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

September 14, 2017

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 estimates by how much the number of treatments demanded is reduced by a higher waiting time. 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.

Keywords: Hospitals, Waiting times, Demand

JEL Classification Numbers: H42, I11.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1002/hec.3610](https://doi.org/10.1002/hec.3610)

1 Introduction

How can public service providers ration services where they are provided free of charge? In the absence of the price mechanism, government-funded hospital emergency departments (EDs) around the world 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), Canada (New Democrats 2016) 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 are a form of ‘ordeal’ mechanism (Nichols and Zeckhauser 1981) which rations available treatments 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. The empirical model is 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 estimated coefficients 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.

This rationing function of waiting times is an example of the ‘ordeal mechanism’ in public economics (Nichols and Zeckhauser 1982). A recent example has shown how small waiting and administrative burdens can help to ‘select’ the most needy recipients of aid programs (Alatas et al 2016). Similarly, queueing can help efficiently target subsidised consumer goods to low income groups (Alexeev and Leitzel 2001). Further, a classic paper by Parsons (1991) shows rigorous enforcement and delays in applications for disability benefits reduces applications substantially. In this paper I investigate whether hospital ED waiting times play a similar role, for example by restraining demand from patients with the least urgent symptoms.

The existing literature on waiting time and demand for hospital services has focused on *elective* rather than emergency care (Martin and Smith 1999, Gravelle et al 2002, Windmeijer et al 2005, Martin et al 2007, Sivey 2012) and 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. 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 allow me to employ an innovative identification strategy. Previous studies on hospital waiting times for elective care (Martin and Smith 1999, Gravelle et al 2002, Windmeijer et al 2005, Martin et al 2007, Sivey 2012) have used 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 test this hypothesis formally 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 a valuable additional validation test of the baseline model.

2 Data

2.1 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. Hospital emergency departments in Victoria are larger than the national average (13% higher average presentations per year at hospitals in Victoria, AIHW 2011a), but waiting time performance is very close to the national average.

The data only covers major public hospital emergency departments, which covers most of the presentations to EDs in Victoria. Approximately 94% of patients in Australia are treated in public hospital EDs (AIHW 2011b). Patients are treated at no financial cost 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

¹<http://www.health.wa.gov.au/emergencyactivity/edsv/index.cfm> [accessed 19/06/16]

²<http://www.edwaittimes.ca/WaitTimes.aspx> [accessed 19/06/16]

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. 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

2.2 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

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

urgent or clinico-administrative problems' and the response time target is 120 minutes. There are three categories in-between with response time targets of 10 minutes, 30 minutes and 60 minutes respectively.

2.3 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 form a substantial proportion of presentations in category 4 (4.0 per day or 9%) and category 5 (1.75 per day or 12%).

3 Econometric Model

3.1 Quantity of Treatments Demanded

The empirical approach is a reduced-form estimation of the relationship hypothesised in Appendix A, Sections A.1 and A.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 :

$$Q_{it} = \beta_1 + \beta_2 W_{it} + X_{it}\beta_3 + u_i + t_t + \varepsilon_{it}. \quad (1)$$

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:

$$W_{it} = \delta_1 + \delta_2 Z_{it} + X_{it} \delta_3 + \mu_i + \tau_t + v_{it} \quad (2)$$

where the instrumental variable Z_{it} explains waiting time W_{it} but can be excluded from the demand equation (1). After estimating this first-stage model, waiting times are predicted by the coefficient estimates and inserted into equation (1) 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 A.4 .

3.2 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 A.3, where walk-outs WO_{it} are measured for hospital i and time period t , I can specify a model as follows:

$$WO_{it} = \lambda_1 - \beta_2 w_{it} + X_{it} \lambda_3 + \pi_i + v_{it} \quad (3)$$

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 (1) and (2)).

3.3 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 A.2). For this reason, I construct the primary measure of the quantity of treated patients at a hospital, in a time period, using only

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

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-threatening patient presenting at a hospital ED represents a supply shift affecting low-urgency patients. Doctors and nurses are immediately diverted to treating the life-threatening patient, reducing the supply of treatments to low-urgency patients. This shift results in a higher waiting time for low-urgency patients, thus identifying the slope of the demand curve (see section A.4).

One threat to the validity of the instrument is that presentations by low-urgency patients and life-threatening 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-threatening patients in the previous period) or lead (the presentation of a life-threatening 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-threatening 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 to observe substantial changes in waiting time caused by presentations of life-threatened patients during the period. 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. In the baseline (three-hour periods) models, there are 6.5% missing values, increasing to 11.6% for the two-hour period models and reducing to only 4.4% for the four-hour time period models.

4 Results

4.1 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 A.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.

4.2 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 by 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 A.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 (1) and (3)). The difference between the waiting time coefficient in the treated patients model compared to the walk outs model is not statistically significant at conventional levels ($z=1.10$).

[INSERT TABLE 5 HERE]

4.3 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 (1) and (3)).

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 waiting 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 by 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.

4.4 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. These results allow for nonlinearity in the waiting time effect: first through a log-log specification and a Poisson 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. I also estimate a count data model, a Poisson regression model with log waiting time as the key explanatory variable (the Poisson-log model), using the two-stage residual inclusion approach to instrumental variables. The coefficient estimate from this model can be interpreted as an elasticity (Cameron and Trivedi 2005, section 20.2.3) and the estimation yields a comparable elasticity of -0.339.

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. Some additional robustness checks are described in Appendix B.

[INSERT TABLE 8 HERE]

5 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.

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.

This result reflects findings from the ‘ordeal mechanism’ literature which have emphasised how rationing by waiting (or administrative delays) can improve efficiency or improve targeting of programs provided free of charge

(Alatas et al 2016, Alexeev and Leitzel 2001, Parsons 1991). In our context, the fact that patients in the lowest urgency category are most affected by waiting times may be interpreted as ‘targeting’ emergency care at those who need it most.

I also 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.

The lesson of my results for policy is that waiting times play an important role in rationing demand for care in hospital EDs, especially in the long-run. 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. Therefore, as hospital capacity is progressively increased, the effect in reducing waiting times will become progressively smaller. This paper therefore suggests that substantial increases in capacity will be necessary to reduce waiting times and it may be hard to reduce ED wait times beyond a certain level.

As to other policies aimed at reducing ED waits, international efforts to divert low-urgency patients (eg through walk-in centres) seem to be unlikely to achieve this goal, partly because low-urgency patients are not the primary determinant of ED waits (Salisbury et al 2007, Schull et al 2007). My re-

sults also highlight a potential downfall of waiting time targets as a policy lever. While targets have been shown to be effective in reducing some waiting time measures in elective care (Propper et al 2009, Kelman and Friedman 2009) waiting time targets may lead to gaming of the distribution of waits (eg Nikolova et al 2015) especially if overall waiting time cannot be reduced because of the additional demand response.

This study extends the literature on waiting time and demand for health-care 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. Alatas V, Banerjee A, Hanna R, Olken BA, Purnamasari R, Wai-Poi M. 2013. Ordeal mechanisms in targeting: Theory and evidence from a field experiment in Indonesia. NBER Working Paper No. 19127
2. Alexeev M, Leitzel J. 2001. Income distribution and price controls: Targeting a social safety net during economic transition. *European Economic Review* 45: 1647-1663
3. Australian Institute for Health and Welfare. 2011a. Australian Hospital Statistics 2009-10. Health Services Series No 40. Canberra: Australian Institute for Health and Welfare. [URL: <http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=10737418865> accessed 21/07/2017]
4. Australian Institute for Health and Welfare. 2011b. Australia's Hospitals 2009-10 at a glance. Health Services Series No 39. Canberra: Australian Institute for Health and Welfare. [URL: <http://www.aihw.gov.au/publication-detail/?id=10737418683> accessed 25/07/2017]
5. Barzel Y. 1974. A theory of rationing by waiting. *Journal of Law and Economics* 17:73-96
6. Cameron AC, Trivedi PK, 2005. Microeconometrics: methods and applications. Cambridge University Press.
7. 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]
8. Friedman A. 2014. Firm response to low-reimbursement patients in the market for unscheduled outpatient care. PhD Dissertation, University of Pennsylvania.
9. 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.

10. Goodacre S, Webster A. 2005. Who waits longest in the emergency departments and who leaves without being seen. *Emergency Medicine Journal* 22: 93-96.
11. 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.
12. 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.
13. Kelman S, Friedman JN. 2009. Performance Improvement and Performance Dysfunction: An Empirical Examination of Distortionary Impacts of the Emergency Room Wait-Time Target in the English National Health Service. *Journal of Public Administration Research and Theory* 19 (4): 917-946.
14. Kyriakou DN. 1999. A 5-year time study analysis of emergency department patient care efficiency. *Annals of Emergency Medicine* 34 (3): 326-335.
15. Labour Party. 2015. Britain can be better: The UK Labour Party Manifesto 2015. [URL: <http://www.labour.org.uk/page/-/BritainCanBeBetter-TheLabourPartyManifesto2015.pdf> accessed 04/06/15]
16. Lindsay CM, Feigenbaum B. 1984. Rationing by waiting lists. *American Economic Review* 74(3):404-417
17. 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.
18. Martin S, Smith P. 1999. Rationing by waiting lists: an empirical investigation. *Journal of Public Economics* 71(1): 141-164.
19. 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): 263-285.

20. Meacock R, Anselmi L, Kristensen SR, Doran T, Sutton M. 2017. Higher mortality rates amongst emergency patients admitted to hospital at weekends reflect a lower probability of admission. *Journal of Health Services Research and Policy* 22(1): 12-19.
21. New Democrats. 2016. Broten's ER pledge: waits measured in minutes, not hours. [URL: <http://www.saskndp.ca/broten-s-er-pledge-waits-measured-in-minutes-not-hours> accessed 21/03/16]
22. Nichols AL, and Zeckhauser RJ. 1982. Targeting Transfers through Restrictions on Recipients. *American Economic Review* 72 (2):372-377
23. Nikolova S, Sinko A, Sutton M. 2015. Do maximum waiting times guarantees change clinical priorities for elective treatment? Evidence from Scotland. *Journal of Health Economics*, 41: 72-88.
24. Parsons DO. 1991. Self-screening in targeted public transfer programs. *Journal of Political Economy* 99(4):859-876
25. 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.
26. Propper C, Sutton M, Whitnall C, Windmeijer F. 2008. Did 'targets and terror' reduce waiting times in England for hospital care? *The BE Journal of Economic Analysis and Policy* 8(2).
27. Salisbury C, Hollinghurst S, Montgomery A, Cooke M, Munro J, Sharp D, Chalder M. 2007. The impact of co-located NHS walk-in centres on emergency departments. *Emergency Medicine Journal* 24:265-269.
28. Schull MJ, Kiss A, Szalai JP. 2007. The effect of low-complexity patients on emergency department waiting times. *Annals of Emergency Medicine* 49(3):257-264
29. Sivey P. 2012. The effect of waiting time and distance on hospital choice for English cataract patients. *Health Economics* 21 (4):444-456
30. 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: 971-985.

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

Variable	Treated Patients (per hospital-day)		Walk-outs (per hospital-day)		Waiting time (mins)	
	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 fibrillation, 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** (0.005)	0.001 (0.002)	-0.004** (0.002)	-0.003* (0.002)	-0.033*** (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 10 %. 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

Explanatory variables	Dependent Variable: Walk-outs				
	(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-threatening 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 10 %. 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 10 %. 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	Poisson- log (2SRI)	Low wait time	High wait time
Waiting time coefficient	-0.033*** (0.008)	-0.394*** (0.083)	-0.339*** (0.036)	-0.038*** (0.011)	-0.029*** (0.010)
Mean wait time	54.831	54.831	54.831	42.986	70.885
Mean treated patients	7.318	7.318	7.318	7.569	6.977
First-stage F-stat	94.43	81.72	81.72	72.134	40.94
Waiting time elasticity	-0.247	-0.394	-0.339	-0.216	-0.295
Observations	207883	207883	207883	119626	88256

Notes: *** indicates significance at 1 %, ** significance at 5 % and * significance at 10 %. All models include the covariates and fixed effects specified in column (5) of Table 4 (the baseline model). All standard errors are adjusted for clustering at the hospital level except for the Poisson-log model. The implied waiting time elasticity is calculated at the mean waiting time and number of treated low urgency patients. For the log-log model and the Poisson-log model the elasticity is equal to the waiting time coefficient. All models are two-stage least squares regressions (two-stage residual inclusion for the Poisson-log model) 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 10 %. All models include the covariates and fixed effects specified in column (5) of Table 4 (the baseline model). 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.

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Sivey, P

Title:

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

Date:

2018-03-01

Citation:

Sivey, P. (2018). Should I stay or should I go? Hospital emergency department waiting times and demand. *Health Economics*, 27 (3), pp.e30-e42. <https://doi.org/10.1002/hec.3610>.

Persistent Link:

<http://hdl.handle.net/11343/293890>