the post-threshold movers, where the weights can be constructed from the observed and counterfactual wait time distributions.

Estimating counterfactual outcomes

We now revisit the bunching estimator and show it can be used to obtain the counterfactual outcomes in Equation (4.10). We require two assumptions for this purpose.

Assumption 2 (Local outcome effects)**.** *Outcomes outside of an 'exclusion window', defined locally around the threshold* w^**, are unaffected by the target:*

$$E[y_1 \mid w_t] = E[y_0 \mid w_t] \qquad \forall w \notin [w^-, w^* + \varepsilon]. \tag{4.11}$$

Assumption 2 rules out distortion effects outside of the pre-threshold period. It is the parallel of Assumption 1 for the conditional expectation function. In this case the exclusion window ends at $w^* + \varepsilon$, where ε is a small 'overhang period' that extends past the four-hour threshold.

The overhang period allows for the empirical fact that the bunching in outcomes extends slightly past the threshold (see Figure 4.2). We interpret the overhang as being a case of distortion effects for patients that are narrowly discharged or admitted after the threshold. For example, it may be that doctors admit additional patients in attempts to meet the target but not all of the excess admits occur prior to the threshold as some patients may be delayed for unexpected reasons. We determine the size of the overhang period visually, setting $\varepsilon = 20$ in the baseline analysis, and note that our findings are robust to more conservative (larger) overhang periods.[25]

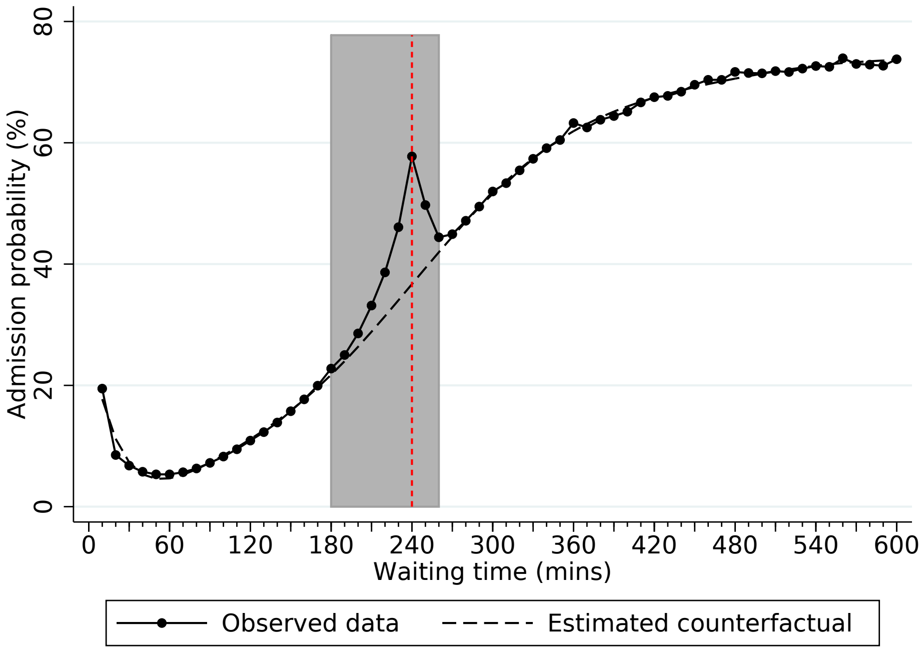
Assumption 3 (No-selection)**.** *Non-targeted regime outcomes conditional on the wait time are comparable for post-threshold movers and post-threshold non-movers:*

$$E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+] = E[y_0 \mid \underline{w}_1^+, \underline{w}_0^+] \tag{4.12}$$

Assumption 3 rules out composition effects in the post-threshold period. It states that after conditioning on the wait time, there is no selection when the post-threshold movers are assigned. This assumption is consistent with doctors randomly selecting which patients get a shorter wait time in response to the target. While this is strong assumption we believe it is plausible. For example, doctors routinely work with incomplete information and this will be exacerbated when they are forced to make earlier admission or discharge decisions (e.g. they may not yet have conducted all tests, or received all test results) and, as a result, may not be able to systematically select which patients to move forward. Importantly, we are also able to evaluate this assumption empirically using placebo tests and discuss this further below.

[25] Our estimates of the distortion effect, which relate to the pre-threshold period, do not capture distortions in the overhang period. These omitted effects are small: the number of patients in the overhang period is 1.3% of the number of patients in the pre-threshold period.

Figure 4.6: Estimated counterfactual admission probability conditional on wait times



Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy; (4) The estimated counterfactual is obtained from a polynomial regression that omits the exclusion window shown in grey.

Together Assumptions 2 and 3 imply that there are no composition or distortion effects outside of the exclusion window $[w^-, w^* + \varepsilon]$. We can therefore apply the bunching estimator in the same way as before but to the conditional expectation function $E[y_1 \mid w_1]$. The estimated counterfactual delivered by the bunching estimator is then $E[y_0 \mid w_0]$. This directly gives us $E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]$ and, given Assumption 3, also provides us with $E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+]$, which are the two terms required to construct Equation (4.10).

Figure 4.6 presents an example showing the observed data and our estimated counterfactual for the likelihood of inpatient admission, where the exclusion window is highlighted in grey and we have set $\varepsilon = 20$.

Testing for distortion effects

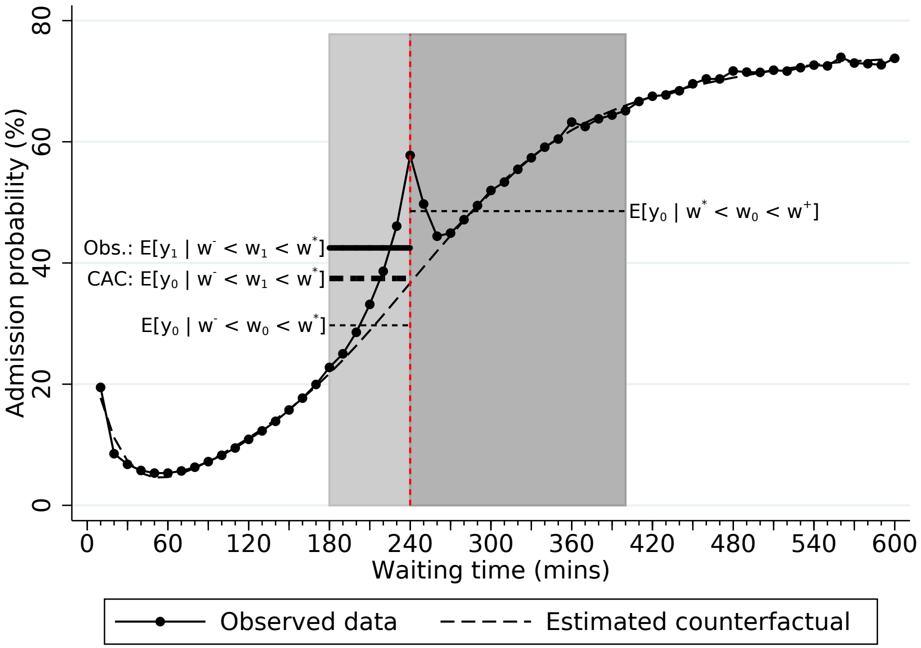
Recalling the definition $\Delta_D = E[y_1 \mid \underline{w}_1^-] - E[y_0 \mid \underline{w}_1^-]$, and noting that this can now be constructed from the observed data and Equation (4.10), the test for distortion effects is simply a hypothesis test that $\Delta_D = 0$. Estimates of this difference and tests of this null hypothesis form the central results of this chapter. We compute statistical significance for the test using non-parametric bootstrapped standard errors clustered at the hospital organisation level.[26]

Figure 4.7 provides a visual example of how we construct the CAC and the test of distortion effects for the probability of inpatient admission. The pre- and post-threshold periods are shown in different shades of grey. In each of these periods the horizontal thin dashed line gives the conditional expectation in Equation (4.10). The CAC, which is a weighted average of these two conditional expectations, is shown in the horizontal thick dashed line in the pre-threshold period.[27] In comparison, the horizontal thick solid line in the pre-threshold period is the mean observed outcome in the pre-threshold period. Finally, the difference between the thick solid and dashed line is the distortion effect, Δ_D, which shows that the observed admission probability in the pre-threshold period is too high to be explained by the composition effect alone. In this case we can reject the null hypothesis that $\Delta_D = 0$.

[26] Throughout the analysis we cluster results at the trust (organisation) level. NHS trusts include groups of one or more hospitals in close geographical proximitiy that share common management. We do not use hospital site codes due to some organisations entering data only at the trust level. All results are robust to clustering at the site level.

[27] The weights are obtained from the wait time distributions shown in Figure 4.3.

Figure 4.7: Constructing the composition-adjusted counterfactual for admission probability



Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy; (4) The horizontal thin dashed lines in the light grey (dark grey) region give the counterfactual outcome in the pre-threshold (post-threshold) period, $E[y_0 \mid w_0]$; (5) The horizontal thick dashed line in the pre-threshold period is the composition-adjusted counterfactual, $E[y_0 \mid \underline{w}_1^-]$; (6) The horizontal thick solid line in the pre-threshold period is the observed observed admission probability, $E[y_1 \mid \underline{w}_1^-]$; (7) The distortion effect is the gap between the thick solid and dashed line,
$\Delta_D = E[y_1 \mid \underline{w}_1^-] - E[y_0 \mid \underline{w}_1^-]$.

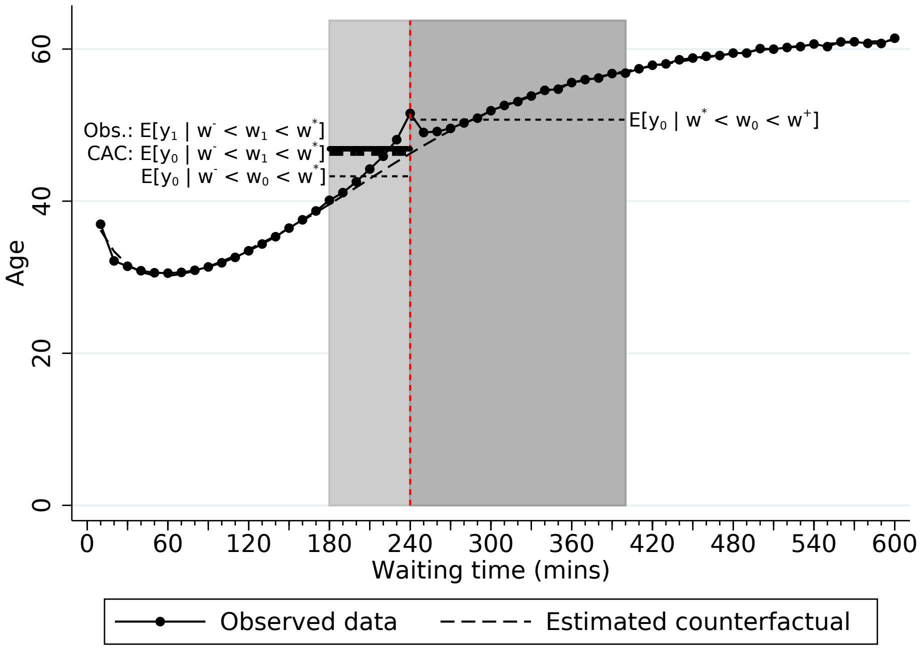
Testing the no-selection assumption

In Assumption 3 we rule out the possibility of non-random selection of post-threshold movers. By adapting our test for distortion effects, it is straightforward to generate placebo tests of this assumption based on observable patient characteristics. The key insight that motivates this is that observed demographics, such as age or gender, are by definition subject to composition effects but not distortion effects. Testing the hypothesis that $\Delta_D = 0$ for any demographic variable is therefore equivalent to testing the no-selection assumption.

This 'demographic test' acts as a placebo test, since we are testing for effects in situations where it is known that none should exist. To the extent that these tests indicate that the no-selection assumption does not hold, our estimated distortion effects will be a combination of distortion and composition effects.

Figure 4.8 provides a visual example of the demographic test using age, which follows the same format as Figure 4.7. There is again bunching at the four-hour threshold but in this case it cannot be explained by any distortion effects because patient age is unaffected by hospital treatment decisions. Comparing the observed data and the CAC shows that these now lie very close to one another and indeed a hypothesis test cannot reject the null hypothesis that $\Delta_D = 0$. This is consistent with the no-selection assumption: the mean age of post-threshold movers is comparable to the mean age of all post-threshold patients.

Figure 4.8: Demographic test of the no-selection assumption using age



Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy; (4) The horizontal thin dashed lines in the light grey (dark grey) region give the counterfactual outcome in the pre-threshold (post-threshold) period, $E[y_0 \mid \underline{w}_0^-]$; (5) The horizontal thick dashed line in the pre-threshold period is the composition-adjusted counterfactual, $E[y_0 \mid \underline{w}_1^-]$; (6) The horizontal thick solid line in the pre-threshold period is the observed observed admission probability, $E[y_1 \mid \underline{w}_1^-]$; (7) The distortion effect is the gap between the thick solid and dashed line, $\Delta_D = E[y_1 \mid \underline{w}_1^-] - E[y_0 \mid \underline{w}_1^-]$.

4.4 Results

We begin this section by first presenting the wait time results. We then present results from the placebo tests of the no-selection assumption, and finally turn to the results concerning treatment decisions and mortality outcomes. We explore the mechanisms behind the mortality outcomes in Section 4.5.

4.4.1 Wait times

Figure 4.3 shows the observed wait time distribution and our estimated counterfactual distribution. The shaded panel is the exclusion window where we estimate the effects of the policy, covering the period between 180 and 400 minutes. The solid line is the observed distribution of patients that exit at each interval and the dashed line is the estimated counterfactual distribution. The effect of the target on exit times is clear: a large proportion of patients from the post-threshold period (240 to 400 minutes) are moved to the pre-target period (180 to 240 minutes); these are the patients we refer to as post-thresholder movers. By comparing the observed wait time distribution with our counterfacutal we can compute the impact of the target on average wait times.

The results indicate that the target is successful in achieving its primary aim of reducing wait times. We estimate that the target reduces mean wait times by 7 minutes. This is equivalent to 4% of the estimated counterfactual mean. For patients affected by the target (i.e. in the exclusion window), we estimate that the target reduces wait times by 19 minutes, or 8% of their estimated counterfactual mean.

4.4.2 Demographic tests

Table 4.2 presents the results of the demographic tests. Column (1) presents estimates of the distortion effect and column (2) presents estimates of the distortion effect as a proportion of the counterfactual mean. Panel A presents results using individual demographic variables, where we test using age, a male indicator, and an indicator for whether the patient arrived via ambulance. We cannot reject the hypothesis of no distortions for age and ambulance-arrival, which supports the plausibility of the no-selection assumption. In contrast, we reject the hypothesis of no distortions for the male indicator. This result indicates that post-threshold movers are more likely to be female than the post-threshold non-movers. However, the extent of this selection effect is small: the difference between the observed and composition-adjusted counterfactual proportion of females in the pre-target period is 0.5 percentage points (1.1% of the baseline).

Panel B in Table 4.2 presents results for variables that are linear combinations of the three individual demographic variables. We use predicted admission and predicted mortality, where the predictions are obtained from linear regressions of the outcome on a fully interacted set of male, age-category, and ambulance indicators. The R^2 statistic from these predicted regressions is 0.21 and 0.06. The demographic tests for these predicted variables cannot reject the hypothesis of no distortion. The implication of these tests is that even though the gender test rejects the hypothesis, the contribution of gender to salient medical outcomes, as measured by predicted admission and mortality, is low.

Together these results indicate that, with only gender as a minor exception, the demographic tests support the no selection assumption. In practice this means that patients observed with wait times in excess of 240 minutes

Table 4.2: Demographic tests of the no-selection assumption

	Distortion effect (Δ_D)		CAC mean
	Level	%	Level
	(1)	(2)	(3)
Panel A: Individual characteristics			
Age	0.417	0.009	46.468
	(0.284)	(0.006)	
Male	-0.005^{***}	-0.011^{***}	0.487
	(0.001)	(0.003)	
Ambulance	-0.002	-0.005	0.440
	(0.004)	(0.010)	
Panel B: Predicted characteristics			
Predicted admission	0.002	0.006	0.323
	(0.002)	(0.007)	
Predicted mortality	0.000	0.015	0.019
	(0.000)	(0.015)	

Notes: (1) CAC mean is measured over the pre-threshold period, $E[y_0 \mid \underline{w}_1^-]$; (2) Predicted admissions and mortality use regressions with fully interacted variables from Panel A; (3) Bootstrapped standard errors clustered at the hospital trust level (199 repetitions).

(post-threshold non-movers) are comparable to those patients that would have had wait times in excess of 240 minutes in the absence of the target (post-threshold movers), and we can therefore use these post-threshold non-movers as the counterfactual for the post-threshold movers.

4.4.3 Treatments and mortality outcomes

Table 4.3 presents results of the distortion test for a range of treatment decisions and costs. Each row shows results for a separate outcome. Column (1) presents estimates of the distortion effect and column (2) presents estimates of the distortion effect as a proportion of the counterfactual mean.

Panel A presents estimates for treatment decisions in the ED. We find that, controlling for compositional changes, there is an increase in the odds

Table 4.3: Estimated distortion effects of the target on treatment decisions and costs

	Distortion effect (Δ_D)		CAC mean
	Level	%	Level
	(1)	(2)	(3)
Panel A: ED treatment decisions			
Pr(admission)	0.046***	0.122***	0.379
	(0.008)	(0.022)	
Pr(discharge)	−0.033***	−0.070***	0.472
	(0.007)	(0.014)	
Pr(referral)	−0.013***	−0.089***	0.150
	(0.003)	(0.020)	
ED investigations	0.108**	0.046**	2.369
	(0.048)	(0.021)	
ED treatments	−0.033	−0.016	2.070
	(0.028)	(0.014)	
Panel B: Inpatient treatment decisions			
Length of stay (days)	0.035	0.015	2.302
	(0.048)	(0.021)	
Inpatient procedures	0.000	0.001	0.290
	(0.006)	(0.020)	
Panel C: Hospital costs			
30-day ED cost	3.040***	0.016***	192.950
	(0.911)	(0.005)	
30-day inpatient cost	125.793***	0.052***	2,414.087
	(33.992)	(0.015)	
30-day total cost	128.833***	0.049***	2,607.037
	(34.389)	(0.014)	

Notes: (1) CAC mean is measured over the pre-threshold period, $E[y_0 \mid \underline{w}_1^-]$; (2) Predicted admissions and mortality use regressions with fully interacted variables from Panel A; (3) All inpatient variables (e.g. length of stay, costs) take on the value zero for patients that are not admitted; (4) Bootstrapped standard errors clustered at the hospital trust level (199 repetitions).

of admission of 4.6%. This is 12.2% of the baseline composition-adjusted counterfactual value, which is sizeable. The results for discharges and referrals out of the ED to specialist clinics or hospitals offset these admission effects, with roughly three-quarters of the effect coming from decreased discharges, and one-quarter from decreased referrals, although as a percentage of the baseline these responses are of comparable magnitude.

We also show target affects on the number of investigations performed in the ED, such as x-rays, blood tests and CT scans. We find that investigations rose by 0.1 per patient, or 4.6% of the baseline. We do not, however, find any effect on the number of treatments performed in the ED. This suggests that doctors perform more tests in order to speed up the admission decision for individuals (i.e. they perform an extra test instead of monitoring the patient for a longer period of time) but has little effect on the treatments that they provide in the ED.

Panel B examines inpatient treatment decisions. For inpatient treatments, in order to avoid selection, we include all ED patients, even those who did not end up being admitted. As a result, the coefficient represents the incremental amount of treatment due to the four-hour target. We find no evidence of any statistically significant increases in length of stay or the number of procedures. This suggests that the extra admissions do not receive substantial amounts of care in the hospital. That is, these admissions appear to be largely placeholders in order to avoid the four-hour target.

Nevertheless, the additional admits are costly. Panel C of Table 2 examines the impact of the four-hour target on 30-day patient costs. There is a small rise in ED costs of $3, or 2% of ED costs. But there is a significant increase in inpatient costs of $126, which is 5% of inpatient costs. That is, even though most patients appear to be only housed in inpatient depart-

Table 4.4: Estimated distortion effects of the target on mortality

	Distortion effect (Δ_D)		CAC mean
	Level	%	Level
	(1)	(2)	(3)
30-day mortality	-0.004^{***}	-0.138^{***}	0.029
	(0.001)	(0.019)	
90-day mortality	-0.004^{***}	-0.079^{***}	0.048
	(0.001)	(0.019)	
1-year mortality	-0.003^{*}	-0.031^{*}	0.090
	(0.002)	(0.017)	

Notes: (1) CAC mean is measured over the pre-threshold period, $E[y_0 \mid \underline{w}_1^-]$; (2) Bootstrapped standard errors clustered at the hospital trust level (199 repetitions).

ments as a way of avoiding the four-hour target, these admissions generate transfers from the government to hospitals. Total costs rise by roughly 5% relative to the baseline.

Table 4.4 then extends our analysis to look at patient mortality outcomes. We consider mortality at a variety of time frames, ranging from 30 days after entering the ED to 1 year later. We find significant short term declines in mortality. Mortality over 30 days declines by 0.4 percent, or 14% of baseline. This effect fades slightly over time and falls as a share of the baseline, so that at one year it is only 3.1% of baseline. This pattern suggests that the health benefits of the four-hour policy are seen in the short term.

This is a sizeable mortality decline given the modest increase in costs documented in Table 4.3. We find that total costs over 30 days from admission to the ER rise by 5%, while mortality falls by 3.1% over a year. Calculating the cost per year of life saved by the policy requires assumptions on how long-lasting is the impact on mortality and on any subsequent costs past 30 days. Assuming no subsequent costs, but also assuming that the mortality impact only lasts one year, this implies a cost per year of life saved of \$43,000.[28] This is low relative to standard valuations of a life-year in the U.S., where typical benchmarks are around \$100,000 (Cutler, 2003), and at the upper end of valuations in the U.K., where the national benchmarks are set at \$28,000 to \$42,000 (McCabe et al., 2008).

In summary, then, our analysis of the four-hour target shows that it led to shorter wait times, more admission, only marginal additional costs (due to little use of inpatient care for those admitted), and significant reductions in mortality. That is, it appears that constraining hospitals did save lives.

4.5 Using Patient Heterogeneity to Identify Mechanisms

Our results so far show a number of effects of the wait time target on patient treatment – on wait times, admission probabilities, and treatment costs more generally. We also show a significant effect on patient mortality. Ideally we would like to uncover the mechanism through which the four-hour target impacts patient mortality. This is difficult since we essentially have one instrument (the target) and multiple changes in patient treatment.

To address this issue we turn to considering heterogeneous impacts across types of patients. That is, we examine whether there are groups of patients where there are differential effects of the four-hour target. If those groups have effects that are focused along one channel (e.g. wait times) but not another (e.g. admits), then we can use this to separate the effect of the two channels on outcomes.

[28] This reflects the cost to the government of the policy due to the increase in HRG transfers to hospitals. The actual cost in terms of resource-use will be even lower if the marginal admissions due to the policy use fewer resources than the average HRG cost.

In particular, we consider two natural sources of heterogeneity. The first is differences across diagnosis. In particular, we divide patients into 36 diagnosis groups.[29] It seems likely that the largest wait time impacts of the target will show up for those who have the most severe diagnoses, since they are the most likely to hit the wait time target. Indeed, Figure 4.9 graphs the proportion of patients hitting the wait time target (in the counterfactual wait time distribution) against the severity of the diagnosis. Severity is measured by mean predicted 30-day mortality for patients within each diagnosis. In fact, we see that the odds of hitting 240 minutes are much higher for the most severe diagnoses.

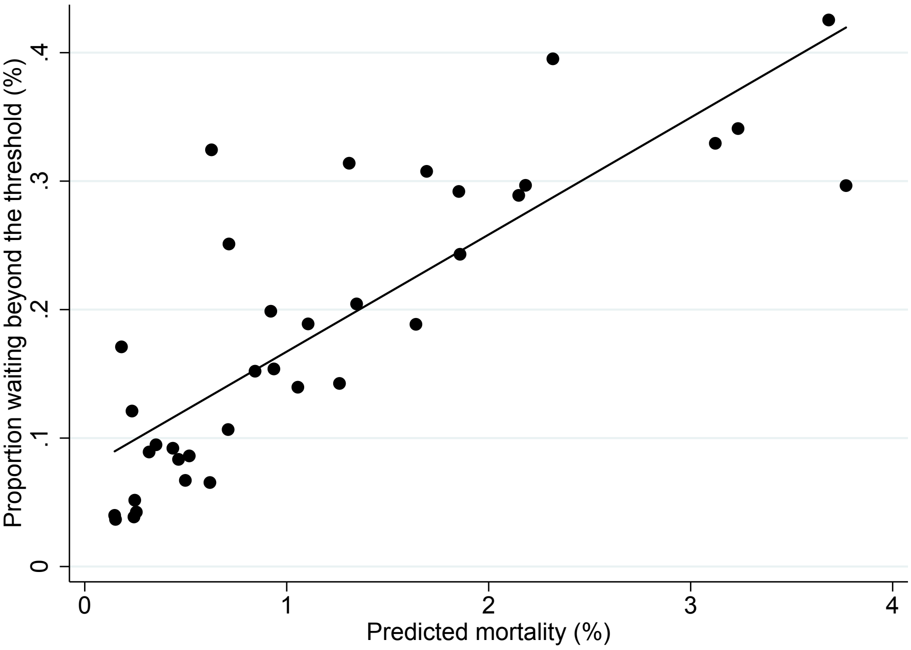
We therefore separately compute the wait time reduction effects, and distortion effects for admissions and 30-day mortality for each diagnosis group. We then assess how the heterogeneity across diagnosis groups translates to each of these outcomes.

The results of this exercise are shown in Figure 4.10. Panel A shows that higher severity diagnoses have larger wait time effects. This is sensible since they are most likely to wait the longest without the four-hour policy. But Panel B shows that the effects of the target on hospital admissions is no higher for more severe diagnoses. That is, the more severe diagnoses are getting treated sooner, but are no more likely than others to have that treatment resolve in an extra hospital admission.

Panel C shows the differential treatment effect on mortality by diagnosis category, where black circles correspond to actual mortality outcomes and red triangles correspond to predicted mortality outcomes. The y-axis shows the absolute value of mortality reduction, so that a larger value means a

[29] The data assign patients to 40 diagnosis categories, including a 'missing' category. We exclude four diagnoses (nerve injuries, electric shock, near drowning and visceral injury) as small samples do not allow us to separately estimate the impact of the target for these groups.

Figure 4.9: Proportion wait beyond the threshold vs. predicted mortality by diagnosis groups

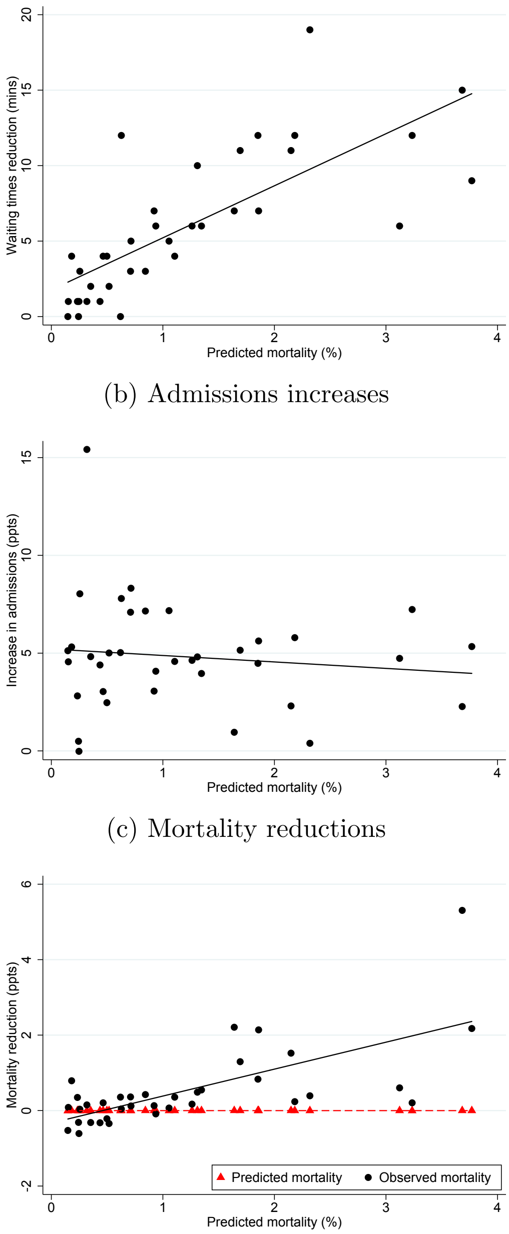


Notes: (1) Each data point corresponds to a diagnosis group average; (2) Proportion waiting beyond the threshold defined using the counterfactual distribution of wait times; (3) Predicted mortality defined using a regression of 30-day in-hospital mortality on a fully interacted set of age, gender, ambulance arrival fixed effects and diagnosis fixed effects.

larger mortality reduction. Looking at the black circles, there is a clear upward slope showing that the mortality effect of the four-hour target is strongest for the most severe diagnoses. To ensure that selection is not driving our result, the graph also repeats this exercise for predicted mortality. If our assumption of no-selection (Assumption 3) holds, these effects should not be statistically different from zero. The red triangles shows that this is indeed the case, with all estimates clustered around zero and no systematic relationship between the effects of the target on predicted mortality and the severity of the diagnosis.

Figure 4.10: Estimated effects of the target vs. predicted mortality by diagnosis groups

(a) Wait times reductions



(b) Admissions increases

(c) Mortality reductions

Notes: (1) Each data point corresponds to a diagnosis group average; (2) Predicted mortality defined using a regression of 30-day in-hospital mortality on a fully interacted set of age, gender, ambulance arrival fixed effects and diagnosis fixed effects.

Given that there is an effect on wait times, but not admissions, this suggests that it is wait time reductions and not increased admissions that are driving the results. Of course, this set of corresponding facts do not prove this causal mechanism because there may be other factors that cause the effects to differ by diagnosis. So to further test this conclusion we consider a second source of heterogeneity.

We next turn to heterogeneity by the degree of inpatient crowding. In times where the inpatient department is more crowded, EDs may be less able to address their wait time targets by admitting patients because the inpatient wards have less spare capacity for these patients to be sent. But it is unclear that inpatient crowding would much affect the marginal wait time impacts of the target. Inpatient crowding therefore provides an opposite test of the diagnosis heterogeneity: an opportunity to observe heterogeneity that drives admission probabilities but not wait times.

To assess this, we divide the data into 50 quantiles depending on how busy the hospital inpatient department is on the day of admission. For each hospital-day, we calculate the daily number of inpatients treated by the hospital, and use this to assign each hospital-day to one of 50 groups in the hospital-specific distribution of inpatient crowding. Patients are then assigned to each of these groups depending on their day of arrival.[30] To address differences in casemix during busy and quiet periods, we also split patients into two severity groups. 'Major' diagnoses are defined as those with a 30-day mortality rate above the overall 30-day mortality rate (1.6%). Interacting the 50 inpatient crowding groups with severity yields 100 groups. For 95 of these groups we have sufficient sample size to independently compute the effects of the target, and therefore across which to examine heterogeneity

[30] We calculate the inpatient census at the daily level as the data do not contain information on time of arrival at, or discharge from, the inpatient department.

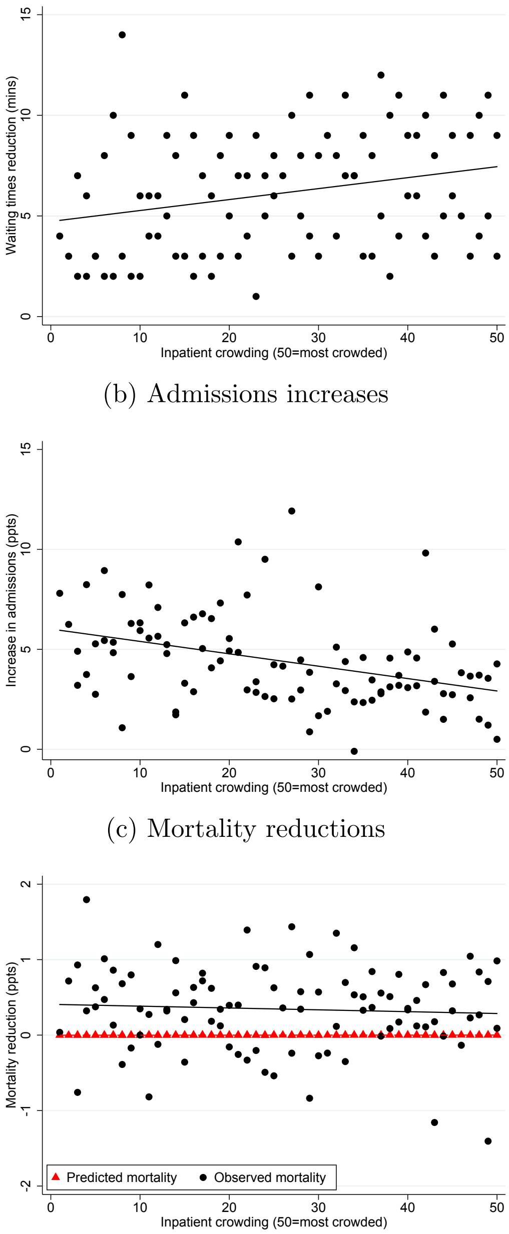
in effects.

Figure 4.11 presents the results of this second heterogeneity test. The figure shows the results for these observations, ranked from least crowded to most crowded. Panel A shows that inpatient crowding has a weak, positive relationship with wait times. Panel B shows a strong, negative relationship between crowded inpatient departments and smaller increases in admission. So this source of heterogeneity gives the opposite results of what we saw for severity: a small effect on wait times and a large effect on admissions. Therefore, if our earlier supposition is correct that it is wait times and not admissions that drives our mortality effects, we should see little differential impact on mortality across these groups.

In fact, that is exactly what we see in Panel C in the black circles: there is no significant relationship between the degree of inpatient crowding and the estimated mortality effect. As in Figure 4.10c, we repeat this analysis with estimated reductions in predicted mortality (which should be unaffected by the target once we adjust for the composition of patients) to show that these results are not driven by selection. The red triangles show that the predicted mortality effects are again all close to zero, with no significant relationship between predicted mortality reductions and inpatient crowding.

We formalize these graphical results in Table 4.5. The unit of analysis in this table is either diagnosis groups (columns 1-3) or inpatient crowding by severity (columns 4-6). The dependent variable is the distortion effect on mortality in absolute value for each group. The independent variables are the estimated wait time reduction and the distortion effect for admission probability. Essentially, these regressions report associations between the estimated impact on mortality and the estimated impact on wait times and admissions, using a grouping estimator with groups defined by severity or

Figure 4.11: Estimated effects of the target vs. inpatient crowding by crowding-severity groups



(a) Wait times reductions

(b) Admissions increases

(c) Mortality reductions

Notes: (1) Each data point corresponds to a crowding-severity group average; (2) Predicted mortality defined using a regression of 30-day in-hospital mortality on a fully interacted set of age, gender, ambulance arrival fixed effects and diagnosis fixed effects.

inpatient crowding. A positive coefficient in these regressions can be interpreted as that margin being associated with a larger policy effect on 30-day mortality.

Column 1 shows that across the 36 diagnosis groups, those groups with larger wait time effects have larger mortality effects. The estimated coefficient suggests that each additional minute of wait time reduction increases the mortality reduction by 0.001 percentage points. Earlier, we estimated that wait times fell by 19 minutes on average. This suggests a mortality reduction of 2.2 percentage points. This is of a similar magnitude to our reduced form estimate in Table 4.4 of 3-4 percentage points. Column 2, however, shows that there is no impact of the increase in admissions on mortality. And column 3 shows it is still the case that groups with larger wait time effects, but not larger admit effects, have larger mortality effects when we consider both variables together.

Columns 4-6 repeat this exercise using the estimates by inpatient crowding and patient severity. Once again, we have a highly significantly relationship between the wait time reduction and mortality reduction, with a coefficient that is similar to column 1. In this case, in column 5, we do see a significant effect of the admissions effect on mortality, albeit with a wrong signed coefficient suggesting that a larger admissions effect leads to a smaller mortality effect. But when both are included in column 6 only the wait time effect persists.

These results are not surprising given the graphical evidence shown above. The bottom line is that heterogeneity associated with wait time variation appears associated with mortality variation, while heterogeneity associated with admissions variation does not. This does not prove that the wait time reductions are driving our mortality reductions, but it is highly

Table 4.5: OLS regressions of the estimated 30-day mortality reductions on other effects of the target

	Diagnosis groups			Crowding-severity groups		
	(1)	(2)	(3)	(4)	(5)	(6)
Wait time	0.118^{***}		0.115^{***}	0.083^{***}		0.066^{***}
	(0.034)		(0.034)	(0.018)		(0.022)
Admission probability		−0.059	−0.029		-0.088^{***}	−0.037
		(0.065)	(0.058)		(0.024)	(0.028)
N	36	36	36	95	95	95

Notes: (1) Dependent variable is the absolute value of the target impact on 30-day mortality measured as % of the CAC mean over the pre-threshold period; (2) Independent variables are the absolute value of the target impact on the respective variable, measured as a % of the CAC mean over the pre-threshold period.

suggestive.

4.6 Conclusion

The Emergency Department is a central node of health care delivery in developed countries around the world. It is the entry point into the hospital for a large share of patients and decisions made rapidly by ED staff have fundamental impacts on the entire course of care. Despite the complicated nature of these decisions, there remains dissatisfaction in most health care systems with the level of crowding in EDs and the speed with which cases are resolved. This has led in recent years to both open competition on ED wait times and to regulatory interventions to reduce those times.

We study one type of regulatory intervention, the four-hour wait target policy enacted in England. We find that this target had an enormous effect on wait times, as illustrated vividly by the spike in the wait times distribution at the four-hour mark. We use well-established bunching methodologies to estimate that this represents a significant reduction of almost 20 minutes, or 8%, in the average wait time of impacted ED patients.

We then turn to assessing how this change in wait times impacted patient care and outcomes. We do so by introducing an econometric framework that allows us to separate the compositional impacts of individuals shifting from after to before the four-hour target from the distortionary effect of the four-hour target on medical decisions. We find this target led to a significant rise in hospital admissions. These admissions do not appear to involve much new treatment, suggesting that they may just be ‘placeholders’ to meet the target. But there is nonetheless a significant rise in inpatient spending of about 5% of baseline.

At the same time, we find striking evidence that the target is associated with lower patient mortality. There is a 0.4 percentage point reduction in patient mortality that emerges within the first 30 days, amounting to a large 14% reduction in mortality in that interval. This reduction fades slightly over time, so that after one year it amounts to a 3.1% mortality reduction. While modest, this effect is large relative to the extra spending, suggesting a cost of extending life by one year of $43,000. Finally, we exploit heterogeneity across patient types to show that this effect arises through reduced wait times, not through increased inpatient admissions.

The implications of our finding is that, unconstrained, EDs in England are not making optimal decisions on patient wait times. By reducing wait times, the four-hour target induced cost-effective mortality reductions. This is of likely a lower bound on the welfare gains due to the target, as it does not value the other benefits to consumers from waiting shorter times, although there may be welfare costs from the extra admissions (as discussed at length in Chapter 2).

Of course, this result only applies to the specific target studied here, and does not necessarily imply that other limits would have equal effects. It is also unclear how this result applies to other nations with different means of rewarding or incentivizing EDs. More work is clearly needed to understand the proper set of rules and incentives for delivering cost-effective ED care.

Chapter 5

Concluding remarks

As nations aim to improve their health care systems, by increasing quality of care while simultaneously managing costs, setting appropriate incentives for health care providers is critical. As the preceding chapters have illustrated, achieving this aim requires a deep understanding of hospital production. Economic research has a key role to play in developing this knowledge, and this thesis has made three new empirical contributions.

Two major themes stand out from the thesis. The first is that time itself is often an important part of the production process in hospitals. This stands in stark contrast to a production function in traditional settings such as manufacturing. In Chapter 2, for example, I show the importance of recognising the trade-off between hospital crowding and the time patients wait for hospital appointments. Even though the waiting time does not impact health outcomes in this setting, it plays an important role in the welfare analysis and I show that patients' preferences for waiting times are being undervalued by current economic policies.

The role of time is also central to Chapters 3 and 4. In these cases, spending more or less time physically in the hospital can positively or negatively impact patient health outcomes. Chapter 3 shows that longer stays in inpatient departments reduce the risk of needing further treatment. In sharp contrast, Chapter 4 shows that policies that limit the time spent by patients in the ED actually saves lives. The delicate balance then of how patients transition through the hospital can have a major impact on quality of care. While these aspects of health care are regulated, it is often informed largely by medical considerations, and incorporating economic considerations looks to be an interesting area for future research.

Following on from this, the second theme from the thesis is how to integrate economic analysis into a medical setting. A traditional cost-benefit analysis in such settings revolves around the monetary cost of an intervention and the quality-of-life benefits it delivers, the latter often monetized according to certain benchmarks. While useful as a starting point, this process can neglect aspects of welfare, such as preferences over the service received, or important interactions with adjacent areas of health care provision. For example, the thesis illustrates cases where several policies interact to regulate an underlying trade-off (Chapter 2), and where policies have unintended consequences that can be either negative (Chapter 3) or positive (Chapter 4).

The study in Chapter 2 is especially interesting, since readmissions and waiting are unlikely to affect long term health outcomes, yet these preferences are the basis upon which efficient policies should be set. Chapters 3 and 4 also illustrate the interaction between economics and medicine, where I use economic techniques to evaluate medical relationships that are central to questions of economic policy. Further integration between these fields of research appears critical to the success of economic policies in health care.

As I pursue my research agenda in the future, I expect these themes to play a central role and I look forward to developing them further.

Appendix A

Identification of average crowding effects

This appendix derives Equation (5) in Section 4.2.1. The derivation follows Angrist & Krueger (1999). By definition the population regression coefficient for a regressor Q (emergency admissions) in a regression of Y (patient health outcomes) on a constant, a discrete covariate D (diagnosis-age-emergency combinations) and the variable Q can be written

$$\beta^{OLS} = \frac{\mathbb{E}\left[(Q - \mathbb{E}[Q|D])\,\mathbb{E}[Y|D,Q]\right]}{\mathbb{E}\left[(Q - \mathbb{E}[Q|D])^2\right]}. \tag{A.1}$$

Labelling values of Q by $p = 0, ..., P$, the values of D by d, and abbreviating $\mathbb{E}[Y|D = d, Q = p]$ as $\mathbb{E}[Y|d,p]$, it is possible to write

$$\mathbb{E}[Y|d,p] = \mathbb{E}[Y|d,0] + \sum_{r=1}^{p} \left\{ \mathbb{E}[Y|d,r] - \mathbb{E}[Y|d,r-1] \right\} \tag{A.2}$$

$$= \mathbb{E}[Y|d,0] + \sum_{r=1}^{p} \Delta\beta_{dr} \tag{A.3}$$

where $\Delta\beta_{dr} \equiv \mathbb{E}[Y|d,r] - \mathbb{E}[Y|d,r-1]$. Substituting Equation (A.3) into (A.1) gives

$$\beta^{OLS} = \frac{\mathbb{E}\left[(Q - \mathbb{E}[Q|D]) \sum_{r=1}^{P} \beta_{dr}\right]}{\mathbb{E}\left[(Q - \mathbb{E}[Q|D])^2\right]} \tag{A.4}$$

$$= \frac{\mathbb{E}\left\{\, \mathbb{E}\left[(Q - \mathbb{E}[Q|D]) \sum_{r=1}^{P} \beta_{dr} \mid D\right]\right\}}{\mathbb{E}\left\{\, \mathbb{E}\left[(Q - \mathbb{E}[Q|D])^2 \mid D\right]\right\}} \tag{A.5}$$

where $\mathbb{E}[Y|d,0]$ cancels from the first line because it is uncorrelated with Q and the second line follows by iterating expectations over D in the numerator and the denominator. Writing out the expectations first with respect to Q and then D gives

$$\beta^{OLS} = \frac{\mathbb{E}\left\{\, \sum_p (Q - \mathbb{E}[Q|D]) \Pr(Q = p|D) \sum_{r=1}^{p} \beta_{dr}\right\}}{\mathbb{E}\left\{\, \sum_p (Q - \mathbb{E}[Q|D])^2 \Pr(Q = p|D)\right\}} \tag{A.6}$$

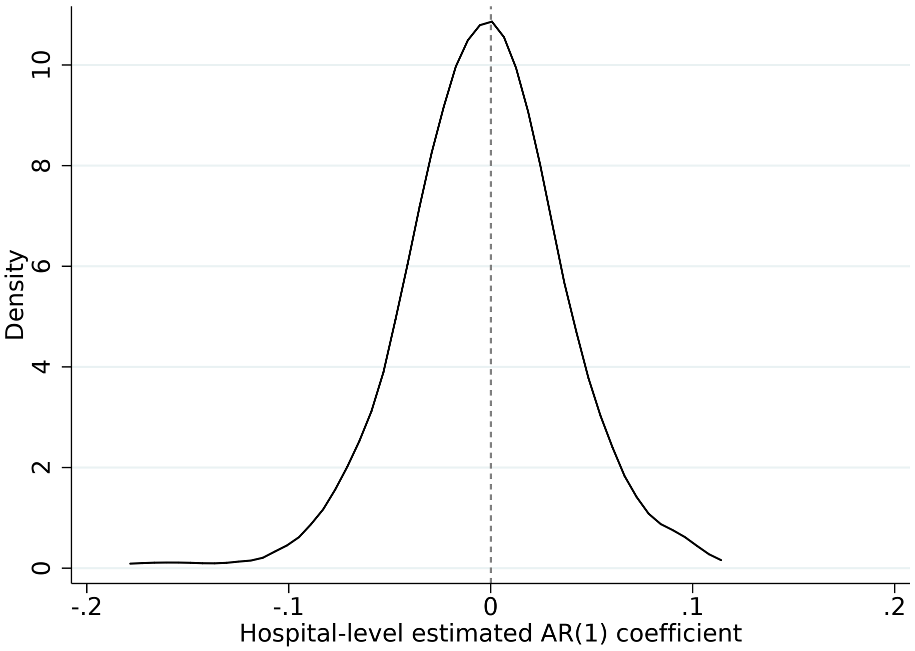
$$= \frac{\sum_d \sum_p (Q - \mathbb{E}[Q|D = d]) \Pr(Q = p|D = d) \Pr(D = d) \sum_{r=1}^{p} \beta_{dr}}{\sum_d \sum_p (Q - \mathbb{E}[Q|D = d])^2 \Pr(Q = p|D = d) \Pr(D = d)}. \tag{A.7}$$

Replacing each term in the final equation with its sample counterpart and rearranging gives Equation (5) in Section 4.2.1.

Appendix B

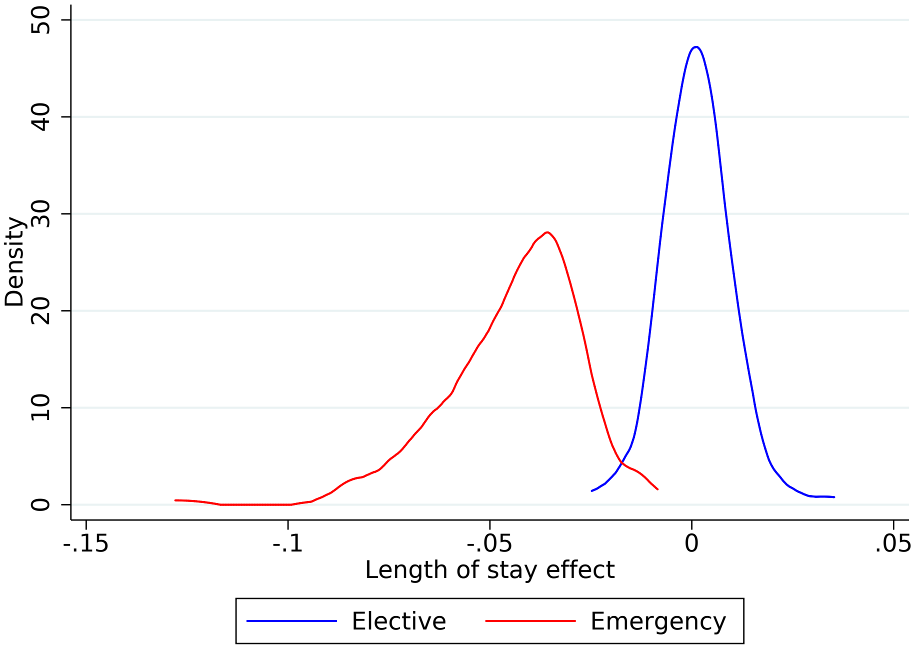
Additional charts and tables

Figure B.1: Hospital-level tests of first-order serial correlation



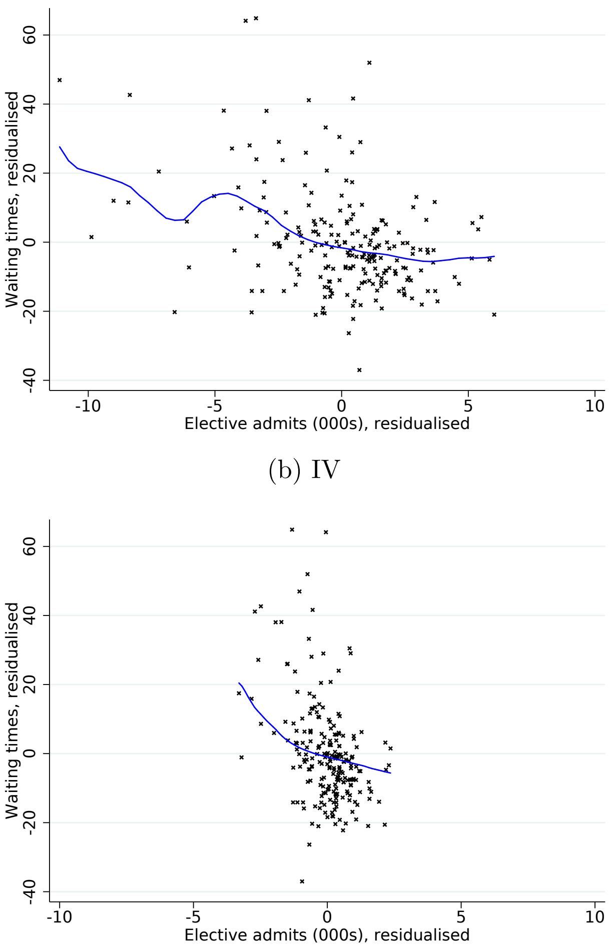
Notes: (1) Figure shows the density of estimated AR(1) coefficients from regressions of emergency shocks on their lag for each hospital separately; (2) Emergency shocks are defined as residuals from a regression of daily emergency admissions on hospital-specific year, weekly seasonal, and day-of-the-week fixed effects.

Figure B.2: Distribution of estimated hospital-level effects of emergency admissions on length of stay by patient type



Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable in the baseline specification estimated separately for each hospital.

Figure B.3: Non-parametric estimates of the effect of elective admissions on elective waiting times

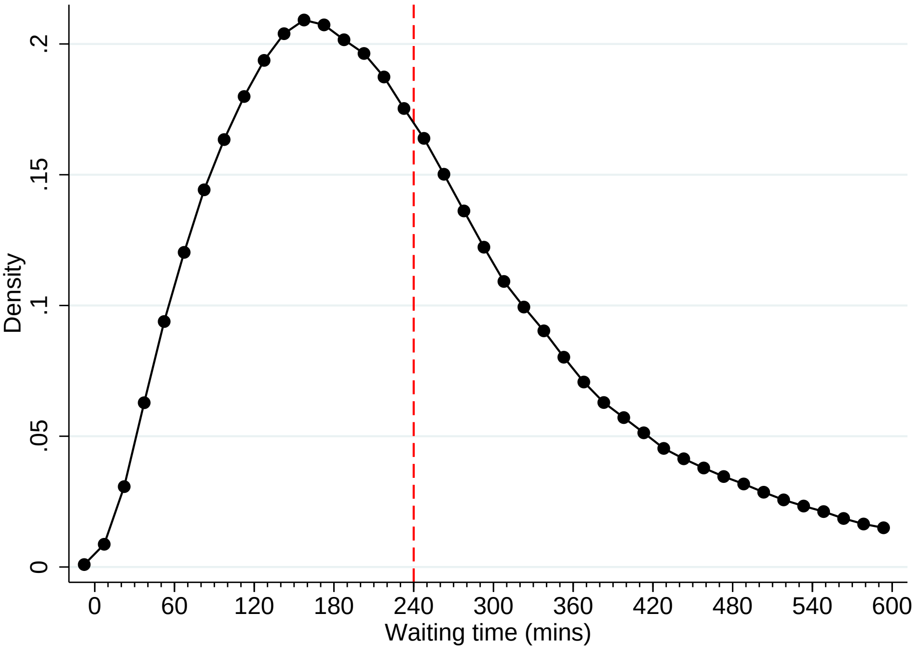


(a) OLS

(b) IV

Notes: (1) Kernel regression estimates of residualised elective waiting times at general acute hospitals on residualised elective admissions at general acute hospitals; (2) Residuals computed from a regression of the dependent variable on regional fixed effects and elective admissions at non-general acute hospitals.

Figure B.4: Distribution of wait times at a large hospital in California



Notes: (1) The English data displays a sharp discontinuity in the wait time distribution at four hours (see Figure 4.1). Here we present the wait time distribution from a large hospital in California to illustrate that the discontinuty in the English data is unlikely to naturally occur, and is instead induced by the target; (2) We thank David Chan for providing the data for this chart.

Table B.1: Summary statistics for the panel dataset

	Mean	Std dev	Min	Max	N
Elective admissions	7.1	7.9	0.0	86.0	338, 746
Emergency admissions	4.6	2.8	0.0	43.0	338, 746

Table B.2: Summary statistics for the inpatient dataset

	All patients
Patient characteristics	
Age	52.9
Male, %	48.4
White, %	85.4
Emergency, %	39.4
Diagnosis count	3.4
Charleson co-morbitidity index	1.7
ED admissions within past 12 months	0.8
Waiting time, days	84.8
Inpatient outcomes	
Daycase operation, %	22.7
Delayed operation, %	35.2
Number of operations	1.1
Length of stay, days	4.2
Transfers out, %	2.9
Home discharge, %	93.7
7-day unplanned readmission, %	2.8
30-day in-hospital death, %	1.1
N	$3,940,878$

Notes: (1) An operation is classified as a daycase if length of stay is zero in over 50% of cases; (2) An operation is classified as delayed if the patient did not receive their primary operation on the day of admission.

Table B.3: Summary statistics for the ED dataset

	All patients
Patient characteristics	
Age	39.7
Male, %	50.3
Ambulance pickup, %	29.4
Injury at or near home, %	54.0
ED outcomes	
Time in the ED, mins	149.9
Attended nearest hospital with T&O dept., %	83.8
Distance to nearest hospital with T&O dept., km	6.3
Inpatient admission, %	24.1
N	$22,519,392$

Notes: (1) T&O departments are defined as those contained in the panel dataset (i.e. general acute hospitals with an active ED).

Table B.4: Estimated effects of emergency status on length of stay

	(1)	(2)	(3)	(4)
Emergency	0.938***	0.559***	0.482***	0.463***
	(0.02)	(0.014)	(0.014)	(0.013)
Age category		✓	✓	✓
Age category			✓	✓
Gender			✓	✓
Ethnicity			✓	✓
Co-morbidity index			✓	✓
Past ED admissions			✓	✓
Hospital fixed effects				✓
Year fixed effects				✓
N	3,940,878	3,940,878	3,940,878	3,940,878

Notes: (1) Dependent variable is log(length of stay+1); (2) Diagnosis, age category, gender, ethnicity are specified as fixed effects, with fully interacted diagnosis and age categories; (3) Co-morbidities and past ED admissions enter the specific as linear terms; (4) Standard errors clustered at the hospital-level; (5) ***/**/* indicates statistical significance at the 1/5/10% level.

Table B.5: Top 10 causes of unplanned readmission

Diagnosis	Length of stay	N
Mechanical complication of internal joint prosthesis	12.9	4,538
Infection following a procedure	11.2	4,027
Infection and inflammation due to internal joint prosthesis	21.6	1,955
Haemorrhage and haematoma complicating a procedure	5.3	1,885
Other complications of internal devices	4.3	1,759
Cellulitis of axilla, hip or shoulder	5.2	1,442
Follow-up care involving removal of device	3.1	1,056
Disruption of operation wound	8.1	1,039
Infection and inflammation due to internal device	14.5	872
Other complications of procedures	1.5	871
All readmissions	7.3	77,392

Notes: (1) Diagnosis descriptions from ICD-10 codes; (2) Excludes diagnosis codes that are the same as the index admission.

Table B.6: Regression estimates of emergency shocks on admissions at other hospital departments

	(1)		(2)	
General surgery	0.614	(0.885)	−0.226	(1.324)
General medicine	0.222	(0.265)	−0.065	(0.550)
Cardiology			−0.963	(2.358)
Urology			1.351	(2.555)
Gastroenterology			−1.701	(1.927)
Paediatrics			0.731	(1.112)
Obstetrics			3.733*	(2.009)
Gynaecology			−1.498	(2.068)
N	317,958		141,640	

Notes: (1) Standard errors clustered at the hospital-level; (2) ***/**/* indicates statistical significance at the 1/5/10% level.

Table B.7: Estimated effects of emergency admissions on inpatient care and health outcomes using different horizon outcomes

Dependent variable	Coeff	Std error	N
Panel A: Admission cohorts			
7-day unplanned readmission	0.011***	(0.003)	3,940,878
15-day unplanned readmission	0.016***	(0.004)	3,940,878
30-day unplanned readmission	0.015***	(0.005)	3,940,878
7-day in-hospital mortality	0.003	(0.002)	3,940,878
15-day in-hospital mortality	0.002	(0.002)	3,940,878
30-day in-hospital mortality	0.003	(0.002)	3,940,878
Panel B: Discharge cohorts			
7-day unplanned readmission	0.047***	(0.007)	3,940,878
15-day unplanned readmission	0.047***	(0.006)	3,940,878
30-day unplanned readmission	0.040***	(0.006)	3,940,878
7-day in-hospital mortality	−0.002	(0.002)	3,940,878
15-day in-hospital mortality	−0.001	(0.002)	3,940,878
30-day in-hospital mortality	−0.001	(0.002)	3,940,878

Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable; (2) All specifications include a fully interacted set of diagnosis, age category, and emergency status fixed effects, and hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) Standard errors clustered at the hospital-level (149 clusters); (4) ***/**/* indicates statistical significance at the 1/5/10% level.

Table B.8: Estimated effects of emergency admissions by inpatient care and health outcomes with different sets of control variables

Dependent variable	(1)	(2)	(3)	(4)
Panel A: Admission cohorts				
Daycase operation	0.019^{*}	0.018^{*}	-0.004	-0.004
Delayed operation	0.218^{***}	0.185^{***}	0.203^{***}	0.204^{***}
Number of procedures	-0.002^{***}	-0.001^{***}	-0.001^{***}	-0.001^{***}
7-day unplanned readmission	0.011^{***}	0.011^{***}	0.011^{**}	0.012^{**}
30-day in-hospital mortality	0.005^{*}	0.003	0.004	0.005
Panel B: Discharge cohorts				
Length of stay	-0.009^{***}	-0.009^{***}	-0.008^{***}	-0.008^{***}
Transfers to other hospitals	-0.003	0.001	0.001	0.002
Discharges to home	0.001	0.009	0.005	0.065^{***}
7-day unplanned readmission	0.065^{***}	0.047^{***}	0.031^{***}	0.032^{***}
30-day in-hospital mortality	-0.009^{***}	-0.001	-0.002	0.000
Diagnosis-age-emergency FEs		✓	✓	✓
Gender			✓	✓
Ethnicity			✓	✓
Local area deprivation			✓	✓
Diagnosis count				✓
Co-morbidities				✓
Past ED admits				✓

Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable; (2) All specifications include a fully interacted set of diagnosis, age category, and emergency status fixed effects, and hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) N = 3,940,878 in all specifications; (4) Standard errors clustered at the hospital-level (149 clusters); (5) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

Table B.9: Estimated effects of emergency admissions on inflows of emergency patients

Dependent variable	Coeff	Std error	N
Panel A: Admission cohorts			
Attended nearest hospital with T&O department	0.516	(0.339)	22,519,392
Time spent in the ED	0.104^{***}	(0.025)	22,519,392
Inpatient admission	−0.001	(0.005)	22,519,392
Panel B: Admission cohorts–predicted T&O patients			
Attended nearest hospital with T&O department	0.492	(0.345)	16,397,024
Time spent in the ED	0.105^{***}	(0.025)	16,397,024
Inpatient admission	−0.001	(0.005)	16,397,024
Panel C: Admission cohorts–predicted T&O patients, home-ambulance pickups			
Attended nearest hospital with T&O department	0.899^{**}	(0.355)	2,531,304
Time spent in the ED	0.172^{***}	(0.044)	2,531,304
Inpatient admission	0.012	(0.014)	2,531,304

Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable (time spent in ED, inpatient admission) or the daily emergency admissions at the nearest hospital (attended nearest ED) on the day prior to the inflows of emergency patients; (2) All specifications include a fully interacted set of diagnosis, age category, and ambulance arrival fixed effects, and hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) The nearest ED is defined according to straight-line distances from the patient's home to the set of general acute hospitals in the panel dataset; (4) N = 22,519,392 / 16,397,024 / 2,531,304 in panels A / B / C; (5) Standard errors clustered at the hospital-level (149 clusters); (6) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

Table B.10: Estimated effects of emergency admissions on the characteristics of admitted elective patients

	Age	Male	White	Diagnosis count	Past ED admits	Co-morbidities	Waiting time
Emergency admits, t	−0.007	0.009	−0.002	−0.003***	0.000	−0.002**	0.028
	(0.006)	(0.015)	(0.01)	(0.001)	(0.000)	(0.001)	(0.031)
Emergency admits, t-1	0.003	0.005	0.015	−0.001	0.000	0.001	0.058**
	(0.006)	(0.014)	(0.011)	(0.001)	(0.000)	(0.001)	(0.024)
Emergency admits, t-2	−0.003	−0.005	0.006	−0.002***	0.000	0.000	−0.046
	(0.006)	(0.013)	(0.011)	(0.001)	(0.000)	(0.001)	(0.029)
N	2,369,066	2,369,066	2,369,066	2,369,066	2,369,066	2,369,066	2,369,066

Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable; (2) All specifications include hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) Standard errors clustered at the hospital-level (149 clusters); (4) $^{***}/^{**}/^{*}$ indicates statistical significance at the 1/5/10% level.

Table B.11: Estimated effects of emergency admissions on inpatient care and health outcomes for emergency patients by expected mortality risk

	Low risk		High risk	
Dependent variable	Coeff	Std err	Coeff	Std err
Panel A: Admission cohorts				
Daycase operation	0.000	(0.008)	−0.004	(0.007)
Delayed operation	0.394***	(0.023)	0.306***	(0.027)
Number of procedures	−0.002***	(0.000)	−0.001**	(0.001)
7-day unplanned readmission	0.033***	(0.008)	−0.008	(0.014)
30-day in-hospital mortality	−0.001	(0.001)	0.018	(0.016)
Panel B: Discharge cohorts				
Length of stay	−0.026***	(0.001)	−0.008***	(0.001)
Transfers to other hospitals	−0.002	(0.006)	0.024	(0.026)
Discharges to home	0.011	(0.010)	0.001	(0.031)
7-day unplanned readmission	0.108***	(0.011)	0.075***	(0.029)
30-day in-hospital mortality	0.000	(0.001)	−0.014	(0.016)

Notes: (1) Reported coefficients are parameter estimates on the daily emergency admissions variable; (2) All specifications include a fully interacted set of diagnosis, age category, and emergency status fixed effects, and hospital-specific year, weekly-seasonal, and day-of-week fixed effects; (3) N = 1,061,432 and 478,695 for elective and emergency patients; (4) Standard errors clustered at the hospital-level (149 clusters); (5) ***/**/* indicates statistical significance at the 1/5/10% level.

Bibliography

Acemoglu, Daren and Finkelstein, Amy. (2008). 'Input and Technology Choices in Regulated Industries: Evidence from the Health Care Sector', *Journal of Political Economy* 116(5), 837–880.

Ackerberg, Daniel, Benkard, C. Lanier, Berry, Steven and Pakes, Ariel. (2007). 'Econometric Tools for Analyzing Market Outcomes', *Handbook of Econometrics* 6A, 4171–4276.

Almond, Douglas and Doyle, Joseph J. (2011). 'After Midnight: A Regression Discontinuity Design in Length of Postpartum Hospital Stays', *American Economic Journal: Economic Policy* 3(3), 1–34.

Almond, Douglas, Doyle, Joseph J., Kowalski, Amada E. and Williams, Heidi. (2010). 'Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborns', *Quarterly Journal of Economics* 125(2), 591–634.

Angrist, Joshua D. and Krueger, Alan B. (1999). 'Empirical Strategies in Labor Economics', *Handbook of Labor Economics* 3(23), 1277–1366.

Attanasio, Orazio P., Maro, Vincenzo Di and Vera-Hernandez, Marcos. (2012). 'Community Nurseries and the Nutritional Status of Poor Children. Evidence from Colombia', *Economic Journal* 123(571), 1025–1058.

Axon, R. Neal and Williams, Mark V. (2011). 'Hospital Readmission as an Accountability Measure', *Journal of the American Medical Association* 305(5), 504–505.

Bartel, Ann P., Beaulieu, Nancy D., Phibbs, Ciaran S. and Stone, Patricia W. (2014). 'Human Capital and Productivity in a Team Environment: Evidence from the Healthcare Sector', *American Economic Journal: Applied Economics* 6(2), 231–259.

Beckert, Walter and Kelly, Elaine. (2017). 'Divided by Choice? Private Providers, Patient Choice and Hospital Sorting in the English National Health Service', *IFS Working Paper W17/15* .

Best, Michael, Cloyne, James, Ilzetzki, Ethan and Kleven, Henrik J. (2017). ‘Estimating the Elasticity of Intertemporal Substitution Using Mortgage Notches’. Working Paper.

Best, Michael and Kleven, Henrik J. (2018). ‘Housing Market Responses to Transaction Taxes: Evidence from Notches and Stimulus in the UK’, *Review of Economic Studies* (85), 157–193.

British Medical Association. (2017), ‘State of the health system’.

Card, David, Dobkin, Carlos and Maestas, Nicole. (2009). ‘Does Medicare Save Lives?’, *Quarterly Journal of Economics* 124(2), 597–636.

Chan, David. (2016). ‘Teamwork and Moral Hazard: Evidence from the Emergency Department’, *Journal of Political Economy* 124(3).

Chan, David. (2017). ‘The Efficiency of Slacking Off: Evidence from the Emergency Department’, *Econometrica* . Forthcoming.

Chandra, Amitabh, Finkelstein, Amy, Sacarny, Adam and Syverson, Chad. (2016). ‘Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector’, *American Economic Review* 106(8), 2110–2144.

Chetty, Raj, Friedman, John N., Olsen, Tore and Pistaferri, Luigi. (2013). ‘Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records’, *Quarterly Journal of Economics* 126(2), 749–804.

Competition and Markets Authority. (2014), ‘Private healthcare market investigation: final report’.

Cooper, Zack, Gibbons, Stephen, Jones, Simon and McGuire, Alistair. (2011). ‘Does Hospital Competition Save Lives? Evidence from the English NHS Patient Choice Reforms’, *Economic Journal* 121(554), 228–260.

Correia, Sergio. (2016), Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator. Working Paper.

Cutler, David. (1995). ‘The Incidence of Adverse Medical Outcomes Under Prospective Payment’, *Econometrica* 63(1), 29–50.

Cutler, David. (2003), *Your Money Or Your Life: Strong Medicine for America’s Health Care Sytem*, Oxford University Press.

Cutler, David M. and Zeckhauser, Richard J. (2000). ‘The Anatomy of Health Insurance’, *Handbook of Health Economics* 1(11), 563–643.

Department of Health. (2012), 'A simple guide to Payment by Results'.

De Vany, Arthur. (1975). 'Capacity Utilization under Alternative Regulatory Constraints: An Analysis of Taxi Markets', *Journal of Political Economy* 83(1), 85–94.

De Vany, Arthur. (1976). 'Uncertainty, Waiting Time and Capacity Utilization: A Stochastic Theory of Product Quality', *Journal of Poltiical Economy* 84(3), 523–541.

De Vany, Arthur and Frey, N.G. (1982). 'Backlogs and the Value of Excess Capacity in the Steel Industry', *American Economic Review* 72(3), 441–451.

De Vany, Arthur and Saving, Thomas. (1977). 'Product Quality, Uncertainty and Regulation: The Trucking Industry', *American Economic Review* 67(4), 583–594.

Diamond, Rebecca and Persson, Petra. (2016), The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests. Working Paper.

Doyle, Joseph J. (2005). 'Health Insurance, Treatment and Outcomes: Using Auto Accidents as Health Shocks', *Review of Economics and Statistics* 87(2), 256–270.

Doyle, Joseph J. (2011). 'Returns to Local-Area Health Care Spending: Evidence from Health Shocks to Patients Far From Home', *American Economic Journal: Applied Economics* 3(3), 221–243.

Dranove, David. (2012). 'Health Care Markets, Regulators, and Certifiers', *Handbook of Health Economics* 2(10), 639–690.

Dranove, David, Kessler, Daniel, McClellan, Mark and Satterthwaite, Mark. (2003). 'Is More Information better? The Effects of "Report Cards" on Health Care Providers', *Journal of Political Economy* 111(3), 555–588.

Einav, Liran, Finkelstein, Amy and Polyakova, Maria. (2018). 'Private Provision of Social Insurance: Drug-specific price elasticities and cost sharing in Medicare Part D', *American Economic Journal: Economic Policy* . Forthcoming.

Einav, Liran, Finkelstein, Amy and Schrimpf, Paul. (2015). 'The Response of Drug Expenditure to Non-Linear Contract Design: Evidence from Medicare Part D', *Quarterly Journal of Economics* 130(2), 841–899.

Einav, Liran, Finkelstein, Amy and Schrimpf, Paul. (2017). 'Bunching at the kink: implications for spending responses to health insurance contracts', *Journal of Public Economics* 146, 27–40.

Ellis, Randall P. and McGuire, Thomas G. (1986). 'Provider Behavior Under Prospective Reimbursement', *Journal of Health Economics* 5(2), 129–151.

Eriksson, Carl O., Stoner, Ryan C., Eden, Karen B., Newgard, Craig D. and Guise, Jeanne-Marie. (2017). 'The Association Between Hospital Capacity Strain and Inpatient Outcomes in Highly Developed Countries: A Systematic Review', *Journal of General Internal Medicine* 32(6), 686–696.

Fiedler, Matthew Aaron. (2016), How does hospital congestion affect care and outcomes for patients with acute coronary illness? Doctoral dissertation.

Freedman, Seth. (2016). 'Capacity and Utilization in Health Care: The Effect of Empty Beds on Neonatal Intensive Care Admission', *American Economic Journal: Economic Policy* 8(2), 154–185.

Friedman, Bernard and Pauly, Mark V. (1981). 'Cost Functions for a Service Firm with Variable Quality and Stochastic Demand: The Case of Hospitals', *Review of Economics and Statistics* 63(4), 620–624.

Friedman, Bernard and Pauly, Mark V. (1983). 'A New Approach to Hospital Cost Functions and Some Issues in Revenue Regulation', *Health Care Financing Review* 4(3), 105–114.

Friedrich, Benjamin U. and Hackmann, Martin B. (2017). 'The Returns to Nursing: Evidence from a Parental Leave Program', *NBER Working Paper No. 23174* .

Gaynor, Martin and Anderson, Gerard F. (1995). 'Uncertain demand, the structure of hospital costs, and the cost of empty hospital beds', *Journal of Health Economics* 14(3), 291–317.

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

Gowrisankaran, Gautam, Joiner, Keith A. and Léger, Pierre-Thomas. (2017). 'Physician Practice Style and Healthcare Costs: Evidence from Emergency Departments', *NBER Working Paper No. 24155* .

Gruber, Jonathan and Kleiner, Samuel A. (2012). 'Do Strikes Kill? Evidence from New York State', *American Economic Journal: Economic Policy* 4(1), 127–57.

Gupta, Atul. (2017). 'Impacts of performance pay for hospitals: The Readmissions Reduction Program', *BFI Working Paper No.2017-07* .

Heckman, James J., Cunha, Flavio and Schennach, Susanne. (2010). 'Estimating the Technology of Cognitive and Noncognitive Skill Formation', *Econometrica* 78(3), 883–931.

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

Hoot, Nathan R. and Aronsky, Dominik. (2008). 'Systematic Review of Emergency Department Crowding: Causes, Effects and Solutions', *Annals of Emergency Medicine* 52(2), 126–137.

Hughes, David and McGuire, Alistair. (2003). 'Stochastic demand, production responses and hospital costs', *Journal of Health Economics* 22(6), 999–1010.

Joskow, Paul L. (1980). 'The Effects of Competition and Regulation on Hospital Bed Supply and the Reservation Quality of the Hospital', *Bell Journal of Economics* 11(2), 421–447.

Kaboli, Peter J., Go, Jorge T., Hockenberry, Jason, Glasgow, Justin M., Johnson, Skyler R., Rosenthal, Gary E., Jones, Michael P and Vaughan-Sarrazin, Mary. (2012). 'Associations Between Reduced Hospital Length of Stay and 30-Day Readmission Rate and Mortality: 14-Year Experience in 129 Veterans Affairs Hospitals', *Annals of Internal Medicine* 157(12), 837–845.

Kangovi, Shreya and Grande, David. (2011). 'Hospital Readmissions–Not Just a Measure of Quality', *Journal of the American Medical Association* 306(16), 1796–1797.

Keeler, Theodore E. and Ying, John S. (1996). 'Hospital Costs and Excess Bed Capacity: A Statistical Analysis', *Review of Economics and Statistics* 78(3), 470–481.

Kehlet, Henrik. (2013). 'Fast-track hip and knee arthroplasty', *The Lancet* 381(9878), 1600–1602.

Kelly, Elaine and Stoye, George. (2015). 'New joints: private providers and rising demand in the English National Health Service', *IFS Working Paper W15/22* .

Kessler, Daniel and McClellan, Mark. (1996). 'Do Doctors Practice Defensive Medicine?', *Quarterly Journal of Economics* 111(2), 353–390.

Kizer, Kenneth W. and Jha, Ashish K. (2014). 'Restoring Trust in VA Health Care', *New England Journal of Medicine* 371(4), 295–297.

Kleven, Henrik J. (2016). 'Bunching', *Annual Review of Economics* (8), 435–464.

Kleven, Henrik J. and Waseem, Mazhar. (2013). 'Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan', *Quarterly Journal of Economics* 128(2), 669–723.

Kocher, Robert P. and Adashi, Eli Y. (2011). 'Hospital Readmissions and the Affordable Care Act: Paying for Coordinated Quality Care', *Journal of the American Medical Association* 306(16), 1794–1795.

Kolstad, Jonathan T. (2013). 'Information and Quality when Motivation is Intrinsic: Evidence from Surgeon Report Cards', *American Economic Review* 103(7), 2875–2910.

Kristensen, Soren Rud and Sutton, Matt. (2016). 'Financial Penalties for Readmissions in the English NHS'. Working Paper.

Laing & Buisson. (2013), 'Laing's Healthcare Market Review 2012-13'.

Lindsay, Cotton M. and Feigenbaum, Bernard. (1984). 'Rationing by Waiting Lists', *American Economic Review* 74(3), 404–417.

Macartney, Hugh, McMillan, Robert and Petronijevic, Uros. (2015), Incentive Design in Education: An Empirical Analysis.

McCabe, Christopher, Claxton, Karl and Culyer, Anthony J. (2008). 'The NICE Cost-Effectiveness Threshold: What it is and What that means', *Pharmacoeconomics* 26(9), 733–744.

Morrisey, Michael A., Sloan, Frank A. and Valvona, Joseph. (1988). 'Shifting Medicare Patients Out of the Hospital', *Health Affairs* 7(5), 52–64.

Mortimore, Andy and Cooper, Simon. (2007). 'The '4-hour target': emergency nurses' views', *Emergency Medicine Journal* 24(6), 402–404.

National Audit Office. (2004). 'Improving Emergency Care in England'. HC 1075 Session 2003-2004.

NHS Digital. (2017), 'Patient Reported Outcome Measures (PROMs) in England'.

OECD. (2015), 'Health at a Glance 2015: OECD Indicators'.

Olivella, Pau and Vera-Hernández, Marcos. (2013). 'Testing for Asymmetric Information in Private Health Insurance', *Economic Journal* 123(567), 96–130.

Propper, Carol. (1990). 'Contingent Valuation of Time Spent on NHS Waiting Lists', *Economic Journal* 100(400).

Propper, Carol, Sutton, Matt, Whitnall, Carolyn and Windmeijer, Frank. (2008). 'Did 'Targets and Terror' Reduce Waiting Times in England for Hospital Care', *B.E. Journal of Economic Analysis & Policy* 8(2).

Saez, Emmanuel. (2010). 'Do Taxpayers Bunch at Kink Points?', *American Economic Journal: Economic Policy* 2(3), 180–212.

Shonick, William. (1970). 'A Stochastic Model for Occupancy-Related Random Variables in General-Acute Hospitals', *Journal of the American Statistical Association* 65(332), 1474–1500.

Shonick, William and Jackson, James R. (1973). 'An Improved Stochastic Model for Occupancy-Related Random Variables in General-Acute Hospitals', *Operations Research* 21(4), 952–965.

Silver, David. (2016), Haste or Waste? Peer Pressure and the Distribution of Marginal Returns to Health Care. Working Paper.

Vorhies, John S., Wang, Yun, Herndon, James, Maloney, William J. and Huddleston, James I. (2011). 'Readmission and Length of Stay After Total Hip Arthroplasty in a National Medicare Sample', *The Journal of Arthroplasty* 26(6), 119–123.

Windmeijer, Frank, Gravelle, Hugh and Hoonhout, Pierre. (2005). 'Waiting lists, waiting times and admissions: an empirical analysis at hospital and general practice level', *Health Economics* 14(9), 971–985.

Young, John P. (1962), A queuing theory approach to the control of hospital inpatient census. Doctoral dissertation.