HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 16/03

Should I stay or should I go? Hospital emergency department waiting times and demand

Peter Sivey

February 2016

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

Should I stay or should I go? Hospital emergency department waiting times and demand

Peter Sivey*

Department of Economics and Finance, La Trobe University;

February 29, 2016

Abstract

In the absence of the price mechanism hospital emergency departments rely on waiting times, alongside prioritisation mechanisms, to restrain demand and clear the market. This paper aims to estimate the relationship between waiting times and demand: by how much is the number of treatments demanded reduced by a higher waiting time, other things equal? I use variation in waiting times for low-urgency patients caused by rare and resource-intensive high-urgency patients to estimate the relationship. I find that when waiting times are higher, more low-urgency patients are deterred from treatment and leave the hospital during the waiting period without being treated. The waiting time elasticity of demand for low-urgency patients is approximately -0.25, and is highest for the lowest-urgency patients and when more substitute forms of care are available. The results imply waiting times play a substantial role in reducing demand from low-urgency patients and large increases in hospital capacity will be necessary to reduce emergency department waiting times.

Keywords: Hospital emergency departments, Waiting times, Demand

JEL Classification Numbers: I11.

*e-mail: *p.sivey@latrobe.edu.au*. The author is also an Honorary Senior Fellow at the Melbourne Institute for Applied Economic and Social Research, University of Melbourne. The author is grateful to Tony Scott, Vijaya Sundararajan and Jongsay Yong who are chief investigators of the grant that initially funded this research and to Hugh Gravelle for advice and suggestions, and to Jon Evans, Daniel Donnelly, Anna Burgess, Anne-Maree Kelly, Russell Thomson, Buly Cardak, David Prentice, David Johnston, Peter Siminski, Bronwyn Croxson, Nils Gutacker, Rita Santos and seminar participants at Monash University, University of York, University of Technology Sydney, Queensland University of Technology, University of Melbourne, RMIT University and the Department of Health and Human Services (Victoria) for helpful comments. This research was funded by NHMRC Partnership grant 567217 with the Department of Health and Human Services (Victoria) as a funding partner.

1 Introduction

In the absence of the price mechanism, government-funded hospital emergency departments (EDs) rely on waiting times, alongside prioritisation mechanisms, to restrain demand and clear the market. An estimate of the causal effect of waiting time on quantity of treatments demanded in hospital EDs is necessary for predicting the effects of policies to reduce waiting times, and for evaluating the characteristics of waiting times versus other rationing mechanisms. For example, recent policy announcements in Australia (Department of Treasury and Finance 2015) and in the UK (Labour Party 2015) propose additional investment in hospital capacity through new facilities and increased staffing to reduce ED waiting times. The effects of these policies depend on how many extra patients will demand treatment as waiting times start to fall. For example, when demand is highly elastic to waiting times, increases in hospital capacity will only result in small decreases in waiting time.

Waiting times act like a price by ensuring that only patients prepared to bear the cost of waiting will be treated. For a given increase in waiting time, the number of patients willing to bear the opportunity cost of waiting for treatment will fall, indicating a movement along the demand curve. Low-urgency patients will not be prioritised, so face the highest waiting times, and have a higher opportunity cost of time, so their decision to demand treatment is most likely to be affected by high waiting times. The effects of hospital ED waiting times are therefore a complex interplay of prioritisation and patient heterogeneity. Despite a well developed literature on the economics of elective care waiting times (eg Martin et al 2007), economists' study of hospital ED waiting times is still in its infancy (Gaudette 2014, Friedman 2014). In particular, a primary question to be resolved is: what is the role of waiting time in restraining ED demand?

This paper aims to estimate for the first time the effect of waiting times on the quantity of treatments demanded in hospital emergency departments. I develop a theoretical framework where the demand curve is derived from the urgency of patients' symptoms and their opportunity cost of time spent waiting, to motivate the empirical specification. The empirical

model is then estimated using detailed panel data from Victoria, Australia with time period and hospital fixed effects. I use an instrumental variables approach where waiting time for low-urgency patients is predicted by the rare presentations of life-threatened patients.

The model estimates produce a waiting time elasticity of demand for low-urgency patients of approximately -0.25, implying that a reduction in waiting times of 10 minutes (approximately 20% of the mean wait) will increase the quantity of treatments demanded by low urgency patients by 5 %. I show the size of the elasticity appears to be related to the availability of substitute care from general practitioners (GPs) as the elasticity is lower at weekends and outside of office hours, when fewer substitute forms of care are available. The elasticity is also higher for those patients in the lowest triage category, suggesting patients with less severe symptoms have a higher opportunity cost of time spent waiting. I also find some evidence of a larger effect of waiting times on quantity of treatments demanded when waiting times are relatively low. The results are robust to a number of alternative specifications and sensitivity checks.

The existing literature on waiting time and demand for elective hospital care (Martin and Smith 1999, Gravelle et al 2002, Windmeijer et al 2005, Martin et al 2007, Sivey 2012) has generally estimated relatively small waiting time elasticities of demand in the range of -0.1 to -0.4. A smaller literature has attempted to estimate the effect of waiting time on demand for non-emergency physician appointments (Lourenco and Ferreira 2005, Pizer and Prentice 2011). These studies find 'appointment delay' reduces quantity of care demanded (Lourenco and Ferreira 2005) or increases demand for substitute care (Pizer and Prentice 2011). Friedman (2014) has analysed the determinants of hospital ED waiting times in a US context with the focus on the level of reimbursement (through insurance generosity) for different patient groups, and the market entry of 'urgent care clinics' which can substitute for EDs for relatively low-urgency patients.

In contrast, I analyse the relationship between the demand for hospital emergency department treatments and waiting times for the first time. A key difference is that in hospital

emergency departments, patients experience waiting time and the symptoms of health conditions in a 'queue' (Barzel 1974) where they have to wait in a particular location and forgo alternative activities rather than a list where activity is unrestricted (Lindsay and Feigenbaum 1984). As waiting times in hospital EDs are relatively short compared to elective care waiting times (minutes and hours rather than weeks or months) this implies the primary cost imposed by ED waits is the opportunity cost of time rather than the diminished present value of the service (Gaudette 2014).

In another contrast with the literature on elective care waiting times, the rich emergency department data also allows me to employ an innovative identification strategy. To date, the hospital waiting time and demand literature has used panel data models (Martin and Smith 1999, Martin et al 2007), dynamic panel methods (Windmeijer et al 2005), instrumental variables (Gravelle et al 2002) and multinomial choice models (Sivey 2012). Due to the nature of the data used, these papers use variation between hospitals over relatively long time periods (year to year or quarter to quarter) which makes claims to causality in the coefficient estimates difficult. The hospital ED data I use exhibits a large degree of *intra-day* variation in the number of treated patients and waiting times which I exploit to improve identification.

Using variation in the data over such short time periods allows the identification of 'shocks' which cause short-term increases in waiting times for patients. One such shock is when 'life threatened' patients present to a hospital ED. This relatively rare event (less than one per hospital per day in the data) shifts medical and nursing resources away from less urgent patients, raising their waiting time. I use this plausibly exogenous variation in waiting time for low-urgency patients to estimate the causal effect of waiting times on the number of low-urgency patients demanding treatment.

A further novel aspect of the identification strategy is through using data on patients who 'walk out' of the hospital ED during the waiting period. Descriptive studies in the health services research literature have studied walk-outs and shown that patients tend to leave

without being seen when waiting times are longest (Kyriakou 1999, Goodacre and Webster 2005). I formalise this relationship in the theoretical framework and show that waiting times should have the opposite effect on 'walk outs' as on treated patients. I then test this hypothesis in the empirical analysis and find, indeed, that increases in waiting time cause an increase in walk-outs which is nearly equal to the fall in treated patients. Therefore, as an alternative dependent variable, the 'walk-outs' measure provides an valuable additional validation test of the theoretical framework.

2 Theoretical framework

2.1 Set-up

This theoretical framework aims to illustrate the nature and determinants of the waiting time demand curve and provide some intuition for empirical identification. I assume patient i at time t experiences symptoms S_{it}, which are distributed according to the distribution function $F_S(s)$. The higher is a patients draw of S_{it}, the more severe are her symptoms. The quantity $Y_{it} = \alpha^{-1} S_{it}$ represents the maximum time patient i at time period t is prepared to wait at a given hospital ED. Y is therefore distributed according to the distribution function $F_Y(s) = F_S(\alpha s)$, is increasing in the symptoms experienced, and decreasing in α.

The intuition of this set-up is that patients with more severe symptoms receive a higher benefit from treatment and have a lower opportunity cost of waiting for treatment. Their opportunity cost of waiting is lower as their symptoms (eg. pain or restriction to physical activity) hinder other activities which could be undertaken as an alternative to waiting.

The parameter α captures other factors which increase the opportunity cost of treatment. I have two main factors in mind that may affect α: Firstly, the availability of substitute treatment options (such as general practitioner appointments or alternative hospital emergency departments) should increase α. If similar treatment could be sought elsewhere at a shorter waiting time, then these alternatives increase the opportunity cost of waiting for treatment

at the present hospital. Secondly, the value of time to the individual patient also increases α. The value of time will capture the alternative activities available to the individual rather than seeking treatment, and how much she values them. This parameter could include a market measure of time such as wages or the value of family caring responsibilities.

This framework has parallels with the seminal work of Lindsay and Feigenbaum (1984). In their paper, each patient has a 'critical waiting time' $\hat{t}$ which depends on personal characteristics, including health and a discount rate representing their focus on 'list' waiting where the main costs of waiting are in discouting the future health benefits of treatment. Here, as the waiting costs are mainly related to the opportunity cost of time spent in the queue (Barzel 1974), the critical waiting time Y_{it} will depend on symptom severity, available substitutes and the patients value of time.

2.2 The effect of waiting time on quantity of patients demanding treatment

Consider the case with a single waiting time, W, offered to all patients. The number of patients demanding treatment will be given by $Q_D(W) = 1 - F_Y(W) = 1 - F_S(\alpha W)$, the number of patients who have symptoms severe enough that they are prepared to wait at least W minutes to be treated. Patients with more severe symptoms ($S > \alpha W$) will be treated with less severely ill patients ($S < \alpha W$) remaining untreated. The slope of the demand curve is given by

$$Q'_D(W) = -\alpha F'_S(\alpha W) < 0 \tag{1}$$

showing that an increase in waiting times reduces the quantity of treatments demanded, by reducing the number of patients prepared bear the opportunity cost of waiting. The slope of the demand curve depends on the slope of the distribution function of symptoms F'_S, and the other determinants of the opportunity cost of waiting time α. The effect of α on $Q'_D(W)$

depends on $F''(.)$, but in the simple case where $F''(.) = 0$ (such as the uniform distribution), higher α will increase the effect of waiting time on quantity demanded. This will inform the empirical work where I will test for the effect of the availability of substitutes (a determinant of α) on the effect of waiting time on quantity demanded.

2.3 Untreated patients

Another important quantity for the empirical analysis is the quantity of *untreated* patients $U(W) = F_Y(W) = F_S(\alpha W)$ which is the group of 'potential' patients who would be treated if the waiting time was zero, but choose not to be treated at the waiting time W. We can derive the effect of waiting time on this quantity: $U'(W) = 1 - Q'_D(W) = \alpha F'_S(\alpha W)$. So intuitively, a increase in waiting time causes an decrease in the number of treated patients and an equal and opposite increase in the number of untreated patients. This is a prediction which I will test in the empirical analysis by using both treated patients and a proxy of untreated patients as alternative dependent variables and comparing the key coefficient estimates.

2.4 Supply shifts identify the slope of the demand curve

Assuming a supply of treatments Q_S, is available, A single waiting time W^*, that sets demand equal to supply will be given by the solution to:

$$Q_S = Q_D(W^*) = 1 - F_S(\alpha W^*) \tag{2}$$

A change in the quantity of treatments supplied from Q_S^0 to $Q_S^1 > Q_S^0$ will identify the slope of the demand curve by the relative change in quantity of patients treated to waiting time $\frac{Q_S^1 - Q_S^2}{W^1 - W^0}$(where W^1 and W^0 are defined implicitly by $Q_S^1 = F_S(\alpha W^1)$ and $Q_S^0 = F_S(\alpha W^0)$).

2.5 Waiting times with prioritisation

A key characteristic of waiting time systems to include in the model is prioritisation. In emergency care, just as in elective care (Gravelle and Siciliani 2008, Gravelle and Siciliani 2009), patients are not all offered the same waiting time, but more severely ill patients are treated first. In this section I show that prioritisation in the form of third-degree waiting time discrimination will leave the equilibrium waiting time unchanged as long as prioritised patients are those who would choose to be treated under the non-prioritisation, single waiting time solution.

Taking as a starting point the single waiting time solution, W^* defined by equation (2), we allow a group of patients $Q^P = 1 - F_S(\alpha S^P)$ with symptoms $S > S^P$ to be offered a lower waiting time, $W^P < W^*$. All of the prioritised patients will choose to be treated at the prioritised waiting time $(1 - F(Y^P))$. However the prioritised patients are a subset of the $1 - F(W^*)$ patients who would have been treated anyway under non-prioritisation as $Y^P > W^*$. The remaining, non-prioritised patients can be offered the waiting time W^*, and $F(Y^P) - F(W^*)$ non-prioritised patients will be treated. Overall, $1 - F(Y^P) + F(Y^P) - F(W^*) = 1 - F(W^*)$ patients demand treatment, which is equal to supply S by equation (2). Non-prioritised patients will still be treated at the same waiting time as in the single waiting time case, W^*. The key assumption is that the prioritised patients would be treated under prioritisation or no prioritisation, they are not the 'marginal patients', so offering them a lower waiting time does not change the overall level of number of treatments demanded. This result is similar to that outlined for elective care by Gravelle and Siciliani (2009).

3 Data

3.1 Institutional setting

I use data on presentations to hospital emergency departments in the state of Victoria, Australia. Victoria is Australia's second most populous state with 5.7 million inhabitants of whom 4.3 million live in the Melbourne metropolitan area. The data only covers major public hospital emergency departments, which covers most of the presentations to emergency departments (EDs) in Victoria. Patients are treated at no financial cost in public hospital EDs, and private health insurance plays no role in determining treatment priority.

The level of information patients have about waiting times at alternative hospitals may be important in determining the waiting time elasticity of demand. During the time period of the dataset in Victoria, there is no 'real time' information available about waiting times at alternative hospital EDs such as that available in other states of Australia,[1] or in some Canadian provinces.[2] Some historical waiting time information is available at the hospital level through state government reports. In the hospital EDs, some hospitals provide screens with estimates of the current waiting time, and the nurse at the time of triage may also give an estimate of the current waiting time. In sum, we may expect that patients will have a general expectation, but no specific knowledge about waiting times before visiting hospital, which will be revised with more accurate information if and when they present at the hospital ED.

Availability of substitutes will clearly be important in determining the waiting time elasticity of demand. The main substitute to hospital care is general practitioner (GP) consultations. Most GPs in Victoria require booking at least a few hours in advance for appointments. Opening hours vary but many practices are only open during office hours on weekdays, or extended a few hours into the evening. There is a limited 'home visit' GP service available 24 hours a day but which is also subject to substantial waiting times.

[1] http://www.health.wa.gov.au/emergencyactivity/edsv/index.cfm
[2] http://www.edwaittimes.ca/WaitTimes.aspx

In general it is fair to say the availability of GP care is quite restricted overnight and at weekends. The other substitute to public hospital EDs is a small number of private hospital EDs in Victoria which charge large out-of-pocket fees.

Victoria had a system in use at the time of the dataset called"Ambulance Bypass" in which hospitals in the metropolitan area can elect not to receive ambulance patients for a certain time period due to overcrowding. This system might be an important factor in our analysis if it was a significant feature of the system. However, data from 2009 and 2010 shows that hospitals were "on bypass" for only 2-3 % of the time during the period of the data.[3]

3.2 The Dataset

I use the Victorian Emergency Minimum Dataset (VEMD), an administrative dataset with unit-record information from financial years 2008/9 and 2009/10 which comprises 2.7 million presentations to 38 hospital emergency departments. The dataset contains information on age, gender, the time and date of presentation, triage category, whether the patient left without being seen (walked out), waiting time between presentation and being seen by a doctor, and up to three diagnosis codes (using the ICD10-AM coding protocol). I divide the presentations into two groups: (1) treated patients, who present and are treated (after some waiting time), (2) walk-outs, who present but leave without being seen (during the waiting period).

The triage process is a form of prioritisation which plays an important role in my analysis. Patients presenting to hospital emergency departments in Victoria are triaged at presentation by a triage nurse into one of five categories which determines the priority assigned to their treatment. Category one patients are deemed to have an 'immediately life threatening' medical condition and should be treated immediately. Category five patients have 'less urgent or clinico-adminstrative problems' and the response time target is 120 minutes. There

[3]http://performance.health.vic.gov.au/downloads/hospital-bypass-hews.pdf accessed 12/06/15

are three categories in-between with response time targets of 10 minutes, 30 minutes and 60 minutes respectively.

3.3 Exploratory Descriptive Statistics

Before defining variables for the formal analysis, I start by providing some general descriptive statistics to motivate my analytical approach. Firstly, Table 1 presents means and standard deviations of the number of treated patients, number of walk-outs, and waiting time for treated patients, with data aggregated to the hospital/day level.

[INSERT TABLE 1 HERE]

As the triage category forms an important part of the variable definitions and identification strategy, it is informative to consider the characteristics of patients across triage categories. Table 2 presents the summarised descriptors of ICD10-AM codes for the most common five primary diagnoses (the first diagnosis field filled out) for presentations in each of the five triage categories.

[INSERT TABLE 2 HERE]

The most common diagnoses in triage category 1, such as multiple injuries, cardiac arrest, and pulmonary oedema, are largely serious acute heart, lung and brain conditions. It is intuitive that such life-threatening conditions are given the highest priority, immediate treatment. Referring to Table 1 we can see that category 1 patients are almost all treated immediately (mean waiting time of 0.65 mins) and that presentations in this category are relatively rare, less than one per hospital, per day on average.

The most common diagnoses in categories 2 and 3 include severe heart and lung conditions related to those in category 1 as well as potentially less urgent symptoms and conditions such as abdominal pain, gastroenteritis and urinary tract infection. Categories 2 and 3 have a minority of ED volume with an average of eight and 29 presentations per hospital per day

respectively, and relatively short waiting times with an average of 7 minutes and 26 minutes respectively. Category 4 and 5 presentations include abdominal pain and viral infections as well as wounds and fractures among the most common diagnoses. These categories make up the majority (59 %) of presentations with an average of 42 and 12 presentations per hospital per day respectively. Follow-up examinations and checking of dressings, sutures and plaster casts are also common in category 5. Average waiting times in these categories are much higher at 55 and 52 minutes respectively. Walk-outs are very uncommon in the first three triage categories ($<$ 1 per day) but are form a substantial proportion of presentations in category 4 (4.0 per day or 9%) and category 5 (1.75 per day or 12%).

4 Econometric Model

4.1 Quantity of Treatments Demanded

The empirical approach is a reduced-form estimation of the relationship hypothesised in sections 2.1 and 2.2. I model the number of treated patients Q_{it} at hospital i in time period t as a function of waiting times W_{it}, other explanatory variables X_{it}, time period and hospital fixed effects u_i and t_t:

$$Q_{it} = \beta_1 + \beta_2 W_{it} + X_{it}\beta_3 + u_i + t_t + \varepsilon_{it}. \tag{3}$$

The coefficient of interest is β_2, which measures the effect of changes in the hospital waiting time on the number of treated patients. To identify this effect, I want to allow for a vector of potential confounding variables X_{it}, which may be correlated with waiting time and also affect the number of treatments demanded through a vector of coefficients β_3. Hospital fixed effects u_i capture time-invariant unobserved hospital factors such as location and hospital size, which may cause some hospitals to have higher waiting times and a higher number of treated patients. Time fixed effects t_t allow the model to capture

shocks to demand which are common over all hospitals, such as weather events, seasonality and episodes of infectious illness.

Despite controlling for potential confounding variables, hospital and time fixed effects, there is still potential for correlation between ε_{it} and W_{it} if there are time and hospital specific shocks that affect both waiting time and the quantity of treatments demanded. An example of such a shock would be a hospital-specific demand shift, an incident that increases demand at hospital i in time period t, but not at other hospitals.

A potential solution for this endogeneity problem is to use an instrumental variables approach. This approach proceeds with a first-stage regression:

$$W_{it} = \delta_1 + \delta_2 Z_{it} + X_{it}\delta_3 + \mu_i + \tau_t + \upsilon_{it} \tag{4}$$

where the instrumental variable Z_{it} explains waiting time W_{it} but can be excluded from the demand equation (3). After estimating this first-stage model, waiting times are predicted by the coefficient estimates and inserted into equation (3) in a conventional two-stage least squares procedure.[4] The ideal variable to specify as the instrument, Z, is an exogenous supply shift, which produces an increase in waiting time identifying the slope of the demand curve as shown in section 2.4 .

4.2 Walk-outs

An important attribute of my analysis is that an alternative dependent variable is available, walk-outs, that proxies 'untreated patients', the number of potential patients deterred from seeking care by waiting time. Following section 2.3, where walk-outs WO_{it} are measured for hospital i and time period t, we can specify a model as follows:

$$WO_{it} = \lambda_1 - \beta_2 w_{it} + X_{it}\lambda_3 + \pi_i + \upsilon_{it} \tag{5}$$

[4] I use the xtivreg2 package in STATA 13 for the estimations.

The model is identical to the model for quantity of treatments, but waiting time has the opposite effect, so the model identifies $-\beta_2$. In other words, *increases* in walk-outs, are interpreted as equal and opposite *decreases* in treated patients. I estimate this model in exactly the same way as the model for treated patients in a two-stage least squares procedure with hospital and time period fixed effects (as described by equations (3) and (4)).

4.3 Empirical specification of variables

Quantity of treated patients

The theoretical framework predicts that the patients most likely to be affected by waiting time (the 'marginal patients') are those with lower urgency conditions (see section 2.2). For this reason, I construct the primary measure of the quantity of treated patients at a hospital, in a time period, using only patients in the two lowest triage categories (categories four and five). This includes patients with less urgent symptoms (see Table 2) but is actually a majority (58%) of the presentations in the data (see Table 1). Presentations are included in the measure of quantity of treated patients for hospital i and for time period t if the patient presented during the time period, and the patient was eventually treated (i.e. they didn't walk out before being treated). It is important that the variable for quantity demanded is defined by the patient's time of presentation, not the time of treatment by a doctor to avoid conflating quantity demanded with quantity supplied.

Walk-outs

Our measure of walk-outs is defined analogously to the definition of number of treated patients. A presentation is included in the measure of walk-outs for hospital i for time period t if the patient presented during the time period but subsequently 'walked out' (left the ED without being seen by a doctor).

Waiting time

As the dataset is aggregated I cannot use the actual waiting time for each patient in the regression models but must use a summary measure at the hospital-time period level. For the

primary measure of waiting time, I use the mean waiting time for all presentations included in the measure of treated patients for hospital i for time period t. In other words, the mean waiting time for patients presenting at hospital i in time period t who are eventually treated.

Instrumental Variable: Immediately life-threatened patients

I use an instrumental variable to explain variation in waiting time for treated category 4 and 5 patients, W_{it} but which is excluded from explaining the number of these patients treated (Q_{it}). This instrumental variable is 'life-threatened patients', a binary variable equal to one if there is at least one presentation categorised into triage category 1 for hospital i for time period t. Table 3 presents the number of observations, as well as the means and standard deviations of the treated patients, walk-outs and waiting time variables for hospital-time period observations when there is at least one category 1 presentation ("life-threatened patients") and for when there are no category 1 presentations. Firstly, note that hospital-time period observations with life-threatened patients are relatively rare, only 7.3% of observations. Waiting times are 17 minutes higher, the number of treated low-urgency patients is higher, and the number of walk-outs higher when life-threatened patients are present.

[INSERT TABLE 3 HERE]

The intuition of this IV approach is that the relatively rare event of a life-threatened patient presenting at a hospital ED represents a supply shift to low-urgency patients. Doctors and nurses are immediately diverted to treating the life-threatened patient, reducing the supply of treatments to low-urgency patients. This shift results in a higher waiting time for low-urgency patients identifying the slope of the demand curve (see section 2.4).

One threat to the validity of the instrument is that presentations by low-urgency patients and life-threatened patients may be directly correlated (not through the effect on low-urgency waiting time). I develop a robustness check for validity by exploring alternative timing of the instrument. I test using the instrument as a lag (the presentation of life-threatened

patients in the previous period) or lead (the presentation of a life-threatened patient in the next period). If the instrument is valid we would expect the lead of the instrument to be insignificant in predicting low-urgency waiting time in the current period. In contrast, the lag of the instrument may predict low-urgency waiting time in the current period if there is a persistent effect of presentation of life-threatened patients on low-urgency waiting time.

Control variables

The scope to include control variables at the hospital level is limited by the anonymisation of hospitals in the dataset. This is one reason why the use of time and hospital fixed effects is particularly important in this study. All characteristics of hospitals which don't change over time, and all 'common shocks' which affect all hospitals in specific time periods are accounted for in the two-way fixed effects specification.

The dataset does include characteristics of the patients being treated, including their age, gender, diagnoses, procedures performed, details of travel to the hospital and exit from the ED (eg admitted to hospital or returned home). However most of these characteristics would be expected to change with movements along the demand curve. For example, as waiting times fall, more patients with lower-urgency diagnoses may be treated, hence it seems inappropriate to control for the diagnosis mix of patients. For this reason the control variables used in this main specifications are quite limited: the age and gender mix of patients presenting to the hospital in the time period. These are specified as proportions, with age split into 5-year age groups. I test the robustness of the results to also including the proportion of patients in different diagnosis categories as additional controls.

Time period

The dataset for estimation is collapsed to the hospital-*three hour* time period level such that each hospital has eight observations per day: 12am-3am, 3am-6am, 6am-9am, 9am-12pm, 12pm-3pm, 3pm-6pm, 6pm-9pm and 9pm-12am. The choice of time period for the estimation balances two considerations. Firstly I need the time period to be short enough to capture short-run changes in waiting time and the responses of patients to those changes. In

particular, the time period should be short enough to capture the increases in waiting time caused by the presentation of life-threatened patients to the hospital.

Secondly, the time period should not be so short as to cause many 'zero' observations where hospitals do not have any low-urgency patients presenting in a given time period. Where there are 'zero' observations, the waiting time measure (the mean of treated patients in the time period) will be missing, rendering the panel of hospital-time period observations very unbalanced, with many hospitals dropping out of the sample from period to period. The three-hour time period balances these two considerations, it is short enough that we observe substantial changes in waiting time caused by presentations of life-threatened patients during the period, but there are only approximately 10% of hospital-time period observations that are missing due to 'zeroes'. I test the robustness of the results to collapsing the data with a slightly longer (four-hour) and a slightly shorter (two-hour) time period.

5 Results

5.1 Dependent variable: quantity of treated patients

Table 4 presents coefficient estimates for the effect of waiting time on the number of treated patients for five alternative models. The models begin with a linear regression with no controls (model (1)) and progress by adding hospital fixed effects (model (2)), time fixed effects (model (3)), control variables (model (4)), and finally a 2SLS approach using the life-threatened patients instrumental variable to predict waiting time in the first stage (model (5)). The results from model (1) show that there is a small positive correlation between waiting time and number of treated patients: with no control variables or fixed effects the coefficient estimate is positive. As fixed effects and control variables are progressively added in models 2, 3 and 4, this positive correlation is reduced, then becomes negative and marginally statistically significant. The largest change is between model 1 and model 2 when hospital fixed effects are added and the positive correlation becomes very small and

insignificant.

[INSERT TABLE 4 HERE]

The full specification is model 5, where life-threatened patients (category 1 patients) are used as an IV for the waiting time for treated category 4 and 5 patients. The instrument is a strong predictor of waiting time in the first stage with a F-statistic on the excluded instrument of 94.25. In the second stage, the coefficient estimate for waiting time becomes more strongly negative and significant, suggesting that waiting times reduce the quantity of low-urgency patients demanding treatment.

My use of the life-threatened patients as an IV for waiting time for low urgency patients clearly has an important impact on the results. The models which don't use the IV find a small positive or near-zero relationship between waiting times and treated low-urgency patients. Only the IV model finds a relatively large negative effect. My interpretation of these results is that the models without IV identify a mixture of demand shifts and supply shifts which both imply different relationships between waiting time and quantity. In contrast, the IV model (in combination with the two-way fixed effects) manages to identify a supply shift to low urgency patients (see section 2.4) caused by nursing and medical resources being shifted unexpectedly to the life-threatened patient(s) which clearly identifies a relatively strong negative effect of waiting time on patients demanding treatment.

5.2 Dependent variable: walk-outs

Table 5 presents coefficient estimates for models 6 to 10 which are equivalent to those in Table 4 but with the walk-outs dependent variable. The results show that waiting time and walk-outs are positively correlated: when waiting times are higher, more patients leave hospital emergency departments without being seen be a doctor (model (6)). The size of the coefficient becomes smaller after adding control variables and hospital and time period fixed effects (models 7-9). The coefficient is larger when using the life-threatened patients

IV approach (model 10) suggesting that waiting times increase the number of walk-outs. The coefficient is similar, but slightly smaller than the negative coefficient in the IV treated patients model (model 5). This comparison of model 5 and model 10 provides a test of convergent validity as suggested in section 2.3: If the reduction in treated patients caused by a waiting time increase are all 'walk outs', we would expect these coefficients to be identical (see equations (3) and (5)).

[INSERT TABLE 5 HERE]

5.3 Elasticity estimates

Tables 4 and 5 also present estimates of the waiting time elasticity of demand, $\frac{\Delta Q_D/Q_D}{\Delta W/W} = \hat{\beta}_2 \frac{W}{Q_D}$, using the waiting time coefficient estimate and the mean of waiting time (for W) and number of treated low-urgency patients (for Q_D) in the estimation sample. For models (6) to (10) using walk outs, the coefficient on waiting time is multiplied by -1 and applied to the same mean values as for the treated patients models. In other words, for the walk-outs elasticity it is assumed that quantity of treated patients is only reduced by waiting time through an increase in walk-outs (see equations (3) and (5)).

The elasticities mirror the coefficient results in sign and relative magnitude. In the preferred specification (model (5)) using the IV approach, the elasticity estimate is -0.25, implying a 10% increase in waiting time would reduce the number of treated patients by 2.5%. The estimate using the coefficient estimated using the walk outs variable (model (10)) is -0.2, implying a slightly smaller effect.

A potential explanation for the lower elasticity using the walk-outs coefficient is that there may be an effect of qaiting time on the number of treated patients other than through walk outs. Even short-term changes in waiting time, such as those caused by the presentation of life-threatened patients, could reduce the number of presentations to the ED as well as increasing walk-outs. Patients presenting at the ED at a time of high waiting times (such as those caused by the presence of a life-threatened patient) could be informed of the triage

nurse, or through their own observations become aware of the current high waiting time and decide against seeking treatment before being triaged. Such patients would reduce the number of treated patients without increasing the number of walk-outs.

5.4 Heterogeneity and Robustness

Table 6 presents estimates of the IV treated patients model (model 5 in Table 4) for category 4 and category 5 patients separately. The results show waiting time reduces the number of both category 4 and category 5 patients demanding treatment. The coefficient for category 4 patients is the larger, showing that for a given increase in waiting time, there is a larger reduction in the number of category 4 patients than category 5 patients. However, the waiting time elasticity of demand for category 5 patients (-0.43) is actually higher than for category 4 patients (-0.20), representing the size of the coefficients in proportion to the smaller number of patients in category 5 on average.

[INSERT TABLE 6 HERE]

Table 7 presents additional estimates to illustrate further heterogeneity in the effect of waiting time on patients demanding treatment. The first two columns of results allow for nonlinearity in the waiting time effect: first through a log-log specification and secondly through 'high wait' and 'low wait' subsamples. The log-log model produces an elasticity of -0.39, the same order of magnitude but larger than the linear model estimate at the mean of -0.25. The log-log model imposes a constant elasticity of demand implying a convex demand curve. The next two columns show the model estimated for two sub-samples: for a low wait time sub-sample and for high wait time sub-sample. An observation is defined as 'high wait time' if the previous observation (three-hour period) at that hospital had a waiting time higher than the mean waiting time. The 'lag' of waiting time is used in this way to define the high and low wait sub-samples so as not to make the inclusion of an observation conditional on the current value of the key explanatory variable. The results show that the

absolute size of the effect is higher for the low wait time subsample than for the high wait time subsample. The elasticity, however, is higher for high wait time observations where waiting time changes are low proportionately and enter on the denominator of the elasticity calculation.

[INSERT TABLE 7 HERE]

Table 8 presents results for two subsamples: weekdays (178,027) and weekends (29,856). Results show a slightly higher effect of waiting time on treated patients, and a higher elasticity, on weekdays (-0.26) compared to weekends (-0.23). Similarly the following four columns show heterogeneity over four different six-hour time-of-day subsamples: 12am to 6am (44,072) 6am to 12pm (54,218) 12pm-6pm (55,123) and 6pm-12am (54,470). The results show a larger effect and elasticity in the afternoon and evening (-0.45 and -0.59) compared to overnight and in the morning (-0.216 and -0.158). These results could represent differing availability of substitutes where there is less availability overnight and at weekends.

[INSERT TABLE 8 HERE]

Table 9 presents robustness checks based on variations to model (5) in Table 4. Firstly I check that the inclusion of several 'specialist' hospitals in the data does not qualitatively affect the results. I remove observations from two women-only hospitals, one children's hospital and one eye and ear hospital. The resulting elasticity estimate of -0.26 shows that these specialist hospitals have little impact on the main result.

[INSERT TABLE 9 HERE]

The next two columns present results using the lag or lead of the life-threatened presentation variable as instruments for waiting time, in place of the contemporaneous value of the life-threatened patient variable. Firstly, when the lag is used as an instrument, we can see that the IV is still strongly predictive of waiting time in the current period (F-stat 73.245)

and the estimated coefficient on waiting time in the outcome equation is still significant and is actually larger than in the baseline model. In contrast, when the leading value is used as the instrument, this does not predict waiting time in the current period (F-stat 2.48 and the instrument is statistically insignificant in the first stage), and waiting time is not statistically significant in the second-stage equation. The finding that the lagged value but not the leading value of the instrumental variable is a significant predictor of waiting time is suggestive that the relationship between the life-threatened patient presentations and waiting time for low-urgency patients is causal and not just associative.

The fourth column of Table 9 presents results from a model where extra controls for patients diagnoses are included. I noted in section 4.3 that the ability and necessity to add control variables to the model is limited due to the data available and our IV approach. In this model I add an additional set of controls: the proportions of patients presenting in different diagnosis categories in each hospital-time period observation. The diagnosis categories are defined by the first letter of the ICD10-AM code for the first primary diagnosis. Including these additional control variables may help control for additional 'demand shifts' that occur between hospitals over time (and which may be correlated with our instrumental variable), therefore removing bias in the estimates. However, they come at a cost of potentially reducing useful variation in how patients respond to changes in waiting time (movements *along* the demand curve), for example if patients with less serious diagnoses are more susceptible to changes in waiting time then we would observe a reduction in patients of these diagnoses being treated when there is a positive shock to waiting times (such as that caused by the instrumental variable). This mechanism could therefore introduce some unwanted confounding of the waiting time variable into the model which would *introduce* bias into the estimates (following the concept of a 'bad control' as discussed by Angrist and Pischke, 2009). The results of this model show a smaller estimate for the waiting time elasticity (-0.17) which may be due to this reduction in useful variation (the 'bad control' mechanism).

The fifth column of Table 9 presents a model which uses data only from the first three

months of the sample time period (1 July to 30 September 2008). This model aims to test if the main result is caused by long-run changes at the hospitals in my sample over the two-year time period of the data. The tables show a similar result (elasticity of -0.23) to the baseline model, underlining the fact that this model relies on short-term (intra-day) variation in waiting time caused by the instrumental variable, to estimate the effect and so using a much shorter time period does not affect the result.

The sixth and seventh columns of Table 9 show the model's robustness to using a different length of time period in collapsing the data. Instead of using a three-hour window, here I test a two-hour or four-hour window. In both cases the waiting time coefficient is still negative, statistically significant and of a similar order of magnitude. Shortening the time period to two hours does substantially reduce the size of the mean elasticity from -0.247 to -0.157 whereas increasing the time period from three to four hours increases the mean elasticity somewhat from -0.247 to -0.300. This pattern may be explained by the shorter two-hour window not allowing enough time for the instrument (the presentation of a life-threatened patient) to affect the waiting time of low-urgency patients and for them to react to this shock by walking out. Hence I find a weaker effect when using this shorter time period. This rationale could also explain the slightly larger effect when the window is lengthened to four hours.

6 Discussion

This paper estimates the effect of waiting times on the number of patients demanding treatment in hospital emergency departments. My approach uses an innovative identification strategy with two components: Firstly I have two related dependent variables available. One measures patients presenting and being treated and the second measures patients who initially present for treatment but then walk out during the waiting period. I expect when identifying the demand curve to find waiting times reduce the number of treated patients

and increase the number of 'walk-outs'. Secondly, I use an instrumental variables strategy to overcome the endogeneity of waiting time. I use variation in waiting time for low-urgency patients caused by rare and resource-intensive life-threatened patients. These life-threatened patients produce a negative supply shift for low-urgency patients, causing a plausibly exogenous increase in the waiting time they face.

The results show changes in waiting time have a significant effect on the quantity of treatments demanded. The point estimate of the waiting time elasticity of demand for low-urgency patients is approximately -0.25. To interpret this effect, consider a reduction in waiting times of 10 minutes from a level of 50 minutes, a 20% change from the mean wait for low-urgency patients. The estimated elasticity would predict an increase in demand for treatments of 5.0 % or 1.65 low-urgency patients per day on average.

As hypothesised, I find a similar and opposite effect of waiting time on the number of walk-outs per hospital time-period, implying a slightly smaller elasticity of -0.20 when using the walk-outs dependent variable. The interpretation of this result is that the increase in walk-outs caused by an 10 % increase in waiting time is equivalent to a 2% fall in the number of patients treated. The finding of a smaller effect using the walk-outs variable may indicate that short-run increases in waiting time (such as those caused by the IV) can reduce the number of patients presenting at hospital EDs (for example just before triage) as well as increasing walk-outs from patients who have already been triaged.

These results can be interpreted in terms of predicting the effect of changes in hospital emergency department capacity on waiting times, a potentially important exercise for policymakers. So, for the baseline elasticity estimate of -0.25, the results predict an investment in increasing capacity of 10% (approximately 6 patients per hospital per day) would reduce waiting times for low-urgency patients by 36% (from a mean of 55 minutes to 36 minutes).

Estimates of heterogeneity in the effects of waiting time on the number of low-urgency patients demanding treatment helps to test other elements of the theoretical framework. The lowest urgency patients are the most responsive to changes in waiting time, with the

elasticity rising to -0.43 for the patients in the lowest triage category. This accords with my expectation that 'marginal' patients are likely to be those who have less serious symptoms and therefore have a higher opportunity cost of time spent waiting. Where triage category is an imperfect signal of symptom severity, we expect lower triage categories to have more 'marginal' patients whose patients would decide not to be treated for a given increase in waiting time.

I show the size of the elasticity appears to be driven by the availability of substitute care from GPs as the elasticity is lower at weekends and outside of office hours, when fewer substitute forms of care are available.

We must interpret the elasticities estimated in this paper as local-average treatment effects (LATE) due to the particular IV approach used. As such we can interpret the effect of waiting time changes caused by the presence or absence of life-threatened patients in the same hospital-time period. This should influence the interpretation of our results as the variations in waiting time caused by the instrument are very short-term. An increase in waiting time caused by the presentation of a category 1 patient is only likely to affect patients already in the waiting room of the hospital, or patients at the point of triage. As with many elasticities (for example the price elasticity of demand for normal goods), we may expect the elasticity to be larger in the long-run. For long term changes in hospital ED waiting times patients will update their priors about waiting times before deciding to present at a hospital ED. The effect of substitute forms of treatment, such as GP appointments, will also be larger in the long-run. The elasticity point estimate of -0.25 must be regarded as a lower bound and as such, for an increase of capacity of 10 % I expect my model's prediction of a fall in waiting times of 36% to be an upper bound of the long run effect.

My results also suggest the effect of waiting times on the number of low-urgency patients demanding treatment is higher at lower waiting times than at higher waiting times. This also implies caution in interpreting the point estimate of the elasticity for policy analysis. As hospital capacity is progressively increased, the effect in reducing waiting times will become

progressively smaller.

This study extends the literature on waiting time and demand for healthcare to an important new context. Perhaps surprisingly, my models provide similar elasticity estimates to the studies on elective care waiting times and demand (Martin and Smith 1999, Gravelle et al 2002, Windmeijer et al 2005, Martin et al 2007, Sivey 2012). While their elasticity estimates are in the range of approximately -0.1 to -0.4, in this study where waiting times are best characterised as 'queues' rather than 'lists', my estimates fall in the same range. These elasticities can be generally thought of as modest, perhaps reflecting the lack of close substitutes for publicly provided elective and emergency hospital care. However, in this study, the elasticity estimates are likely to be a lower bound for the long-run elasticity, suggesting waiting times play a substantial role in restraining demand from low-urgency patients.

References

1. Barzel Y. 1974. A theory of rationing by waiting. *Journal of Law and Economics* 17:73-96

2. Department of Treasury and Finance. 2015. Budget media release: $1.38 billion extra for Victoria's health. [URL: http://www.dtf.vic.gov.au/State-Budget/2015-16-State-Budget/Budget-media-releases/138-billion-extra-funding-for-Victoria%E2%80%99s-health, accessed 4/6/2015]

3. Friedman A. 2014. Firm response to low-reimbursement patients in the market for unscheduled outpatient care. PhD Dissertation, University of Pennsylvania.

4. Gaudette E. 2014. Health care demand and impact of policies in a congested public system. CESR-Shaeffer Working paper 2014-005, University of Southern California.

5. Goodacre S, Webster A. 2005. Who waits longest in the emergency departments and who leaves without being seen. *Emergency Medicine Journal* 22: 93-96.

6. Gravelle H, Dusheiko M, Sutton M. 2002. The demand for elective surgery in a public system: time and money prices in the UK National Health Service. *Journal of Health Economics* 21(3): 423-449.

7. Gravelle H, Siciliani L. 2009. Third-degree waiting time discrimination: Optimal allocation of a public sector healthcare treatment under rationing by waiting. *Health Economics* 18:977-986.

8. Kyriakou DN. 1999. A 5-year time study analysis of emergency department patient care efficiency. *Annals of Emergency Medicine* 34 (3): 326-335.

9. Labour Party. 2015. Britain can be better: The UK Labour Party Manifesto 2015. [URL: http://www.labour.org.uk /page/-/ BritainCanBeBetter-TheLabourParty Manifesto2015.pdf accessed 04/06/15]

10. Lindsay CM, Feigenbaum B. 1984. Rationing by waiting lists. *American Economic Review* 74(3):404-417

11. Lourenco OD, and Ferreira PL. 2005. Utilization of public health centres in Portugal: effect of time costs and other determinants. Finite mixture models applied to truncated samples. *Health Economics* 14: 939-953.

12. Martin S, Smith P. 1999. Rationing by waiting lists: an empirical investigation. *Journal of Public Economics* 71(1): 141-164.

13. Martin S, Rice N, Jacobs R, Smith P. 2007. The market for elective surgery: joint estimation of supply and demand.*Journal of Health Economics* 26(2): 263285.

14. Pizer SD, Prentice JP. 2011. Time is money: Outpatient waiting times and health insurance choices of elderly veterans in the United States. *Journal of Health Economics* 30: 626-636.

15. Sivey P. 2012. The effect of waiting time and distance on hospital choice for English cataract patients. *Health Economics* 21 (4):444-456

16. Windmeijer F, Gravelle H, Hoonhout P. 2005. Waiting lists, waiting times and admissions: an empirical analysis at hospital and general practice level. *Health Economics* 14: 971985.

Table 1: Treated patients, walk-outs and waiting time per hospital, per day by triage category

	Treated Patients		Walk-outs		Waiting time	
Variable	Mean	S.D.	Mean	S.D.	Mean	S.D.
All	92.84	42.91	6.47	7.37	42.42	20.51
in triage cat 1	0.65	1.33	0.00	0.00	0.08	0.25
in triage cat 2	8.37	8.35	0.02	0.15	7.29	6.18
in triage cat 3	28.87	20.05	0.71	1.42	25.96	16.84
in triage cat 4	42.33	20.89	4.00	5.52	54.72	29.06
in triage cat 5	12.44	15.81	1.75	2.47	52.21	36.96

Table 2: Top Primary Diagnoses by Triage Category

Triage category	Top five diagnoses	Proportion of treated patients in top five diagnoses
1	Unspecified multiple injuries, cardiac arrest, pulmonary oedema, convulsions, acute myocardial infarction	26.6 %
2	Chest pain, angina, atrial fibrilation, AMI, stroke	28.2%
3	Chest pain, abdominal pain, syncope and collapse, viral infection, gastroenteritis	14.6%
4	Abdominal pain, viral infection, open wound of wrist/hand, fracture of wrist/hand, gastroenteritis	13.0%
5	Follow-up examination, attention to dressings or sutures, eye disorders, change/checking/removal of plaster cast, open wound of wrist/hand	22.3%

Notes: Summarised descriptors of the ICD10-AM codes of the most common 5 primary diagnoses in each triage category. Also shown is the proportion of all presentations in each triage category which fall into the most common 5 diagnoses.

Table 3: Mean and S.D. of key variables by the presence of an immediately life-threatened patient in the same hospital-time period

	All observations		LT patients = 0		LT patients > 0	
Hospital-time period observations	207,883		192,687		15,196	
	Mean	S.D	Mean	S.D	Mean	S.D
Number of treated low-urgency patients	7.318	5.586	7.245	5.163	8.242	5.140
Waiting time for low-urgency patients	54.831	46.999	53.600	46.161	70.433	54.187
Walk-outs from low-urgency patients	0.854	1.581	0.802	1.523	1.507	2.078

Notes: Observations are in the "LT patients> 0" column if there is at least one category 1 (life-threatened patient) presentation in the hospital-time period observation. I use three hour time periods 12am-3am, 3am-6am, 6am-9am, 9am-12pm, 12pm-3pm, 3pm-6pm, 6pm-9pm and 9pm-12am. Statistics are presented for low-urgency (category 4 and 5) patients.

Table 4: Linear regressions and IV model of the number of treated low-urgency patients

Dependent Variable: Treated low-urgency patients					
Explanatory Variables	(1)	(2)	(3)	(4)	(5)
Waiting time	0.009**	0.001	-0.004**	-0.003*	-0.033***
	(0.005)	(0.002)	(0.002)	(0.002)	(0.008)
Hospital fixed effects		Yes	Yes	Yes	Yes
Time period fixed effects			Yes	Yes	Yes
Age and gender proportions				Yes	Yes
IV: life-threatened patients					Yes
First-stage F-statistic					94.25
Implied waiting time elasticity	0.07	0.01	-0.03	-0.02	-0.25
Obs			207883		

Notes: *** indicates significance at 1 %, ** significance at 5 % and * significance at 1 %. All standard errors are adjusted for clustering at the hospital level. The implied waiting time elasticity is calculated at the mean waiting time and mean number of treated low urgency patients. Model (5) is a two-stage least squares regression where a binary variable indicating the presence of a category 1 presentation in the hospital-time period is used as an instrument for waiting time.

Table 5: Linear regressions and IV model of the number of walk-outs

	Dependent Variable: Walk-outs				
Explanatory variables	(6)	(7)	(8)	(9)	(10)
Waiting time	0.012*** (0.002)	0.010*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	0.027*** (0.006)
Hospital fixed effects		Yes	Yes	Yes	Yes
Time period fixed effects			Yes	Yes	Yes
Age and gender proportions				Yes	Yes
IV: life-threatened patients					Yes
First-stage F-statistic					94.25
Implied waiting time elasticity	-0.09	-0.07	-0.06	-0.06	-0.20
Obs	207883				

Notes: *** indicates significance at 1 %, ** significance at 5 % and * significance at 1 %. All standard errors are adjusted for clustering at the hospital level. The implied waiting time elasticity is calculated at the mean waiting time and mean number of treated low urgency patients. Model (10) is a two-stage least squares regression where a binary variable indicating the presence of a category 1 presentation in the hospital-time period is used as an instrument for waiting time.

Table 6: IV models of number of treated patients split by triage category

	Dependent Variable: Treated low-urgency patients	
	Cat 4	Cat 5
Waiting time	-0.021*** (0.006)	-0.013*** (0.005)
Hospital fixed effects	Yes	Yes
Time period fixed effects	Yes	Yes
Age and gender proportions	Yes	Yes
IV: life-threatened patients	Yes	Yes
Implied waiting time elasticity	-0.20	-0.43
Obs	207883	

Notes: *** indicates significance at 1 %, ** significance at 5 % and * significance at 1 %. All standard errors are adjusted for clustering at the hospital level. The implied waiting time elasticity is calculated at the mean waiting time and number of treated low urgency patients. Both models are two-stage least squares regressions where a binary variable indicating the presence of a category 1 presentation in the hospital-time period is used as an instrument for waiting time.

Table 7: IV models of number of treated patients: Nonlinearity in the waiting time effect

	Baseline model	Log-log	Low wait time	High wait time
Waiting time coefficient	-0.033***	-0.394***	-0.038***	-0.029***
	(0.008)	(0.083)	(0.011)	(0.010)
Mean wait time	54.831	54.831	42.986	70.885
Mean treated patients	7.318	7.318	7.569	6.977
First-stage F-stat	94.43	81.72	72.134	40.94
Waiting time elasticity	-0.247	-0.394	-0.216	-0.295
Observations	207883	207883	119626	88256

Notes: *** indicates significance at 1 %, ** significance at 5 % and * significance at 1 %. All standard errors are adjusted for clustering at the hospital level. The implied waiting time elasticity is calculated at the mean waiting time and number of treated low urgency patients. All models are two-stage least squares regressions where a binary variable indicating the presence of a category 1 presentation in the hospital-time period is used as an instrument for waiting time.

Table 8: IV models of number of treated patients: Heterogeneity in the waiting time effect over time

	Baseline model	Week days	Week-end	12am-6am	6am-12pm	12pm-6pm	6pm-12am
Waiting time coefficient	-0.033*** (0.008)	-0.035*** (0.010)	-0.030** (0.013)	-0.012*** (0.004)	-0.031* (0.017)	-0.081*** (0.022)	-0.062*** (0.010)
Mean wait time	54.831	54.775	55.161	52.629	42.603	55.97	67.598
Mean treated patients	7.318	7.345	7.153	2.929	8.362	9.994	7.121
First-stage F-stat	94.43	88.725	52.333	44.678	48.069	31.169	54.447
Waiting time elasticity	-0.247	-0.261	-0.231	-0.216	-0.158	-0.454	-0.589
Observations	207883	178027	29856	44072	54218	55123	54470

Notes: *** indicates significance at 1 %, ** significance at 5 % and * significance at 1 %. All standard errors are adjusted for clustering at the hospital level. The implied waiting time elasticity is calculated at the mean waiting time and number of treated low urgency patients. All models are two-stage least squares regressions where a binary variable indicating the presence of a category 1 presentation in the hospital-time period is used as an instrument for waiting time.

Table 9: IV models of number of treated patients: Robustness checks

Description	No specialist hosps	Lag IV	Lead IV	Diagnoses controls	First three months	Two-hour window	Four-hour window
Waiting time coefficient	-0.035*** (0.007)	-0.059*** (0.010)	0.317* (0.175)	-0.023*** (0.007)	-0.030** (0.015)	-0.021*** (0.006)	-0.040*** (0.013)
First-stage F-stat	99.269	73.245	2.48	85.682	34.47	83.62	81.06
Waiting time elasticity	-0.262	-0.442	2.375	-0.172	-0.225	-0.157	-0.300
Observations	185778	201968	201976	207883	26051	294734	159329

Notes: *** indicates significance at 1 %, ** significance at 5 % and * significance at 1 %. All standard errors are adjusted for clustering at the hospital level. The implied waiting time elasticity is calculated at the mean waiting time and number of treated low urgency patients. All models are two-stage least squares regressions where a binary variable indicating the presence of a category 1 presentation in the hospital-time period is used as an instrument for waiting time.