Saving Lives by Tying Hands:

The Unexpected Effects of Constraining Health Care Providers

Jonathan Gruber, Thomas P. Hoe and George Stoye*

February 11, 2019

Abstract

In this paper we study how emergency department (ED) doctors respond to incentives to reduce wait times. We use bunching techniques to study an English policy that imposed strong incentives to treat patients within four hours. The policy reduced affected wait times by 21 minutes, yet caused doctors to increase the intensity of treatment and admit more patients. We find a striking 14% reduction in mortality. Analysis of patient severity and hospital crowding strongly suggests it is the wait time reduction that saves lives. We conclude that, despite distorting medical decisions, constraining ED doctors can induce cost-effective reductions in mortality.

*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). Author affiliations: Gruber (MIT and NBER); Hoe (Cornell University and Institute for Fiscal Studies); Stoye (University College London and Institute for Fiscal Studies).

1 Introduction

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 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.[1] 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 directly compensated for shortening wait times. Moreover, while health-maximizing ED personnel may internalize the costs of waiting to the extent that they impact patient outcomes, this may only be partial if physicians have incomplete knowledge or are imperfect agents for their patients. 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 decrease wait times for different types of care, and came into force in all English public hospitals in 2004.[2] The policy set arbitrary targets for wait times,

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

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

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 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 paper 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 21 minutes (8%) lower for patients affected by the policy, and for those patients that move from after-to-before the four hour position, the wait time reductions are large and average 59 minutes.

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 little meaningful 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. As a final check, we examine whether the reductions in mortality occur among patients with potentially time-sensitive conditions, and find that the majority of these reductions are found among conditions which are known to benefit from rapid treatment.

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.

Our paper proceeds as follows. Section 2 provides background information on emergency care in England and on the four-hour target policy. Section 3 describes the data. Section 4 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 6 describes our results for wait times, treatment decisions and health outcomes. Section 7 explores heterogeneity and mechanisms. Section 8 concludes.

2 Background

2.1 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.[3] In 2015/16, there were 11 separate tariffs for ED treatment depending on the severity of the patient and the type of treatments administered.[4] These tariffs ranged from $77 to $272 per

[3]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.

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

visit.[5] Revenue from the ED accounted for 5.3% of total hospital income in 2015/16.[6]

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 seriousness 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.[7] 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 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

[5] 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).

[6] 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

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

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.

Importantly, the objective of much of the ED care provided in this setting is either to treat simple conditions, or to provide early diagnostic information that can be used to send patients to the correct specialist inpatient ward for future treatment (while also stabilising patients and providing basic treatment as these tests are carried out). Table A1 in Appendix A shows the most common ED investigations and treatments across 40 ED diagnosis categories, distinguishing between the first and subsequent investigations and treatments.[8] X-rays and blood tests are the most common primary investigations, while in a quarter of (more minor) diagnoses the majority of patients receive no specific investigations. Most treatments are simple: providing guidance or advice, or treating minor cases (e.g. wound closures for lacerations, plaster of paris for fractures). For the more serious cases, common treatments include inserting an intravenous cannula or observing patients (including taking an ECG or recording patients' pulses) while diagnostic tests such as x-rays, CT scans and blood tests are carried out. This is further reflected in Table A2, which shows the most common ED investigations and treatments for admitted (and therefore likely more serious) patients only.

Patients who require further specialist treatment are then admitted as an inpatient. Table A3 shows the most common first and subsequent inpatient procedure across each ED diagnosis. Initial inpatient treatment is also often diagnostic in nature, as shown by the frequency of the use of CT and MRI scans. More comprehensive treatment of the condition then follows.

Taken together, these tables demonstrate that the ED provides an important first stage of treatment, solving more minor problems in the department itself, while collecting important diagnostic information that is important in ensuring that more complex cases receive the correct inpatient treatment further along the treatment pathway.

2.2 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

[8]There are 24 ED investigation categories, including 'none'. There are 57 ED treatments. Details of these can be found in the HES Data Dictionary (Accident and Emergency), available here: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary

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) 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

[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.

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

[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

of all patients against the clock, and to alert physicians that an admission decision is required soon.

3 Data

3.1 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, where the arrival time is recorded as patients enter the ED.[16,17]

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.[18] We 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

[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] For non-ambulance patients, this time is recorded when they first speak with the receptionist.

[17] Hospitals may attempt to manipulate wait times to meet the target. We evaluated one possibility in this regard, namely that hospitals simply miscode the timing of the admission decision, such that the total wait time is 4 hours or less. Following Locker and Mason (2006), we analyzed the distribution of 'final digits' in wait times (e.g. the digits 0 to 9 at the end of each wait time value) which in the absence of manipulation should be uniformly distributed. Relative to this benchmark, we found that less than 1% of records were likely to be miscoded and that this would have a negligible impact on our analysis.

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

the government through a system of national tariffs for each HRG.[19] 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.

3.1.1 Sample construction

Our analysis focuses on a sample of emergency patients treated in 'major' emergency departments.[20] We exclude patients treated at specialist clinics that treat only particular diagnoses (e.g. dental) and minor injury ('walk 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%.[21] 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.

3.1.2 Summary statistics

Table 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

[19] 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.

[20] 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.

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

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 1 also shows summary statistics separately for visits that result in an inpatient admission. As expected, these case are typically more severe, 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 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 Online Appendix Figure A1, that such a spike is not present in data on ED wait times from a major U.S. hospital.[22]

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 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

[22]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.

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 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.

4.1 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 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 into two channels.

The first channel is the 'composition effect'. As Figure 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 paper.

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*.[23] 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 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

[23]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. Our approach is similar in spirit to Diamond and Persson (2016) and Gerard et al. (2018).

that the no-selection assumption has been violated. We pass these placebo tests for several demographic variables.

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

4.2 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

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

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. This regression fits a polynomial to the wait time distribution in periods outside of the exclusion window, where the window is captured by the indicator variables which then do not feature in $\hat{f}_0(w)$.

Equation (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:*

$$f_0(w) = f_1(w) \qquad \forall w \notin [w^-, w^+]. \tag{2}$$

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.[24] We discuss this as-

[24] 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

sumption at length in the next section.

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 $(w^*, w^+]$.[25] An advantage of this iterative approach is that we make no assumption about w^+ and let the data determine where the effects on the wait time distribution end. 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 $w^+ = 400$. Our results are robust to variations in the choice of polynomial and w^- (see Online Appendix Tables A4 and A5).

The observed data and our estimated counterfactual distribution are shown in Figure 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.

4.2.1 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 prioritization, 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 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.

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.

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

4.3 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.

4.3.1 Composition and distortion effects

In a potential outcomes framework, let y_t be an outcome (treatment decision or mortality outcome) and w_t be the wait time in regime $t \in \{0, 1\}$. We then 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

$$E[y_1 \mid \underline{w}_1^-] = \rho E[y_1 \mid \underline{w}_1^-, \underline{w}_0^-] + (1 - \rho) E[y_1 \mid \underline{w}_1^-, \underline{w}_0^+], \tag{3}$$

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:*

$$\Delta_C \equiv \rho\big(E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]\big) + (1-\rho)\big(E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]\big) \tag{4}$$

$$= (1-\rho)\big(E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]\big). \tag{5}$$

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:*

$$\Delta_D \equiv \rho\big(E[y_1 \mid \underline{w}_1^-, \underline{w}_0^-] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]\big) + (1-\rho)\big(E[y_1 \mid \underline{w}_1^-, \underline{w}_0^+] - E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+]\big). \tag{6}$$

Note that the distortion effects may impact both pre-target non-movers (first term) or post-target movers (second term). With these definitions the observed outcomes in the pre-threshold period can be written as

$$\underbrace{E[y_1 \mid \underline{w}_1^-]}_{\text{Targeted regime (observed)}} = \underbrace{E[y_0 \mid \underline{w}_0^-]}_{\text{Non-targeted regime}} + \underbrace{\Delta_C}_{\text{Composition effect}} + \underbrace{\Delta_D}_{\text{Distortion effect}} \tag{7}$$

which can be verified by substituting in Equations (3), (5) and (6) and rewriting the non-targeted regime outcome as $E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-]$.

4.3.2 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 effect only:*

$$E[y_0 \mid \underline{w}_1^-] \equiv E[y_0 \mid \underline{w}_0^-] + \Delta_C \tag{8}$$

$$= \rho E[y_0 \mid \underline{w}_1^-, \underline{w}_0^-] + (1-\rho)E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+]. \tag{9}$$

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 (9) 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.

4.3.3 Estimating counterfactual outcomes

We now revisit the bunching estimator and show it can be used to obtain the counterfactual outcomes in Equation (9). 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:*

$$E[y_1 \mid w_t] = E[y_0 \mid w_t] \qquad \forall w \notin [w^-, w^* + \varepsilon]. \tag{10}$$

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 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.[26]

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

$$E[y_0 \mid \underline{w}_1^-, w_0] = E[y_0 \mid \underline{w}_1^+, w_0] = E[y_0 \mid w_0] \qquad \forall w_0 \in \underline{w}_0^+. \tag{11}$$

[26] 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.

Assumption 3 rules out composition effects in the post-threshold period. It states that after conditioning on the non-targeted wait time, there is no selection when the post-threshold movers are assigned. The assumption is consistent with doctors randomly selecting which patients get a shorter wait time in response to the target, and in that sense it is equivalent to an unconfoundedness assumption in traditional IV terminology.[27] While this is strong assumption we believe it is plausible and, most importantly, we are also able to evaluate the assumption empirically using placebo tests. We discuss this assumption and the results of these tests in detail shortly.

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 (9).[28]

Figure 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 $\varepsilon = 20$.

4.3.4 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 (9), 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 paper. We compute statistical significance for the test using non-parametric bootstrapped standard errors clustered at the hospital organisation level.[29]

Figure 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

[27] Similarly, in IV terminology, the post-threshold movers would be compliers, the post-threshold non-movers would be never-takers, and the pre-threshold non-movers would be always-takers. We implicitly make the assumption that there are no defiers.

[28] Note that when constructing $E[y_0 \mid \underline{w}_1^-, \underline{w}_0^+]$ it is necessary to use the distribution of movers, which is equal to the difference between the counterfactual and observed distribution of patients in the post-threshold period.

[29] 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.

the conditional expectation in Equation (9). 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.[30] 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 $\Delta_D = 0$.

4.3.5 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 8 provides a visual example of the demographic test using age, which follows the same format as Figure 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.

5 Validity of key assumptions

Our methodology rests on our assumptions about local effects (Assumptions 1 and 2) and selection (Assumption 3). We discuss these assumptions and supporting evidence below.

[30] The weights are obtained from the wait time distributions shown in Figure 3.

5.1 Local effects assumption

The local effects assumption is that wait times and treatment decisions prior to w^- are unaffected by the target, and we set w^- at 180 minutes in the baseline analysis. As noted earlier this will not hold if hospitals substitute time or resources between patients that exit before w^- ('early exit patients') and after w^- ('late exit patients'). We view our assumption as very reasonable in this setting with little potential for bias. To support this claim, we first discuss the potential concerns, and then describe the relevant institutional factors in our setting and a series of robustness tests.

As an example, a violation of this assumption would occur if the target induced physicians to substitute from early exit patients (often high severity patients with unambiguous symptoms, e.g. knife attack victims) to late exit patients (often high severity patients with uncertain diagnoses, e.g. headaches). By making this type of substitution, physicians may be able to delay some patients 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.

The example above is a general type of concern (which we refer to as 'substitution') and would arise in a dynamic model where physicians allocate their time across patients. In such a model, imposing any target would cause physicians to reallocate their time across all patients, and change the entire wait time distribution. A second type of concern in a general dynamic setting is that physicians may not know on arrival which patients are at risk of breaching the target, and may therefore speed up care for some patients that ultimately would leave the ED prior to w^- irrespective of the target (an 'uncertainty' effect). Both of these responses, substitution and uncertainty, raise the prospect that wait times and treatment decisions are affected prior to 180 minutes.

It is possible to analyze the direction of bias from violations of Assumptions 1 and 2. For the wait time analysis, this task is relatively straightforward: substitution will cause us to overstate the wait time reduction (since our analysis omits any delays prior to w^-) while uncertainty will cause us to understate the reduction (we omit any reductions prior to w^-).[31] The potential bias in our analysis of treatment decisions and health outcomes is more difficult to characterize

[31] A related concern in both cases is that our counterfactual could be substantially misspecified if the observed distribution in the pre-exclusion window is entirely uninformative. The impact of this on our estimates is harder to predict. Given the institutional factors discussed below, however, we view a major misspecification of this type as unlikely.

because it depends on composition and distortion effects and may involve non-random selection. As we set out below, however, it is possible to characterize how and where any bias is most likely to be present, and we run a series of tests that evaluate our assumption on this basis.

5.1.1 Institutional factors

There are three factors that mitigate concerns about a violation of local effects. The first is simply that there is far less scope for dynamic responses than Figure 1 would suggest. In particular, this figure aggregates across many hospitals, days, and time periods. The actual scope for substitution between patients is much more limited, since there is no incentive to substitute between patients that are not within the same four hour window. Looking at the data, there are on average only 33 patients who arrive at a hospital within a given four-hour window. This limits the potential extent of dynamic responses.

Second, ED staff are organized in a way that further limits scope for substitution. Physicians and nurses in English EDs are generally separately assigned to minor or major units within the ED, and this physical separation limits the prospect of substitution between early and late exit patients. It is of course possible that as a major unit becomes busy, staff could be diverted from the minors unit to assist and this could violate Assumption 1 by delaying early exit patients. However, our understanding is that there are often slack resources in the majors unit, such as a senior physician who will step in at busy times rather than pulling resources across from minors. This will again limit the likelihood of substitution.[32]

Finally, hospitals are likely to be maximizing an objective function that, at least in part, contains patient mortality. This will naturally place limits on their willingness to substitute between different types of patients. For example, patients with clear and life-threatening injuries (e.g. knife wounds) will always be treated immediately, and for a similar length of time, irrespective of the target. Similarly, patients with very minor injuries will always be sent home shortly after initial assessment. These unambiguously high and low severity patients are likely to account for a significant proportion of exits prior to the exclusion window, and suggest that if there are substitution or uncertainty responses, then these are more likely to occur away from the very early exits and nearer to w^{-}.

[32] A limitation of our data is that it does not record the outcome of any minor or major classification, nor does it record staff or patient locations.

5.1.2 Robustness test I: Variation in w^-

In the baseline specification we use $w^- = 180$ on the basis of visual inspection. As we note above, based on institutional factors, if there are violations of Assumptions 1 and 2, then they will be most evident in the neighborhood of w^-. We therefore vary our choice of w^- and assess how sensitive our results are. If there are major shifts in our estimates then this would be a sign that there are significant substitution or uncertainty effects prior to 180 minutes. In Online Appendix A (Table A4) we show that this is not the case and our results are robust to the choice of w^- as it varies between 160 and 200 minutes.

5.1.3 Robustness test II: Disaggregated analysis

In the baseline specification we pool the data for all patients together. As a robustness check, we can separate out patients according to their ED diagnosis and run the same baseline specification. Doing this eliminates any bias from substitution effects across patients with different ED diagnoses (e.g. if hospitals shorten wait times for stroke patients at the expense of patients with broken arms). We present and discuss the details of these results in Section 7.2 (see Table A6), but note here that the weighted average of the estimated impacts across diagnoses (weighted using the number of patients in each diagnosis) provide very similar estimates to the aggregate baseline results.

5.1.4 Robustness test III: Planned substitution effects

We now turn to two tests that evaluate specific types of substitution effects. First, we test for evidence of 'planned substitution', by which we refer to when hospitals anticipate that the target will bind and in response change the relative priority given to certain patients to improve target performance. While we would ideally evaluate this using data from before the target, given that is unavailable we instead exploit variation in expected volumes of ED arrivals. The idea is that this variation changes how tightly the target binds since higher volumes of arrivals make the target more challenging to meet in relative terms. If planned substitution effects are present, we anticipate it should be noticeable in the prioritization of patients across periods that are expected to be more or less busy.[33]

[33] 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 prioritization that we compare in this test. 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 plots the proportion of patients who exit within 180 minutes 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 a smaller proportion of higher severity patients leave the ED within 180 minutes. It also shows that busier periods have longer wait times for patients of all severity, with a smaller proportion of patients across the severity distribution leaving the ED within 3 hours as the department becomes busier. Most importantly for our purposes, the relative probabilities of exits within 180 minutes for high and low severity patients are very similar in both types of period. The same is also true if you repeat this exercise for other waiting periods.[34] These results suggest that as the target binds more or less tightly, hospitals maintain the same prioritization of patients and there are no planned substitution effects.

5.1.5 Robustness test IV: Temporary substitution effects

Turning to the second substitution test, we now consider 'temporary substitution', by which we refer to the case when hospitals experience a demand shock–such as the ED being momentarily overrun with patients–and this causes them to make short-term deviations from their planned priorities. A specific example could be that the hospital has a build up of patients that are close to breaching the target and they temporarily substitute resources away from newly-arriving patients to clear their backlog of patients and therefore increase the wait time of these new arrivals.

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

[34] We also examined the proportion of patients who exit the ED after 60 minutes, 120 minutes, 240 minutes and 400 minutes. This produces similar results in all cases, with a broadly parallel shift downwards in the proportion of exits when moving from a non-busy to busy period. This shift is very small when examining the proportion of patients exiting within 240 and 400 minutes as the vast majority of patients have exited the ED by this point.

patients near the four-hour threshold on the wait times of new patients.

We examine four groups of newly arriving patients on the basis of their predicted waiting times: those predicted to have wait times below 150 minutes; between 150 and 180 minutes; 180-210 minutes; and 210-240 minutes. For each group, we regress wait times on the volume of existing patients waiting ahead of them at each 10-minute internal of the queue. We then compare the results between the early exit patients (those in the first two groups with predicted wait times below 180 minutes, such that they exit prior to the exclusion window) and late exit patients (those predicted to have wait times above 180 minutes). The late exit groups act as control groups in the sense that Assumption 1 allows for temporary substitution effects to occur for these groups (inside the exclusion window) but not for the early exit groups (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

$$w_{ht}^{g} = \sum_{k} \beta_k q_{h,t-k} + \mu_{hw} + \delta_{hd} + \gamma_{hp} + e_{ht} \tag{12}$$

where w_{ht}^{g} is the mean wait time for newly arriving patients of group g (as per the four categories described above) 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 5 presents the estimated β_k coefficients from Equation (12). 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. The results for each group are shown separately in panels (a) - (d). Looking first at the early exit groups (panels (a) and (b)), 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 in either group.[35] Looking now at the late exit groups (panels (c) and (d)), there is again evidence of longer queues increasing wait times but for these groups there is clear evidence of a

[35] We would expect to see a spike around the target in panel (b) if the exclusion window should start prior to 180 minutes. The similarities between panel (a) and panel (b) therefore suggest that 180 minutes is a reasonable choice for the lower bound of the exclusion window.

discontinuity at the four-hour threshold. This indicates that, for the late exit groups, doctors actively substitute resources away from newly arriving patients towards those patients that are at risk of breaching the target. The spike is largest for the group who are predicted to wait between 210 and 240 minutes (panel (d)), indicating that the greatest substitution occurs between the most similar patients. 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 and 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 and our choice of lower bound for the exclusion window. The robustness tests based on w^- and our disaggregated analysis further support this claim.

5.2 No selection assumption

We set out our methodology for a test of Assumption 3, based on observable demographic variables, in Section 4.3.5. Table 2 presents the results of the relevant 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, an indicator for whether the patient arrived via ambulance, and the Charlson co-morbidity index based on the 12 months of hospital admissions prior to the beginning of our ED data ('past-CCI'). Each of these variables should be unaffected by decisions made in the ED, and thus allow us to test our selection assumption.[36]

For age and ambulance-arrival we cannot reject the no-selection hypothesis. In contrast, we reject the no-selection hypothesis for gender and past-CCI. The gender result suggests 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).[37] With regard to the past-CCI results, the positive estimate suggests

[36] We considered several other variables for the demographic test. Unfortunately a number of variables have missing data around the threshold (e.g patient ethnicity).

[37] Exploring the male indicator more carefully shows that, unlike the other variables we study, it is poorly correlated with ED wait times. This causes the polynomial regression, which we use to determine the counterfactual outcomes, to fit the data less robustly (i.e. it is sensitive to the choice of polynomial). The demographic test is therefore not reliable for this variable. The same is also true for other variables that we considered for the test, including whether the patient lives in an urban area and the deprivation level of the local area.

post-threshold movers are on average less healthy than post-threshold non-movers, with a past-CCI score that is 4% higher. This is consistent with physicians responding to the target by prioritizing patients with a worse health record.

Panel B in Table 2 presents results for variables that are linear combinations of the individual demographic variables. We use predicted admission and predicted mortality, where the predictions are obtained from linear regressions of the outcome on a flexible specification of the demographic variables (past-CCI score, and a fully interacted set of age, gender and ambulance-arrival fixed effects). The R^2 statistic from these predicted regressions is 0.22 and 0.06. An advantage of using these predicted variables is that they weight individual demographic variables according to their relative importance for clinical outcomes. Weighting factors on this basis is useful because selection on factors which do not impact these outcomes is unlikely to bias our estimates. Looking at the estimates, the demographic tests for these predicted variables cannot reject the hypothesis of no-selection. So even though the gender and past-CCI tests reject the hypothesis, the contribution of these variables to salient medical outcomes and thus the likelihood of bias is low.

As a direct test of whether gender and past-CCI introduce meaningful bias to our estimates, we computed estimates conditional on these observables and compared them to our baseline estimates that we present below. The two sets of results were very similar, suggesting that any selection does not introduce substantive bias to the estimates.[38] We also note that any bias from selection on (unobservable) severity, if it mirrors the past-CCI result, would attenuate our estimates towards zero and thus make our mortality estimates conservative.

As a final probe of the no-selection assumption, we simulated how selection of different degrees would manifest itself in the observed data. To do this we built a simulated dataset using the counterfactual wait time and age distributions and then artificially assigned post-threshold movers using different selection rules. We describe this process in more detail in Online Appendix B. The simulation highlights three facts about selection in our setting. First, the observed data on age looks very similar to the simulated data with a random selection rule. Second, even very modest selection is predicted to have a clear impact on the data, by creating a spike in outcomes in the pre-threshold period and a very pronounced 'dip' in outcomes in the post-threshold period, and neither of these features of selection are present in our data.

[38]To compute the conditional estimates, we apply our methodology to subgroups of patients defined by gender and past-CCI, and then aggregate these results up to be comparable to the baseline estimates using the sample weights associated with each subgroup.

Third, an advantage of our test is that it has potential to detect selection-on-unobservables even though it relies purely on observables. This follows since a test based on age, for example, would reveal selection on another unobservable variable as long as it is sufficiently correlated with age.

Together these results indicate that the no-selection assumption is plausible in this setting. On its face, this is perhaps a surprising finding. While patients themselves do not make the selection decision, hospitals do make these choices and selecting certain patients may be in their interest. But on a day-to-day basis, hospitals are treating patients at different times and we find this limits the scope for selection. If the data is segregated into hospital-hour periods, for example, then the number of patients approaching the target at any given point in time is actually small, at around 3 to 4. This compares to an average of 3.5 physicians that are on shift in a typical ED, suggesting physicians rarely have a choice between multiple 'potential breach' patients.[39] Rather than being a result of selection on patient characteristics, we view breaches of the target as more likely to occur due to idiosyncratic events and delays (e.g. staff shortages). While we cannot rule out that such events could be correlated with patient characteristics, our demographic tests suggest that this is not the case.

In practice, we therefore treat those patients observed with wait times in excess of 240 minutes (post-threshold non-movers) as 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.

6 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 7.

[39] We do not have detailed staffing data but obtained this number from a data request sent to all hospital trusts. We received responses from roughly 40% of trusts, and average physician figure is an average of physician numbers across all hospital-hours.

6.1 Wait times

Figure 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 21 minutes, or 8% of their estimated counterfactual mean. Moreover, if we restrict our attention to those patients moved to the pre-threshold period from the post-threshold period (the post-threshold movers), then the average wait time reduction is 59 minutes.[40]

6.2 Treatments and mortality outcomes

Table 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 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.

[40]To obtain estimates of the distribution of wait time reductions would require further assumptions on the ordering of patients. We do not impose these assumptions, but note that the maximum wait time reduction could be as large as 200 minutes (i.e. a patient moved from 390 minutes to 190 minutes).

We also show target effects on the number of investigations performed in the ED, such as x-rays, blood-test 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 two percent 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 departments 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 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. The CAC for 30-day mortality is shown in Figure 9; here, after adjusting for the composition effect, we find that the observed data is lower than the CAC and this is what produces the negative estimate. 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 2. 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.[41] 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.

7 Mechanisms

7.1 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.

In particular, we consider two natural sources of heterogeneity. The first is differences across diagnosis. In particular, we divide patients into 36 diagnosis groups.[42] 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

[41] 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. The calculation also omits the fines levied on hospitals for breaching the target. Incorporating these fines into the cost calculation reduces the cost per year of life saved (as the government effectively recoups some of its additional expenditure through collecting the fines) but this effect is very small because performance was close to the 95% target in the period we analyze.

[42] 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. We also tested whether the missing data had an impact on the results, by conducting the same analysis for hospitals that had fewer missing data points and we found similar results to those presented here.

10 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.[43] We then assess how the heterogeneity across diagnosis groups translates to each of these outcomes.

The results of this exercise are shown in Figure 11. 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 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.

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

[43] We adjust the polynomial choice for each diagnosis to ensure that it both fits the data well and meets the condition that the excess and missing mass are equal. To do this we use an approach that maximizes the adjusted-R^2 of Equation (1) for each outcome. We use the same approach for the crowding analysis that follows.

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.[44] 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 in effects.

Figure 12 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 11c, 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 close to zero. There is a positive relationship between predicted mortality reductions and inpatient crowding but this is small in magnitude.[45]

[44] 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.

[45] This means that our results may actually understate the mortality reductions in the most crowded periods.

We formalize these graphical results in Table 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 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 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, which 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 suggestive.

Given that these periods are also those with the smallest increases in admissions, this would strengthen the conclusion that mortality reductions are associated with reductions in wait times and not additional admissions.

7.2 Wait times, diagnoses and causes of death

This evidence raises the question of how reductions in wait times could lead to lower mortality rates. The most likely mechanism is that reductions in wait times lead to lower time-to-treatment for patients with severe diagnoses. An extensive medical literature makes it clear that rapid treatment is associated with better mortality outcomes for patients across a range of conditions. For example, Seymour et al. (2017) find a strong positive association between time-to-treatment and survival for ED patients with sepsis and septic shock.[46] However, it may be difficult for physicians to identify these patients as they arrive at the ED: a body of medical evidence suggests that misdiagnosis in the ED is not uncommon, while there is often disagreement between ED physician and subsequent specialist diagnosis.[47] This suggests why the target may have been successful in improving outcomes relative to an unconstrained scenario as it leads doctors to speed up treatment for all patients, which is costly but ensures that hard-to-diagnose and time-sensitive patients ultimately get the correct treatment sooner.

We explore the likelihood that this mechanism is driving our results in two ways. First, we look at how hospitals achieve the reductions in wait times by examining which parts of the treatment pathway they are compressing. This provides some evidence on whether patients start to receive treatment earlier. Second, if wait times are driving mortality reductions we would expect to see the greatest mortality reductions for patients with conditions where outcomes are known to be time-sensitive. We therefore examine variation in mortality reductions across diagnoses and primary causes of death.

To examine how wait times are reduced, we break down the overall impacts on waits into the three separate components that the data allows: time to initial assessment; time between assessment and the beginning of treatment; and duration of ED treatment. The initial assessment is usually conducted by a triage nurse, and includes a relatively basic examination. Treatment begins when the patient is first examined by a doctor (i.e. when the first ED treat-

[46] There are also many examples from other diagnoses. For example, Saver et al. (2013) find significant improvements in mortality and post-hospital outcomes for stroke patients when cutting time-to-treatment. Cannon et al. (2000) also find substantial increases in mortality following a heart attack when patients receive angioplasty more than two hours after arriving at hospital.

[47] Shojania et al. (2003) conducted a systematic review into studies of autoposy-detected diagnostic errors over a 40 year period in the US and found a median error rate of 23.5%, although this rate was decreasing over time. Delays and misdiagnoses are particularly common for neurological and cerebrovascular patients, and many of the existing studies are in this area. For example, Newman-Toker et al. (2014) estimate that between 15,000 and 165,000 cerebrovascular events are misdiagnosed annually in US EDs, while Moulin et al. (2003) found that half of ED patient diagnoses in a large French hospital were changed after neurology consults were obtained, following access to more detailed testing equipment (e.g. CT and MRI scanners) which were not available to the original ED physician.

ment is received, and we document the common first treatments in Table A1), and ends when the ED makes the decision to admit or discharge the patient. Admitted patients will then receive further treatment from a specialist within the hospital. As noted in Section 2.1 and shown in Appendix Tables A1 and A2, most ED treatment in England is aimed at stabilising and diagnosing patients, with more extensive treatments provided by specialists in inpatient wards. Reducing treatment time in the ED therefore means that patients start to receive this specialist treatment sooner.

We repeat the analysis in Section 4.3 using each of the three components as an outcome variable. The results show that the reductions are achieved both by reducing the initial wait for treatment (48% of the overall reduction), and by shortening the duration of ED treatment (45%). The remaining reduction (7%) is explained by the initial time to assessment. These results suggest that the target reduces the wait for both ED and specialist treatment to begin. Patients start to receive treatment from ED physicians sooner and spend less time receiving treatment in the ED.[48] Importantly, shorter periods spent in the ED also mean that admitted patients start to receive specialist inpatient treatment sooner. This specialist treatment often begins with further diagnostic testing (such as CT or MRI scanning, as documented in Table A3), and so reducing ED time means patients are likely to receive a detailed diagnosis sooner.

Next we turn to examining which patients benefit most (in terms of mortality reductions) from the target. If quicker treatment is responsible for improving patient outcomes, we would expect to see the greatest improvements for patients with diagnoses that can be affected by time-sensitive treatments. We therefore examine in which diagnosis groups we see the biggest impact of the target on patient outcomes. Table A6 in Appendix A shows the estimated impact on mortality and wait times within each of the 40 ED diagnoses categories, ordered by the size of the mortality reduction. The largest impacts of the target on mortality rates are found in patients with septicaemia, cerebo-vascular (stroke) and other vascular injuries. These are all areas, as noted above, in which medical evidence suggests benefits to patients from reduced time-to-treatment.

While the impacts are largest among these diagnoses, the total number of patients saved in each diagnosis group will also depend on the number of patients who attend ED with these diagnoses. For example, septicaemia is a relatively rare condition, while respiratory problems

[48]The data do not include more detailed information on the amount of time actually spent with physicians. As a result we cannot test whether the target reduces time spent being treated by an ED physician. However, our results in Table 3 suggest that patients do not receive fewer ED treatments as a result of the target.

are more common. We therefore use the estimates to compute the number of patients for each broad ED diagnosis who survived for at least 30 days following their ED visit as a result of the target.

The last column of Table A6 reports these estimates as the share of total lives saved among patients with complete diagnosis information. The aggregate estimates indicate that in 2012/13, 17,800 patients were saved by the target, or just under one patient per hospital every three days. Among patients with complete diagnosis information, a third of the lives saved are from patients attending the ED with a respiratory problem. Gastrointestinal, cardiac and cerebro-vascular diagnoses also explain substantial shares of the lives saved. While these categories are still relatively broad, they provide reassurance that the majority of the mortality reductions come from serious conditions where timely treatment can plausibly make a difference to patient outcomes.

An alternative way of analyzing which patients are saved by the target is to examine whether we observe reductions in the specific causes of death contained in the official mortality records that are linked to our hospital records (ICD-10 codes of primary cause of death). These data provide a far more granular record of a patient's condition than the ED diagnosis data. However, while the recording of ED diagnosis should be independent of the the treatment received, mortality is an endogenous outcome. Rather than split the sample by cause of death, we therefore use indicators for specific causes of death as outcome variables and test whether the target reduced the prevalence of each cause. In this way we can further test the time-to-treatment mechanism by examining whether the target reduced deaths in time-sensitive conditions, but not in conditions that we would not expect to be time-sensitive in an acute setting.

We begin by classifying deaths into 23 categories according to the first letter of the ICD-10 code, and repeat our analysis with the 23 dummy variables as outcome variables. Table A7 in Appendix A shows the results. We find that 70% of the reduction in mortality can be explained by reductions in deaths related to circulatory (30.1%), respiratory (25.7%) and digestive (15.0%) conditions. These are all categories which include specific conditions that are likely to be time-sensitive.[49] In contrast, there is no significant reduction in mortality attributed to neoplasms (cancers), which include a number of high mortality conditions unlikely to be time-sensitive.

Undertaking the same analysis in yet more detail, we analyze more detailed causes of death

[49] Circulatory conditions in the ICD-10 data include strokes, which are the most common cause of death for patients with the ED diagnosis of cerebro-vascular.

using the first 2 digits from the ICD-10 codes. We examine the ten most common causes of death for ED patients, which together account for 60% of all patient deaths. Table A8 in Appendix A presents the results. The estimates again show that the greatest mortality reductions are found for time-sensitive conditions. The largest effects are found in patients with cerebrovascular diseases, both as a proportion of the overall mortality reduction (column 3) and as a proportion of deaths due to the specific cause (column 4). Deaths from chronic lower respiratory diseases, influenza and pnuemonia, and ischemic and pulmonary heart diseases are also substantially reduced.

In contrast, we again observe no significant changes in mortality associated with cancers of any type. These are all conditions which we would expect to be less sensitive to time-to-treatment in an acute setting, and so act as a convincing placebo test when examining the time-to-treatment mechanism.

While these results do not provide definitive proof that wait time reductions are causing mortality reductions, they do provide reassuring evidence that many of the mortality reductions occur in diagnoses where timely treatment is known to be important, and not in areas where it is less so. Wait times, and specifically time-to-treatment, therefore do appear to play an important role in explaining patient outcomes.

8 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 around 20 minutes, or 11%, 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, with the majority of mortality reductions occuring in diagnoses where rapid treatment is known to benefit patients.

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 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 (Hoe, 2017).

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. A question raised by our results is why physicians and EDs do not optimize wait times in the absence of the policy. One credible explanation is that physicians are simply imperfect agents for their patients, a longstanding concern in medical markets (Arrow, 1963). This seems especially plausible in our setting where physicians are dealing with patients prior to their full diagnosis being revealed. An alternative explanation could come from physicians lacking information on the relative benefits of timely treatment for certain patients. In practice, however, we are unable to separate these two potential explanations in this analysis. More work is clearly needed to understand the proper set of rules and incentives for delivering cost-effective ED care.

References

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

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.

Cannon, Christopher P., Gibson, C. Michael, Lambrew, Costas T., Shoultz, David A., Levy, Drew, French, William J., Gore, Joel M., Weaver, W. Douglas, Rogers, William J. and Tiefenbrunn, Alan J. (2000). 'Relationship of symptom-onset-to-balloon time and door-to-balloon time with mortality in patients undergoing angioplasty for acute myocardial infarction', *Journal of the American Medical Assosciation* 283(22), 2941–2947.

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.

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.

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

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

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.

Gerard, Francois, Rothe, Christoph and Rokkanen, Miikka. (2018). 'Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable, with an Application to Unemployment Insurance in Brazil', *NBER Working Paper No. 22892* .

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* .

Hoe, Thomas P. (2017), Are Public Hospitals Overcrowded? Evidence from Trauma and Orthopaedics in England. Working Paper.

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.

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.

Locker, Thomas and Mason, Suzanne M. (2006). 'Are these emergency department performance data real?', *Emergency Medicine Journal* 23(7), 558–559.

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.

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

Moulin, Thierry, Sablot, Denis, Vidry, Elisabeth, Belahsen, Faouzi, Berger, Eric, Lemounaud, Patrick, Tatu, Laurent, Vuillier, Fabrice, Cosson, Anne, Revenco, Eugeniu, Capellier, Gilles and Rumbach, Lucien. (2003). 'Impact of emergency room neurologists on patient management and outcome', *European neurology* 50(4), 207–214.

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

Newman-Toker, David E., Moy, Ernest, Valente, Ernest, Coffey, Rosanna and Hines, Anika L. (2014). 'Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample', *Diagnosis* 1(2), 155–166.

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

Saver, Jeffrey L., Fonarow, Gregg C., Smith, Eric E., Reeves, Mathew J., Grau-Sepulveda, Maria V., Pan, Wenqin, Olson, DaiWai M., Hernandez, Adrian F., Peterson, Eric D. and H., Lee. (2013). 'Time to treatment with intravenous tissue plasminogen activator and outcome from acute ischemic stroke', *JAMA* 309(23), 2480–2488.

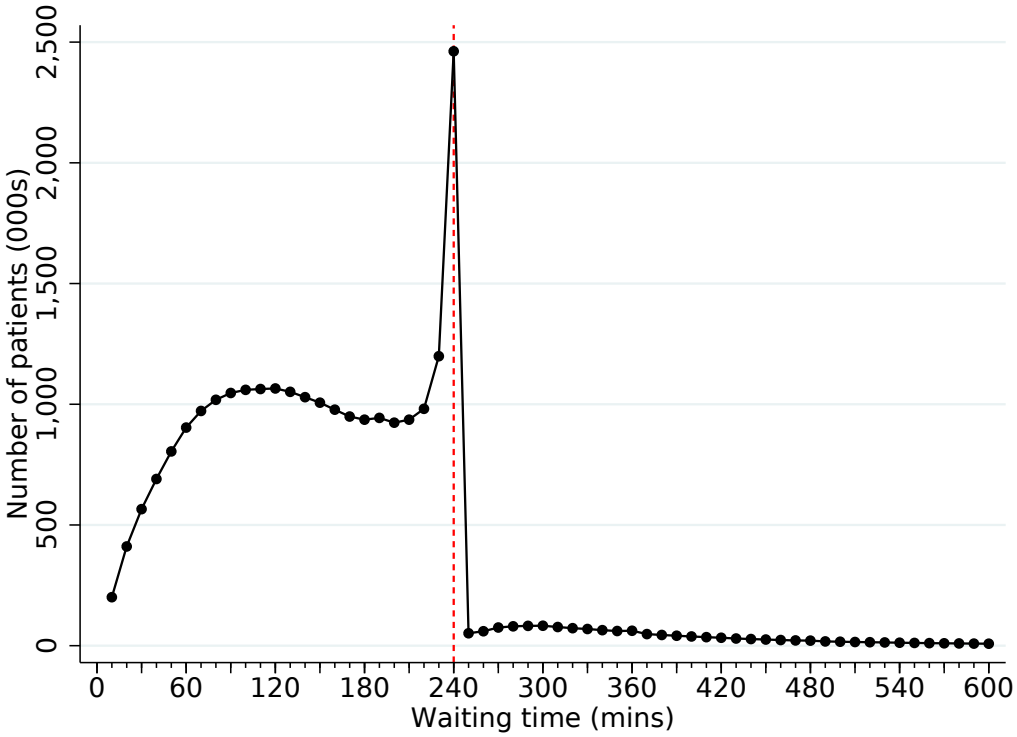
Seymour, Christopher W., Gesten, Foster, Prescott, Hallie C., Friedrich, Marcus E., Iwashyna, Theodore J., Phillips, Gary S., Lemeshow, Stanley, Osborn, Tiffany, Terry, Kathleen M. and Levy, Mitchell M. (2017). 'Time to Treatment and Mortality during Mandated Emergency Care for Sepsis', *New England Journal of Medicine* 376(23), 2235–2244.

Shojania, Kaveh G., Burton, Elizabeth C., McDonald, Kathryn M. and Goldman, Lee. (2003). 'Changes in rates of autopsy-detected diagnostic errors over time: a systematic review.', *Journal of the American Medical Association* 289(21), 2849–56.

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

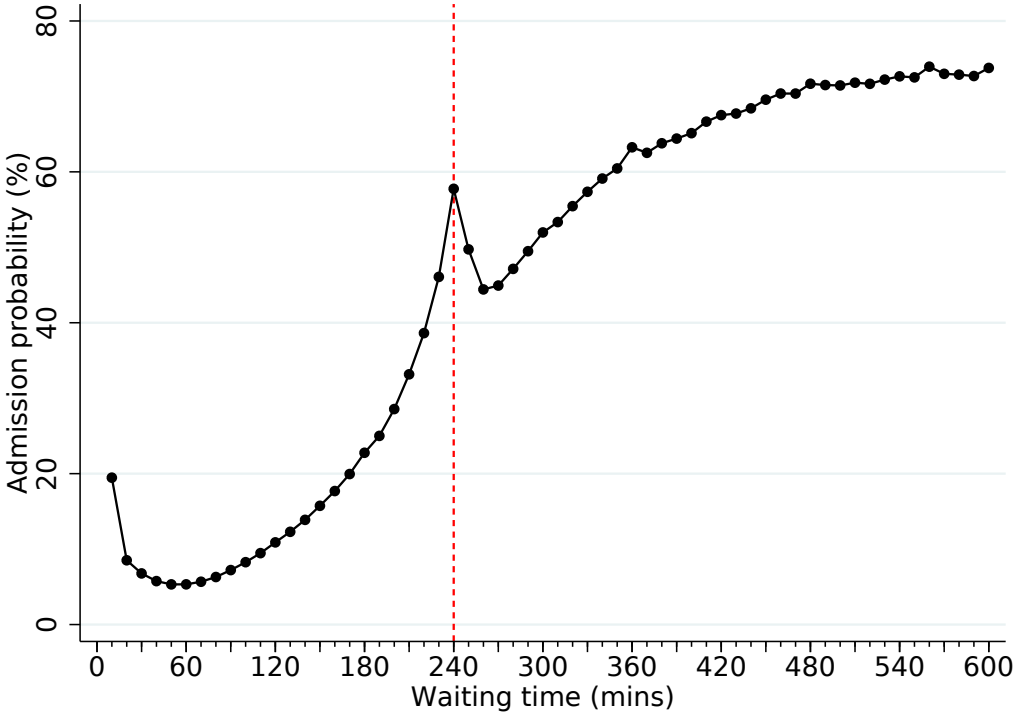
Figures and Tables

Figure 1: Distribution of wait times



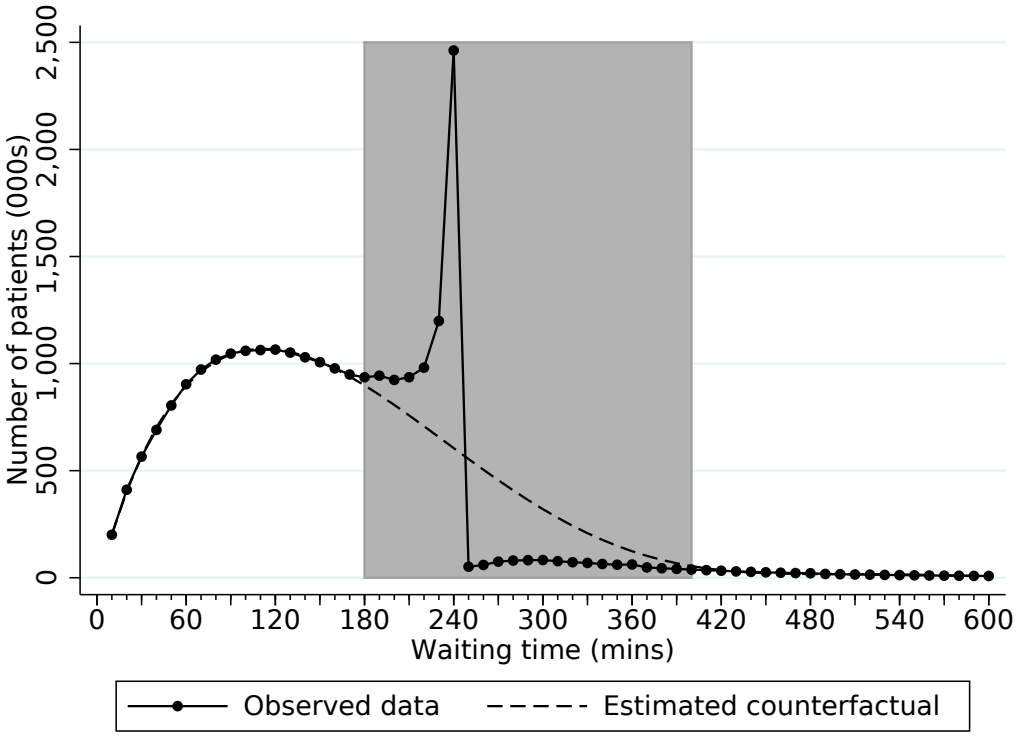
Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy.

Figure 2: Inpatient admission probability conditional on wait time



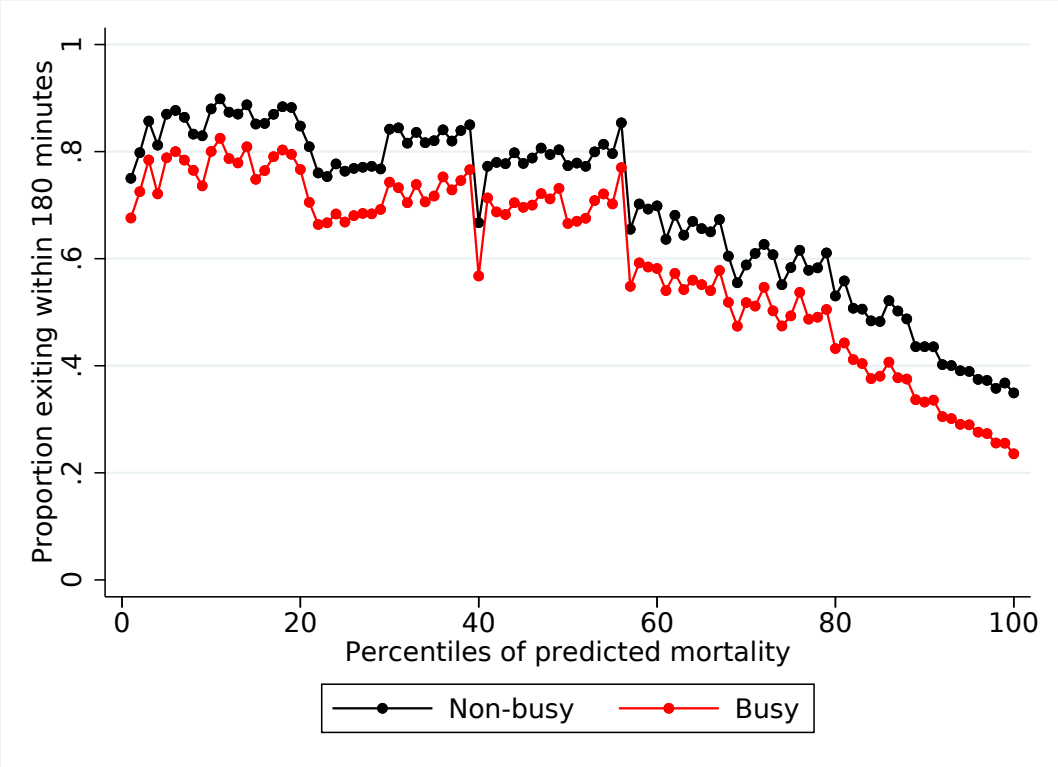
Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy.

Figure 3: Estimated counterfactual wait time distribution



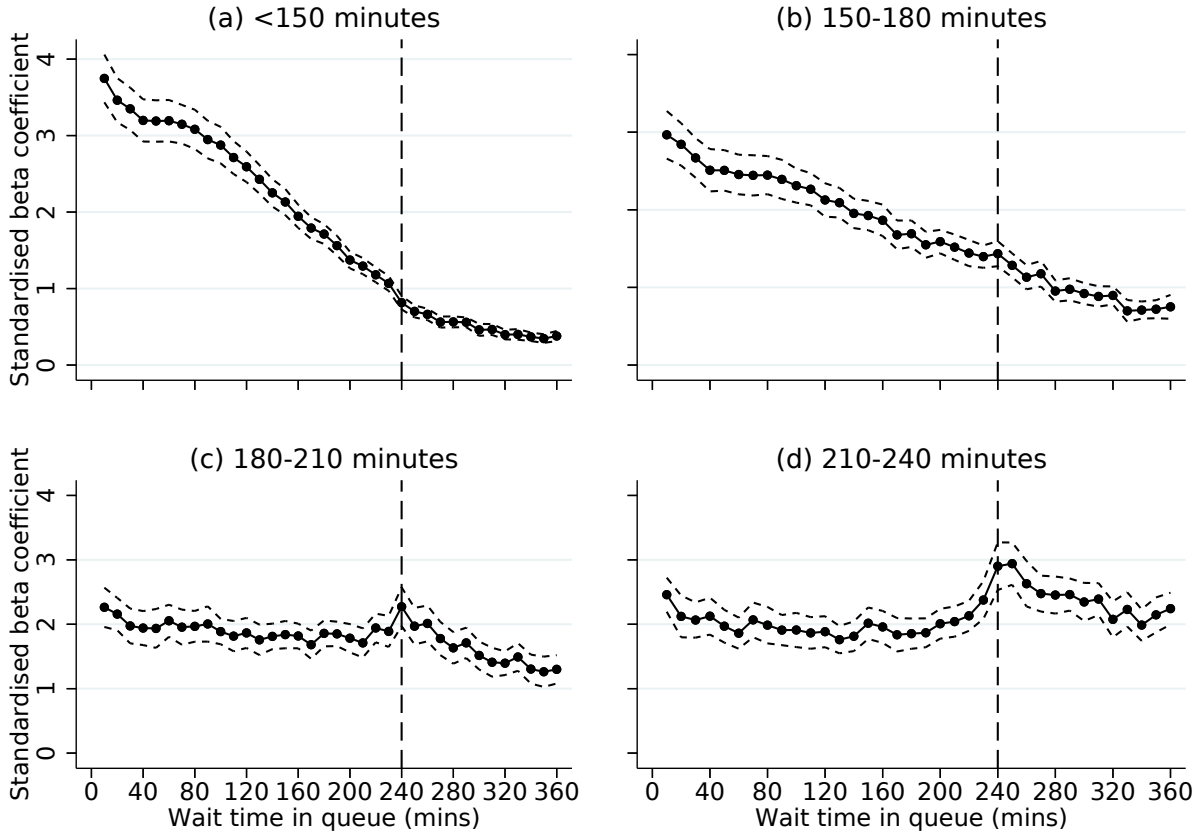
Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy; (4) The estimated counterfactual is obtained from a polynomial regression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

Figure 4: The probability of ED exit within 180 minutes by patient severity and expected volumes of ED arrivals



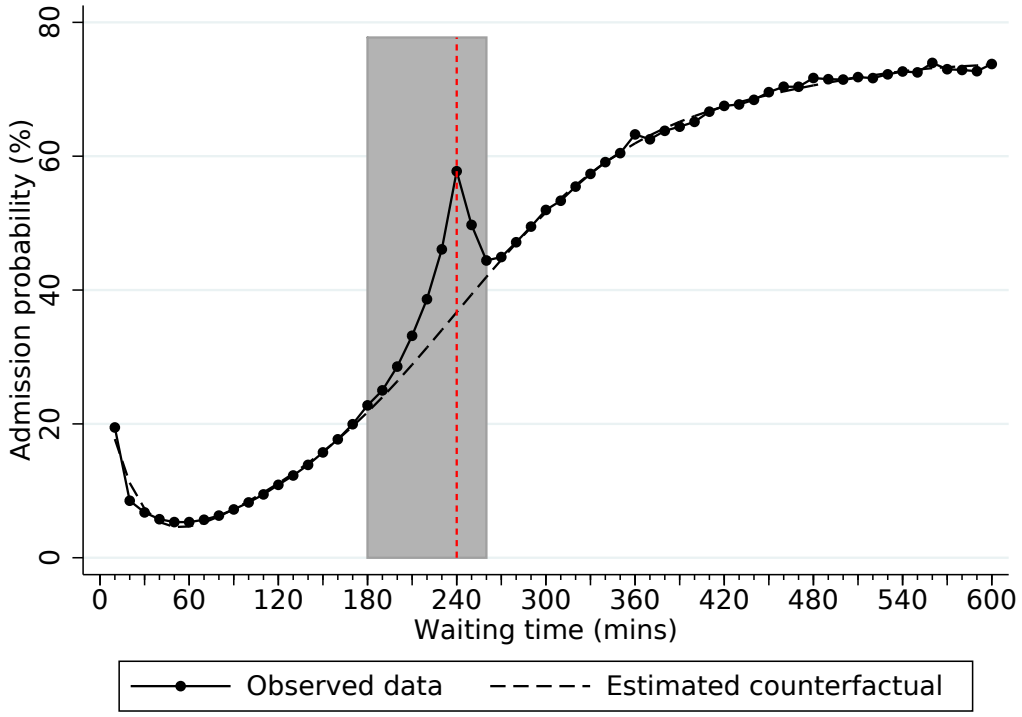
Notes: (1) Wait times defined as the time from arrival in the ED to leaving the ED; (2) Predicted mortality defined using a regression of 30-day in-hospital mortality on a fully interacted set of age, gender, ambulance arrival fixed effects and diagnosis fixed effects; (3) Busy and non-busy periods defined 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).

Figure 5: Impact of queues on wait times for arriving patients by predicted waiting times



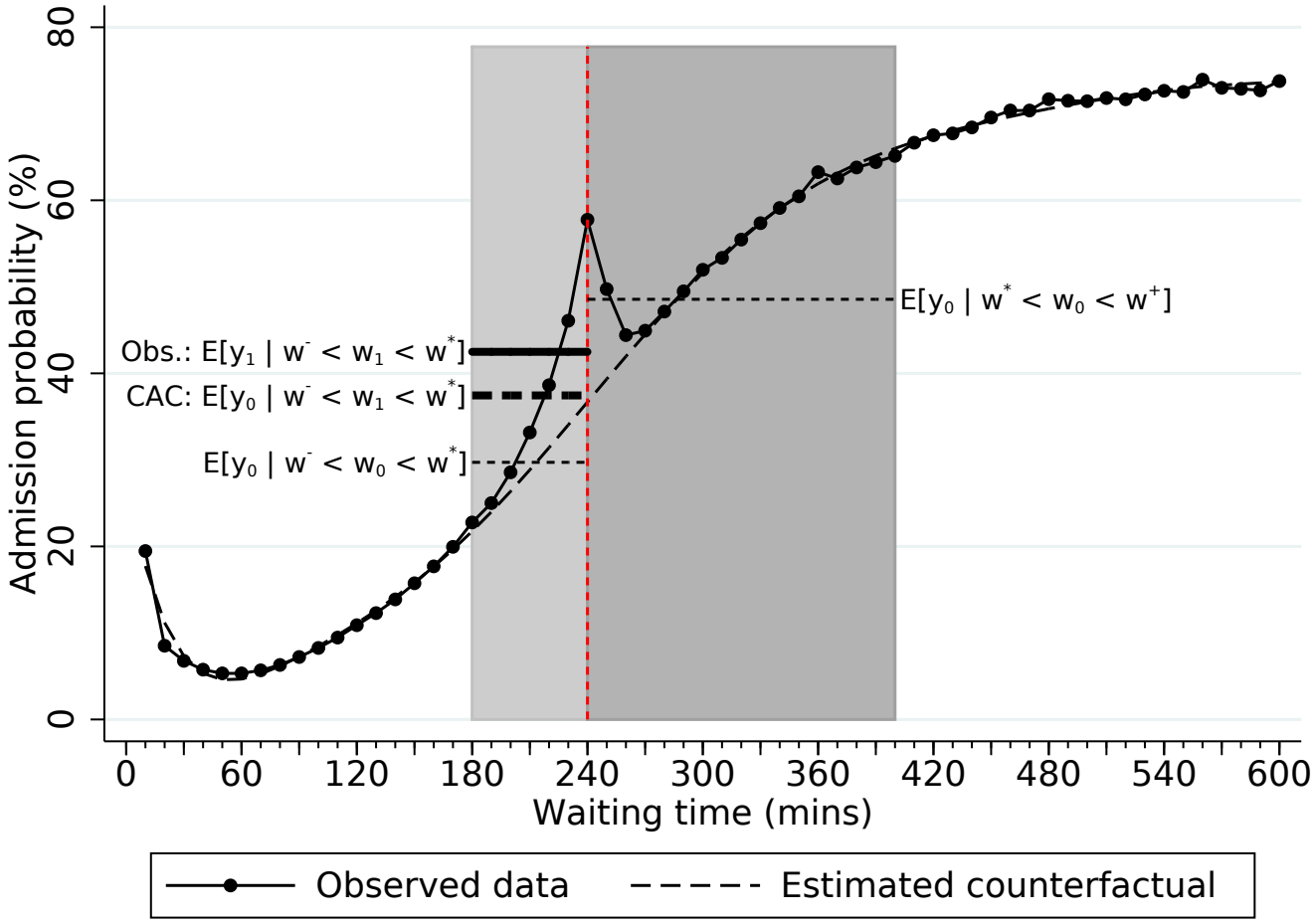
Notes: (1) Wait times defined as the time from arrival in the ED to leaving the ED; (2) We normalise coefficients so they can be interpreted as the impact of a one standard deviation increase in the queue length at each horizon on newly arriving patients' wait times; (3) Predicted wait times are estimated using a regression of wait times on age, gender, diagnosis fixed effects and an ambulance indicator. Panel (a) contains all individuals with predicted wait times below 150 minutes. Panel (b) includes individuals with predicted wait times between 150 and 180 minutes. Panel (c) includes individuals with predicted wait times between 180 and 210 minutes. Panel (d) includes individuals with predicted wait times between 210 and 240 minutes.

Figure 6: Estimated counterfactual admission probability conditional on wait times



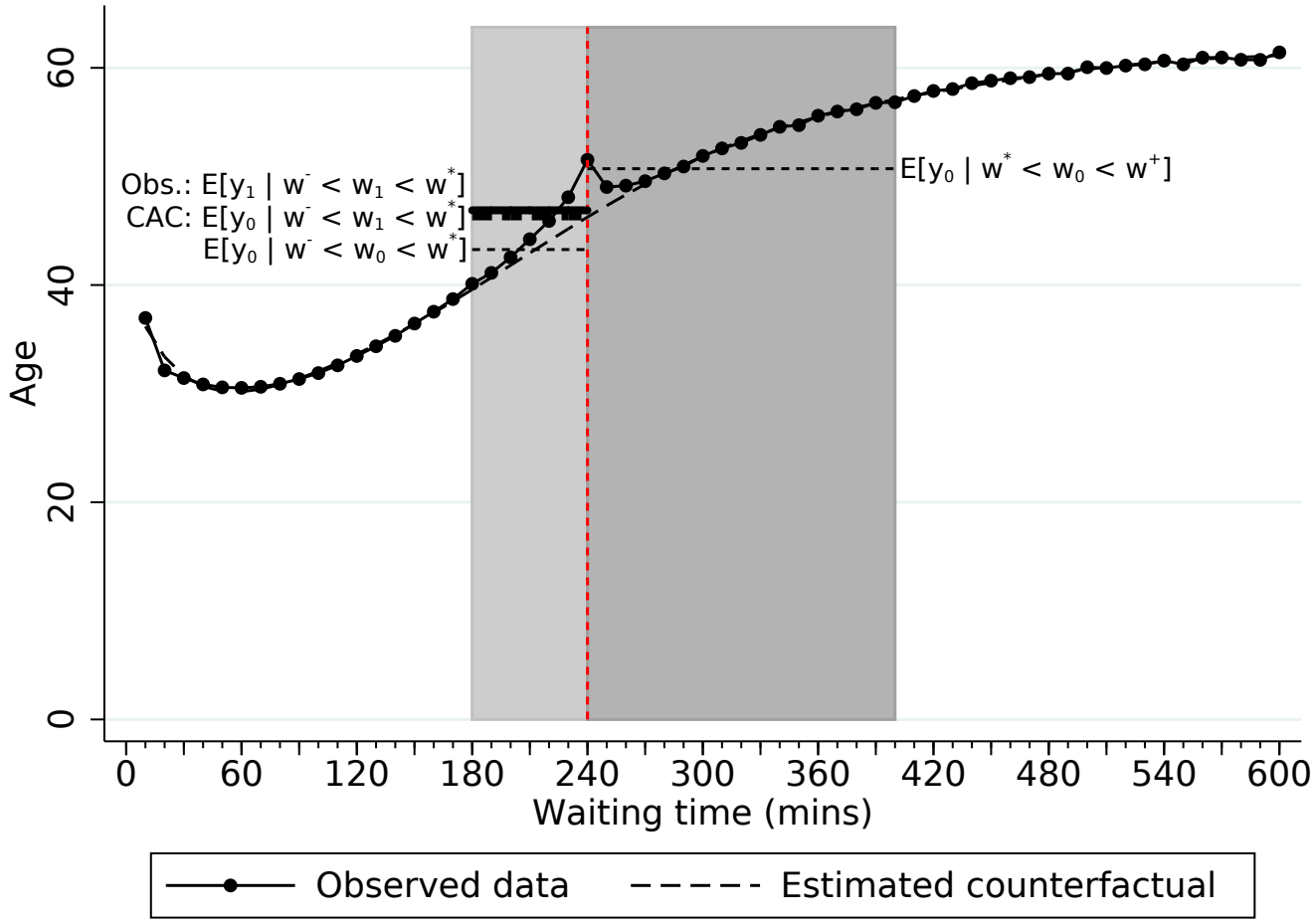
Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy; (4) The estimated counterfactual is obtained from a polynomial regression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

Figure 7: Constructing the composition-adjusted counterfactual for admission probability



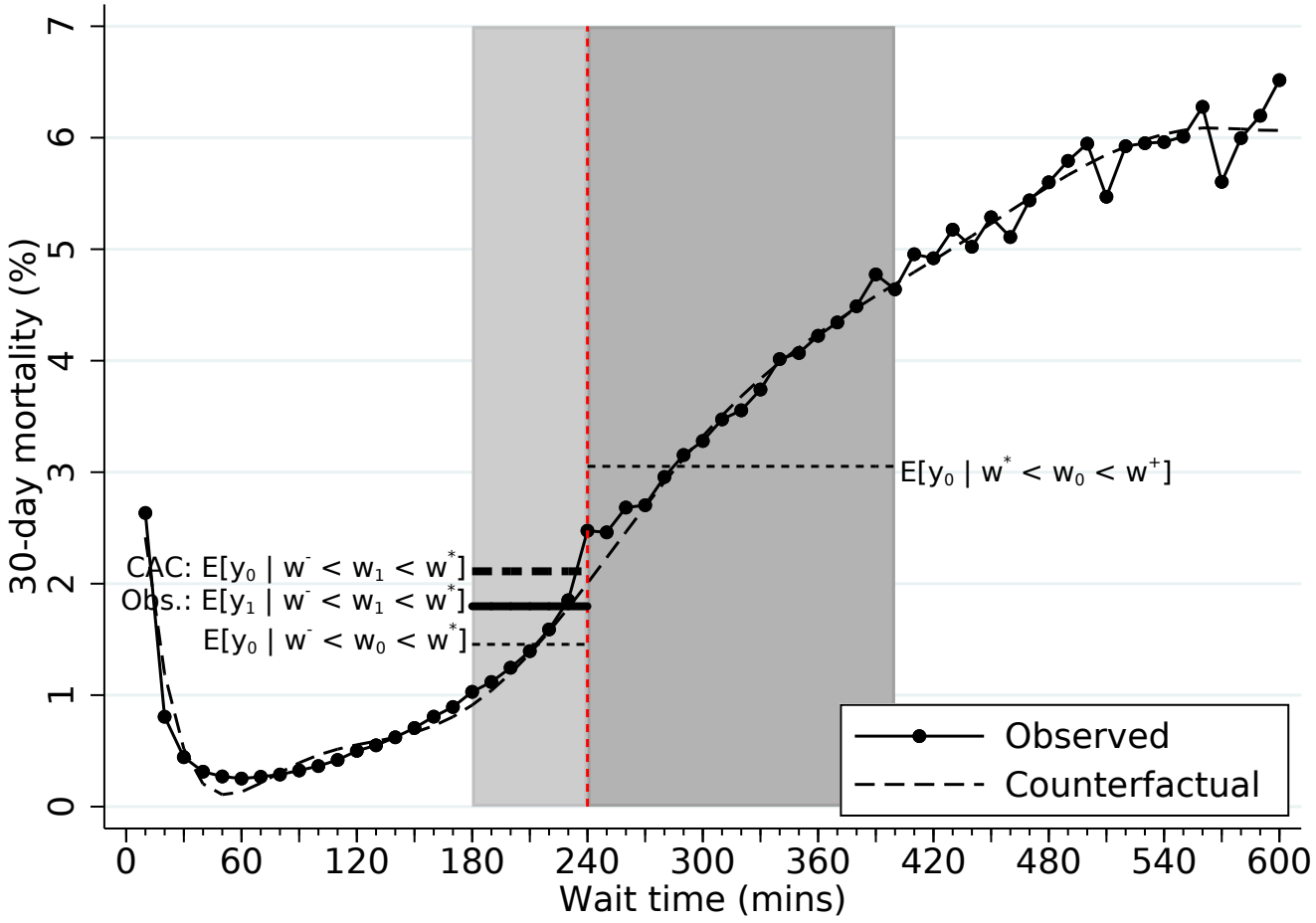
Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy; (4) The horizontal thin dashed lines in the light grey (dark grey) region 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, $\Delta_D = E[y_1 \mid \underline{w}_1^-] - E[y_0 \mid \underline{w}_1^-]$.

Figure 8: Demographic test of the no-selection assumption using age



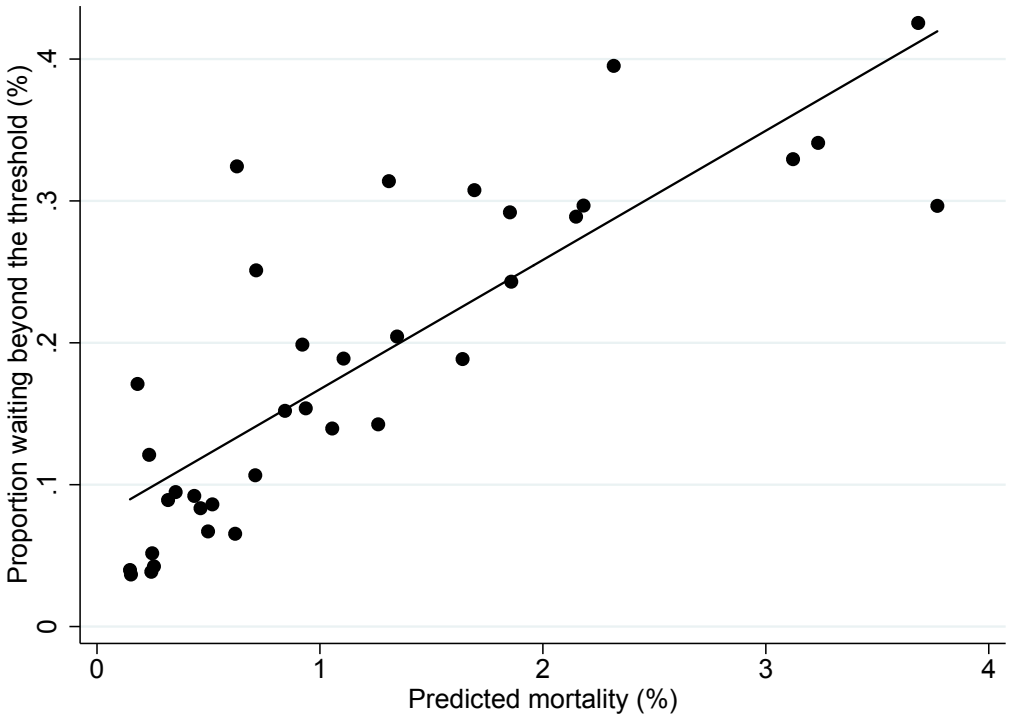
Notes: (1) Wait time intervals are 10-minute periods and defined as the time from arrival in the ED to leaving the ED; (2) Wait times over 600 minutes not shown; (3) 240 minutes is the four-hour threshold specified in the policy; (4) The horizontal thin dashed lines in the light grey (dark grey) region 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, $\Delta_D = E[y_1 \mid \underline{w}_1^-] - E[y_0 \mid \underline{w}_1^-]$.

Figure 9: Constructing the composition-adjusted counterfactual for 30-day mortality



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, $\Delta_D = E[y_1 \mid \underline{w}_1^-] - E[y_0 \mid \underline{w}_1^-]$.

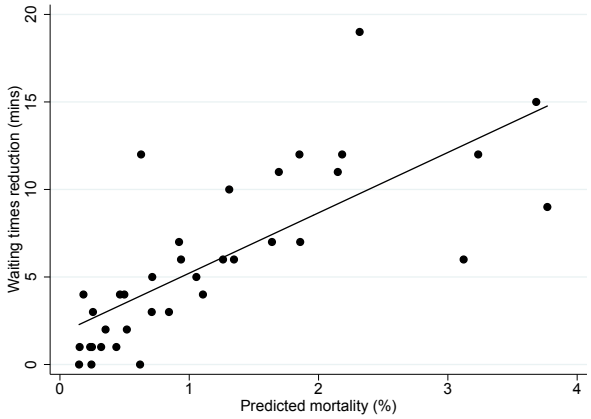
Figure 10: Proportion wait beyond the threshold vs. predicted mortality by diagnosis groups



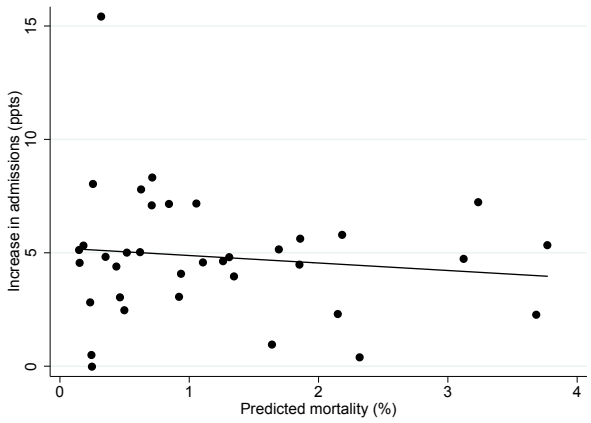
Notes: (1) Each data point corresponds to a diagnosis group average; (2) Proportion waiting beyond the threshold defined using the counterfactual distribution of wait times; (3) Predicted mortality defined using a regression of 30-day in-hospital mortality on past-CCI and a fully interacted set of age,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

Figure 11: Estimated effects of the target vs. predicted mortality by diagnosis groups

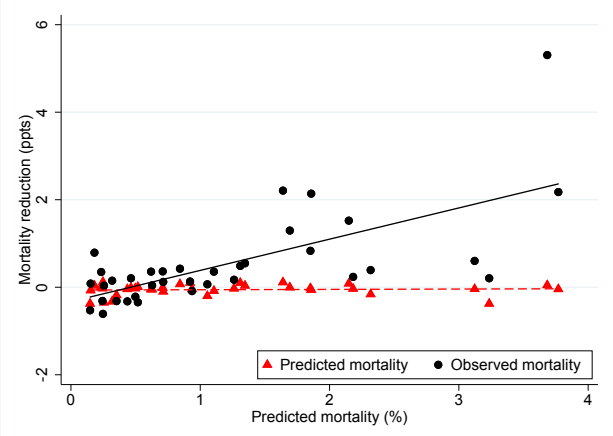
(a) Wait times reductions



(b) Admissions increases



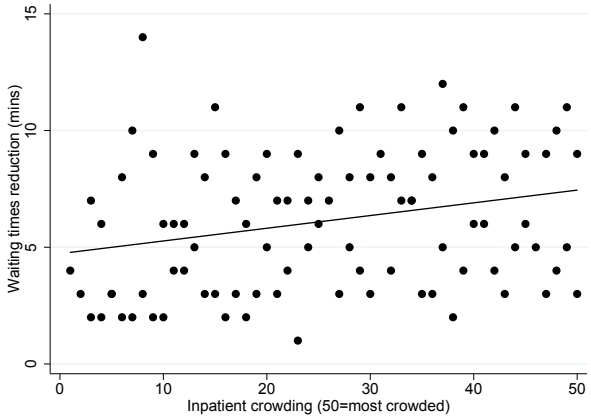
(c) Mortality reductions



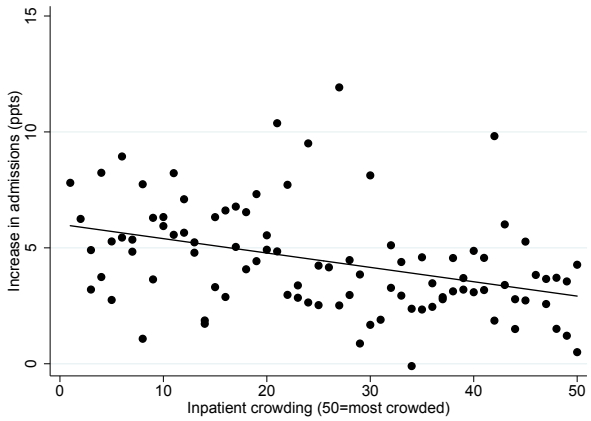
Notes: (1) Each data point corresponds to a diagnosis group average; (2) Predicted mortality defined using a regression of 30-day in-hospital mortality on past-CCI and a fully interacted set of age, gender, and ambulance arrival fixed effects.

Figure 12: Estimated effects of the target vs. inpatient crowding by crowding-severity groups

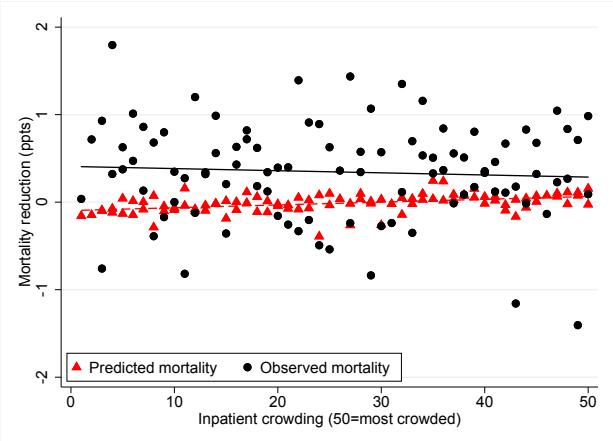
(a) Wait times reductions



(b) Admissions increases



(c) Mortality reductions



Notes: (1) Each data point corresponds to an inpatient crowding-severity group average; (2) Inpatient crowding groups defined according to the number of inpatients treated per hospital-day, which we then use to split into 50 quantiles; (3) Severity is defined as diagnoses with a mean 30-day mortality rate above the mean overall 30-day mortality rate; (4) Predicted mortality defined using a regression of 30-day in-hospital mortality on past-CCI and a fully interacted set of age, gender, and ambulance arrival fixed effects.

Table 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
Past-CCI	0.20	0.78	0.47	1.20
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 treatment count	1.81	1.38	2.22	1.68
ED investigation count	1.54	2.03	3.18	2.50
Inpatient length of stay (days)	1.28	5.63	5.41	10.58
Inpatient procedure count	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.

Table 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)	
Past-CCI	0.013^{***}	0.043^{***}	0.300
	(0.005)	(0.016)	
Panel B: Predicted characteristics			
Predicted admission	0.002	0.007	0.307
	(0.002)	(0.007)	
Predicted mortality	0.000	0.014	0.019
	(0.000)	(0.013)	

Notes: (1) CAC mean is measured over the pre-threshold period, $E[y_0 \mid \underline{w}_1^-]$; (2) Predicted variables defined using a regression of the variable on past-CCI score, and a fully interacted set of age, gender and ambulance-arrival fixed effects; (3) Bootstrapped standard errors clustered at the hospital trust level (199 repetitions).

Table 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 investigation count	0.108**	0.046**	2.369
	(0.048)	(0.021)	
ED treatment count	−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 procedure count	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) All inpatient variables (e.g. length of stay, costs) take on the value zero for patients that are not admitted; (3) Bootstrapped standard errors clustered at the hospital trust level (199 repetitions).

Table 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).

Table 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.

A Online Appendix: Additional Figures and Tables

Tables A1 and A2 show the most common first and subsequent ED investigations and treatments, by ED diagnosis, for all ED and admitted patients respectively. The tables show that ED treatments are often simple, with basic treatment for simple conditions and diagnostic investigations for more complex cases.

Table A3 shows the most common first and subsequent inpatient procedures for all admitted patients, by ED diagnosis. Initial procedures are largely diagnostic, continuing the diagnostic investigations carried out in the ED.

Table A4 shows the estimated impact of the target according to the lower bound of the exclusion window. In our baseline results (as shown by Table 3), the exclusion window began at 180 minutes. Table A4 shows two other scenarios: 170 and 190 minutes. In each case, an iterative procedure is used to automatically pick the end point of the exclusion window, as described in Section 4.2. The results show that the main estimates are robust to the choice of the lower bound of the exclusion window. While the exact magnitude of the estimates varies, the results show that the target is associated with an increased admission probability, increased number of ED investigations, increased costs and reduced 30-day mortality.

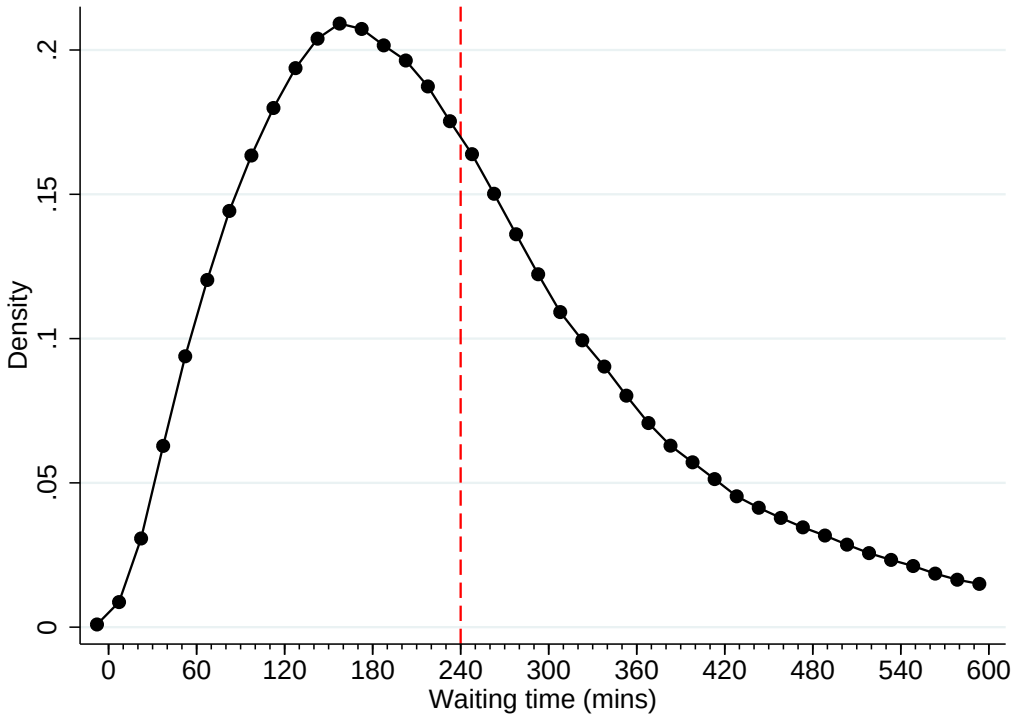
Table A5 presents the results of a second robustness check, showing how the estimated impact of the target varies by the order of polynomial used in estimation. The baseline results (shown in Table 3) use a polynomial of 10. Table A5 presents additional results when using a polynomial of order 6 and 8. In both cases, the results are similar to the baseline results. The final column shows the results from separately picking the polynomial for each outcomes. This is an automated process that maximised the adjusted-R^2 statistic from estimating Equation (1). Again, the results are similar to the baseline estimates presented in the main text.

Table A6 shows the estimated mortality impact of the target by ED diagnosis. We created 36 outcome variables, which take the value of 1 if a patient with a specific ED diagnosis died, and 0 otherwise. We produce bootstrapped standard errors using 199 repetitions. The results show that the largest mortality reductions took place in potentially time-sensitive diagnoses with high baseline mortality rates.

Table A7 shows the estimated mortality impact of the target on broad causes of death. We created 23 outcome variables, which take the value of 1 if the ED patient dies of a specific cause of death, based on the first letter of the ICD-10 chapter recorded on their official death certificate, and 0 otherwise. We produce bootstrapped standard errors using 199 repetitions. The results show that the mortality reductions are focused among potentially time-sensitive conditions: circulatory, respiratory and digestive conditions. The results are discussed in more detail in Section 7.2.

Table A8 shows the estimated mortality impact of the target on the ten most common causes of death, as defined by the first letter and first digit of the ICD-10 code recorded on official death certificates. We create 10 dummy indicators, which take the value of 1 if the ED patient dies of a specific cause within 30 days of visiting an ED, and zero otherwise. Together, these conditions account for 60% of ED patient deaths in 2011/12 and 2012/13. Again, the results show that mortality impacts are focused among the potentially time-sensitive conditions, but not among others such as cancer. The results are discussed in more detail in Section 7.2.

Figure A1: 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 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 A1: The most common ED invesigations and treatments, all patients

	Most common ED investigations				Most common ED treatments			
ED diagnosis	First investigation	% of patients	Subsequent investigation	% of patients	First treatment	% of patients	Subsequent treatment	% of patients
Laceration	None	65.8%	Biochemistry	2.6%	Wound closure	20.1%	Wound closure	18.8%
Contusion/abrasion	X-ray plain film	55.0%	Biochemistry	3.0%	Guidance/advice only	36.8%	None	6.8%
Soft tissue inflammation	X-ray plain film	51.9%	Biochemistry	6.2%	Guidance/advice only	37.2%	Medication administered	7.0%
Head injury	None	61.6%	Biochemistry	6.9%	Guidance/advice only	30.9%	Observation (ECG, pulse oximetry etc)	12.8%
Joint injury/fracture	X-ray plain film	81.6%	Haematology	8.0%	Plaster of paris	17.9%	Plaster of paris	9.4%
Sprain/ligament injury	X-ray plain film	65.0%	Biochemistry	1.7%	Guidance/advice only	38.8%	Medication administered	6.4%
Muscle/tendon injury	X-ray plain film	44.9%	Biochemistry	7.7%	Guidance/advice only	30.5%	Medication administered	9.1%
Nerve injury	None	42.9%	Biochemistry	19.2%	Recording vital signs	36.4%	Medication administered	17.2%
Vascular injury	None	31.9%	Biochemistry	21.5%	Guidance/advice only	16.7%	Observation (ECG, pulse oximetry etc)	16.0%
Burns and scalds	None	84.4%	ECG	1.8%	Dressing	42.4%	Dressing	17.8%
Electric shock	X-ray plain film	50.7%	Urinalysis	2.9%	Guidance/advice only	44.2%	Medication administered	15.1%
Foreign body	None	62.3%	Biochemistry	1.5%	Removal foreign body	26.4%	Removal foreign body	12.2%
Bites/stings	None	78.3%	Biochemistry	2.1%	Medicines prepared to take away	18.9%	Medicines prepared to take away	19.3%
Poisoning (inc overdose)	None	28.4%	Biochemistry	31.0%	Observation (ECG, pulse oximetry etc)	22.5%	Observation (ECG, pulse oximetry etc)	19.1%
Near drowning	None	73.4%	Biochemistry	5.8%	Dressing	28.2%	Dressing	36.9%
Visceral injury	None	38.1%	Biochemistry	20.8%	Guidance/advice only	16.0%	Observation (ECG, pulse oximetry etc)	22.0%
Infectious disease	None	43.2%	Biochemistry	22.3%	Guidance/advice only	18.4%	Observation (ECG, pulse oximetry etc)	15.1%
Local infection	None	49.5%	Biochemistry	16.7%	Medicines prepared to take away	18.5%	Observation (ECG, pulse oximetry etc)	13.8%
Septicaemia	X-ray plain film	44.3%	Haematology	47.3%	Intravenous cannula	21.3%	Observation (ECG, pulse oximetry etc)	27.7%
Cardiac	X-ray plain film	38.7%	Haematology	49.4%	Intravenous cannula	19.2%	Observation (ECG, pulse oximetry etc)	21.5%
Cerebro-vascular	CT scan	26.4%	Biochemistry	43.7%	Intravenous cannula	21.8%	Observation (ECG, pulse oximetry etc)	21.6%
Other vascular	Haematology	20.5%	Biochemistry	33.5%	Medication administered	18.7%	Observation (ECG, pulse oximetry etc)	17.5%
Haematological	X-ray plain film	29.0%	Biochemistry	36.3%	Guidance/advice only	18.7%	Observation (ECG, pulse oximetry etc)	16.8%
Central nervous system	None	24.6%	Biochemistry	33.1%	Observation (ECG, pulse oximetry etc)	20.3%	Observation (ECG, pulse oximetry etc)	18.8%
Respiratory	X-ray plain film	40.7%	Haematology	33.1%	Observation (ECG, pulse oximetry etc)	15.2%	Observation (ECG, pulse oximetry etc)	19.5%
Gastrointestinal	X-ray plain film	21.8%	Biochemistry	36.8%	Observation (ECG, pulse oximetry etc)	15.2%	Observation (ECG, pulse oximetry etc)	18.4%
Urological	Urinalysis	19.3%	Biochemistry	34.7%	Observation (ECG, pulse oximetry etc)	13.5%	Observation (ECG, pulse oximetry etc)	19.3%
Obstetric	None	21.8%	Biochemistry	24.9%	Observation (ECG, pulse oximetry etc)	19.6%	Observation (ECG, pulse oximetry etc)	18.0%
Gynaecological	Haematology	26.4%	Biochemistry	32.1%	Guidance/advice only	19.2%	Observation (ECG, pulse oximetry etc)	14.3%
Diabetes and endocrine	Haematology	24.1%	Biochemistry	44.2%	Intravenous cannula	20.6%	Observation (ECG, pulse oximetry etc)	19.4%
Dermatological	None	68.4%	Biochemistry	9.3%	Guidance/advice only	23.1%	Recording vital signs	11.8%
Allergy (inc anaphylaxis)	None	65.9%	Biochemistry	11.2%	Medication administered	18.6%	Observation (ECG, pulse oximetry etc)	15.7%
Facio-maxillary conditions	None	66.0%	Biochemistry	6.8%	Guidance/advice only	20.1%	Observation (ECG, pulse oximetry etc)	11.9%
ENT	None	61.5%	Biochemistry	11.9%	Guidance/advice only	20.8%	Observation (ECG, pulse oximetry etc)	11.4%
Psychiatric	None	55.8%	Biochemistry	16.5%	Guidance/advice only	28.0%	Observation (ECG, pulse oximetry etc)	11.6%
Ophthalmological	None	58.3%	Biochemistry	1.6%	Guidance/advice only	26.4%	Medication administered	11.9%
Social problems	None	30.8%	Biochemistry	29.5%	Observation (ECG, pulse oximetry etc)	21.1%	Observation (ECG, pulse oximetry etc)	17.1%
Diagnosis not classigiable	None	41.1%	Biochemistry	21.7%	None	18.5%	Observation (ECG, pulse oximetry etc)	12.6%
Nothing abnormal detected	None	53.0%	Biochemistry	15.1%	None	33.1%	None	12.1%
Diagnosis missing	None	39.2%	Biochemistry	19.1%	Guidance/advice only	19.3%	Observation (ECG, pulse oximetry etc)	11.7%

Notes: (1) A full list of investigations and treatments are available from the NHS Digital HES Data Dictionary (Accident and Emergency): https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary; (2) First ED investigation/treatment contains the first recorded investigation/treatment code for a specific ED visit; (3) Subsequent investigations/treatments combined information across all other investigation/treatment codes in a specific ED visit (up to 12 investigations and 8 treatments).

Table A2: The most common ED invesigations and treatments, admitted patients only

	Most common ED investigations				Most common ED treatments			
ED diagnosis	First investigation	% of patients	Subsequent investigation	% of patients	First treatment	% of patients	Subsequent treatment	% of patients
Laceration	X-ray plain film	36.7%	Haematology	25.9%	Observation (ECG, pulse oximetry etc)	11.6%	Observation (ECG, pulse oximetry etc)	16.2%
Contusion/abrasion	X-ray plain film	50.0%	Haematology	37.1%	Observation (ECG, pulse oximetry etc)	14.1%	Observation (ECG, pulse oximetry etc)	19.3%
Soft tissue inflammation	X-ray plain film	40.8%	Biochemistry	40.0%	Intravenous cannula	15.7%	Observation (ECG, pulse oximetry etc)	18.5%
Head injury	CT scan	26.7%	Haematology	27.9%	Observation (ECG, pulse oximetry etc)	25.1%	Observation (ECG, pulse oximetry etc)	22.2%
Joint injury/fracture	X-ray plain film	72.3%	Haematology	39.4%	Intravenous cannula	16.6%	Intravenous cannula	18.8%
Sprain/ligament injury	X-ray plain film	56.8%	Biochemistry	30.6%	Medication administered	19.3%	Recording vital signs	17.0%
Muscle/tendon injury	X-ray plain film	56.5%	Haematology	34.7%	Medication administered	20.5%	Recording vital signs	19.8%
Nerve injury	ECG	36.1%	Biochemistry	58.1%	Recording vital signs	59.1%	Intravenous cannula	47.7%
Vascular injury	X-ray plain film	18.9%	Biochemistry	40.8%	Intravenous cannula	21.0%	Observation (ECG, pulse oximetry etc)	22.9%
Burns and scalds	Haematology	48.5%	ECG	37.0%	Other parenteral drugs	33.0%	Medication administered	9.6%
Electric shock	Haematology	27.0%	Haematology	19.4%	Guidance/advice only	16.9%	Medication administered	16.4%
Foreign body	X-ray plain film	54.2%	Haematology	19.6%	Observation (ECG, pulse oximetry etc)	16.4%	Observation (ECG, pulse oximetry etc)	15.6%
Bites/stings	X-ray plain film	34.5%	Haematology	24.2%	Intravenous cannula	14.4%	Intravenous cannula	16.4%
Poisoning (inc overdose)	Haematology	29.4%	Biochemistry	41.5%	Observation (ECG, pulse oximetry etc)	23.1%	Observation (ECG, pulse oximetry etc)	22.0%
Near drowning	X-ray plain film	56.0%	Haematology	42.0%	Intravenous cannula	18.8%	Observation (ECG, pulse oximetry etc)	22.8%
Visceral injury	X-ray plain film	28.9%	Biochemistry	50.8%	Intravenous cannula	18.2%	Observation (ECG, pulse oximetry etc)	35.6%
Infectious disease	X-ray plain film	33.7%	Biochemistry	45.4%	Other parenteral drugs	17.4%	Observation (ECG, pulse oximetry etc)	23.3%
Local infection	X-ray plain film	32.2%	Biochemistry	41.1%	Intravenous cannula	21.4%	Observation (ECG, pulse oximetry etc)	24.1%
Septicaemia	X-ray plain film	50.8%	Haematology	57.2%	Intravenous cannula	25.8%	Intravenous cannula	31.6%
Cardiac	X-ray plain film	47.4%	Haematology	57.7%	Intravenous cannula	24.9%	Observation (ECG, pulse oximetry etc)	24.3%
Cerebro-vascular	CT scan	36.4%	Haematology	54.0%	Intravenous cannula	29.4%	Observation (ECG, pulse oximetry etc)	26.0%
Other vascular	X-ray plain film	27.9%	Haematology	44.8%	Intravenous cannula	19.0%	Observation (ECG, pulse oximetry etc)	25.0%
Haematological	X-ray plain film	32.5%	Haematology	48.6%	Intravenous cannula	24.5%	Observation (ECG, pulse oximetry etc)	25.3%
Central nervous system	X-ray plain film	21.8%	Haematology	46.9%	Observation (ECG, pulse oximetry etc)	22.0%	Observation (ECG, pulse oximetry etc)	22.8%
Respiratory	X-ray plain film	51.3%	Haematology	48.7%	Intravenous cannula	18.4%	Observation (ECG, pulse oximetry etc)	23.8%
Gastrointestinal	X-ray plain film	32.6%	Biochemistry	46.7%	Intravenous cannula	23.9%	Intravenous cannula	23.6%
Urological	X-ray plain film	24.1%	Biochemistry	48.7%	Intravenous cannula	21.3%	Observation (ECG, pulse oximetry etc)	24.8%
Obstetric	Haematology	28.0%	Biochemistry	35.7%	Observation (ECG, pulse oximetry etc)	24.4%	Observation (ECG, pulse oximetry etc)	23.0%
Gynaecological	Haematology	33.8%	Biochemistry	44.0%	Intravenous cannula	21.5%	Observation (ECG, pulse oximetry etc)	18.1%
Diabetes and endocrine	X-ray plain film	33.6%	Biochemistry	51.0%	Intravenous cannula	27.2%	Intravenous cannula	24.1%
Dermatological	None	25.9%	Biochemistry	34.2%	Observation (ECG, pulse oximetry etc)	17.4%	Recording vital signs	23.1%
Allergy (inc anaphylaxis)	None	31.9%	Biochemistry	30.2%	Observation (ECG, pulse oximetry etc)	16.6%	Observation (ECG, pulse oximetry etc)	24.1%
Facio-maxillary conditions	X-ray plain film	37.7%	Biochemistry	32.8%	Intravenous cannula	17.9%	Observation (ECG, pulse oximetry etc)	29.0%
ENT	Haematology	29.8%	Biochemistry	34.9%	Intravenous cannula	18.6%	Observation (ECG, pulse oximetry etc)	20.4%
Psychiatric	None	22.7%	Biochemistry	40.3%	Observation (ECG, pulse oximetry etc)	23.4%	Observation (ECG, pulse oximetry etc)	19.3%
Ophthalmological	None	23.7%	Biochemistry	30.3%	Observation (ECG, pulse oximetry etc)	17.1%	Observation (ECG, pulse oximetry etc)	27.1%
Social problems	X-ray plain film	34.1%	Biochemistry	46.2%	Observation (ECG, pulse oximetry etc)	23.1%	Observation (ECG, pulse oximetry etc)	20.2%
Diagnosis not classigiable	X-ray plain film	35.4%	Biochemistry	45.0%	Observation (ECG, pulse oximetry etc)	21.5%	Observation (ECG, pulse oximetry etc)	22.1%
Nothing abnormal detected	X-ray plain film	29.8%	Biochemistry	38.2%	None	31.8%	Recording vital signs	21.1%
Diagnosis missing	X-ray plain film	34.5%	Biochemistry	48.1%	Recording vital signs	18.6%	Intravenous cannula	21.9%

Notes: (1) Includes only ED visits which resulted in an inpatient admission; (2) A full list of investigations and treatments are available from the NHS Digital HES Data Dictionary (Accident and Emergency):
https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary; (3) First ED investigation/treatment contains the first recorded investigation/treatment code for a specific ED visit; (4) Subsequent investigations/treatments combined information across all other investigation/treatment codes in a specific ED visit (up to 12 investigations and 8 treatments).

Table A3: The most common inpatient procedures

ED diagnosis	Most common inpatient procedures			
	First procedure	% of patients	Subsequent procedure	% of patients
Laceration	Suture of skin	17.4%	Debridement/cleaning of skin/wound	48.6%
Contusion/abrasion	CT / MRI scan of head, spine or CNT	33.2%	Treatment/examination of pelvis or spine	36.6%
Soft tissue inflammation	CT / MRI scan (site not specified)	12.3%	Treatment/examination of pelvis or spine	22.4%
Head injury	CT / MRI scan of head, spine or CNT	67.8%	CT / MRI scan of head, spine or CNT	49.2%
Joint injury/fracture	Closed reduction of fracture	17.3%	Debridement/cleaning of skin/wound	13.9%
Sprain/ligament injury	CT / MRI scan of head, spine or CNT	23.8%	Treatment/examination of pelvis or spine	23.9%
Muscle/tendon injury	Primary repair of tendon	21.4%	Debridement/cleaning of skin/wound	29.8%
Nerve injury	CT / MRI scan of head, spine or CNT	22.0%	Treatment/examination of pelvis or spine	21.0%
Vascular injury	CT / MRI scan (site not specified)	12.5%	CT / MRI scan (site not specified)	22.2%
Burns and scalds	CT / MRI scan of head, spine or CNT	18.9%	Debridement/cleaning of skin/wound	21.5%
Electric shock	CT / MRI scan of head, spine or CNT	28.8%	Debridement/cleaning of skin/wound	23.1%
Foreign body	Removal of inorganic substance from the skin	17.0%	Debridement/cleaning of skin/wound	28.4%
Bites/stings	Debridement/cleaning of skin/wound	28.1%	Debridement/cleaning of skin/wound	72.3%
Poisoning (inc overdose)	CT / MRI scan of head, spine or CNT	31.3%	CT / MRI scan of head, spine or CNT	35.6%
Near drowning	Ventilation	19.0%	CT / MRI scan of head, spine or CNT	50.0%
Visceral injury	CT / MRI scan (site not specified)	26.5%	Treatment/examination of pelvis or spine	62.3%
Infectious disease	CT / MRI scan of head, spine or CNT	22.3%	CT / MRI scan of head, spine or CNT	32.1%
Local infection	Drainage/incision of lesion or skin	23.1%	Debridement/cleaning of skin/wound	21.8%
Septicaemia	CT / MRI scan of head, spine or CNT	19.0%	Treatment/examination of pelvis or spine	34.5%
Cardiac	Echocardiography	25.3%	Echocardiography	46.6%
Cerebro-vascular	CT / MRI scan of head, spine or CNT	76.5%	CT / MRI scan of head, spine or CNT	73.0%
Other vascular	CT / MRI scan of head, spine or CNT	31.2%	CT / MRI scan of head, spine or CNT	23.0%
Haematological	Blood transfusion (inc blood stem cell transplant)	15.1%	Treatment/examination of pelvis or spine	41.2%
Central nervous system	CT / MRI scan of head, spine or CNT	70.9%	CT / MRI scan of head, spine or CNT	67.8%
Respiratory	Ventilation	23.1%	Treatment/examination of pelvis or spine	24.5%
Gastrointestinal	CT / MRI scan (site not specified)	29.2%	Treatment/examination of pelvis or spine	68.3%
Urological	CT / MRI scan (site not specified)	30.7%	Treatment/examination of pelvis or spine	40.3%
Obstetric	Aspiration/extraction of products of conception from uterus	21.5%	Examination of female genital tract	25.6%
Gynaecological	Examination of female genital tract	22.9%	Treatment/examination of pelvis or spine	28.5%
Diabetes and endocrine	CT / MRI scan of head, spine or CNT	26.0%	Treatment/examination of pelvis or spine	39.4%
Dermatological	Drainage/incision of lesion or skin	17.7%	Treatment/examination of pelvis or spine	29.5%
Allergy (inc anaphylaxis)	CT / MRI scan of head, spine or CNT	16.4%	Treatment/examination of pelvis or spine	22.9%
Facio-maxillary conditions	Operations on tooth and surrounding area	20.9%	Extraction of teeth	30.5%
ENT	Packing of cavity of nose	28.2%	Packing of cavity of nose	70.4%
Psychiatric	CT / MRI scan of head, spine or CNT	32.9%	CT / MRI scan of head, spine or CNT	25.6%
Ophthalmological	CT / MRI scan of head, spine or CNT	35.8%	CT / MRI scan of head, spine or CNT	24.6%
Social problems	CT / MRI scan of head, spine or CNT	52.4%	CT / MRI scan of head, spine or CNT	34.4%
Diagnosis not classigiable	CT / MRI scan of head, spine or CNT	26.6%	Treatment/examination of pelvis or spine	32.0%
Nothing abnormal detected	CT / MRI scan of head, spine or CNT	27.0%	Treatment/examination of pelvis or spine	36.0%
Diagnosis missing	CT / MRI scan of head, spine or CNT	23.3%	Treatment/examination of pelvis or spine	31.5%

Notes: (1) Inpatient procedures are recorded using OPCS4.8 codes. For a mapping of OPCS4.8 codes to procedures, see:http://www.surginet.org.uk/informatics/opcs.php; (2) First procedures contains the first recorded procedure code for a specific inpatient spell; (3) Subsequent procedures combine information across all other procedure codes in a specific ED visit (up to 12 investigations and 8 treatments).

Table A4: Estimated distortion effects of the target, robustness by exclusion window lower bound

	Exclusion window lower bound (mins):				
	160	170	180 (baseline)	190	200
Panel A: ED treatment decisions					
Pr(admission)	0.033***	0.039***	0.046***	0.058***	0.074***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Pr(discharge)	−0.022***	−0.027***	−0.033***	−0.042***	−0.054***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Pr(referral)	−0.011**	−0.012***	−0.013***	−0.016***	−0.020***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
ED investigation count	0.090*	0.098**	0.108***	0.133**	0.169***
	(0.049)	(0.048)	(0.048)	(0.050)	(0.052)
ED treatment count	−0.039	−0.037	−0.033	−0.026	−0.013
	(0.029)	(0.029)	(0.028)	(0.029)	(0.029)
Panel B: Inpatient treatment decisions					
Length of stay (days)	0.022	0.022	0.035	0.087*	0.168***
	(0.047)	(0.047)	(0.048)	(0.050)	(0.052)
Inpatient procedure count	−0.003	−0.002	0.000	0.007	0.017***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Panel C: Hospital costs					
30-day ED cost	2.503***	2.786***	3.040***	3.590***	4.335***
	(0.866)	(0.876)	(0.911)	(0.972)	(1.042)
30-day inpatient cost	62.459*	93.132***	125.793***	183.993***	260.498***
	(34.028)	(33.421)	(33.992)	(35.047)	(36.550)
30-day total cost	64.962*	95.918***	128.833***	187.583***	264.833***
	(34.336)	(33.770)	(34.389)	(35.496)	(37.055)
Panel D: Patient outcomes					
30-day mortality	−0.003***	−0.003***	−0.004***	−0.004***	−0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

Notes: (1) Polynomial order is set to 10 in all specifications; (2) All inpatient variables (e.g. length of stay, costs) take on the value zero for patients that are not admitted; (3) Bootstrapped standard errors clustered at the hospital trust level (199 repetitions).

Table A5: Estimated distortion effects of the target, robustness by polynomial order

	Polynomial order			
	6	8	10 (baseline)	Auto
Panel A: ED treatment decisions				
Pr(admission)	0.033***	0.041***	0.046***	0.046***
	(0.008)	(0.008)	(0.008)	(0.008)
Pr(discharge)	−0.013**	−0.029***	−0.033***	−0.033***
	(0.007)	(0.006)	(0.007)	(0.007)
Pr(referral)	−0.020***	−0.012***	−0.013***	−0.013***
	(0.003)	(0.003)	(0.003)	(0.003)
ED investigation count	0.101**	0.090*	0.108***	0.108***
	(0.046)	(0.048)	(0.048)	(0.048)
ED treatment count	−0.024	−0.031	−0.033	−0.026
	(0.027)	(0.027)	(0.028)	(0.030)
Panel B: Inpatient treatment decisions				
Length of stay (days)	−0.066	−0.006	0.035	0.035
	(0.051)	(0.050)	(0.048)	(0.048)
Inpatient procedure count	−0.014	−0.005	0.000	0.000
	(0.006)	(0.006)	(0.006)	(0.006)
Panel C: Hospital costs				
30-day ED cost	1.651*	2.638***	3.040***	3.080***
	(0.946)	(0.880)	(0.911)	(0.939)
30-day inpatient cost	46.035	95.305***	125.793***	125.793***
	(36.389)	(34.955)	(33.992)	(33.992)
30-day total cost	47.680	97.905***	128.833***	128.833***
	(36.857)	(35.331)	(34.389)	(34.389)
Panel D: Patient outcomes				
30-day mortality	−0.005***	−0.007***	−0.004***	−0.004***
	(0.001)	(0.001)	(0.001)	(0.001)

Notes: (1) Exclusion window begins at 180 minutes in all specifications; (2) 'Auto' selects the polynomial separately for each outcome, by selecting the polynomial that maximizes the adjusted-R^2 statistic from estimating Equation (1); (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).

Table A6: Estimated effects of the target on mortality, by ED diagnosis

ED diagnosis	Wait time reduction (mins)	Mortality reduction (ppts)	Predicted mortality	Affected patients in 2012/13	% of total lives saved
Septicaemia	15	5.3	3.7%	14,510	6.9%
Vascular injury	7	2.2	1.6%	3,381	0.7%
Cerebro-vascular	9	2.2	3.8%	50,257	9.9%
Other vascular	7	2.1	1.9%	29,940	5.6%
Respiratory	11	1.5	2.1%	244,732	32.7%
Haematological	11	1.3	1.7%	15,094	1.7%
Central nervous system	12	0.8	1.9%	108,315	7.7%
Gynaecological	4	0.8	0.2%	38,203	2.7%
Cardiac	6	0.6	3.1%	229,047	12.2%
Gastrointestinal	10	0.5	1.3%	329,935	14.2%
Laceration	0	0.4	0.6%	99,535	3.5%
Local infection	3	0.4	0.8%	63,561	2.3%
Diabetes and endocrine	19	0.4	2.3%	28,252	1.0%
ENT	3	0.4	0.7%	51,964	1.9%
Obstetric	1	0.3	0.2%	12,728	0.3%
Soft tissue inflammation	4	0.2	0.5%	104,420	1.9%
Urological	12	0.2	2.2%	128,363	2.3%
Facio-maxillary	1	0.2	0.3%	9,715	0.2%
Social problems	12	0.2	3.2%	17,378	0.3%
Nothing abnormal detected	6	0.2	1.3%	85,682	1.5%
Head injury	5	0.1	1.1%	84,319	0.8%
Bites/stings	1	0.1	0.2%	7,692	0.0%
Infectious disease	7	0.1	0.9%	47,447	0.4%
Psychiatric	5	0.1	0.7%	46,649	0.4%
Burns and scalds	3	0	0.3%	9,913	0.0%
Poisoning (inc overdose)	12	0	0.6%	72,641	0.0%
Joint injury/fracture	6	-0.1	0.9%	217,152	-1.9%
Contusion/abrasion	4	-0.2	0.5%	69,328	-1.2%
Muscle/tendon injury	2	-0.3	0.5%	49,384	-1.3%
Dermatological	2	-0.3	0.4%	17,143	-0.5%
Allergy	1	-0.3	0.4%	14,942	-0.4%
Ophthalmological	0	-0.3	0.2%	25,701	-0.7%
Foreign body	0	-0.5	0.1%	15,314	-0.7%
Sprain/ligament injury	1	-0.6	0.2%	92,555	-4.9%

Notes: (1) Column 2 and 3 contain the estimated reduction in wait times and 30-day mortality for patients with each ED diagnosis, respectively; (2) There are 40 diagnosis categories. Non-missing data defined as patients without a missing or 'not classifiable' diagnosis.

Table A7: Estimated mortality effects of the target by cause of death (ICD-10 chapter)

Cause of death	Mortality reduction (percentage points)		% of overall mortality impact	% reduction in deaths
Circulatory	0.124***	(0.021)	30.1%	19.6%
Respiratory	0.105***	(0.018)	25.6%	25.5%
Digestive	0.062***	(0.010)	15.0%	35.2%
Unintentional accidents	0.024***	(0.004)	5.8%	47.9%
Mental/behavioural	0.015**	(0.007)	3.6%	15.2%
Genitourinary disease	0.014***	(0.005)	3.5%	27.1%
Infectious disease	0.013***	(0.004)	3.2%	39.0%
Neoplasms	0.011	(0.018)	2.7%	2.8%
Musculoskeletal	0.011***	(0.003)	2.6%	38.4%
Nervous system	0.009*	(0.005)	2.3%	15.2%
Endocrine/metabolic	0.005*	(0.003)	1.3%	20.5%
Disorders of the blood	0.004	(0.003)	0.9%	20.7%
Vehicle and traffic accidents	0.003**	(0.001)	0.8%	45.1%
Skin and subcutaneous tissue	0.003	(0.002)	0.7%	26.7%
Other external causes	0.002	(0.002)	0.6%	27.8%
External cause (e.g. fire, nature)	0.002	(0.003)	0.5%	7.7%
Codes for special purposes	0.002**	(0.001)	0.5%	53.9%
Congenital	0.001	(0.001)	0.2%	14.8%
Parasitic diseases	0.000	(0.001)	0.1%	9.0%
Pre-natal	0.000	(0.000)	0.1%	63.3%
Eye and ear	0.000	(0.000)	0.0%	17.9%
Pregnancy related	−0.000	(0.000)	0.0%	−109.3%
Symptoms not classified	−0.000	(0.000)	−0.1%	−3.0%

Notes: (1) Cause of death categories defined by the first letter of the ICD-10 diagnosis code; (2) Column 2 shows the estimated reduction in 30-day mortality attributed to the cause of death; (3) Column 3 shows the proportion of the overall mortality reduction that is accounted for by the cause of death; (4) Column 4 shows the proportion of deaths due to the specific cause that is avoided because of the target.

Table A8: Estimated distortion effects of the target by cause of death (ICD-10 sub-chapter)

Cause of death	Mortality reduction (percentage points)		% of overall mortality impact	% reduction in deaths
Cerebrovascular diseases	0.071***	(0.010)	17.2%	33.3%
Chronic lower respiratory diseases	0.055***	(0.009)	13.3%	29.0%
Influenza and pneumonia	0.034***	(0.008)	8.2%	23.0%
Ischemic and pulmonary heart diseases	0.030**	(0.012)	7.4%	11.7%
Organic mental disorders	0.013*	(0.006)	3.2%	14.1%
Malignant neoplasms (respiratory, intrathoracic)	0.005	(0.008)	1.2%	4.8%
Malignant neoplasms (lip, oral cavity and pharynx)	0.002	(0.005)	0.5%	3.8%
Malignant neoplasms (digestive)	0.001	(0.005)	0.2%	2.0%
Malignant neoplasms (male genital organs, urinary tract)	0.001	(0.004)	0.2%	1.5%
Malignant neoplasms (breast, female genital organs)	−0.000	(0.004)	−1.1%	−12.4%

Notes: (1) Cause of death categories are defined by the first letter and digit of their ICD-10 code: Cerebrovascular diseases (I6), chronic lower respiratory disease (J4), influenze and pneumonia (J1), ischemic heart diseases and pulmonary heart disease (I2), Organic, including symptomatic, mental disorders (F0), Malignant neoplasms of respiratory and intrathoracic organs (C3), Malignant neoplasms of lip, oral cavity and pharynx (C1), Malignant neoplasms of digestive organs (C2), Malignant neoplasm of breast and female genital organs (C5), and Malignant neoplasm of male genital organs and urinary tract (C6); (2) Column 2 shows the estimated reduction in 30-day mortality attributed to the cause of death; (3) Column 3 shows the proportion of the overall mortality reduction that is accounted for by the cause of death; (4) Column 4 shows the proportion of deaths due to the specific cause that is avoided because of the target.

B Online Appendix: Selection simulation

This appendix sets out the details of a simulation we conducted to evaluate the no-selection assumption. There are two stages in the simulation. The first stage is to produce a simulated 'counterfactual dataset' that is based on the counterfactual wait time and age distribution. The second stage is to use the counterfactual dataset to simulate different responses to the four-hour target, specifically in terms of how the post-threshold movers are selected. We then compare these simulated outcomes to the observed data to learn about selection and the validity of our no-selection assumption. Below we describe the two stages of the simulation and the results.

B.1 Constructing the counterfactual dataset

We take the following steps:

1. We compute the counterfactual wait time distribution as described in Section 4.2.

2. We compute the counterfactual expectation of age conditional on wait times as described in Section 4.3.3. We use this same approach to compute the counterfactual standard deviation of age conditional on wait times.

3. Using outputs from steps 1 and 2, we create a simulated dataset of patients. This dataset has the counterfactual distribution of wait times and an age distribution that is normally distributed with its mean and standard deviation defined according to the results from step 2. As a result, the wait time distribution and the conditional expectation of age are both smooth functions through the four-hour threshold.

4. We generate a random variable, denoted ε_i, where $\varepsilon_i \sim N(0, \sigma^2_{age})$ and σ^2_{age} is the variance of the age variable created in step 3. We normalise this variance by the variance of age to help with interpretation later.

B.2 Simulating the selection of post-threshold movers

We take the following steps:

1. We set up the following selection equation

$$S_i = \beta age_i + \varepsilon_i, \tag{B.1}$$

where S_i is the selection index for patient i, β is a selection parameter to be specified, age_i is the age of patient i, and ε_i is a random term for patient i.

2. For each wait time bin in the post-threshold period, w^* to w^+, we define post-threshold movers as follows

$$M_i = 1\{S_i \geq \tau_w\}, \tag{B.2}$$

where M_i is a binary variable equal to one if patient i is a post-threshold mover and τ_w is a threshold specific to wait time bin w.

3. The selection thresholds τ_w are unknown but we can estimate each threshold by finding the number of post-threshold movers that equates the wait time distributions for that bin in the counterfactual and observed datasets. The post-threshold movers are then identified as those patients with $M_i = 1$.

4. We consider the following different scenarios for β:

   (a) If $\beta = 0$ then there is 'random selection' as post-threshold movers are determined purely by ε_i, which is entirely random.

   (b) If $\beta = 1$ then there is 'selection-on-observables', specifically on age. Note that age and ε_i contribute equally to variation in S_i in this scenario.

   (c) If $\beta \in (0, 1)$ then there is selection-on-observables, but age plays a smaller role than ε_i in determining S_i. Note that this scenario can also be thought of as 'selection-on-unobservables' in the following sense: if $\beta = 0$ but ε_i contains some non-random element that is positively correlated with age, then the resulting selection equation is equivalent to Equation (B.1) with $\beta \in (0, 1)$.

5. To complete the simulation, we need to specify how post-threshold movers are allocated to the pre-threshold period. This allocation is unknown and so we simply adopt the simplest possible rule for the purposes of illustration. When shifting post-threshold movers we maintain their existing wait time ordering, such that those located just above w^* are moved to w^-, and those located at w^+ are moved to w^*.

B.3 Results

Figure B1 first shows the conditional expectation of age in the simulated dataset (Panel a), and then the results of simulating random selection of post-threshold movers (Panel b). Random

selection has three main features: (i) there is a spike at 240 minutes, which is where many of the post-threshold movers are shifted to; (ii) there is an increase in the pre-threshold level which is smaller the further away it is from the 240 threshold; (iii) there is a smooth distribution after the 240 threshold.

Figure B2 compares the random selection case to the observed data. The two conditional expectation functions are very similar, with the observed data exhibiting the same three features described above. The observed data lies marginally above the simulated data, but the gap is small.

Figure B3 now introduces the different selection scenarios and in each case compares the scenario to the random selection case. The scenarios are $\beta = \{0.1, 0.5, 1\}$. As described earlier, the cases where $\beta < 1$ can be thought of as selection-on-unobservables that are correlated with age.

Looking first at the scenario with $\beta = 1$ in Panel (b). The markers of selection are clear: the spike at 240 minutes is large; the pre-threshold period is substantially higher; and there is a pronounced drop in the post-threshold period. The post-threshold drop is more pronounced than the pre-threshold increase because, in the pre-threshold period, selection causes the post-threshold movers to be averaged with the pre-threshold non-movers while, in the post-threshold period, selection leaves behind a very select group of post-threshold non-movers.[50]

Turning now to the cases with $\beta = 0.5$ (Panel c) and $\beta = 0.1$ (Panel d). These share the same characteristics as the previous selection scenario, although the markers of selection become less pronounced as the selection mechanism is weaker. In the final case, where $\beta = 0.1$, the pre-threshold period differences are very small. Yet even in this case, the post-threshold period exhibits the pronounced drop.

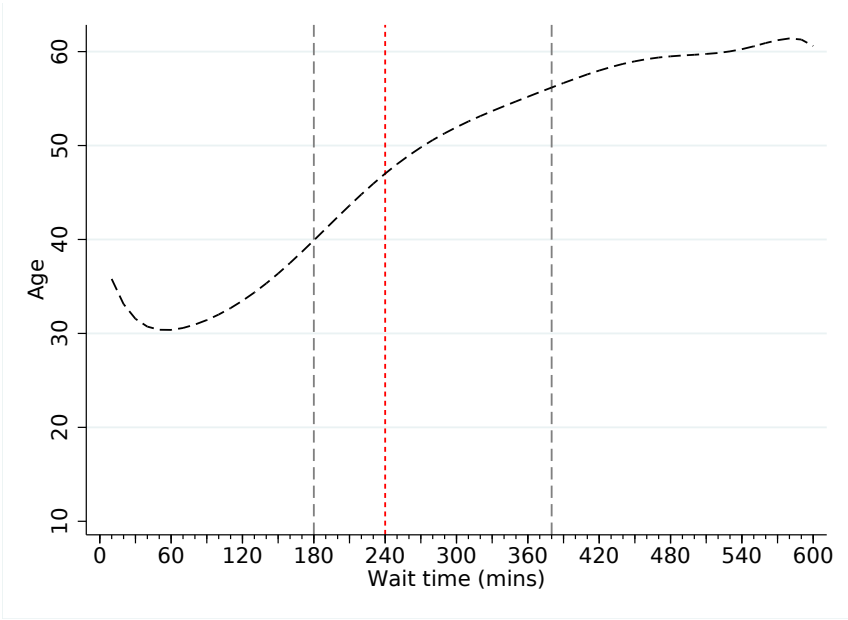
These simulations highlight three key points. First, the data looks very similar to the random selection case, sharing the same three features. Second, to the extent that there is significant selection, for example in the case of $\beta = 1$, then its markers show up clearly in the data. None of these markers are present in the observed data. Third, to the extent that observable and unobservable variables are correlated, then selection-on-unobservables will manifest itself directly in the observables. As a result, concerns about selection-on-unobservables can be thought of directly in terms of this correlation. While it is difficult to quantify the correlation

[50]Note that the specific pattern of the pre-threshold and post-threshold period is influenced by the allocation of post-threshold movers which is somewhat arbitrary in this simulation.

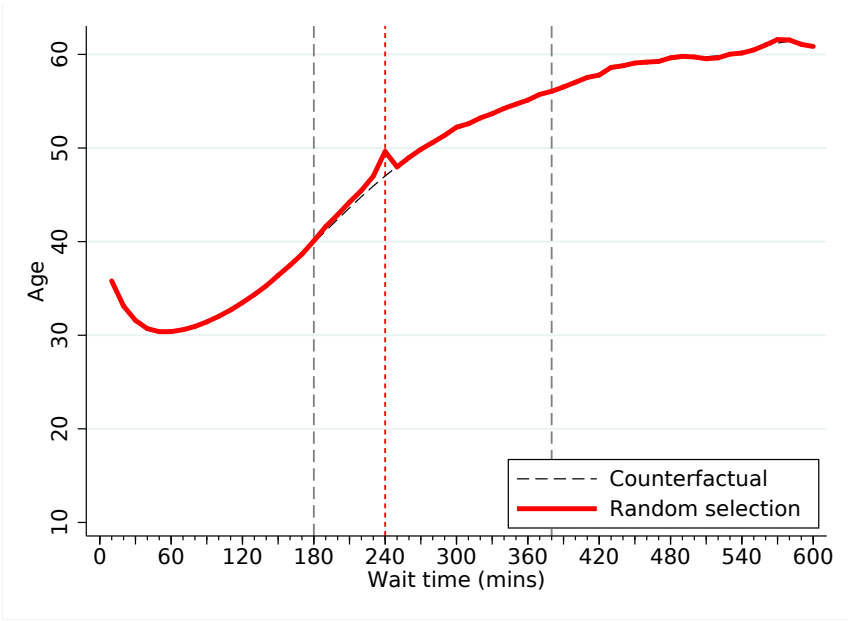
and link it with the power of our demographic test, there is clear evidence linking age and ambulance status to many medically relevant unobservables. For example, medical guidelines for physicians routinely incorporate age-based decision rules and, even after conditioning on age and diagnosis, the likelihood of death is more than 150% higher for ambulance patients than the average patient. These facts, which suggest that the correlation between observables and unobservables is far from negligible, are reassuring given that we find no evidence of selection on these observable variables.

Figure B1: Selection simulation using the counterfactual dataset

(a) Conditional expectation of age

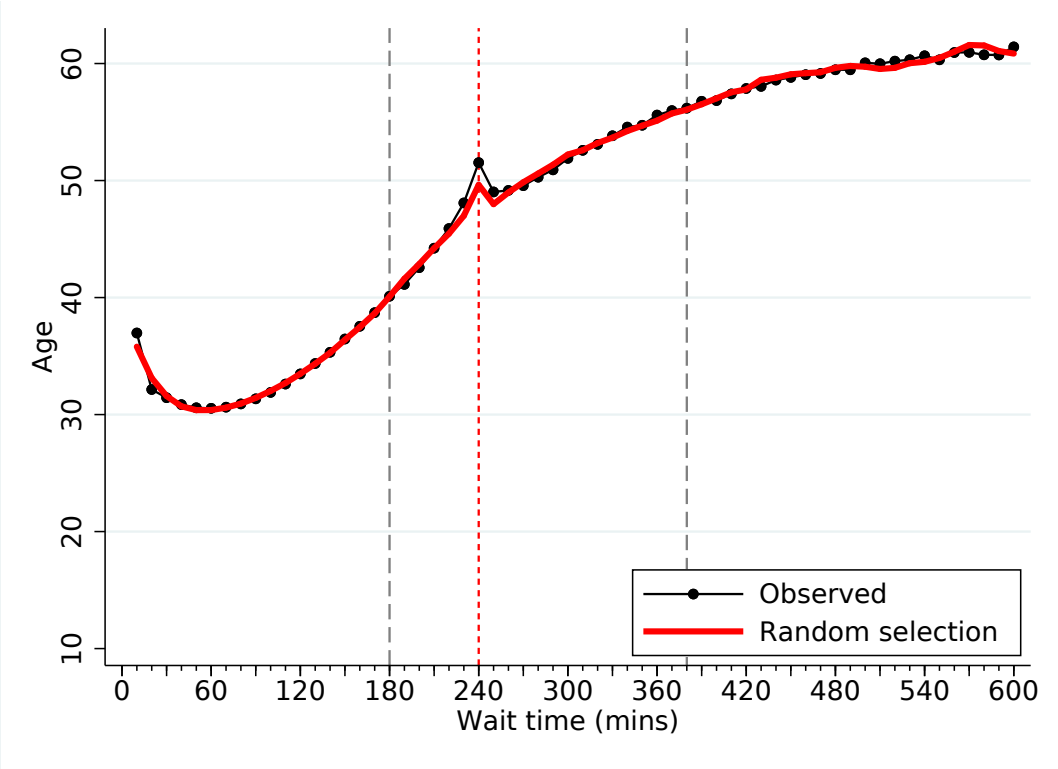


(b) Random selection simulation of post-threshold movers



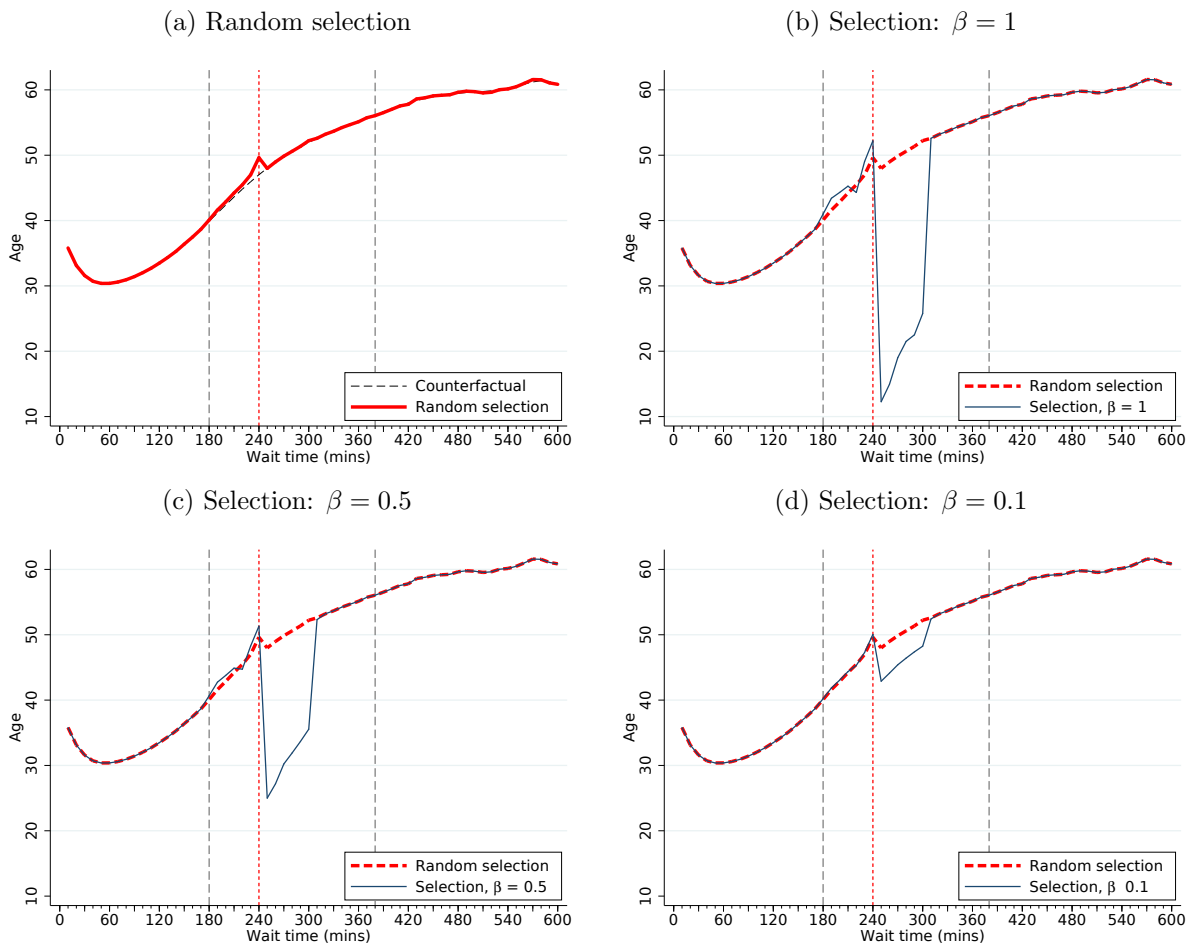
Notes: (1) Panel (a) shows the expectation of age conditional on the wait time bin in the counterfactual dataset ('counterfactual'); (2) Panel (b) shows the simulated data when post-threshold movers are chosen at random and shifted into the pre-threshold period ('random selection'); (3) Vertical dashed lines indicate the exclusion window and the vertical dashed red line indicates the 240 minute threshold.

Figure B2: Comparison of the random selection simulation and the observed data



Notes: (1) This chart compares the simulated data when post-threshold movers are chosen at random and shifted into the pre-threshold period ('random selection') with the observed data ('observed'); (2) Vertical dashed lines indicate the exclusion window and the vertical dashed red line indicates the 240 minute threshold.

Figure B3: Comparison of the random selection and selection simulations



Notes: (1) These charts compares the simulated data when post-threshold movers are chosen at random and shifted into the pre-threshold period ('random selection') with various degrees of selection-on-observables ('selection'); (2) Vertical dashed lines indicate the exclusion window and the vertical dashed red line indicates the 240 minute threshold.