HOSPITAL CHOICE IN A NATIONAL HEALTH SYSTEM COMPETING WITH THE PRIVATE SECTOR: A TALE OF TWO SAMPLES

VALENTINO DARDANONI, MAURO LAUDICELLA, AND PAOLO LI DONNI

Abstract. We study choice in the publicly funded National Health Service in England using a structural model of demand for elective procedures. Patients are allowed to opt out from the market of free-of-charge public hospitals and choose a private provider. The model is identified by using a two-sample strategy making creative use of widely accessible administrative data on public and private providers. We find that the outside option has an important effect on patient choice compared with traditional models ignoring the private sector. Considering heterogeneity in patient preferences, endogeneity of waiting time, and the existence of private sector, we find different policy conclusions compared to traditional hospital demand models.

JEL Classification Numbers D12, I11, I18, H51.

Keywords Hospital Demand, Patient Choice, Quality, NHS.

1. Introduction

There is a growing debate on the introduction of choice in publicly funded health care markets as many European countries are progressively empowering patients with more choice. On the one hand, supporters of choice argue that it will force hospitals to respond to patients' preferences driving competition between providers, which in turn will deliver greater quality and efficiency to the health system. On the other hand, skeptics argue that patients don't respond to quality signals as they are unable to observe or interpret these signals, hence market incentives are too weak to be considered seriously by hospital managers.[1] Producing new evidence is very important and equally challenging due to modelling issues that entails the study of consumer choice with particular references to publicly funded health markets that we discuss in this paper.

Date: June 12, 2019.

⋆ Valentino Dardanoni: Università degli Studi di Palermo, Dipartimento SEAS, 90129, Palermo, Italy, valentino.dardanoni@unipa.it. Mauro Laudicella: Faculty of Health Sciences, University of Southern Denmark, Winsløws Vej 19, 3, 5000, Odense, Denmark, mlaudicella@health.sdu.dk. Paolo Li Donni: Università degli Studi di Palermo, Dipartimento SEAS, Viale delle Scienze, 90129, Palermo, Italy, paolo.lidonni@unipa.it.

[1]On the discussion between "Death by market power" *vs.* "Competition kills" see the debate between Bloom *et al.* (2011) and Pollock *et al.* (2011).

We contribute to this debate by investigating patients' response to choice in the market for elective care in England. Our empirical application focuses on the market for primary hip replacements described in detail in Section 3 of the paper. We estimate a structural model of patient's choice and calculate demand elasticities for waiting time and quality of care allowing for endogeneity of waiting time, heterogeneity in patient preferences and unobserved quality of care.

One of the novelties of this paper is the use of two different administrative data sources to estimate patient's choice allowing for the impact of an outside option, such as private providers for elective care, on patient's choice set of public hospitals operating in the National Health System (NHS). The majority of studies on hospital choice in publicly funded health markets omit private hospitals due to the lack of micro data on the private sector, hence the patient's menu is restricted to public hospitals only. However patients may not choose from this menu; in fact about 20% of primary hip replacements examined by our study were performed by private providers, and the rapid growth of the demand for health care is likely to encourage an expansion of private providers in the future. Ignoring the private option in choice modeling might imply biased estimates of demand elasticities and other features of interest in the hospital care market.

To address these issues we implement a two-sample estimation strategy that allows us to model private hospitals as part of the outside option to the patient's choice set of public hospitals. Our solution makes a creative use of micro and macro data from data sources widely accessible to the researcher, hence it can be replicated without the need for special authorisation to access data on private providers. We find evidence that using a comprehensive model of hospital choice is key for drawing correct policy conclusions on the impact of choice in publicly funded health care system. Omitting private providers may result in serious misspecification of hospital demand, since patients choosing private hospitals are unlikely to be a random sample of the population.

This group of patients is expected to be especially responsive to variation in the supply of elective care opting out of the market of public hospitals when waiting time or quality do not fit their preferences.

Our structural model also allows controlling for waiting time endogeneity. Endogeneity may arise typically since higher quality hospitals have higher demand which increases waiting times; since true hospital quality may be largely unobserved, it gets sucked in the error term, which becomes correlated with waiting time. Failing to provide appropriate control for the endogeneity of waiting time may result in a large underestimation of waiting time demand elasticities, analogous to the bias found in the identification of price demand elasticities in standard industries reported in the industrial organisation literature (Berry *et al*, 1995; Nevo, 2011). A feature of structural models is that the error term is recovered by inverting the demand equation, and can be therefore controlled by second stage linear Instrumental Variable (IV) estimation. This is in contrast with most hopsital demand literature where endogeneity is addressed by hospital fixed effects. We show in our application that using two stage IV estimation combined with hospitals fixed effects implies significantly higher demand elasticiticies compared to just using fixed effect. A byproduct of our approach is that it makes possible to obtain a measure of unobserved hospital quality, which may be of independent interest, for example for studying the relationship between quality and market power in the hospital industry.

The paper is structured as follows. In the next sections we review existing literature and describe the institutional details of the market for hip replacement in the England in the period 2006-2009 explaining our two-sample strategy employed in estimation. Section 4 introduces our hospital demand model and describes our estimating strategy. Section 5 illustrates the estimation results and provide an application of our model to study the impact of reducing choice by simulating a hospital closure in a local market area where the regulator has considered a rationalisation of services. Section 6 concludes.

2. Existing literature

Choice in publicly funded health markets is usually studied by estimating hospital demand functions distinguished into two major branches: the demand for unplanned emergency procedures (e.g. AMI and stroke) and the demand for elective procedures (e.g. hip replacement and cataract). One of the main advantages of working with emergency procedures is that some of the endogeneity issues affecting identification are reduced. For instance, patients' sorting according to preferences for quality and health is mitigated by allocation to the closest hospital available at the time of their acute episode. A large number of studies take advantage of these features and examine the impact that hospital choice introduced in the elective market have on quality produced in the emergency market, arguing that the latter is likely to benefit from positive spillovers from the former (Gaynor *et al.*, 2013; Cooper *et al.*, 2011). However, a disadvantage of such an approach is that in emergency procedure the meaning of patients' choice is quite limited, hence the researcher is restricted to focus on the indirect effects of choice, often leaving an open question on the mechanism through which such effects are channelled from elective to emergency markets.

Studying the demand for elective procedures allows the researcher to focus on the direct effects that choice may have on this market but presents numerous challenges. Hospitals' quality of care and patients' preferences for willingness to travel and willingness to wait for the intervention are likely to be correlated and heterogeneous according to unobservable patients' characteristics. For instance, some patients with strong preferences for health might be willing to travel longer distance to get better quality or lower waiting times, other patients might prefer being treated in their local hospital. Several attributes of quality that might be relevant to the patient are often unobservable to the researcher, e.g. successful rates for small interventions, reputation of the specialist consultant, experience of other people close to the patient, cleanness, level of comfort and parking facilities. Observed indicators of quality, such as hospital mortality, readmissions and infection rates, are largely driven by the hospital performance in emergency care and their ability to capture attributes of quality in elective care is debated (Gutacker *et al.*, 2016; Longo *et al.*, 2017; Papanicolas and McGuire, 2017). Unobservable hospital quality is likely to lead to a large bias in the estimation of waiting time elasticities, analogous to the bias in the estimation of price elasticity when quality is unobservable in standard industry.[2]

Finally, patients have the option of seeking care in the private sector for an increasing number of elective interventions by buying private health insurance or paying out of pockets. However, data from private providers in publicly funded system are often incomplete or not accessible to the researcher, hence they are typically excluded from the analysis. An exception is Kelly and Stoye (2015) who access rarely available micro data on private providers and examine whether the entry of private providers created additional demand for hip replacements by using an aggregate model of demand. The authors reports that data on private providers had several limitations, including poor information on the patient's address.

Other studies adopt different identification strategies to address some of the issues described above. Sivey (2012) estimates demand elasticities for waiting time and travel distance for elective cataract procedures using a multilevel latent class model that allows for patient preference heterogeneity and hospital fixed effects (FE) to control for quality of public hospitals. Beckert *et al.* (2012) use a conditional logit model to estimate the demand for elective hip replacements in public hospitals and illustrate a new method to simulate the impact of hospital mergers on patient's quality elasticities. Gaynor *et al.* (2016) estimate the impact of removing constraints to patient's choice on the demand for elective CABG procedures by using a structural model that allows for patient preference heterogeneity, and controls for endogeneity of quality and waiting time by using hospital FE. Berta *et al.* (2016) estimate hospital demand assuming that information regarding the quality of past treatment received at a specific hospital is transmitted through social interaction among the population living in the same neighbourhood.

[2]For instance, a hospital improving quality by updating its facilities is likely to attract patients from neighbour hospitals, thus increasing its waiting time and reducing the waiting time of its neighbours; this might result in a spurious positive association between waiting time and hospital demand if quality is unobservable to the researcher.

In this paper, compared to existing literature on patients' choice in public funded hospitals' system, we explicitly address the presence of an independent sector private provider. We also argue that control for endogeneity has been less than perfect in current literature, since it has mainly relied on FE that might not fully address the issue. In contrast, our model allows the researcher to complement the FE with an IV approach. A vast literature on choice and competition in the US hospital market stresses the importance of modelling patient preference heterogeneity and allowing for the endogeneity of price and quality (Capps *et al.*, 2003; Ho, 2006-2009). However, the US health market has deep structural differences from publicly funded health markets considered here, including a large penetration of private providers and the presence of a differentiated market for private health insurance.

3. Institutional Details and Data

Our empirical application considers the demand for primary hip replacements in the English hospital market in the period 2006-2009. Elective hip replacement is a relatively simple elective procedure usually performed on elderly patients suffering from arthritis. January 2006 marks the introduction of the freedom of choice reform giving patients the right to choose the hospital for their elective treatment either at their general practitioner (GP) practice or at home using the choose and book website.[3] Normally, GPs offer a choice of four to five hospitals including information on waiting time and distance; similar information is available on the choose and book website. The market for elective hip replacement is self-contained as there is no substitute operation, although the patient can opt for no operation.

The market is served by four types of providers, including NHS public hospitals and NHS treatment centres, independent hospitals and independent sector treatment centres (ISTC). NHS public hospitals are large multi-service organization, while NHS treatment centres are public health centres specialized in few elective procedures performed routinely; we will refer to both as *public providers.* Independent hospitals are

[3]Choose and Book, operating from 2005 to 2015, was an software application to allows patients to choose date and time for an hospital appointment.

privately owned organizations offering few elective procedures to privately insured patients; ISTC are a subgroup of Independent hospitals that can provide services to both privately insured and publicly funded NHS patients. We will refer to these as *private providers*. They had a market share of 22% in 2006 and 15% in 2009 for elective hip replacement (see Table 1). Average time waited (in months) for accessing the procedure in public hospitals decreased over time for a number of reasons, including the introduction of a national system of waiting time targets paired with additional hospital capacity and a more rational allocation of patients through the choice reform. As a consequence, getting the treatment from private providers with low wait but high price lost part of its appeal to patients resulting in loss of market shares.

Virtually all public providers and a large share of private providers are able to offer hip replacement operations, making it an ideal candidate for studying choice and competition in the hospital market. Finally, hip replacement is less likely to be associated with a more complex health situation that may make choice more difficult.

TABLE 1

	2003	2004	2005	2006	2007	2008	2009
NHS Hospitals waiting times	6.177	5.403	4.83	3.817	2.782	2.891	3.010
IS Market share	0.315	0.263	0.224	0.217	0.186	0.173	0.154

3.1. **A Tale of Two Samples.** In this section we introduce our two-sample estimation strategy and the different source of data used. Suppose we want to estimate a parametric model of hospital choice

$$\Pr(\textit{patient } i \textit{ chooses } j \mid \boldsymbol{x}_{ij}, \boldsymbol{\beta}),\ \ j = 0, \ldots, J \tag{1}$$

where $j = 1, \ldots J$ indicizes the set of J NHS hospitals which perform elective hip replacement, $j = 0$ denotes the outside option, $\boldsymbol{x}_{ij}$ denotes the set of variables affecting patient's i choice of option j, and $\boldsymbol{\beta}$ is the parameter to be estimated.

Suppose we have data on $\boldsymbol{x}_{ij}$ and on actual choices (denoted C_{ij}, which takes value one if patient i chooses hospital j) only for $j = 1, \ldots J$, that is, only for the set of patients who have chosen a NHS hospital. In other words, we have a selected sample of the population. Let $S = \{1, \ldots, n\}$ denote this selected sample, and let $U = \{1, \ldots, N\}$ denote the universe of all English over 65 patients in need of elective hip replacement. The problem is that we do not observe U.

Suppose we have external information of the total number of patients in U which did not choose a NHS hospital, say $N_0 = N - n$. In practice, since we are agnostic about the precise nature of the selection process, we generate a synthetic sample of size N –say U'– using administrative data, reproducing the population of over 65 English patients in need of elective hip replacement.

Note that while under sufficiently rich administrative data, U' may contain all the variables $\boldsymbol{x}_{ij}$ needed to calculate the choice probabilities (1), U' of course does not contain real patients, and thus does not contain data on actual choices C_{ij}. In section 4.3 we show how we can use observed sample moments in both the actual NHS sample S and in our synthetic sample U' for parameter estimate.

3.2. **The samples we use.** We use two main data sources: the Hospital Episodes Statistics (HES) including the universe of publicly funded inpatient admissions in NHS hospitals in England, and small-area data reported by the Office for National Statistics (ONS) at the level of Lower-Layer Super Output Areas (LSOAs).

From the HES dataset, we extracted all elective admissions during the fiscal year 2006 to 2009 of patients aged 65 and over receiving a primary hip replacement (HRG codes: H01 H02 H80 H81). Our sample includes 27,962, 29,604, 31,206, 31,875 patients respectively treated in each year from 2006 to 2009. For each patients we collect three key variables: their place of residence (LSOA), treatment hospital, and waiting time, defined as the time between GP referral and inpatient admission. We calculate straight line travel distance (in km) between the centroid of the LSOA where patient lives and each public hospital.

The ONS small-area data provides information on the characteristics of the LSOA where the patient is resident. LSOA are geographical units developed by the ONS to improve the reporting of small-area statistics for the UK (Briggs *et al.*, 2007). There are 32,482 LSOA units in England for the period we consider, with an average total population over 65 of 252 individuals per LSOA.

We use the patient's LSOA income score as a proxy for the patient's socioeconomic status. The LSOA income score is a domain of The Indexes of Multiple Deprivation (IMD), which capture multidimensional aspects of deprivation experienced by the population at the LSOA geographical level. The IMD income domain measures the proportion of the LSOA population living in low-income households reliant on one or more means-tested benefits. We also use the IMD health deprivation and disability domain as a proxy for the patient's need of care. This indicator identifies areas with relatively high rates of people dying prematurely, or whose quality of life is impaired by poor health, or who are disabled (Noble *et al.*, 2006).

Our strategy is to couple the sample extracted from the HES dataset, with a synthetic sample which uses LSOA's data to mimic the patients aged over 65 English population seeking hip replacement surgery. For this purpose we build a synthetic LSOA sample where the total number of patients seeking a hip replacement equals to the number of elective hip replacement procedures performed in all England by the patients aged over 65 in those years, as reported by the Annual reports of the National Joint Registry (see Table 2.3 of the National Joint Registry 2010 for 2009).[4] Table 1 reports the share of hip replacements received by privately funded patients in England. Using this information, the size of the outside option can be easily recovered by adding all patients privately funded to all patients who are NHS funded, but received a hip replacement in a private hospital. We decided to include ISTC treating NHS patients in the outside option for two main reasons: they treat both private and NHS patients; they report micro data on NHS patients only and with very poor quality in our study period (Kelly and Stoye, 2015). The size of the synthetic sample, N_t^S, is equal to 35,716, 36,355, 37,739 and 37,677, for $t = 2006, 2007, 2008, 2009$. For each time t we sample, with replacement, N_t^S units in proportion to the total population over 65 in each LSOA.

[4]The NJR provides free access to aggregated data on all hip replacement operations performed by private providers every year at the regional level. Although the NJR report is an important source of information to appraise the size of private providers in England, the average compliance in the period considered is about 85%, giving an underestimate of the total performed procedures.

For robustness we also built a synthetic sample based on epidemiological studies assuming 5 procedures for 1000 individuals over 65 as an informed guess on the incidence of hip replacement in the population of over 65, based on epidemiological studies on the incidence of hip replacement in England (e.g. Williams *et al.*, 1994) and consultations with various health care professionals. The size of the second synthetic sample is equal to 40,429, 40,797, 41,426 and 42,001.

In the main text we report results based on the first sample. For completeness, results obtained with the Epidemiological sample are reported in Appendix A. Descriptives statistics on observed variables used in the application are reported in Appendix B.

3.3. **Using the two samples.** Both the HES and the synthetic LSOA samples contain the observable variables affecting hospital choice, as specified in the patient's indirect utility function, with a qualification: individual economic and health status are proxied by the IMD income deprivation and IMD health deprivation indices in both samples.

The HES dataset does not include data on the patient's economic status, hence using the IMD income domain as a proxy is a standard approach in the literature. However, HES contains information on patient individual health, e.g. comorbidities at the time of hospital admission. Therefore, a drawback of our two-sample strategy is that we have to proxy patient's need of care by the IMD health deprivation index instead of using individual health information, which are not contained in the LSOA sample. We test the robustness of our approach by comparing Logit estimates using individual level information on patient's health status (in particular, patient's number of comorbidities) with estimates using IMD health deprivation. We obtain similar predictions and policy conclusions.

Two-sample estimation strategies have been used in other studies to combine key variables of interest that were collected by different surveys covering different populations of respondents (Blundell *et al.*, 2008). In contrast, the present application combines samples from different datasets covering the same population, i.e. the total population of England, which makes much easier to assure consistency in the combined sample.

4. Hospital Demand

Patient's choice depends on hospital characteristics such as the distance from the patients' residence, the time she has to wait to get the procedure, and the quality of hospital care. The indirect utility of patient i at time t for NHS hospital $j = 1, \cdots, J_t$ is given by

$$U_{ijt} = \beta_i w_{jt} + \gamma_i d_{ij} + \eta a_{ij} + q_{jt} + \epsilon_{ijt}, \tag{2}$$

where, at time t, q_{jt} denotes hospital j's quality, which is unobserved; w_{jt} is average waiting time (in months at time of admission) for hospital j; d_{ij} is (the log of) the distance (in kilometres) between the residence of patient i and hospital j; a_{ij} is a dummy variable which takes value one if hospital j is in the 'attention area' of patient i (namely, j is in the attention area of patient i if it is either within a distance of 20 km or is one of the 5 closest hospitals to patient i); and ϵ_{ijt} is an i.i.d. extreme value individual preference shifter.

In this formulation, there are two types of heterogeneity in patients' preferences: the purely idiosyncratic shifter ϵ_{ijt}, and the marginal (dis)utility for waiting time and distance β_i and γ_i. We model taste heterogeneity for distance and waiting time in terms of observable patients' characteristics and an idiosyncratic random term:

$$\beta_i = \beta_0 + \beta_I I_i + \beta_H H_i + \sigma_w R_{w,i}$$

$$\gamma_i = \gamma_0 + \gamma_I I_i + \gamma_H H_i + \sigma_d R_{d,i}$$

where I_i and H_i denote patient i economic and health deprivation, and $R_{w,i}$ and $R_{d,i}$ are distributed in the population as standard normal variates. As well known in the discrete choice literature, modelling taste heterogeneity is key for estimating realistic substitution patterns between products, dispensing of the highly restrictive property of independence of irrelevant alternatives (IIA).

Following most of the literature, we assume that patients can choose according to their preferences and their GP acts as a perfect agent advising on the available options.

This is a reasonable assumption in publicly funded health markets as the GP as no financial incentive in referring patients to any particular hospital.

Contrary to most of the literature, we allow patients to choose an outside option, which we denote $j = 0$. As discussed in the introduction, the outside option contains private hospitals, that a patient may choose outside the menu of the J_t public hospitals. The utility of patient i from choosing the outside option ($j = 0$) at time t is

$$U_{i0t} = \alpha_i + \epsilon_{i0t} \tag{3}$$

where ϵ_{i0t} is an i.i.d. extreme value preference shifter and

$$\alpha_i = \alpha_{0t} + \alpha_{I1} I_i + \alpha_{I2} I_i^2 + \alpha_{H1} H_i + \alpha_{H2} H_i^2 + \sigma_0 R_{0,i}$$

with $R_{0,i}$ standard normal distributed.

It turns out that it is quite useful to decompose the utility of choosing a NHS hospital $j = 1, \ldots, J_t$ into: i) a component which does not vary among patients', say $\delta_{jt} = \beta_0 w_{jt} + q_{jt}$; ii) a component $\mu_{ijt} = \big(\beta_I I_i + \beta_H H_i + \sigma_w R_{w,i}\big) w_{jt} + \big(\gamma + \gamma_I I_i + \gamma_H H_i + \sigma_d R_{d,i}\big) d_{ij} + \eta a_{ij}$ which captures individual patients' heterogeneity (excluding the error term); iii) the purely idiosyncratic logit error ϵ_{ijt}. Therefore,

$$U_{ijt} = \delta_{jt} + \mu_{ijt} + \epsilon_{ijt}, \; j = 1, \ldots, J_t, \tag{4}$$

while

$$U_{i0t} = \delta_{0t} + \mu_{i0t} + \epsilon_{i0t}, \tag{5}$$

with $\delta_{0t} = \alpha_{0t}$ and $\mu_{i0t} = \alpha_{I1} I_i + \alpha_{I2} I_i^2 + \alpha_{H1} H_i + \alpha_{H2} H_i^2 + \sigma_0 R_{0,i}$.

This formulation makes it clear that in this choice model the only source of endogeneity is included in the constants δ_{jt}, which are defined for each hospital j at each time t, and so they absorb all hospitals' unobservable characteristics which may be correlated with the observable variables contained in the utility function. In particular, unobservable quality –which is typically strongly correlated with waiting time– has been subsumed into the constants δ_{jt}.

4.1. **Choice Probabilities in the Two Samples.** As discussed above, we use two samples: the HES sample collecting the universe of NHS financed patients which have chosen one of the J_t NHS hospitals, and a (synthetic) LSOA samples which mimicks the over 65 population in England seeking a hip replacement procedure.

Omitting t for simplicity, the probability that individual h in the HES sample chooses hospital j is

$$P_{hj}^{N} = \frac{\exp(\delta_j + \mu_{hj})}{\sum_{k=1}^{J} \exp(\delta_k + \mu_{hk})}, \; j = 1, \ldots, J, \tag{6}$$

while the probability that an individual s in a LSOA sample chooses option j is

$$P_{sj}^{S} = \frac{\exp(\delta_j + \mu_{sj})}{\sum_{k=0}^{J} \exp(\delta_k + \mu_{sk})}, \; j = 0, 1, \ldots, J. \tag{7}$$

Notice that δ_j, which captures hospital characteristics, is common in the two samples.

4.2. **Estimation.** We follow Goolsbee and Petrin (2004) and Train and Winston (2007) and estimate the model in two stages. In the first stage we estimate the mean utilities $\boldsymbol{\delta}$ and the parameters included in μ_{ijt}. These are collected as

$$\boldsymbol{\theta}_t = [\underbrace{\boldsymbol{\delta}_t}_{P^N, P^S}, \overbrace{\alpha_0, \alpha_{I1}, \alpha_{I2}, \alpha_{H1}, \alpha_{H2}}^{P^S}, \underbrace{\beta_I, \beta_H, \gamma, \gamma_I, \gamma_H, \eta}_{P^N, P^S}, \overbrace{\sigma_0}^{P^S}, \underbrace{\sigma_I, \sigma_d}_{P^N, P^S}]. \tag{8}$$

illustrating how the $\boldsymbol{\theta}_t$ parameters belong to P^N and P^S. Estimation of $\boldsymbol{\theta}_t$ is implemented by simulated GMM. In the second stage we use estimated hospitals' mean utilities $\hat{\boldsymbol{\delta}}_t$ to estimate waiting time and observable quality parameters, correcting for endogeneity by using a two stage least square (TSLS) estimator.

4.3. **Moments.** To simplify notation, let $\bar{P}_{sjt}^{S}(\boldsymbol{\theta}_t; \boldsymbol{z}_s^S)$ denote the expected probability for patient s in the LSOA sample to choose hospital j at time t, where $\boldsymbol{z}^S$ collects the variables which enter the utility function in the LSOA sample. $\bar{P}_{sjt}^{S}(\boldsymbol{\theta}_t; \boldsymbol{z}_s^S)$ is the integral of $P_{sjt}^{S}(\boldsymbol{\theta}_t; \boldsymbol{z}^S)$ over the distribution of the random variables R_w, R_h, R_0. In practice we approximate $\bar{P}_{sjt}^{S}(\boldsymbol{\theta}_t; \boldsymbol{z}^S)$ by simulation, using 100 antithetic Halton draws of the standard normal variables R_w, R_h, R_0. Similarly, $\bar{P}_{hjt}^{N}(\boldsymbol{\theta}_t; \boldsymbol{z}_h^N)$ denotes the expected probability of patient h in the HES sample to choose hospital j at time t.

Let S_{jt}^S and S_{jt}^N denote the share of hospital j at time t with and without the outside option respectively. We use the BLP share equation and two sets of moments:

(1) *The share equations*: we equate the observed aggregate hospital shares to the average probabilities in the LSOA sample:

$$S_{jt}^S = \frac{1}{N_t^S} \sum\nolimits_s \bar{P}_{sjt}^S(\boldsymbol{\theta}_t; \boldsymbol{z}_s^S). \tag{9}$$

Berry (1994) shows that the predicted shares can be inverted to get the vector $\boldsymbol{\delta}_t$, for any value of the remaining parameters in $\boldsymbol{\theta}_t$.

(2) *The HES Moments*: in the HES sample we set standard observation-specific moments:

$$\boldsymbol{g}_{hjt}^N(\boldsymbol{\theta}_t) = \Big(C_{hjt} - \bar{P}_{hjt}^N(\boldsymbol{\theta}_t; \boldsymbol{z}^H)\Big)(\boldsymbol{z}_{hjt}^H, \boldsymbol{v}_{jt}), \tag{10}$$

where C_{hjt} is the choice variable which takes value one if individual h at time t choose hospital j and $\boldsymbol{v}$ is an appropriate vector of hospital characteristics. In our application we use hospital dummies for teaching, acute and London hospitals, and hospital capacity variables (such as number of beds, doctors, qualified and unqualified nurses and health practitioners), with squares and interactions.

(3) *The LSOA Moments*: we link the HES and LSOA samples by matching observed attributes in the HES sample with those in the LSOA.[5] Consider for example income deprivation I. From observed choices in the HES sample, we derive the average income deprivation of patients using hospital j at time t, namely $\frac{1}{N_t^N} \sum_{h=1}^{N_t^N} I_h C_{hjt}/S_{jt}^N$, and, using Bayes' rule, we match this with the expected deprivation of patients using hospital j in the LSOA sample under

[5] Imbens and Lancaster (1994) discuss using macro moments in micro models with choice based samples (see also Petrin (2002) and Berry, Levinsohn and Pakes (2004)). We create a synthetic sample representing the universe of patients, reproducing at the micro level patients' characteristics for all options (including the outside one).

the theoretical choice probabilities $\bar{P}^S_{sjt}(\boldsymbol{\theta}_t; \boldsymbol{z}^S_s)$:

$$\boldsymbol{g}^I_{sjt}(\boldsymbol{\theta}_t) = \left(\frac{1}{N^N_t}\sum\nolimits_{h=1}^{N^N_t} I_h C_{hjt}/S^N_{jt} - I_s \bar{P}^S_{sjt}(\boldsymbol{\theta}_t; \boldsymbol{z}^S_s)/S^S_{jt}\right)\bar{\boldsymbol{v}}_{jt}, \tag{11}$$

where the S^S_{jt} denotes the estimated aggregate market share and $\bar{\boldsymbol{v}}$ denotes the vector of hospital characteristics $\boldsymbol{v}$ above, plus a constant.

By a similar reasoning, we derive another set of sample moments by matching LSOA's health deprivation H

$$\boldsymbol{g}^H_{sjt}(\boldsymbol{\theta}_t) = \left(\frac{1}{N^N_t}\sum\nolimits_{h=1}^{N^N_t} H_h C_{hjt}/S^N_{jt} - H_s \bar{P}^S_{sjt}(\boldsymbol{\theta}_t; \boldsymbol{z}^S_s)/S^S_{jt}\right)\bar{\boldsymbol{v}}_{jt}. \tag{12}$$

4.4. **First Stage Estimation.** Stack these individual moments to get

$$\boldsymbol{h}^N_{jt}(\boldsymbol{\theta}) = \sum\nolimits_h \boldsymbol{g}^N_{hjt}(\boldsymbol{\theta}_t); \quad \boldsymbol{h}^S_{jt}(\boldsymbol{\theta}) = \sum\nolimits_s \left(\boldsymbol{g}^I_{sjt}(\boldsymbol{\theta}_t), \boldsymbol{g}^H_{sjt}(\boldsymbol{\theta}_t)\right);$$

$$\boldsymbol{S}^N(\boldsymbol{\theta}) = \sum\nolimits_t \sum\nolimits_j \boldsymbol{h}^N_{jt}(\boldsymbol{\theta}_t)\boldsymbol{h}^N_{hjt}(\boldsymbol{\theta}_t)'; \quad \boldsymbol{S}^S(\boldsymbol{\theta}) = \sum\nolimits_t \sum\nolimits_j \boldsymbol{h}^S_{jt}(\boldsymbol{\theta}_t)\boldsymbol{h}^S_{hjt}(\boldsymbol{\theta}_t)'.$$

and define

$$\boldsymbol{h}(\boldsymbol{\theta}) = \left(\boldsymbol{h}^N(\boldsymbol{\theta}), \boldsymbol{h}^S(\boldsymbol{\theta})\right), \quad \boldsymbol{S}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{S}^N(\boldsymbol{\theta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}^S(\boldsymbol{\theta}) \end{pmatrix}. \tag{13}$$

To estimate $\boldsymbol{\theta}$ we use two-step Simulated GMM: in the first step we estimate $\boldsymbol{\theta}_1 = argmin\ \boldsymbol{h}(\boldsymbol{\theta})'\boldsymbol{h}(\boldsymbol{\theta})$; in the second step we find $\hat{\boldsymbol{\theta}} = argmin\ \boldsymbol{h}(\boldsymbol{\theta})'\boldsymbol{S}(\boldsymbol{\theta}_1)^{-1}\boldsymbol{h}(\boldsymbol{\theta})$. At each step, and within each iteration of the minimization problem, we also use the aggregate constraints (9) which equate observed market shares with theoretical ones. Estimated standard errors may be corrected using Pollard and Pakes (1989) procedure taking into account the simulation. For robustness we report bootstrapped standard errors, which are similar to the ones obtained using the Pollard and Pakes procedure.

4.5. **Second stage.** In the second stage we recover the parameters which enter the mean utility vector $\boldsymbol{\delta}$. Let $\hat{\delta}_{jt}$ denote the estimated mean utilities from the first stage. Decompose hospital quality q_{jt} into a time and hospital fixed effect Δ_t and Δ_j and observed and unobserved residual components of hospital quality, to get the regression equation

$$\hat{\delta}_{jt} = \beta_0 w_{jt} + \boldsymbol{x}'_{jt}\boldsymbol{\phi} + \Delta_t + \Delta_j + \xi_{jt} \tag{14}$$

where $\boldsymbol{x}_{jt}$ denotes a vector of observed quality indicators.

Quality of care is generally multidimensional and only partially observable to the researcher, hence a common strategy is to use all available indicators with a logical association to quality to capture it. However, there is not a definite view on how available indicators may effectively capture quality (Gutacker *et al.*, 2016; and Gravelle *et al.* 2017). Moreover, existing indicators are mostly based on hospital performance in emergency care, rather than in elective care. We include the following indicators: the Care Quality Commission (CQC) quality rating, the incidence of Methicillin-Resistant Staphylococcus Aureus (MRSA) infections, the standardized mortality rate (SMR) for hip fracture, and an indicator of hospital's predicted performance on readmission (READ) after a hip fracture (see Laudicella *et al.*, 2013).

Clearly equation (14) is subject to a serious endogeneity problem, since waiting times are strongly correlated with unobserved hospital quality. Using hospitals FE only might not fully address the endogeneity problem as documented in the application below (see e.g. Table 3), hence we estimate equation (14) by using a TSLS estimator with instrumental variables (IV) and hospital FE . As IV, we use a set of exogenous supply shifters that explain variation in waiting times that are not driven by hospital demand. We considered indicators of hospital capacity and input costs following a similar approach in the literature (Martin and Smith, 2003; Riganti *et al.*, 2017; Ho, 2006). Number of hospital beds and hospital sites were used as indicators of capacity, and the Market Force Factor (MFF) was used as an indicator of input costs. The MFF measures differences in input costs between hospitals that are due to their geographical location, and it is used to adjust reimbursement payments to NHS public hospitals (Monitor, 2003). However, this adjustment does not fully cover differences in costs, and hospitals can use it strategically to redistribute resources across different services (Propper and Van Reenen, 2010). The assumption for identification requires that our supply shifters are not correlated with unobserved hospital quality; we believe this is a reasonable assumption with respect to quality attributes that can be relevant to patients in elective care as discussed in the introduction section. The F-test score for our instruments is 9.90 suggesting that they are sufficiently strong.

As argued above, a host of observable measures are used in the literature to measure quality: mortality, infections, readmissions, patient satisfaction etc. It is well documented that there is often little correlation between these measures, and results may depend on the measure used. An advantage of our structural model is that it allows to estimate overall unobservable quality q_j for each hospital as $\hat{q}_{jt} = \hat{\delta}_{jt} - \hat{\beta}_0 w_{jt}$ (see equations (2) and (14)). Estimated q_j can be seen as an overall index of hospital quality, which includes all observable and unobservable (to the econometrician) factors which affect patient choice (utility) after conditioning on distance and waiting times. Thus, $\hat{q}_j$ may include things which go beyond observed medical quality but are valued by patients, such as parking facilities, room amenities, staff behaviour, etc. which may be important, especially for routine elective procedures.

5. Results

This section presents estimated parameters from two main models:

(1) *The Standard Choice Model.* As a benchmark, we estimate a standard Logit model that is the workhorse of most NHS hospital demand studies. The first stage is estimated by using GMM with the set of moments given by (9) and (10), but omitting random coefficients and outside option.[6] The second stage is estimated by using two alternative regression models to highlight the effect of waiting time endogeneity: an OLS model controlling for endogeneity by hospital FE, and a TSLS model controlling for endogeneity with IV and hospital FE. The Logit model with hospital FE can be considered the standard model used by several papers on hospital demand in publicly funded health markets (e.g. Sivey, 2012; Beckert *et al.*, 2012; Moscelli *et al.* 2016; Moscone *et al.*, 2012).

[6] Results are robust to Maximum Likelihood estimation.

(2) *The Two-Sample Model (2SM).* We estimate our model using the two-sample estimation strategy described in Section 4. As a robustness check, we estimate two versions of the 2SM based on two alternative LSOA synthetic samples: one based on NJR data, and the other based on Epidemiological studies, described in Section 3.2. Results from using the Epidemiological Sample are reported in Appendix A, and are very similar in substance policy conclusions to the results obtained with the NJR sample.

5.1. **First stage.** Table 2 shows estimates of the $\boldsymbol{\theta}$ parameters in equation (8). Estimates from the Logit model are in the first two columns of table 2, while the remaining columns report the parameters of the 2SM. Similarly to other studies on hospital demand, we find that distance strongly affects patient choice in both models; patients are significantly likely to choose a closer hospital, with the 'attention area' dummy, *"a"*, being strongly significant. The 2SM shows significant heterogeneity in patients' disutility for distance, *"d*I", "d*H"* , which suggests that individuals living in more income deprived areas prefer closer hospitals, while individuals in more health deprived areas are willing to travel further away to get a hospital that fits their preferences. The 2SM also reports evidence of heterogeneity in patients' disutility for waiting time by socioeconomic status, *"w*I"* , with patients coming from more income deprived areas experiencing higher disutility from waiting time. This can be explained by poorer patients being unable to avoid long waits by choosing a private hospital in the outside option. The 2SM enable us to estimate idiosyncratic waiting time and distance taste heterogeneity, "$R_{0,s}$", "$R_{w,s}$", "$R_{d,s}$", and heterogeneity in the probability of choosing the outside option, *"I", "H"* . The latter is found to be higher in patients coming from wealthier and more health deprived areas. Previous studies reporting that private hospitals admit a greater share of patients from rich areas than public hospitals, but with less comorbidities (Sivey, 2012). It is worth noticing that the indicator of health deprivation used in our analysis captures area-level need for health care, rather than comorbidities in patients seeking a hip replacement. Hence, our result for health deprivation might be explained by a higher demand in health deprived areas that is not fully met by public hospitals, thus increasing the likelihood of choosing the outside option for people living in those areas.

TABLE 2. First Stage

	Logit		2SM	
	Coef.	SE	Coef.	SE
wt*I	0.0037	0.0222	-0.0898	0.0434
d*I	-0.0114	0.0364	-0.7894	0.0523
wt*H	-0.0346	0.0198	0.0190	0.0565
d*H	0.0175	0.0365	0.8584	0.0756
d	-3.3068	0.0144	-8.3739	0.2719
a	1.6274	0.0262	2.9380	0.0557
Outside option				
I			-3.3548	0.0198
I^2			1.1276	0.0923
H			5.2416	0.0662
H^2			-0.4306	0.1610
Preference Heterogeneity				
$R_{0,s}$			3.6250	0.0513
$R_{w,s}$			0.5445	0.3525
$R_{d,s}$			2.9146	0.0882

Bootstrapped standard errors (SE) are based on 500 replications.

5.2. **Second stage.** Table 3 reports estimated coefficients for the second stage regression described in section 4.5 and using two different strategies to control for endogeneity of waiting times, i.e. hospital FE only (OLS model), and hospital FE and IV (2SLS model). Waiting time has a significant negative effect on patients' utility, hence hospitals with longer waiting time are less likely to be chosen. However, the estimated effect is more than 3 times larger after controlling for endogeneity of waiting time by using both hospital FE and IV, rather than FE only as a common practice in this literature. Moving to quality, only MRSA infections have a significant effect across models. This is not surprising since MRSA infections are a serious complications that might occur after an hip replacement operation negatively affecting the outcomes ()Senthi *et al.*, 2011). In contrast, other indicators of quality are driven by hospital performance in emergency care, thus it is not surprising that their impact on choice for elective hip replacements is weak.

Table 3. Second Stage

	Hospital FE (OLS)		Hospital FE + IV (TSLS)	
	Coef.	SE	Coef.	SE
Logit				
CQC	0.0430	0.0522	0.0161	0.0532
SMR	0.0076	0.0219	0.0090	0.0186
Read	-0.6820	0.8227	-1.1123	0.7264
MRSA	-0.2212	0.1543	-0.2811	0.0969
Waiting Time	-0.1176	0.0349	-0.4059	0.1117
Fixed Effects	Yes		Yes	
2SM				
CQC	A.1083	0.1926	0.0244	0.2168
SMR	0.0442	0.0776	0.0486	0.0744
Read	-1.7540	2.8844	-2.7163	2.8827
MRSA	-0.8596	0.4573	-1.0426	0.3927
Waiting Time	-0.6494	0.1491	-1.5315	0.4546
Fixed Effects	Yes		Yes	

Bootstrapped standard errors (SE) are based on 500 replications.

5.3. **Elasticities.** For each hospital market t, we first calculate the $J_t \times J_t$ matrix of elasticities, with the J_t-sized diagonal containing the hospital own elasticity. We then report the mean and standard deviation of these J_t own elasticities for each time t.

Table 4. Average Waiting Time Elasticities

Year	Hospital FE (OLS)			Hospital FE+IV (TSLS)		
	Mean	S.D.	% Diff	Mean	S.D.	% Diff
Panel A: Logit						
2006	-0.1917	0.0664	36.9%	-0.6738	0.2278	38.5%
2007	-0.1540	0.0643	42.5%	-0.5387	0.2238	40.8%
2008	-0.1168	0.0528	35.7%	-0.3999	0.1780	38.4%
2009	-0.1278	0.0603	31.3%	-0.4261	0.2024	38.2%
Panel B: 2SM						
2006	-0.3039	0.2836		-1.0949	0.3505	
2007	-0.2677	0.2549		-0.9100	0.3826	
2008	-0.1817	0.2000		-0.6488	0.2797	
2009	-0.1859	0.1888		-0.6893	0.2948	

Table 4 compares predicted hospital elasticities for waiting times from the Logit model and the 2SM described in Section 4. Again we use two different strategies to control for endogeneity of waiting times, i.e. hospital FE (OLS model) and hospital FE and IV (2SLS model). Estimated elasticities decrease over time in all model specifications. Patients become less sensitive to differences in hospital waiting times as the average waiting time drops markedly over time during our study period (see Table 4). Estimated elasticities from the standard logit model using hospital FE only to control for endogeneity range between -0.13 (in 2009) and -0.19 (in 2006), and are in the ballpark of previous studies. In contrast, corresponding estimates from the 2SM are larger ranging between -0.19 (in 2009) and -0.30 (in 2006). Hence, omitting the outside option of private hospitals result in underestimating waiting time elasticities by about 35% when endogeneity is controlled by using hospital FE only. Estimated own elasticities from the standard logit model controlling for endogeneity by using FE and IV are between -0.43 (in 2009) and -0.67 (in 2006), suggesting that imperfect control for endogeneity of waiting time results in underestimating elasticities by about 70%. Corresponding estimates from the 2SM are between -0.70 (in 2009) and -1.10 (in 2006), hence omitting the outside option of private hospitals result in underestimating waiting time elasticities by about 40% when controlling for endogeneity by using IV and FE.

TABLE 5. Average MRSA Elasticities

Year	MRSA		
	Mean	S.D.	% Diff
Panel A: Logit			
2006	-0.1567	0.0860	53.7%
2007	-0.1144	0.0711	55.1%
2008	-0.1147	0.0708	54.2%
2009	-0.1127	0.0671	55.8%
Panel B: 2SM			
2006	-0.3388	0.1584	
2007	-0.2548	0.1545	
2008	-0.2502	0.1426	
2009	-0.2549	0.1480	

Table 5 reports estimated elasticities for quality of care measured by MRSA infection rates, i.e. the only indicator of observable hospital quality that is statistically significant in our analysis. The standard Logit model predicts elasticities between -0.11 (in 2009) and -0.16 (in 2006), in contrast corresponding predictions from the 2SM range between -0.25 (in 2009) and -0.34 (in 2006). Therefore, omitting private hospitals result in underestimating quality elasticities by about 50%.

5.4. **Hospital closure simulation.** A recurrent policy issue in the health care market is the "rationalisation" of the NHS hospital industry by mergers and closures of underperforming hospitals. Evaluating the change in the local demand after a closure is key to understand its effects on patient flows, competition and supply conduct. Using a structural model allows the researcher to compute the counterfactual size of market share of each hospital, and thus to simulate the potential effects of an hospital closure. In contrast, using a standard logit model to calculate these counterfactuals may result in an unaccurate picture for at least two main reasons: (1) omitting the outside option implies that the estimated increase in the market shares of other hospitals is overestimated; (2) if preferences are not homogeneous, the logit structure, imposing the restrictive IIA property may give biased estimates of the substitution substitution patterns between hospitals, e.g. closing an hospital in London should have a barely noticeable impact to the demand of hospitals in Manchester 300 Km away.

Table 6. Hospital Closure Simulation: Estimated number of relocated patients

Hosp.	Before	Logit			2SM			Waiting Time	Teach.	MRSA	Tot. Beds
		After	Increase	% Increase	After	Increase	% Increase				
H	137	0	-	-	0	-	-	2.72	1	1.62	1449
A	98	134.16	36.16	0.37	168.12	70.12	0.72	3.03	1	1.44	901
B	35	48.91	13.91	0.40	46.62	11.62	0.33	2.88	0	1.46	431
C	78	98.52	20.52	0.26	84.98	6.98	0.09	3.23	0	1.32	663
D	86	93.7	7.7	0.09	90.49	4.49	0.05	3.73	1	1.52	1109
E	603	617.11	14.11	0.02	604.78	1.78	0.01	2.64	1	2.31	824
All other NHS hospitals			44.60			9.01					
Outside Option			-			33					

Note: H is the closing hospital.

Recent news in UK newspapers reported that the NHS considered cutting expenses by downgrading some hospitals in London, i.e. closing the Acute and Emergency hospitals and some other departments. We draw inspiration from this news to illustrate an application of our model to estimate the change in hospital market shares and patient flows after an hypothetical hospital closure. We consider the closure of the orthopaedic department of a teaching hospital in a highly competitive market area served by public and private providers, i.e. hospital H in Figure 1. In 2009, the number of over-65 patients treated for elective hip replacement in hospital H was 137.

Table 6 shows estimated increments in the market shares of five of the major competitors of hospital H after its closure (i.e. Hospital A, B, C, D, E). Although both the standard Logit model and the 2SM model identify hospital A as having the greatest increase in demand after closure, the magnitude and substitution patterns are sharply different. The increment in the demand for hospital A predicted by the 2SM is almost double the size predicted by the Logit model; this reflects the fact that hospital A is very similar to the closing hospital H, and hence a close substitute for patients (both are teaching hospitals of similar size, quality and waiting time). In contrast, the Logit model reallocate patient flows more smoothly across the five competitors, including hospitals that are quite different with respect to distance and quality, such as hospital E, and predicts that a substantial portion of patients (about 45) will relocate to other hospitals, compared to the 2SM model which predicts that only 9 patients will go to hospitals different than A, B, C, D, E. The Logit model also fails to predict the flow of patients that will choose the outside option after hospital closure, and spreads them across public hospitals (about 33 patients using the 2SM predictions).

6. Conclusions

This study presents a structural model of hospital choice that allows for outside option, endogeneity of waiting time and patient preference heterogeneity. One of the original contributions of our model is to use a two-sample estimation strategy that makes creative use of widely available administrative data. This solution can be adopted for studying choice in elective care markets where patients can choose between public and private hospitals, but micro data on the latter are not easily available to the researcher.

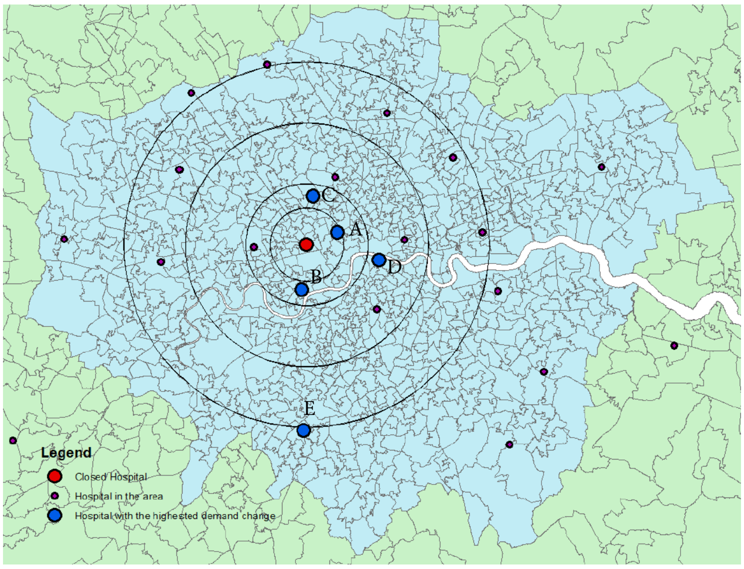

FIGURE 1. Geographic Distribution of Hospitals in Hypothetical Closure

We examine data on all elective admissions during the fiscal year 2006 to 2009 of patients aged 65 and over receiving a primary hip replacement. Using information on patients' LSOA place of residence and the size of private sector, we build a synthetic sample to mimic the population of patients seeking hip replacement surgery. To match the observed with the synthetic sample, we set micro and macro moments based on observed patient's characteristics - e.g. patient's socioeconomic and health status measured at LSOA geographical level. Although using an area level indicator to capture patient's health might be seen as a limitation of our approach, we find that it produces similar elasticities and policy conclusions to individual level indicators. In contrast, omitting the private sector results in serious bias in estimated elasticities and substantially different policy conclusions.

We find evidence that using a rich structural model of hospital choice is key for understanding the relevance of choice in a publicly funded health system. In contrast, predictions from standard choice models, typically used in this literature, might lead to biased conclusions on hospital demand estimation. Patient's response to choice, as measured by hospital demand elasticities, is noticeably underestimated when private providers are omitted from the patient's menu, a simplification adopted by the vast majority of studies. Demand elasticities predicted by a standard choice model omitting private providers are underestimated by 35-40% for waiting time and 50% for MRSA (an indicator of hospital quality) as compared to predictions from our model.

Our results also show that using an appropriate control for endogeneity is key for the correct identification of waiting time elasticities preventing bias from the correlation between waiting time and unobserved quality. By using hospital FE and an IV approach in a nonlinear context, we find waiting time elasticities that are three times larger than by using hospital FE only, which is an approach widely adopted in the literature (e.g. Beckert, *et al.*, 2012; Sivey, 2012; Gaynor *et al.*, 2016; Gutacker *et al.* 2016; Moscone *et al.*, 2012). Our findings are in line with evidence in the IO literature on the identification of price elasticity when quality is partially observable by the researcher (Berry *et al.* 1995; Nevo 2000). We also find evidence that patients respond to differences in hospital quality as measured by MRSA infection rates. Other studies that had access to a larger basket of quality indicators found evidence that Patient Reported Outcome Measure (PROMs) is a relevant indicator for patient choice (Gutacker *et al.*, 2016).

The key policy implication emerging from our contribution is that, in a publicly funded health care system, policy relevant decisions involving the hospital industry should take into account the existence of private providers, the presence of unobservable hospital quality, and heterogeneity in patients' preferences. Omitting these factors is likely to produce poorly informed policy measures to regulate this industry.

The relevance of these factors is highlighted in our hospital closure simulation analysis. In particular, predictions from a standard choice model might result in overestimating the effect on flows and market shares of hospitals that are distant alternatives to the closing hospital, while underestimating the impact on close substitute hospitals.

Our model offers a range of potential applications to future research, including investigating the relationship between market structure and quality in the hospital industry, and simulating the impact of hospital mergers to inform decisions made by the competition authority.

References

Beckert, W., M. Christensen, and K. Collyer (2012). Choice of NHS-funded hospital services in england. *The Economic Journal 122*(560), 400–417.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.

Berry, S., J. Levinsohn, and A. Pakes (2004). Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy 112*(1), 68–105.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics 25*(2), 242–262.

Berta, P., G. Martini, F. Moscone, and G. Vittadini (2016). The association between asymmetric information, hospital competition and quality of healthcare: evidence from italy. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 179*(4), 907–926.

Bloom, N., Z. Cooper, M. Gaynor, S. Gibbons, S. Jones, A. McGuire, R. Moreno-Serra, C. Propper, J. Van Reenen, and S. Seiler (2011). In defence of our research on competition in England's National Health Service. *The Lancet 378*(9809), 2064–2065.

Blundell, R., L. Pistaferri, and I. Preston (2008, December). Consumption inequality and partial insurance. *American Economic Review 98*(5), 1887–1921.

Briggs, D., D. Fecht, and K. De Hoogh (2007). Census data issues for epidemiology and health risk assessment: experiences from the Small Area Health Statistics Unit. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170*(2), 355–378.

Capps, C., D. Dranove, and M. Satterthwaite (2003). Competition and market power in option demand markets. *The RAND Journal of Economics 34*(4), 737–763.

Cooper, Z., S. Gibbons, S. Jones, and A. McGuire (2011). Does hospital competition save lives? evidence from the english nhs patient choice reforms. *The Economic Journal 121*(554), F228–F260.

Gaynor, M., R. Moreno-Serra, and C. Propper (2013). Death by market power: Reform, competition, and patient outcomes in the National Health Service. *American Economic Journal: Economic Policy 5*(4), 134–66.

Gaynor, M., C. Propper, and S. Seiler (2016). Free to choose? Reform, choice, and consideration sets in the English National Health Service. *American Economic Review 106*(11), 3521–57.

Goolsbee, A. and A. Petrin (2004). The consumer gains from direct broadcast satellites and the competition with cable TV. *Econometrica 72*(2), 351–381.

Gutacker, N., L. Siciliani, G. Moscelli, and H. Gravelle (2016). Choice of hospital: Which type of quality matters? *Journal of Health Economics 50*, 230 – 246.

Ho, K. (2006). The welfare effects of restricted hospital choice in the US medical care market. *Journal of Applied Econometrics 21*(7), 1039–1079.

Ho, K. (2009). Insurer-provider networks in the medical care market. *The American Economic Review 99*(1), 393–430.

Imbens, G. W. and T. Lancaster (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies 61*(4), 655–680.

Kelly, E. and G. Stoye (2015). New joints: private providers and rising demand in the English National Health Service. Working Papers W16/15, Institute for Fiscal Studies.

Laudicella, M., P. Li Donni, and P. C. Smith (2013). Hospital readmission rates: signal of failure or success? *Journal of health economics 32*(5), 909–921.

Longo, F., L. Siciliani, H. Gravelle, and R. Santos (2017). Do hospitals respond to rivals' quality and efficiency? A spatial panel econometric analysis. *Health Economics 26*, 38–62.

Martin, S. and P. C. Smith (2003). Using panel methods to model waiting times for national health service surgery. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 166*, 369–387.

Moscelli, G., L. Siciliani, N. Gutacker, and H. Gravelle (2016). Location, quality and choice of hospital: Evidence from England 2002-2013. *Regional Science and Urban*

Economics 60(Supplement C), 112 – 124.

Moscone, F., E. Tosetti, and G. Vittadini (2012). Social interaction in patients' hospital choice: evidence from italy. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(2), 453–472.

National Joint Registry (2010). National Joint Registry for England and Wales: Annual report. Technical report, The National Joint Registry Centre, Hemel Hempstead.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy 9*(4), 513–548.

Nevo, A. (2011). Empirical models of consumer behavior. *Annual Review of Economics 3*(1), 51–75.

Noble, M., G. Wright, G. Smith, and C. Dibben (2006). Measuring multiple deprivation at the small-area level. *Environment and planning A 38*(1), 169–185.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*(5), 1027–1057.

Papanicolas, I. and A. McGuire (2017, 2). Measuring and forecasting quality in english hospitals. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 180*(2), 409–432.

Petrin, A. (2002). Quantifying the benefits of new products: The case of the minivan. *Journal of Political Economy 110*(4), 705–729.

Pollock, A., A. Majeed, A. Macfarlane, I. Greener, G. Kirkwood, H. Mellett, S. Godden, S. Boyle, C. Morelli, and P. Brhlikova (2011). In defence of our research on competition in England's National Health Service–authors' reply. *The Lancet 378*(9809), 2065–2066.

Propper, C. and J. Van Reenen (2010). Can pay regulation kill? panel data evidence on the effect of labor markets on hospital performance. *Journal of Political Economy 118*(2), 222–273.

Riganti, A., L. Siciliani, and C. V. Fiorio (2017). The effect of waiting times on demand and supply for elective surgery: Evidence from Italy. *Health Economics 26*(S2), 92–105.

Senthi, S., J. T. Munro, and R. P. Pitto (2011). Infection in total hip replacement: Meta-analysis. *International orthopaedics 35*(2), 253–260.

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

Train, K. E. and C. Winston (2007). Vehicle choice behavior and the declining market share of U.S. automakers. *International Economic Review 48*(4), 1469–1496.

Williams, M., J. Newton, S. Frankel, F. Braddon, E. Barclay, and J. Gray (1994). Prevalence of total hip replacement: how much demand has been met? *Journal of Epidemiology & Community Health 48*(2), 188–191.

Appendix A. Epidemiological Two Samples Model Estimates

This appendix is intended as supplementary material to this article. Appendix A reports the full set of results from the second version of the 2SM where the LSOA synthetic sample is based on Epidemiological studies (Epi-2SM).

Table 7. First Stage of Epi-2SM

	Coef.	SE
wt*I	-0.0377	0.0439
d*I	-0.8898	0.0587
wt*H	0.0596	0.0346
d*H	0.8917	0.0406
d	-8.1910	0.2074
a	2.9304	0.1057
Outside option		
I	-3.3379	0.5871
I^2	1.0687	0.1267
H	5.2135	0.2741
H^2	-0.2364	0.1012
Preference Heteroegenity		
$R_{0,s}$	3.5686	0.8471
$R_{w,s}$	0.1700	0.0411
$R_{d,s}$	2.9928	0.1024

Bootstrapped standard errors (SE) are based on 500 replications.

Table 8. Second Stage of Epi-2SM

	Hospital FE (OLS)		Hospital FE + IV (TSLS)	
	Coef.	SE	Coef.	SE
CQC	0.0421	0.1010	-0.0554	0.1045
SMR	0.0417	0.0268	0.0468	0.0292
Read	-4.2399	1.3174	-5.7979	1.4458
MRSA	-1.0388	0.1580	-1.2555	0.1805
Waiting Time	-0.4041	0.0588	-1.4476	0.3059
Fixed Effects	Yes		Yes	

Bootstrapped standard errors (SE) are based on 500 replications.

Table 9. Average Waiting Time Elasticities of Epi-2SM

Year	Hospital FE (OLS)		Hospital FE+IV (TSLS)	
	Mean	S.D.	Mean	S.D.
2006	-0.2961	0.0989	-1.3231	0.4373
2007	-0.2439	0.1041	-1.0649	0.4328
2008	-0.1740	0.0756	-0.7783	0.3443
2009	-0.1839	0.0794	-0.8283	0.3373

Table 10. Average MRSA Elasticities of Epi-2SM

Year	Hospital FE (OLS)		Hospital FE+IV (TSLS)	
	Mean	S.D.	Mean	S.D.
2006	-0.3353	0.1555	-0.4051	0.1878
2007	-0.2461	0.1393	-0.2974	0.1682
2008	-0.2439	0.1323	-0.2947	0.1597
2009	-0.2490	0.1368	-0.3009	0.1653

Appendix B. Descriptives statistics

Variable	Label	Mean	S.D.
wt	Average waiting time at time of admission	3.58	2.33
d	Log of distance in km	2.11	0.99
I	Index of income deprivation	0.13	0.09
H	Index of health deprivation	-0.14	0.82
CQC	CQC quality rating	2.98	0.84
SMR	Standardized mortality rate	9.07	1.91
Read	Indicator of hospital's predicted performance on readmission	-0.49	0.05
MRSA	Incidence of Methicillin-Resistant Staphylococcus Aureus	1.26	0.53
bed	Total number of beds	877.77	410.81
sites	Total number of sites	1.85	1.08
MFF	Market Force Factor	1.09	0.07
doc	Number of doctors	520.28	286.42
qual	Number of qualified nurses	1695.63	896.76
unqual	Number of unqualified nurses	1916.73	959.55
allied	Number of Allied health profession	213.81	124.16
teaching	1=teaching hospital, 0 otherwise	0.12	
lon	1=London hospital, 0 otheriwse	0.06	

Note: Obs. = 104,606.