Free to Choose?
Reform and Demand Response in the English National Health Service*

Martin Gaynor
Carnegie Mellon University
University of Bristol
NBER

Carol Propper
Imperial College
University of Bristol
CEPR

Stephan Seiler
Stanford University
Centre for Economic Performance

This draft: May 20, 2013

*We are grateful to the UK Department of Health for financial support. No endorsement by the Department of Health is intended or implied. Thanks are due for helpful comments to participants at seminars at a number of universities and conferences. Any errors and all opinions are the responsibility of the authors alone.

Abstract

The impacts of choice in public services are controversial. We exploit a reform in the English National Health Service to assess the impact of relaxing constraints on patient choice. We estimate a demand model to evaluate whether increased choice increased demand elasticity faced by hospitals with regard to clinical quality and waiting time for an important surgical procedure. We find substantial impacts of the removal of restrictions. Patients became more responsive to clinical quality. Sicker and better informed patients were more affected. We leverage our model to calculate potential benefits. We find increased demand responsiveness led to a significant reduction in mortality and an increase in patient welfare. The elasticity of demand faced by hospitals increased post-reform, giving hospitals potentially large incentives to improve their quality of care and find suggestive evidence that hospitals responded strongly to the enhanced incentives due to increased demand elasticity. The results suggests greater choice can enhance quality.

JEL Classification: D12, I11, I18, L13, L30

Keywords: Demand Estimation, Non-price Competition, Health Economics, Patient Choice, Health Care Reform

1 Introduction

Governments facing fiscal pressure have increasingly turned to proposals to create or enhance consumer choice for public services(see, e.g., Besley and Ghatak 2003, Blöchliger 2008, Hoxby 2003, Le Grand 2003). In health care, choice is a popular reform model adopted by administrations of different political orientations in many countries, including the US, the UK, Denmark, Italy (Lombardy), the Netherlands, Germany and Sweden. The belief is that by increasing choice for patients, providers of care or insurers will become more responsive to patient demand which, in turn, will drive greater efficiency in the delivery and funding of health care. These reforms have been controversial, in part due to concerns about adverse impacts on patients' health in general, and inequitable impacts on low income individuals specifically. Whether enhanced patient choice will make hospital choice more responsive to quality is not well established. Consumers may be deterred from exercising choice due to lack of information about medical care providers, because the measures produced to help consumers choose may be noisy and difficult to interpret for consumers, or simply because consumers do not value quality in health care. The consequences of poor quality in health care can be dire. Patients' health can be severely compromised by poor quality care, including, as we show below, an increased risk of death. Thus there is a need to understand the responses of health care consumers when they are offered more choice.

This is exactly the research question we address in this paper: we quantify changes in the elasticity of demand with respect to quality of service when patients are offered more choice. Furthermore, we analyze how the changes in patients' choices translate into changes in the competitive environment hospitals are facing. To answer these questions we exploit a reform which introduced patient choice in the English National Health Service (NHS). Beginning in January 2006, the reform mandated that patients in the NHS had to be offered a choice of 5 hospitals when referred by their physician to a hospital for treatment. At the same time removal of selective contracting allowed the referring physician flexibility in the choice of hospital for the patients treatment. Finally, patients and physicians were provided with greater information about hospitals, primarily through a webpage containing easily accessible performance information across hospitals. Together these three factors changed the way referrals were made by removing institutional and informational constraints on the patients' ability to choose between hospitals. The reform thus provides us with exogenous variation in the ability to exercise choice over time and as the choice set of hospitals is almost constant around the introduction of the choice reform, it allows us to cleanly identify the impact of choice while holding the underlying market structure fixed.

We propose a hospital choice function to estimate the elasticity of demand with respect to various dimensions of hospital service pre- and post-reform. In the NHS patients face zero price at point of use, therefore quality of service, waiting times, and distance serve to allocate patients among hospitals in the absence of a price mechanism. We estimate our model using choice data for Coronary Artery Bypass Graft (CABG) surgery. We

find that the reform had a significant impact on patients' elasticity of demand with respect to the quality of service (captured by a case-mix adjusted mortality rate).[1] We also find substantial heterogeneity in the impact of the reform on responsiveness to quality. Sicker patients react more strongly to the reform as well as patients that were likely to know about the ability to exercise choice. Interestingly, lower income groups did not benefit less than the average patient, which was a widespread concern when the pro-choice policy was introduced. The reform had no significant impact on the elasticity of demand with respect to waiting times for most patients.

Using the estimates from the choice model we estimate that 10 more patients per year would have survived their CABG surgery (a 3 percent decline in mortality) had patients had free choice in the pre-reform period (i.e., had demand been as elastic with respect to quality pre-reform as it was post-reform). When we aggregate patient-level elasticities to the hospital level, our simulation shows that the competitive environment changed substantially. For the average hospital an increase in mortality by one standard deviation led to a 5 percent larger drop in market share post- compared to pre-reform. This lends support to the notion that hospitals did have stronger incentives to improve quality due to the introduction of patient choice.

Relative to other papers that analyze the impact of more choice/stronger competition on health outcomes,[2] we analyze choice more directly by structurally estimating the elasticity of demand with respect to two key dimensions of hospital service, waiting times and mortality rates. We thereby contribute to a growing literature analyzing consumer choice in health care markets such as Luft, Garnick, Mark, Peltzman, Phibbs, Lichtenberg, and McPhee (1990), Tay (2003) and Howard (2005) for the US and Sivey (2008), Beckert, Christensen, and Collyer (2012), Varkevisser, van der Geest, and Schut (2012) and Moscone, Tosetti, and Vittadini (2012) for Europe.

Analyzing hospital choice is a challenging task for two reasons. First, finding an appropriate measure of quality is difficult, as health outcomes are affected by patients selecting into hospitals based on their health status. Second, unobserved hospital quality will influence aggregate demand, which in turn has an impact on waiting times, creating an endogeneity problem.[3] To address the first concern, we introduce an appropriate measure of quality by explicitly estimating the causal effect of visiting a particular hospital on patient survival. To this end we estimate a model of health outcome (i.e. patient survival) conditional on visiting a certain hospital. We use an instrumental variables approach to deal with the issue of patients selecting into hospitals, following Gowrisankaran and Town (1999) and Geweke, Gowrisankaran, and Town (2003). In this way we are able to back out a measure of hospital mortality that is not contaminated by patient selection, i.e. case-mix. The estimated quality measure is then used as a hospital characteristic in the demand model. To deal with the possible endogeneity of waiting times, we control for unobserved heterogeneity in a very flexible way.

[1] Our measure of quality of service is a hospital's case-mix adjusted mortality rate, therefore throughout the paper we will use those terms interchangeably.

[2] See Kessler and McClellan (2000) for evidence from the US Medicare program and Gaynor, Moreno-Serra, and Propper (2010) and Cooper, Gibbons, Jones, and McGuire (2011) for evidence from the English NHS.

[3] Note that price endogeneity is not an issue in the UK setting, since health care is free at the point of service under the NHS.

Specifically, in a first stage we estimate a large set of hospital and time-period specific fixed effects. This enables us to rigorously control for the effect of unobserved hospital quality and to recover parameters that are identified by patient-level variation. In a second stage we recover the average effects of the key parameters by projecting the recovered hospital/time-period fixed effects on hospital characteristics (see Goolsbee and Petrin 2004, for a similar approach).

The paper is structured as follows. Section 2 describes the model, Section 3 explains the institutional background, Section 4 describes the data we use, followed by some reduced-form evidence in Section 5. Section 6 presents econometric issues and estimation methods and Section 7 presents the results, including parameter estimates and elasticities. Section 8 presents the results of some policy analyses using our estimates. The final section contains concluding remarks.

2 Modeling Approach

We discuss CABG surgery and the institutional setting in Section 3. Here we lay out our framework for analyzing the impact of the reforms on hospital demand. But before presenting this we need to note three key features of the institional set-up. First, patients and physicians jointly choose the the hospital for the patient's surgical treatment. The physician makes this referral based on their information set and their clinial judgement. This is standard in a medical setting where the role of the physician is to act as the patient's agent and to provide advice on appropriate treatment. Second, in the UK physicians are salaried, therefore the referring physician does not benefit financially from the referral decision. In fact, a poor clinical choice may have a negative impact on utility, in terms of loss of reputation and possibly increased workload with no increased financial payment. Third, the patient incurs no financial cost for medical treatment. This means that the choice of hospital is the decision of both the patient and the physician but separately identifying patient and physician preferences is not possible here (as is the case for the health demand literature in general). This is not an issue in evaluating the impact of the reform on hospital choice and for ease of exposition we maintain the standard terminology and refer to patient utility and preferences when describing the model.

There are two key ingredients to this model: (1) a model of patient choice as a function of hospital characteristics/dimensions of service, and (2) a hospital-level production function for the quality of clinical care. While the key focus of the paper is the choice model which allows us to pin down how sensitive patients' decisions are to different dimensions of service. But one of these is the quality of care at the hospital and we therefore need to find an appropriate measure for the hospital-level quality of care. A model for the production of clinical quality will help us in that respect. We describe the choice model first.

2.1 Choice Model

Consumers choose between hospitals based on their utility from this choice. Price in the NHS is zero for the consumer, so indirect utility is only a function of patient and hospital characteristics. The key factors which are likely to affect hospital choice are the quality of care, the amount of time a patient has to wait for surgery and patient distance from the hospital. We allow for preference heterogeneity across different patient demographics as well as unobserved heterogeneity in preferences over hospital characteristics. We assume that all people who require a CABG are sick enough that they get one (after a wait). As a consequence, there is no outside good. Finally, in order to capture effects of loosening of choice constraints due to the reform, we allow patients' responsiveness to hospital characteristics to differ pre- and post-reform.

Let a patient i choose between hospitals j in time period t (defined as a quarter) based on the following indirect utility function,

$$V_{ijt} = \beta_{w,it} W_{jt} + \beta_{z,it} Z_{jt} + f(D_{ij}) + \xi_{jt} + \varepsilon_{ijt} \tag{1}$$

where W_{jt} denotes the average waiting time for a CABG in time period t at hospital j, Z_{jt} denotes the quality of clinical care at the hospital, and D_{ij} is the distance from patient i's location to the location of hospital j. The function $f(D_{ij})$ is a transformation of D_{ij} that reflects the (non-linear) preference for distance to the hospital. ξ_{jt} denotes unobserved hospital quality and ε_{ijt} is an idiosyncratic shock that is iid extreme value.

To allow the impact of quality of care to differ across patients as well as before and after the reform let the parameter $\beta_{z,it}$ be defined as follows,

$$\begin{aligned} \beta_{z,it} &= [\overline{\beta}_{z,0} + \beta_{z,0} X_i + \sigma_{z,0} v_{z,i}] \cdot \mathbf{1}(t=0) \\ &\quad + [\overline{\beta}_{z,1} + \beta_{z,1} X_{it} + \sigma_{z,1} v_{z,i}] \cdot \mathbf{1}(t=1). \end{aligned} \tag{2}$$

where $(t=0)$ denotes the pre-reform time period, $(t=1)$ denotes the time period after the reform, and $\mathbf{1}(t)$ is an indicator function. $\overline{\beta}_{z,t}$ denotes the average effect across consumers of quality of care on patient utility in period t. X_i is a vector of observable patient demographics and $\beta_{z,t}$ is a vector of coefficients that capture differences from the average effect across consumers (implemented as various demographic groups). Finally, $\sigma_{z,t}$ captures unobserved heterogeneity in preferences for clinical quality across patients. It is the standard deviation of a normal distribution with mean zero and variance $\sigma^2_{z,t}$. In our specific context, patient health status (which is only partially observable) will most likely affect preferences over the quality of care. The impact of unobserved health status is therefore captured by the realization of $v_{z,i}$. Note that we allow for different average effects pre- and post-reform as well as demographic specific deviations from the average effect pre- and post-reform. In other words, we allow for overall (average) effects of the freeing of choice and for the freeing of choice to differentially affect different demographic groups.

We capture the effect of the reform and consumer heterogeneity on the impact of waiting times on choice in a similar fashion,

$$\begin{aligned} \beta_{w,it} &= [\overline{\beta}_{w,0} + \beta_{w,0} X_i + \sigma_{w,0} v_{w,i}] \cdot \mathbf{1}(t=0) \\ &\quad + [\overline{\beta}_{w,1} + \beta_{w,1} X_i + \sigma_{w,1} v_{w,i}] \cdot \mathbf{1}(t=1) \end{aligned} \tag{3}$$

This allows us to rewrite the utility function in the following way,

$$\begin{aligned} u_{ijt} &= \delta_{jt} \\ &\quad + [\beta_{w,0} X_i \cdot \mathbf{1}(t=0) + \beta_{w,1} X_i \cdot \mathbf{1}(t=1)] \cdot W_{jt} \\ &\quad + [\sigma_{w,0} v_{w,i} \cdot \mathbf{1}(t=0) + \sigma_{w,1} v_{w,i} \cdot \mathbf{1}(t=1)] \cdot W_{jt} \\ &\quad + [\beta_{z,0} X_i \cdot \mathbf{1}(t=0) + \beta_{z,1} X_i \cdot \mathbf{1}(t=1)] \cdot Z_{jt} \\ &\quad + [\sigma_{z,0} v_{z,i} \cdot \mathbf{1}(t=0) + \sigma_{z,1} v_{z,i} \cdot \mathbf{1}(t=1)] \cdot Z_{jt} \\ &\quad + f(D_{ij}) + \varepsilon_{ijt} \end{aligned} \tag{4}$$

where δ_{jt} captures the effect of hospital characteristics on the average patient pre- and post-reform as well as the unobserved quality term

$$\begin{aligned} \delta_{jt} &= [\overline{\beta}_{w,0} \cdot \mathbf{1}(t=0) + \overline{\beta}_{w,1} \cdot \mathbf{1}(t=1)] \cdot W_{jt} \\ &\quad + [\overline{\beta}_{z,0} \cdot \mathbf{1}(t=0) + \overline{\beta}_{z,1} \cdot \mathbf{1}(t=1)] \cdot Z_{jt} \\ &\quad + \xi_{jt} \end{aligned} \tag{5}$$

A main concern that drives our choice of estimation method is the possible endogeneity of waiting times in the utility function (1). Specifically, it is possible that unobservably better hospitals may have longer waiting times because they attract more patients. By increasing aggregate demand, higher unobserved quality ξ_{jt} will lead to longer waiting times for a given level of capacity, so $Corr(W_{jt}, \xi_{jt}) \neq 0$, which implies that we will be unable to obtain a consistent estimate of the effect of waiting times on hospital choice ($\beta_{w,it}$) without addressing this issue. The issue is very similar to the endogeneity of the price coefficient commonly encountered in the empirical IO literature. Products with higher unobserved quality will have greater demand, which in turn leads to higher prices. An analogous mechanism will drive waiting times up in the fixed price (and capacity constrained) environment of the English NHS. In other words, rationing through waiting times plays a similar role to the price mechanism in other markets. This will lead to waiting times being positively correlated with unobserved hospital quality.

In principle this endogeneity problem can be addressed either by using instrumental variables or by controlling for unobserved heterogeneity via fixed effects to absorb the variation in ξ_{jt}. Since there is no obvious good instrument for waiting times, we use a fixed effects approach.[4] Specifically, we estimate a separate hospital fixed effect for every quarter, represented by δ_{jt} in Equations (4) and (5). We therefore control more rigorously for unobserved hospital quality than approaches that use only hospital (but not time period) specific fixed effects, i.e. that assume $\xi_{jt} = \xi_j$ for all t. We think that in our context, the assumption of time-invariant fixed effects is less tenable than in other applications. This is due to the fact that we are analyzing the effect of a reform that (in principle) could have led to hospitals improving along different dimensions of service. Any improvement on the unobserved quality dimension would therefore violate an assumption that $\xi_{jt} = \xi_j$.

This greater degree of flexibility is costly for us in two ways: first, some parameters cannot be identified when a full set of hospital-quarter fixed effects are included (as can be seen in equation (5)), and second, we have to estimate a set of almost 300 fixed effects (for about 30 hospitals and 10 quarters) within a non-linear model, which leads to a considerable increase in the computational burden. We provide details on how we deal with both issues in the estimation section.

2.2 Production Function for the Clinical Quality of Care (Mortality)

The second component of our modeling strategy involves an empirical model of the hospital's production of clinical quality. In the case of CABG where there is a non-negligible risk of dying the patient's chances of survival are a primary dimension of quality. Therefore we focus on a "production function of patient survival." Here, in contrast to the demand model described above, we take a more reduced form approach to modeling the production function. This is motivated by the fact that the main purpose of this production function estimation is simply to deliver an appropriate measure of quality of care to use in the demand model.

Specifically, we specify a linear probability model of patient mortality in which we regress an indicator for whether the patient died after the surgery (conditional on visiting hospital j) on patient characteristics and a set of hospital/time period fixed effects. Let the mortality of patient i in period t at hospital j be determined as follows,

$$M = JT\psi + H^{obs}\gamma_{obs} + H^{unobs}\gamma_{unobs} + \eta \tag{6}$$

where M is a vector of indicator variables. An entry corresponds to a particular patient i receiving a CABG in time period t and is equal to one if the patient died after receiving treatment in hospital j. JT is a matrix of hospital-time period dummy variables and the ψ are a set of coefficients. As previously, we define a time period

[4] One could consider using waiting times for other procedures at the same hospital as instruments. However, it is likely that unobserved quality is correlated across procedures. For instance, general hospital reputation might affect demand similarly across procedures. In addition, NHS hospitals do not operate in multiple, widely dispersed locations. Therefore the common strategy of using values of the endogenous variable from a distant market therefore is not available here. Note also **?** recommends the use of fixed effects if there are enough alternatives.

to be a quarter. H^{obs} is a matrix of patient (i.e. i-specific) characteristics that capture observable health status. H^{unobs} represents the patient's unobserved health status. γ_{obs} is a vector of coefficients that captures the impact of observed patient characteristics on the probability of death following the surgery, while γ_{unobs} captures the impact of unobservable health status. η is a vector of iid normal error terms. The corresponding expression for a specific observation is $M_{it}(j) = \psi_{jt} + H_i^{obs}\gamma_{obs} + H_i^{unobs}\gamma_{unobs} + \eta_i$.

We use this reduced-form production function regression to estimate the causal impact of visiting a particular hospital on the patient's probability of dying. However, simply estimating the relationship by OLS does not allow us to uncover the true causal relationship. Hospital choice will likely be correlated with unobserved patient health status, which will be subsumed in the empirical error term when estimating equation (**??**). More specifically, if $Corr(v_{z,i}, H_i^{unobs}) \neq 0$ then hospital choice will be endogenous. From the demand model described above, we know that $v_{z,i}$ will effect hospital choice, therefore any arbitrary column of the hospital dummy matrix, JT_i^{jt} will be a function of $v_{z,i}$, and so $Corr(JT_i^{jt}(v_{z,i}), H_i^{unobs}) \neq 0$.[5]

This endogeneity problem is very closely related to the fact that the hospital's mortality rate is "contaminated" by differences in patient case-mix. Indeed, when running the above regression without any controls for observed patient health status, the fitted hospital fixed effects $\widehat{\psi}$ will be equal to the hospital-specific mortality rates.[6] The production function equation therefore recasts the issue of case-mix affecting the mortality rate as an endogeneity problem.

To obtain the true causal effect of visiting a particular hospital on the survival probability we need to deal with this endogeneity problem. We rely on the demand model to give us some guidance. In particular, we know that distance affects hospital choice. Together with the additional assumption that distance is uncorrelated with unobserved health status $Corr(D_{ij}, H_i^{unobs}) = 0$, we can use distance as an instrument. This amounts to assuming that people do not choose where they live relative to CABG hospitals based on their unobservable health status.[7] We provide some supporting evidence for this later in the paper. This IV approach is very closely related to Gowrisankaran and Town (1999) as well as Geweke, Gowrisankaran, and Town (2003). Other papers in the health literature such as Luft, Garnick, Mark, Peltzman, Phibbs, Lichtenberg, and McPhee (1990), Howard (2005), and Tay (2003) use the difference between expected and observed mortality, where expected mortality is computed based on observed patient case-mix. Our IV approach goes further than this and also deals with selection based on unobserved health status.

Having recovered the causal impact of visiting hospital j on the patient's chances of survival we use this set

[5] We write JT_i^{jt} to denote the entry in column i and row jt of the JT matrix, i.e. the realization of a certain hospital-time period dummy for patient i.

[6] In a linear regression model without a constant the fixed effects are equal to the hospital-specific means of the dependent variable $M_{it}(j)$. The average of $M_{it}(j)$ for a particular hospital j (and time-period t) is therefore simply equal to the number of death divided by the total number of admissions, i.e. the mortality rate.

[7] This assumption is universally employed in estimating models of hospital choice, e.g., Kessler and McClellan (2000), Gowrisankaran and Town (1999), Capps, Dranove, and Satterthwaite (2003), Gaynor and Vogt (2003), Ho (2009), Beckert, Christensen, and Collyer (2012).

of estimated fixed effects in order to capture hospital quality of care in the demand model. We will refer to these effects as case-mix adjusted mortality rates. This constitutes a slight abuse of terminology as the fixed effects do not constitute mortality probabilities but rather the hospital's impact on mortality conditional on observed case-mix.

3 Institutional Details

3.1 CABG: Medical Background

A coronary artery bypass graft (CABG) is a surgical procedure widely used to treat coronary heart disease. It is used for people with severe angina (chest pain due to coronary heart disease) or who are at high risk of a heart attack. It diverts blood around narrowed or clogged parts of the major arteries, to improve blood flow and oxygen supply to the heart. It involves taking a blood vessel from another part of the body, usually the chest or leg, and attaching it to the coronary artery above and below the narrowed area or blockage. This new blood vessel, known as a graft, diverts the flow of blood around the part of the coronary artery that is narrowed or blocked.[8] Successful bypass surgery improves symptoms and lowers the risk of heart attack.

We focus on CABG for three reasons. First, it is a commonly performed procedure. About 13,500 patients per year receive elective CABGs in England, making CABG one of the most frequently performed elective treatments.[9] The fact that it is commonly performed provides us with statistical power and means that CABG is quantitatively important. Second, CABG is mostly performed on an elective, as opposed to an emergency, basis. Therefore, patients can exercise choice among alternatives, which is not usually the case for emergency treatments. Third, patients who receive heart bypass surgery are very sick, so CABG is among the most risky elective treatments and mortality is a fairly common outcome.[10] The relatively high frequency of death means mortality is a reliable and easily observed measure of quality. Other dimensions of quality which characterize other medical procedures are harder to observe and may be less reliably recorded.

Patients in the NHS who present with symptoms of coronary artery disease or angina are referred to a cardiologist who will perform tests and may perform a non surgical procedure to unblock the artery (angioplasty or PCI). If this fails the patient then will be referred for a CABG to be performed by a cardiac surgeon and put on an elective waiting list for this treatment. The referral is typically made by the cardiologist but in some cases may be made by the patient's primary care physician (the General Practitioner). Cardiologists operate in almost all short term general NHS hospitals but CABGs are performed only at a limited number of hospitals.

[8] http://www.nhs.uk/Conditions/Coronary-artery-bypass/Pages/Introduction.aspx

[9] In the US the number is 415,000, making CABG one of the top 10 most common non-obstetric surgical procedures (National Hospital Discharge Survey 2010, http://www.cdc.gov/nchs/nhds/nhds_products.htm).

[10] Other procedures commonly used in the health economics literature, such as AMI (acute myocardial infarction) treatment have higher mortality rates, but are primarily emergency treatments. They are therefore not directly relevant for an analysis of patient choice.

3.2 The Choice Reform

In the UK health care is tax financed and free at the point of use. Almost all care is provided by the National Health Service. Primary care is provided in the community by publicly funded physicians known as General Practitioners (GPs). Patients have little choice of GP.[11] These GPs also act as gatekeepers for hospital-based (known as secondary) care.[12] Secondary care (e.g., cardiology, cardiac surgery) is provided in publicly funded (NHS) hospitals. NHS hospitals are free standing public organizations (known as NHS Trusts). In these hospitals, the physicians are salaried employees and generally employed only in a single NHS hospital. Publicly funded bodies covering specific geographic areas, called Primary Care Trusts (PCTs), have the task of buying hospital-based health care for their population.

In the pre-reform period, buyers (PCTs) and sellers (the NHS Trusts) negotiated over price, service quality (but mainly waiting times and not clinical outcomes) and volume on an annual basis. The majority of contracts were annual bulk-purchasing contracts between the buyers and a limited number of sellers. Patients requiring secondary care were generally referred by their GPs to the local hospital that provided the service they required and were not offered choice over which hospital they went to. Instead the hospital to which a patient was sent was determined by the selective contracting arrangements negotiated by the PCTs.

In late 2002 the government initiated a reform package to bring about hospital competition from 2006 onwards. There were several elements to this policy. After January 2006 patients had to be offered a choice of five providers for their hospital care (Farrar, Sussex, Yi, Sutton, Chalkley, Scott, and Ma 2007, Gaynor, Moreno-Serra, and Propper 2010, Cooper, Gibbons, Jones, and McGuire 2011), and GPs were required (and paid) to ensure that patients were made aware of, and offered, choice.[13]

There were two other pieces in the reform package to facilitate choice. First, the government introduced a new information system that enabled paperless referrals and appointment bookings and provided information on the different dimensions of service (waiting times and some measures of clinical quality) to help patients make more informed choices. This system, known as "Choose and Book," allows patients to book hospital appointments online, with their GP, or by telephone. The booking interface gave the person booking the appointment the ability to search for hospitals based on geographic distance and to see estimates of each hospital's waiting time. From 2007 the government also introduced a website designed to provide further information to help patients' choices. This included information collected by the national hospital accreditation bodies, such as risk-adjusted mortality rates and detailed information on waiting times, infection rates and hospital activity rates for particular procedures, as well as information on hospital accessibility, general visiting hours and parking

[11] Patients almost always have to choose a GP located near where to they live and practices can choose not to accept new patients.
[12] GPs do not practice in hospitals and therefore have no equivalent of admission rights.
[13] The mandated choice was actually for the hospital in which the patient was initially referred to a specialist. In practice, this generally meant choice of hospital in which treatment took place. In the case of specialised treatments which are not provided in all hospitals (including CABGs), the patient would then be referred by the hospital specialist to another hospital for some or all of their treatment.

arrangements (http://www.chooseandbook.nhs.uk/). Thus, patients and their GPs had greater information on which to make these choices.

Second, from 2006 onwards the NHS adopted a payment system in which hospitals were paid fixed, regulated, prices for treating patients (a regulated price system that is similar to the Medicare hospital payment system in the US). This fixed price system covered around 70% of hospital services including CABG.[14] This change in the remuneration system meant that GPs (and hospital specialists making referrals for treatment at to a hospital other than their own) were no longer restricted in these referral decisions by their PCT's contractual arrangements with individual hospitals.

In summary, restrictions on both physicians and patients were removed in order to make referral decisions more flexible. At the same time all parties were provided with additional information in order to help them in the decision making process. While a well-informed patient with a strong preference might have been able to convince the physician to refer her to a specific hospital pre-reform, the effect of the reforms was to make such choice available and far more explicit for all patients.

In the particular case of CABG, which is a specialised treatment provided at only a few hospitals, the reforms allowed the patient to choose the intitial hospital at which they saw a cardiologist and allowed that cardiologist freedom in choice of where they sent patients for a CABG. This choice would have been based on clinical assessment of what was the best for the patient. The reforms did not change finanical incentives for the cardiologist (or the patient) in this referral decision as both pre- and post-reform, cardiologists received no payment for their referral decision and patients do not pay for their medical care.

In terms of timing, the reform did not happen in a discrete way on a certain date for cardiac care. There were two distinct trial/phase-in periods which we need to take into account when defining the pre- and post-reform dates. The 2006 choice reform was preceded by a choice pilot for cardiac patients who were experiencing particularly long waiting times (over 6 months). Between July 2002 and November 2003 such patients were allowed to change their provider to get treatment earlier. Since we do not observe which patients were actually offered the choice to switch providers, it is difficult to analyze this situation explicitly. At the same time, because patients eligible for this scheme had to have waited for a minimum of six months, this situation is quite different from the full choice reform (in which choice was mandated at the point of the referral) as well as from the situation of no-choice pre-reform. Second, choice was first introduced in a limited way in April 2005 and only fully rolled out in January 2006. In the introductory period, choice between only 2 hospitals was offered to patients and decisions were taken locally as to which choice to offer. In order to keep our analysis as clean as possible we therefore exclude both of these phase-ins of the reform from our analysis. We also allow for

[14]The reforms also promoted use of (mainly) new private providers of care. However, use of these was very limited and accounted for less than 1% of all NHS care during the period in which we analyse. The main services purchased in the private sector were simple elective services (primarily hip and knee replacements and cataract removal) rather than complex interventions such as CABG or cardiac care more generally.

some time (one year) for the reform to settle in and therefore use only data from January 2007 onwards when analyzing the post-reform time period. These restrictions mean we use the period January 2004 to March 2005 as the pre-reform period and January 2007 to March 2008 as the post-reform period.

3.3 Market Structure

In contrast to many other procedures, CABGs are only offered by a small set of hospitals. Of around 160 short term general (acute) public hospitals within the NHS just under 30 hospitals offer bypass operations.[15] There was almost no change in market structure around the time of the policy reform.[16] The choice set faced by patients is therefore very similar before and after the reform which allows us to separate the impact of greater choice from a possible change in market concentration.[17] Figure 1 provides a map of locations of CABG-performing hospitals.

In our econometric modeling, we do not impose a particular geographic market a priori. Post-reform, physicians can refer to any hospital in England that performs CABG. Pre-reform, choice was constrained by selective contracting but the actual provider used might have been some distance from the patient since few hospitals provided CABG surgery. We do not know which hospitals were contracted with by each PCT and cannot infer this from the data. Furthermore, we cannot be sure that the contractual constraints were always binding for every individual referral. We see that a non-negligible fraction of patients travel a substantial distance in order to receive treatment even before the introduction of choice. Both pre- and post-reform about 20 percent of patients traveled more than 50km and 10 percent traveled further than 70km. A small set of patients traveled several hundred kilometers in both time periods. We do not, therefore, restrict patients' choices ex ante by imposing a geographic market but simply allow patients to choose any CABG hospital in England, while controlling for the impact of distance on choice in a flexible way. Thus we allow the data to tell us where patients want to go rather than imposing a priori restrictions on choice. Finally, while in principle patients could choose privately funded treatment, in practice they did not.[18] For this reason we do not include an outside option in the model (this is standard in the health care demand literature).

[15] These hospitals differ from the average short term general hospital. Two-thirds are teaching hospitals (compared to 15 percent for all hospitals) and are also are substantially larger, with about 50 percent more yearly admissions across all departments and 80 percent more medical staff than other short term general hospitals in the NHS.

[16] The only changes are: (i) the merger of Hammersmith Hospital and St. Mary's Hospital that became part of Imperial College Healthcare NHS Trust in 2007, (ii) the opening of the Essex Cardiothoracic Centre at Basildon and Thurrock University Hospitals in July 2007, and (iii) Royal Wolverhampton Hospital started performing a significant number of CABGs only in the second half of 2004 and is therefore excluded from the choice set before that.

[17] Our demand estimation is capable of handling hospital entry and exit but the stable market structure means we isolate the effect of the change in choice without any potential contamination from change in market structure.

[18] During our study period four private providers of CABG surgery operated (all located in London). However, the cost of CABG surgery is such that any patients who might choose to use a private provider would have to have purchased private insurance before they were diagnosed with a heart problem. The four private providers only performed a very small number of CABGs compared to public hospitals (for example, only 67 CABG proceedures were undertaken in the four private hospitals in 2007). Therefore, we think that our data captures the full choice set of patients.

4 Data and Descriptive Statistics

We use data from the UK Department of Health's Hospital Episode Statistics (HES) dataset, which is a standard administrative discharge dataset on every English NHS health episode.Overall, HES data contain approximately 13 million inpatient discharges per year at around 240 hospitals per year. The data contain details of the medical procedures which the patient received (classified according to OPCS codes[19]) and up to 14 diagnoses, classified according to the ICD-10 classification.[20] We have data on the universe of inpatient discharges receiving CABG surgery from every hospital in the NHS in England from April 2003 to March 2008, which corresponds the the UK financial years 2003 to 2007. About 25% of all CABGs are performed as part of an emergency treatment and are excluded from the main analysis. This gives us approximately 13,500 elective CABG discharges performed at 29 hospitals per year. As explained earlier (Section 3.2) we define January 2004 until March 2005 as the pre-reform time period and January 2007 to March 2008 as the post-reform time period and therefore do use only a smaller time-window for our main estimation.

HES contains information on the postal code of the neighborhood in which the patient lives and patient characteristics such as age, sex, and co-morbidities.[21] At the patient-level we observe the time elapsed between the referral and the actual treatment, i.e. the patient's waiting time. We also observe whether the patient died (in the hospital) within 30 days of the treatment. We can therefore compute CABG-specific waiting times and mortality rates by aggregating the data at the hospital level (over a suitable time period). Finally, from the hospital location and the patient's postal code, we can compute the distance to the hospital.

We provide a detailed list of sources for the data used in this paper in Table B in the appendix.

4.1 Hospital Characteristics

In Table 1 we report descriptive statistics for hospitals by (financial) year over the period 2003 - 2007. The average hospital treated about 500 CABG patients per year, but there is substantial variation in admission rates between hospitals. The number of admissions decreases slightly over time as does the variance across hospitals.[22] Waiting times fell dramatically over the period. In 2003 and 2004 they were quite long, with averages over 100 days. They decreased substantially in 2005 due to a government policy enforcing waiting time targets (see Propper, Sutton, Whitnall, and Windmeijer 2008).[23] There is considerable variation in waiting times between hospitals, although somewhat less after 2005. The average mortality rate is approximately 1.9 percent for most

[19] OPCS is a procedural classification for the coding of operations, procedures and interventions performed in the NHS. It is comparable to the CPT codes used for procedural classification in the US.

[20] These are the 10th version of the International Classification of Disease (ICD) codes and are the standard codes used internationally for diagnoses.

[21] Co-morbidities are additional diagnoses associated with greater sickness, for example, a CABG patient who is also a diabetic.

[22] The total number of CABGs in the UK undertaken in our time period fell due to the increased use of angioplasty (PCI).

[23] Note that the drop in waiting times was primarily due to efficiency improvement and did not have any detrimental effect on health outcomes (see Propper, Sutton, Whitnall, and Windmeijer 2008).

years with a slight decline towards the end of the sample period. There is substantial variation in mortality rates across hospitals in all years. We also report an adjusted mortality rate which is purged of differences across hospitals and over time in patients' severity of illness (case-mix).[24] Similar to the raw rate, we can observe a (slightly stronger) downward trend over time. It is worth noting that the adjusted mortality rate shows a larger between hospital variance, which is consistent with the notion that better hospitals attract sicker patients (i.e., a worse case-mix). Since sicker patients go to better hospitals, this selection effect will compress the distribution of the raw mortality rate relative to the case-mix adjusted one.

When using the (case-mix adjusted) mortality rate and waiting times in the demand estimation, we aggregate the patient-level data to the hospital-quarter level. This provides us with variation over time as well as across hospitals. Using January 2004 until March 2005 as the pre-reform time period and January 2007 to March 2008 as the post-reform time period, we exploit 10 quarters of data, 5 in the pre- and 5 in the post-reform time-period. Descriptive statistics of the quarterly variation for this time period are reported in Table 2.

4.2 Patient and Area Characteristics

We can determine the neighborhood of the patient, using the patient's postal code to merge data. We define the neighborhood as a small area containing around 7,000 individuals (the MSOA).[25] We measure patient socio-economic status as the (negative of) the Index of Multiple Deprivation (IMD). The IMD is a measure of income deprivation of the patient's neighborhood. This variable ranks a patient's local neighborhood from richest to poorest. The range is 0 to 1, with higher values implying higher deprivation.[26] Going forward, we simply refer to this variable as "income." Also at a local level, we have data on how well informed patients are about choice of NHS provider. From an NHS patient satisfaction survey,[27] reported at the PCT level, we observe the proportion of patients (of the interviewed sample) who responded positively to a question as to whether they were offered a choice of hospital at the point of referral. While the question is not asked with respect to a specific procedure (such as CABG) but for any referral for elective treatment we see no reason why information about choice should vary by procedure and we think that this is a reasonable proxy for the likelihood that a CABG patient is informed about their possible choices. Finally, HES provides the number of co-morbidities at the patient-level. For the purpose of estimation we cap the co-morbidity count at 6, as hospitals differ in the maximum number

[24] See Section 6.1 for details on the adjustment.

[25] The neighborhood is the Middle Layer Super Output Area (MSOA). These are generally smaller in population than US zip codes. MSOAs are defined to ensure maximum homogeneity of population type. In England each of the 6,780 MSOAs has a minimum population of 5,000 residents and had an average population of 7,200 residents in 2010. For more information see http://neighborhood.statistics.gov.uk/dissemination/Info.do?page=userguide/moreaboutareas/furtherareas/further-areas.htm.

[26] The IMD is computed by the government for each lower layer super output area (LSOA) in the UK. We aggregate from the LSOA to the MSOA. Our data are from England only, which on average is richer than the rest of the UK. Effectively in England the IMD varies over a small range, with most of the sample lying between 0.04 and 0.31 (the 10th and 90th percentile). For more information, see http://www.communities.gov.uk/communities/research/indicesdeprivation/deprivation10/.

[27] See http://www.dh.gov.uk/en/\&Publicationsandstatistics/Publications/PublicationsStatistics/DH_094013 for more details.

of co-morbidities they report and there are only a few cases with a co-morbidity count larger than 6.[28]

Table 3 presents descriptive statistics on the patient characteristics described above. As can be seen, most patients are male and over 60 years of age. There is also considerable variation in patients' general health status, with a large fraction of patients for which several co-morbidities are reported. The mean value of the IMD is relatively low, although there is a fair amount of variation. The average probability of being informed about choice is about 50 percent, showing that not all physicians did offer choice as mandated by the reform. The IMD, the (capped) co-morbidity count and the probability that the patient was informed about choice are all used in the demand estimation to analyze how the reform differentially affected different groups of patients.

Table 4 contains descriptive statistics on the distances patients travel for their CABG treatments. We see that the average patient traveled a substantial distance (over 30 kilometers) and that there is a great deal of variation across patients in how far they traveled for care. It is also notable that there is very little difference in any of the descriptives for distance between the pre- and post-reform time periods. Patients didn't travel any farther post-reform on average, nor were they more likely to bypass the closest hospital. This could occur if patients sorted themselves to better hospitals post-reform within approximately the same distance. In the next section we provide some reduced-form evidence that this is the most likely explanation for the lack of a change in distance traveled.

Finally, we present some preliminary evidence that patients/physicians select hospitals based on patient observed health status. Table 5 shows the results of a regression of hospitals' case-mix adjusted mortality rates on patients' co-morbidity counts. Patients with a higher count, i.e. sicker patients, are significantly more likely to visit a hospital with a lower mortality rate. This is very precisely estimated. The effects are also economically important: on average, a patient with one additional co-morbidity goes to a hospital with a mortality rate about 0.2 percentage points lower than the mean (1.9 percent).We think that the presence of selection based on observable characteristics indicates that selection on *unobservables* is likely to also be an important factor in our data. The latter is of importance to our demand model and the way we measure quality in the estimation.

5 Reduced-Form Evidence

Before proceeding to our formal analysis, we look at patterns in the data to provide some simple empirical evidence on whether patients became more responsive to the hospital mortality rates after the reform. We start by running a simple linear regression of aggregate market shares on the case-mix adjusted mortality rate. We aggregate the patient-level data to the hospital-quarter level[29] and estimate separate OLS regressions for the

[28] Up to 14 co-morbidities fields are available in the HES data set. We also examined the Charlson index, which is a measure of morbidity for heart treatment. But since there is little variation in this variable, many patients having a value of zero, we do not use it in estimation.

[29] This allows us to illustrate some of the main patterns in the data in a simple OLS setup but we lose data on patient distance to the hospital in this aggregation.

pre- and post-reform time periods. This allows us to examine the impact of the introduction of choice on the responsiveness of market shares to the case-mix adjusted mortality rate. The results are reported in Table 6 in columns (1) and (2). These show that, pre-reform, higher quality hospitals did not have significantly larger market shares. In the post-reform period, however, a lower mortality rate is significantly associated with a higher market share. We replicate these regressions using (separate pre- and post-reform) hospital fixed effects in columns (5) and (6) and find the same qualitative results. This provides some initial suggestive evidence that the elasticity of demand with respect to quality rose due to the introduction of choice.

It is possible that this relation has nothing to do with choice but is an artifact of the distribution of market shares and mortality rates, which are unrelated to the introduction of patient choice. We test this by implementing a placebo test in which we replicate the same regression as in columns (1) and (2) but use emergency CABG cases instead of elective ones. Choice does not play a role for emergency admissions – patients are simply taken to the nearest suitable facility. Therefore, if we see a change in the correlation of market shares with mortality for emergency admissions it should not be due to the reform. Examining the results in Table 6, columns (3) and (4), we see that hospital mortality rates have no statistically significant impacts on emergency CABG market shares either pre- or post-reform. We replicate the analysis with a full set of hospital fixed effects and our results are robust to this (columns (7) and (8)).

An alternative way of analyzing the issue of an increased sensitivity of demand is to look directly at the expected (hospital-level) mortality rate that the average patient faces at the hospital of his choice. In Table 7 we report the average mortality rate a patient faces pre- and post-reform (the top row in each of the two panels). We find a substantial fall (about 20 percent for the raw mortality rate and about 13 percent for the case-mix adjusted mortality rate). This fall might occur for a number of reasons. For example, it could be due to a secular downward trend in the mortality rate across all hospitals, to hospitals in high population areas improving more (so more patients are treated at better facilities without necessarily exercising choice), or to patients deliberately choosing higher quality hospitals. To try to identify the impact of choice we report the change in the mortality rate separately for patients that visited the nearest hospital and patients that bypassed the nearest hospital and traveled further. If the drop in average mortality is simply due to an overall downward secular trend, we should not see differences in mortality between patients who visited the nearest hospital and those who bypassed it. Similarly, if the decrease in mortality is due to the fact that patients simply had better hospitals closer by after the reform, we should see most of the drop explained by the group of patients that visited the nearest hospital. Examining the patterns in Table 7 we find that the opposite is true. The drop in mortality among patients bypassing the nearest hospital is more than twice as large as the drop for patients that visit the nearest hospital. This is true both for raw and case-mix adjusted mortality rates. In other words, we observe larger declines in mortality for patients that decide not to use their local hospital. This supports the idea that these patients were

not simply lucky that the local hospital improved its quality but, rather, that they started to deliberately seek out better hospitals once they were allowed a choice of provider.

These patterns in the data provide some initial evidence suggesting that the introduction of patient choice via the reform increased the responsiveness of demand to cross-hospital differences in quality. We now analyze patient choice in a structural demand framework.

6 Estimation Methods

To estimate both the choice model and the production function for quality of care we proceed sequentially. We first estimate the production function relationship and then use the recovered quality measure in the demand estimation. We relegate most of the description of how we recover the case-mix adjusted mortality rate to appendix A and focus here on our main object of interest, the choice model.

6.1 Computing a Case-Mix Adjusted Mortality Measure: Estimation of the Production Function for Clinical Quality

In order to estimate the mortality equation (6) we need to instrument the hospital fixed effects, which we do using distance to the hospital (D_{ij}). This regression provides us with fitted values of the hospital-quarter dummies, $\widehat{\psi}_{jt}$, which can be interpreted as the causal effect of visiting a particular hospital on the probability of death.

$$M = JT\psi + H^{obs}\gamma_{obs} + H^{unobs}\gamma_{unobs} + \eta$$

Appendix A provides the details on the methods and empirical results, including tests of the validity of our IV strategy. These give us confidence that we are able to recover an appropriate measure of quality of service and that controlling for case-mix matters. In what follows, the fitted values of the hospital-quarter dummies $\widehat{\psi}_{jt}$ are used as the quality of care measure in the utility function, i.e. we set $Z_{jt} = \widehat{\psi}_{jt}$ in equation (1).

6.2 Estimating the Choice Model

Having described the methods for estimating our empirical quality of care measure, the adjusted mortality rate, we now turn to the estimation of the demand model. The primary concern that drives our empirical strategy is the possible endogeneity of waiting times in the utility function (1) due to correlation with unobserved quality (ξ_{jt}). To deal with this concern, we control for the effect of unobserved quality by estimating a separate hospital fixed effect for every quarter. This greater degree of flexibility is costly for us in two ways: 1) some parameters cannot be identified when a full set of hospital-quarter fixed effects are included, and 2) we have to estimate a set of almost 300 fixed effects (for about 30 hospitals and 10 quarters) within a non-linear model, which leads to

a considerable increase in the computational burden. To deal with the first issue, we employ a 2-step estimation approach that allows us to recover further parameters in a second step (e.g., Goolsbee and Petrin 2004). To deal with the latter concern we leverage a "BLP-style" contraction mapping ("BLP" denotes Berry, Levinsohn, and Pakes 1995) that allows us to concentrate out the vector of fixed effects. We therefore do not have to engage in a computationally difficult non-linear search over the set of fixed effects. We elaborate on both issues in what follows.

6.2.1 First Step: Estimating Effects with Heterogeneity and Fixed Effects

In the first step we estimate patient-specific parameters as well as a full set of hospital-quarter fixed effects based on equation (4),

$$\begin{aligned} u_{ijt} &= \delta_{jt} \\ &\quad +[\beta_{w,0} X_i \cdot \mathbf{1}(t=0) + \beta_{w,1} X_i \cdot \mathbf{1}(t=1)] \cdot W_{jt} \\ &\quad +[\sigma_{w,0} v_{w,i} \cdot \mathbf{1}(t=0) + \sigma_{w,1} v_{w,i} \cdot \mathbf{1}(t=1)] \cdot W_{jt} \\ &\quad +[\beta_{z,0} X_i \cdot \mathbf{1}(t=0) + \beta_{z,1} X_i \cdot \mathbf{1}(t=1)] \cdot Z_{jt} \\ &\quad +[\sigma_{z,0} v_{z,i} \cdot \mathbf{1}(t=0) + \sigma_{z,1} v_{z,i} \cdot \mathbf{1}(t=1)] \cdot Z_{jt} \\ &\quad +f(D_{ij}) + \varepsilon_{ijt} \end{aligned}$$

We allow distance to enter non-linearly into the utility function in a flexible way,

$$\begin{aligned} f(D_{ij}) &= \alpha_{d1} D_{ij} + \alpha_{d2} (D_{ij})^2 + \alpha_{d3} (D_{ij})^3 \\ &\quad +\alpha_{d4} Closest_{ij} + \alpha_{d5} Closest10k_{ij} \\ &\quad +\alpha_{d6} Closest20k_{ij}, \end{aligned}$$

where $Closest_{ij}$ is a dummy equal to one if hospital j is the closest one in the choice set of patient i. The variable $Closest10k_{ij}$ is a dummy equal to one if the hospital is either the closest one or not more than 10 kilometers further than the closest one. $Closest20k_{ij}$ is defined in a similar way for a 20 kilometer radius. In what follows, when we write out the utility function we continue to use $f(D_{ij})$ to economize on notation.

After controlling for δ_{jt}, there is still variation left to estimate the β parameters, which utilize variation across patients. Similarly, distance depends on the location of both the patient and the hospital, so the parameters of the distance function $f(\cdot)$ can be recovered. However, at this point we cannot identify the effect of hospital waiting times and quality on utility for the average patient: we only know whether certain types of patients are more sensitive to waiting times or quality than the average patient and this interaction effect is relative to a yet

unknown average effect.[30]

Because the set of fixed effects in δ is very large, searching over them non-linearly would be very computationally burdensome. However, for a given guess of the distance parameters, there is a unique vector δ that matches the predicted aggregate market shares to the market shares in the data. This is essentially the insight provided in BLP. We therefore use their contraction mapping on the aggregate market shares in order to recover the vector δ.

Since we are using patient-level data, we first need to aggregate the data at the hospital-quarter level before applying the contraction mapping. We form the hospital-quarter market shares by aggregating over the patient population within each quarter,

$$\widehat{s}^0_{jt} = \frac{1}{N_t} \sum_{i \in I_t} Pr_{it}(j|\widehat{\beta}_{w,it}, \widehat{\beta}_{z,it}, \widehat{\alpha}_d, \delta^0) \tag{7}$$

with

$$Pr_{it}(j|\widehat{\beta}_{w,it}, \widehat{\beta}_{z,it}, \widehat{\alpha}_d, \delta^0) = \int_{A_{ij}} dF(v_w, v_z, \varepsilon_{it}) \tag{8}$$

where $A_{ij} = (v_w, v_z, \varepsilon_{it} | u_{ijt} > u_{ikt} \forall k \neq j)$ and $F(v_w, v_z, \varepsilon_{it})$ is the joint density of the random coefficient draws (v_w, v_z), which are assumed to be independent standard normal variables, as well as the iid extreme value shocks ε_{it}.

$Pr_{it}(j|\cdot)$ is the probability that patient i chooses hospital j in period t. It is computed by integrating over the distribution of the random coefficient and the shocks ε_{it}. The integration over the two dimensions of unobserved preference heterogeneity is implemented using a 2-dimensional Gauss-Hermite quadrature with 9 nodes. $\widehat{s}^0_{jt}$ is the predicted market share for hospital j in period t, given the initial guess for the fixed effects vector δ^0. N_t denotes the number of patients visiting any hospital in a given time period and I_t denotes the set of patients visiting any hospital in a given time period. $\widehat{\alpha}_d$ denotes the current guess of the parameters entering $f(D_{ij})$, and similarly $\widehat{\beta}_{w,it}$ and $\widehat{\beta}_{z,it}$ denote the current guesses of the interactions of waiting time and mortality with patient demographics (including the unobserved heterogeneity terms). In other words, we are averaging over individual choice probabilities for a given guess of the patient-specific parameters.

We can now employ the contraction mapping

$$\delta^{k+1} = \delta^k + [\log(s) - \log(\widehat{s}^k(\widehat{\beta}_{w,it}, \widehat{\beta}_{z,it}, \widehat{\alpha}_d, \delta^k))]$$

[30] This is particularly relevant when assessing the effect of the reform on different demographic groups. This change will be composed of a change in the average effect and a change in the interaction effect. Knowing only the latter does not allow us to make any statement about the direction of the impact of the reform on a particular patient group. We deal with this below.

where k denotes the k^{th} iteration of the contraction mapping, s is the observed market-share, and $\widehat{s}$ denotes the predicted market-share conditional on the parameters of the model. This is iterated until convergence.

The likelihood function to be maximized is given by

$$L = \prod_{it} \prod_{j} [Pr_{it}(j)]^{y_{ijt}}$$

where y_{ijt}, with $(j \in J)$, is a variable that takes the value one for the decision actually taken in a particular period and zero otherwise.[31]

6.2.2 Second Step: Estimating Average Effects

The second step is the estimation of the average impact of hospital characteristics on patient utility. When estimating a full set of hospital-quarter fixed effects, all the variation at the hospital-quarter level is absorbed in the estimated fixed effects. Therefore, we project the fixed effects on hospital characteristics in this second step. Based on our utility specification we can write δ_{jt} as a function of hospital characteristics (equation (5))

$$\begin{aligned} \delta_{jt} = & [\overline{\beta}_{w,0} * \mathbf{1}(t=0) + \overline{\beta}_{w,1} * \mathbf{1}(t=1)] * W_{jt} \\ & + [\overline{\beta}_{z,0} * \mathbf{1}(t=0) + \overline{\beta}_{z,1} * \mathbf{1}(t=1)] * Z_{jt} \\ & + \xi_{jt} \end{aligned}$$

This is a linear relationship between δ_{jt} and the hospital characteristics. In contrast to the first step, where the fixed effects are time-period specific, in this baseline specification we control for a set of time invariant hospital fixed effects. Formally, we estimate

$$\begin{aligned} \delta_{jt} = & [\overline{\beta}_{w,0} * \mathbf{1}(t=0) + \overline{\beta}_{w,1} * \mathbf{1}(t=1)] * W_{jt} \\ & + [\overline{\beta}_{z,0} * \mathbf{1}(t=0) + \overline{\beta}_{z,1} * \mathbf{1}(t=1)] * Z_{jt} \\ & + \overline{\xi}_{j} + \widetilde{\xi}_{jt}, \end{aligned}$$

where $\overline{\xi}_j$ is the fixed effect for hospital j and $\widetilde{\xi}_{jt}$ is the quarter-specific deviation from the average unobserved hospital quality. The latter is the econometric error term and we assume that it is not correlated with any of the observed hospital characteristics.[32]

[31] For each individual i that visits a hospital in time period t we can compute the theoretical probabilities for each option j in the choice set. Multiple CABGs for the same patient are extremely rare, therefore we treat all observations as independent. In other word, there is a unique t associated to every patient i. With some abuse of notation, multiplying over (it), rather then separately over i and t, represents this feature.

[32] Since our model does not include an outside option, in the first step we can only estimate $(k_t - 1)$ hospital fixed effects in each quarter, where k_t denotes the quarter-specific number of hospitals in the market. We therefore do not estimate the quarter-specific fixed effect for one of the hospitals that is present throughout the whole sample period. To implement the second step we compute

To address the concern that unobserved hospital quality changed after the introduction of choice, we allow for a different set of fixed effects pre- and post-reform in a sensitivity check (Table 9).[33] The robustness of our estimates to using separate pre- and post-reform dummies is reassuring and we use time invariant hospital fixed effects in our baseline as this facilitates the comparison of demand elasticities and market shares pre- and post-reform.

7 Estimation Results

In this section we first report the estimation results from the choice model (production function results are in Appendix section A.3), then report how they translate into patient and hospital level elasticities of demand.

7.1 Choice Model Estimates

7.1.1 First Step: Effects with Heterogeneity

The results from the first step of estimation are reported in Table 8. For economy of exposition, the (large number of) fixed effect estimates are not reported in the table. Also for the purpose of exposition, all patient characteristic variables were standardized in order to make the coefficient magnitudes more easily comparable. Further, due to the standardization, the interactions can be interpreted as deviations in preferences relative to a patient with average characteristics.

Examination of the estimates reveals some dramatic results. There are significant changes in almost all of the interactions between the pre- and post-reform time periods. Poorer patients (higher values of IMD) pay more attention to waiting times after the introduction of choice.[34] The interaction term for waiting times post-reform is significantly different from zero (and from the pre-reform coefficient). More severely ill patients (more co-morbidities) care significantly less about waiting times and significantly more about quality in both time periods. In other words, sicker patients are willing to trade off higher waiting times for a better quality of service, as one would expect. Relatively speaking, they become more sensitive to both dimensions of hospital service post-reform. More informed patients are less sensitive to both measures pre-reform, but become relatively more sensitive post-reform. Post-reform they are significantly more sensitive to waiting times than the average patient but not significantly different from the average patient with respect to quality. The change in sensitivity, however, is statistically significant in the case of both dimensions of service.

the hospital characteristics as the difference relative to the hospital for which we do not have a fixed effect. In other words, we subtract the waiting time (mortality) of this particular hospital from the waiting times (mortalities) of all the other hospitals for each of the quarters. After this normalization we run the linear regression outlined above.

[33] This sensitivity check (and our baseline) control less conservatively for unobserved quality than the first step of our estimation. However, we think that the most important change in hospital quality over time will happen with the relaxation of choice constraints due to the reform.

[34] High waiting times are a negative, so a negative coefficient indicates that patients are responsive to waiting times. Similarly, our measure of quality is mortality, so a negative coefficient indicates greater responsive by avoiding low quality (high mortality) hospitals.

In the case of patient informedness the change is the most relevant measure. The variable is only defined in a meaningful way for the post-reform time period, since it refers to information about something that did not exist pre-reform (being informed about the possibility of choice). We include it in the pre-reform period to strengthen the identification of the effect of information. The pre-reform interaction effects with informedness should capture any factors (e.g., unmeasured population characteristics) that are correlated with informedness across PCTs. This difference-in-difference identification is based on the assumption that the only thing changing with the introduction of the reform is the provision of information, i.e., any other factors that might be cross-sectionally correlated with informedness do not change when patient choice is expanded.

We also find evidence of substantial unobserved heterogeneity in preferences over the adjusted mortality rate, both pre- and post-reform. This unobserved heterogeneity is essentially unchanged pre- to post-reform. Our results show little heterogeneity in preferences over waiting times in both time periods. Finally, we also find that, as expected, distance plays an important role in hospital choice. The effects are clearly nonlinear. Four of our six distance terms are highly significant, showing the importance of controlling flexibly for the effects of distance on choice.

Overall, Table 8 provides evidence of substantial patient heterogeneity in response to the reform. These findings regarding the heterogeneity of the treatment effect are highly relevant for policy purposes. We elaborate more on the policy implications in Section 8 below.

7.1.2 Second Step: Average Effects

The results of the linear regression for the 2nd step of estimating the choice model (obtaining average effects) are reported in Table 9. We find substantial impacts of the reform. In particular, patients became substantially more sensitive to the adjusted mortality rate post-reform. The estimated effect is not significantly different from zero pre-reform and becomes significantly negative post-reform, and approximately twice as large. This provides evidence on the primary question of this paper. The choice reform did have a substantial impact on patients' sensitivity to hospitals' clinical quality of care, and in the expected direction – patients became much more responsive to how well hospitals performed in terms of their mortality rate for CABG surgeries.

The estimates with respect to waiting times are not significant either pre- or post-reform. We therefore conclude that there was no substantial change in the sensitivity with respect to waiting times for the average patient. However, as shown above, some groups of patients did change their behavior with respect to waiting times after the introduction.

We also relax the assumption that hospital fixed effects are constant and allow them to differ pre- and post-reform. These estimates are displayed in the second set of columns in the table. As can be seen, the estimates and conclusions are essentially unchanged – patients are substantially more sensitive to the mortality

rate post-reform.

7.2 Elasticities of Demand

Since the primary focus on the paper is on the quality of care, and we find only weak (mostly insignificant) results for sensitivity with respect to waiting times, we focus on the elasticity of demand with respect to the (case-mix adjusted) mortality rate. We start by computing elasticities for individual patients with respect to the adjusted mortality rate. Analyzing individual-level elasticities is helpful in our context to get a better sense of how strongly different patient groups were affected by the relaxation of the constraint on choice. Secondly, we compute aggregate demand elasticities for hospitals by integrating over the changes in individual choice probabilities for each hospital. Ultimately, for assessing the impact of the choice reform, these hospital-level elasticities are the most crucial factor. If the demand that hospitals face becomes more elastic with regard to quality, then relaxing the constraint on choice was successful in increasing hospitals' incentives to provide higher quality.

7.2.1 Patient-level Elasticities

Table 10 reports the sensitivity of choice probabilities to changes in the mortality rate for different groups of patients. The first column reports the percentage change in the choice probability following an increase in the mortality rate by one standard deviation (about 1.11 percentage points). The second column reports the elasticity of individual-level demand. In order to compute both columns we fix the baseline choice probability of the patient to be equal to 50 percent.[35] A different choice probability would equally scale all values reported in the table, so any relative comparison between different rows in the table is not affected by this assumption. To assess the impact on different income groups, we report the elasticity for a patient whose income is 1 standard deviation lower than the average. Similarly, we report results for more severely ill patients as well as more informed patients.

We find that with restrictions on choice (prior to the reform) a 1 standard deviation increase in mortality leads to about a 3 percent drop in the choice probability of the average patient. After the relaxation of choice constraints (post-reform), we see an effect that is more than twice as large, with a drop of about 7 percent in the choice probability. Interestingly, we find no real difference in how lower income patients reacted to increased choice. There are very small differences between their responses to quality and those of the average patient, both with restricted and free choice (pre- and post-reform). The result is intriguing as there was concern that the choice reform would not benefit lower income households or that they would be harmed by it.

Sicker patients reacted more strongly to choice than the average patient. This is intuitive as these patients have a stronger incentive to select a high quality hospital. Note that even pre-reform choices were relatively

[35] Specifically, we compute the first column as $[100 * \beta_x * Std_x * (1 - Pr)]$. The elasticity is computed as $[\beta_x * x * (1 - Pr)]$.

more responsive for sicker patients. This is likely to be because physicians pay more attention to quality when referring a sicker patient, though it is possible that sicker patients gathered more information and managed to influence their GP to refer them to a better hospital, even before the introduction of patient choice. More informed patients are responsive to hospitals' mortality rates post-reform. The change in responsiveness pre- to post-reform is quite a bit larger than for other groups. (As discussed previously, we focus on the change in elasticity for this group rather than the pre- and post-reform levels.)

Finally, we also report the elasticities for patient-level demand in the second column of the table. While across the board the elasticities are fairly small, they are substantially larger with free choice: the elasticity for the average patient more than doubles pre- to post-reform, while it is substantially larger for sicker and better informed patients.

7.2.2 Hospital-level Elasticities

We examine hospital-level demand sensitivity by calculating market share impacts and elasticities. We simulate a one standard deviation change in mortality for each hospital in the choice set and compute the percentage change in the hospital's market share. For the elasticity we divide by the percentage change in the mortality rate.[36] The market share and elasticity effects will differ across hospitals depending on the density of patients in the local area, the demographic composition of the local population, and the locations of other hospitals.

Table 11 reports the distribution of elasticities across all hospitals. We find that when constraints on choice are relaxed post-reform a one standard deviation increase in the mortality rate leads to a 4.9 percent drop in market share for the average hospital, relative to a much smaller decrease of 0.36 percent before the reform. There is substantial heterogeneity in the impacts across hospitals. The market share loss is 6.25 at the 25th percentile and 2.3 at the 75th percentile of the distribution of elasticity changes.[37] We also compute demand elasticities across hospitals and find the elasticities are generally low, but change substantially due to the freeing up of choice. The elasticity at the average hospital increases from 0.02 (i.e. close to zero) to -0.12 pre- to post-reform.[38]

Overall the results suggest that relaxing the constraint on choice increased hospitals' incentives to improve quality. While the effect is not uniformly large across all hospitals, many hospitals experienced substantial changes in the demand elasticities they faced.

[36] More specifically, we simulate a small change in the mortality rate when computing the elasticity rather than a one standard deviation shift.

[37] Note that we report percentiles for the distribution of *changes*, rather than changes for hospitals at certain percentile in the original distribution of *levels*. As hospitals change their position in the level-ranking with the introduction of the reform, the two things are not the same.

[38] We have to make some modification to the data in order to compute hospital-level elasticities due to the fact that the adjusted mortality rate sometimes takes on negative values. We adjust the distribution of the adjusted mortality rate (by shifting the mean and the variance) so that its minimum and maximum are the same as the minimum and maximum values of the raw mortality rate distribution.

8 Policy Evaluation

We provide an evaluation of the impacts of allowing free choice in several steps. We first estimate the number of lives that were saved by allocating patients to better hospitals post-reform, as well as provide a broader analysis of consumer welfare gains due to the relaxation of choice constraints. Both these calculations evaluate effects of the reform under the assumption that hospitals did not react to the change in demand conditions. The survival and welfare effects we present are purely due to an efficient resorting of patients. We then proceed to an analysis of how much the competitive environment changed with the introduction of the reform. Finally, we provide some suggestive supply-side evidence which shows that hospitals seem to have reacted to the change in demand conditions as intended by policy makers. In all but the last step, we simulate counterfactuals in order to quantify the impact of increased choice. However, contrary to many other applications, we do not simulate changes caused by a hypothetical policy, but rather simulate behavior for the post-reform population under the assumption that the reform had not taken place. In this way we leverage the structure of our model to evaluate and quantify the effects of the policy change.

8.1 The Impact of Choice on Patient Survival

An obvious, and very direct, measure by which to evaluate this policy is the impact on the probability of survival following a CABG. This is of direct importance to patients. We assess this by calculating how many more patients would have died had the reform not been implemented, i.e., if patients in the post-reform time period were still subject to pre-reform choice constraints and therefore choosing according to pre-reform parameters.

Formally, we implement the analysis in the following way. The ex-ante mortality probability[39] of any particular patient i in time-period t is given by

$$Pr_{it}(Mortality, \theta_t) = \sum_j Pr_{it}(j, \theta_t) \cdot E(Mortality | choice = j, PatientCharacteristics)$$

Both terms on the right-hand side of the above equation can be computed based on the estimated parameters. The first term denotes the probability of visiting hospital j, which is estimated within the demand model. The second is given by the estimated probability of death conditional on the observed set of patient characteristics and the choice of a particular hospital. Specifically, equation (6), which we used to compute the case-mix adjusted mortality rate, yields the expected probability of dying.

[39] In this context we think of the ex-ante probability as the probability of death before both the error terms of the choice process and the error term influencing survival are realized, i.e. the patient has not decided which hospital to visit and we do not know the patient-specific shock to his mortality probability yet.

To obtain the expected difference in mortality we simply compute

$$E(\Delta Mortality) = \sum_{it} [Pr_{it}(Mortality, \theta_{post}) - Pr_{it}(Mortality, \theta_{pre})] \cdot \mathbf{1}(t = \text{post-reform})$$

In other words, we sum up the changes in mortality probability for each patient in the post-reform period when choice parameters change from post- to pre-reform estimates. Results from this counterfactual are reported in Table 12. We estimate that 12 fewer patients would have survived had the reform not been implemented in 2005. This number is calculated is over the entire five post-reform quarters used in the estimation (so corresponds to roughly 10 lives saved on a annual basis). The lower panels of the table assess the magnitudes of these changes relative to the total number of admissions and deaths during the relevant time period. The changes amount to about 0.06 percentage points or a 3.1 percent decrease in the mortality rate. If we adopt the $100,000 benchmark of Cutler and McClellan (2001) for the value of a year of life, and assume that CABG survivors' lives are extended by 17 years (van Domburg, Kappetein, and Bogers 2009), the beneficial effects of the pro-competition reforms are about $17 million yearly in terms of value of life-years saved.[40]

8.2 Changes in Patient Welfare

In a similar spirit to the analysis of patient survival, we also compute the percentage welfare changes due to removing restrictions on choice. In order to undertake a welfare calculation we assume that preferences are identified from the data post-reform. Pre-reform choices are constrained and the parameters we estimate are therefore the reduced form of preferences and the constraints. Using these assumptions (which play no role in our estimation) we can assess how much welfare was gained by freeing patient choice from the pre-reform constraints. We compute the change in consumer surplus in the following way

$$E(\%\Delta CS) = \left[\frac{Pr_{it}(j, \theta_1) \cdot u_{it}(j, \theta_1) - Pr_{it}(j, \theta_0) \cdot u_{it}(j, \theta_1)}{Pr_{it}(j, \theta_0) \cdot u_{it}(j, \theta_1)} \right] \cdot \mathbf{1}(t = \text{post-reform})$$

In other words, we compare the utility post-reform patients with choice actually received with the utility they would have received had they chosen according to pre-reform restricted choice parameters but if their utility was based on the "true" post-reform preferences. The latter would have been the situation patients would have found themselves in had the reform not been implemented. We simulate over both the random coefficients and the logit error terms in order to compute the expression above. As the welfare levels are expressed in normalized utils, only the percentage change in utils is an interpretable measure.

We find that the freeing of choice led to a 7.68 percent increase in welfare. This is a substantial, although not

[40] $10 \times 17 \times 100,000 = 17,000,000$

overwhelming, increase. Note that both the welfare analysis and the change in survival only assess the changes that are achieved by reallocating patients, i.e. without any supply side adjustment of hospitals to the new demand conditions. We now turn to the further improvements that can be achieved if the reform also provided incentives for hospitals to improve quality.

8.3 Change in the Competitive Environment

In this section we perform a counterfactual in order to get a sense of the magnitude of the relaxation of choice constraints on hospitals' incentives. We look at how hospital market shares in the pre-reform period would have been different if patient choices had occurred based on estimated post-reform parameters. This exercise allows us to compute how much re-shuffling of market shares would have happened had patients had free choice earlier. When implementing the counterfactual, we hold everything fixed except for the choice parameters. In other words, the same set of patients is exposed to the same set of hospitals as in the "real" pre-reform choice situation. For the purpose of this counterfactual we therefore do not allow hospitals to adjust to the changes in demand caused by the parameter change. The magnitude of the re-shuffling of market shares gives us a sense of how much incentives to improve quality changed for hospitals.

The results from this exercise are reported in Table 13. In line with the results presented above, we find that the introduction of choice has a significant impact on some hospitals. At the extremes of the distribution one hospital would have lost 8.7 percent of its market share and another would have gained 14.7 percent. There is a large heterogeneity in the impacts, however, and for most hospitals the effect is more modest.

8.4 Some Assessment of the Supply-Side Response

Having established that the introduction of choice led to a substantial increase in demand elasticities faced by hospitals, we now provide some evidence for a supply-side response to this change in the competitive environment. In order to analyze the supply-side in a simple way, we rely on the fact that the reform differentially affected hospitals due to differences in population density, population demographics, and the location of competitors. We therefore expect hospitals in areas where demand conditions changed more to improve their quality more than other hospitals.

We test this hypothesis by regressing the change in the case-mix adjusted mortality rate on the change in the aggregate elasticity of demand with respect to quality. This simple approach mirrors the difference-in-difference estimation conducted in Gaynor, Moreno-Serra, and Propper (2010) and Cooper, Gibbons, Jones, and McGuire (2011). In these papers, a change in the mortality rate is regressed on cross-sectional variation in hospital market structure. The argument is that the expansion of choice will have a stronger impact in areas with a higher density of competing hospitals. Using a measure of concentration, like the Herfindahl Index, is a reduced-form way of

capturing that the elasticity of demand is expected to change more in some areas than in others. Here we compute demand elasticities (and changes in elasticities due to the reform) based on model primitives. Because CABGs are not offered in all UK hospitals we have only 27 observations available for estimation. For that reason, and because of the ad hoc nature of the specification, we think of the results presented in this section as only suggestive and complementary to the evidence provided elsewhere.

We estimate the following OLS regression,

$$\Delta Mortality_j = \lambda_0 + \lambda_1 \Delta Elasticity_{j,Mortality} + e_j \tag{9}$$

where $Elasticity_{j,Mortality}$ denotes the elasticity of hospital demand with respect to the case-mix adjusted mortality rate, as reported in Table 11.[41] The results from this regression are reported in Table 14. For ease of interpretation, we use the absolute value of the elasticity in the regression.

We find a negative and significant impact of the change in the demand elasticity on the change in the case-mix adjusted mortality rate. In other words, hospitals whose demand became more responsive to quality improved quality disproportionately more than elsewhere. To get a sense of the magnitude of the coefficient, consider a shift of one standard deviation in the elasticity (roughly 6, as reported in Table 11). This shift implies a drop of 0.7 in the mortality rate, which is roughly equal to 75 percent of the general decrease in the case-mix adjusted mortality rate over the entire time period from 2003 to 2007 (see Table 1). This is a very substantial impact.

Overall, this suggests that freeing up patient choice elicited a supply side response by hospitals that improved patient survival. This impact goes beyond the effects presented in the previous analyses. In terms of magnitude, our point estimate implies that competition could have played a large role in the overall drop in the CABG mortality rate from 2003 through to 2007.

9 Summary and Conclusions

This paper takes advantage of a "natural experiment" in the English National Health Service which introduced patient choice among hospitals. This reform allows us to look at the effect of choice on patient behavior and supplier responses to that behavioral change. We evaluate whether increased choice resulted in increased elasticity of demand faced by hospitals with regard to two central dimensions of hospital service – clinical quality of care and waiting times. We estimate a structural model of patient demand, allowing for a different responsiveness pre- and post-reform in a flexible way. On the methodological side, we show how to deal with the endogeneity of waiting times in the demand model as well as how to obtain a valid quality measure using an IV strategy.

[41] In the regression we use (the change in) the percentage change in market share following a one standard deviation increase in mortality instead of the elasticity.

We find substantial impacts of the removal of restrictions on patient choice. Patients are more responsive to the clinical quality of care at hospitals (measured as the hospital's case-mix adjusted mortality rate) but on average are not more responsive to waiting times. These is heterogeneity in these impacts, however. More severely ill patients are more affected by the reform as are better informed patients. We calculate that the increased demand responsiveness alone led to a significant reduction in mortality and an increase in patient welfare. The elasticity of demand faced by hospitals increased post-reform. This gave hospitals (potentially) large incentives to improve their quality of care and we find suggestive evidence that hospitals responded strongly to the enhanced incentives due to increased demand elasticity.

Overall, this paper provides evidence that a reform that removed constraints on patient choice worked: patient flows were more sensitive to clinical quality and patients went to better hospitals. This suggests that there is potential for choice based reforms to succeed and for competition in health care to enhance quality.

References

Anderson, T. W. (1984): *An Introduction to Multivariate Statistical Analysis, 2nd ed.* Wiley, New York.

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.

Besley, T., and M. Ghatak (2003): "Incentives, Choice, and Accountability in the Provision of Public Services," *Oxford Review of Economic Policy*, 19(2), 235–249.

Blöchliger, H. (2008): "Market Mechanisms in Public Service Provision," *OECD Economics Department Working Papers, No. 626*, Organization for Economic Cooperation and Development, Paris, France.

Capps, C., D. Dranove, and M. Satterthwaite (2003): "Competition and Market Power in Option Demand Markets," *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 Patient Choice Reforms," *Economic Journal*, 121(554), F228–F260.

Cutler, D., and M. McClellan (2001): "Is Technological Change in Medicine Worth It?," *Health Affairs*, 20(5), 11–29.

Farrar, S., J. Sussex, D. Yi, M. Sutton, M. Chalkley, T. Scott, and A. Ma (2007): "National Evaluation of Payment by Results - Report to the Department of Health," Report, Health Economics Research Unit, University of Aberdeen, http://www.abdn.ac.uk/heru/documents/pbr_report_dec07.pdf (accessed April 27, 2010).

Gaynor, M., R. Moreno-Serra, and C. Propper (2010): "Death by Market Power: Reform, Competition and Patient Outcomes in the British National Health Service," unpublished manuscript, Carnegie Mellon University, Imperial College.

Gaynor, M., and W. B. Vogt (2003): "Competition Among Hospitals," *Rand Journal of Economics*, 34(4), 764–785.

Geweke, J., G. Gowrisankaran, and R. J. Town (2003): "Bayesian Inference for Hospital Quality in a Selection Model," *Econometrica*, 71(4), pp. 1215–1238.

Goolsbee, A., and A. Petrin (2004): "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 72(2), 351–381.

Gowrisankaran, G., and R. J. Town (1999): "Estimating the Quality of Care in Hospitals Using Instrumental Variables," *Journal of Health Economics*, 18, 747–767.

Hansen, L. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Ho, K. (2009): "Insurer-Provider Networks in the Medical Care Market," *American Economic Review*, 99(1), 393–430.

Howard, D. H. (2005): "Quality and Consumer Choice in Healthcare: Evidence from Kidney Transplantation," *Topics in Economic Analysis and Policy*, 5(1), Article 24, 1–20, http://www.bepress.com/ bejeap/topics/vol5/iss1/art24.

Hoxby, C. M. (2003): "School Choice and School Productivity: Could School Choice Be a Tide that Lifts All Boats?," in *The Economics of School Choice*, ed. by C. M. Hoxby, pp. 287–342. National Bureau of Economic Research, Cambridge, MA.

Kessler, D. P., and M. B. McClellan (2000): "Is Hospital Competition Socially Wasteful?," *Quarterly Journal of Economics*, 155, 577–615.

Le Grand, J. (2003): *Motivation, Agency, and Public Policy: Of Knights & Knaves, Pawns & Queens*. Oxford University Press, Oxford, UK.

Luft, H. S., D. W. Garnick, D. H. Mark, D. J. Peltzman, C. S. Phibbs, E. Lichtenberg, and S. J. McPhee (1990): "Does Quality Influence Choice of Hospital?," *JAMA: The Journal of the American Medical Association*, 263(21), 2899–2906.

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 Hospital Discharge Survey (2010): *National Hospital Discharge Survey*. National Center for Health Statistics, Centers for Disease Control, Atlanta, GA.

Propper, C., M. Sutton, C. Whitnall, and F. Windmeijer (2008): "Did Targets and Terror Reduce Waiting Times in England for Hospital Care?," *The BE Journal of Economic Analysis & Policy*, 8(2), 5.

Sivey, P. (2008): "The Effect of Hospital Quality on Choice of Hospital for Elective Heart Operations in England," *Unpublished Manuscript, University of Melbourne*.

Tay, A. (2003): "Assessing Competition in Hospital Care Markets: the Importance of Accounting for Quality Differentiation," *RAND Journal of Economics*, 34(4), 786–814.

van Domburg, R. T., A. P. Kappetein, and A. J. J. C. Bogers (2009): "The Clinical Outcome After Coronary Bypass Surgery: A 30-year Follow-up Study," *European Heart Journal*, 30, 453–458.

Varkevisser, M., S. A. van der Geest, and F. T. Schut (2012): "Do patients choose hospitals with high quality ratings? Empirical evidence from the market for angioplasty in the Netherlands," *Journal of Health Economics*, 31(2), 371 – 378.

Table 1: Descriptive Statistics — Hospital Characteristics[41]

	Total Admissions CABGs		Waiting Times (Days)	
	Mean	Std	Mean	Std
2003	502.9	189.4	109.1	32.1
2004	507.5	200.0	100.5	20.7
2005	449.1	170.8	67.8	15.2
2006	425.4	172.7	65.6	17.3
2007	459.9	169.9	64.9	21.4

	Mortality Rate CABGs		Adjusted Mortality Rate, CABGs	
	Mean	Std	Mean	Std
2003	1.88	0.82	1.67	1.39
2004	1.93	0.78	1.46	1.45
2005	1.90	0.57	1.19	1.14
2006	1.95	0.79	1.40	1.18
2007	1.51	0.69	0.73	0.90

Table 2: Descriptive Statistics — Mortality Rate and Waiting Times at the Quarter Level

	Waiting Times (Days)		Adjusted Mortality Rate, CABGs	
	Mean	S.D.	Mean	S.D.
2004q1	113.7	36.3	1.46	1.71
2004q2	106.1	26.8	1.76	3.12
2004q3	102.5	26.6	1.20	1.96
2004q4	100.5	23.6	1.74	2.04
2005q1	93.4	21.6	1.14	1.66
2007q1	66.7	19.0	1.50	2.29
2007q2	66.2	18.5	0.55	1.25
2007q3	65.3	22.7	0.74	1.35
2007q4	63.9	23.9	0.66	0.99
2008q1	66.1	23.3	0.94	2.18

Table 3: Descriptive Statistics — Patient Characteristics

	Mean	Median	Standard Deviation	10th Percentile	90th Percentile
Age	65.76	66	55.04	53	76
Index of Multiple Deprivation	0.14	0.1	0.11	0.04	0.31
Comorbidity Count	5.42	5	2.81	2	9
Capped Comorbidity Count (Cap at 6 Comorbidities)	4.57	5	1.61	2	6
Probability of Informedness About Choice	0.53	0.53	0.07	0.45	0.63
Fraction Male	81.18%				

Table 4: Descriptive Statistics: Distance

		Mean	Median	Standard Deviation	10%	90%	95%
Distance	Pre	34.93	22.34	44.97	4.77	71.40	98.15
	Post	32.24	22.91	32.94	4.93	70.58	92.36
Fraction of Patients Visiting the Closest Hospital	Pre	68.14 %					
	Post	68.67 %					

Table 5: Descriptive Statistics: Hospital Selection Based on Observed Severity

Dependent Variable	Adjusted Mortality Rate	Adjusted Mortality Rate
Co-morbidity Count	-0.220**	-0.180**
	(0.006)	(0.006)
Quarter Fixed Effects (Flexible Time Trend)	No	Yes
Number of Observations	32,715	32,715

Table 6: Reduced-Form Evidence: Regressions using Aggregate Market-Shares

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	Elective CABGs Market-share		Emergency CABG Market-share		Elective CABGs Market-share		Emergency CABG Market-share	
Time Period	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Coefficient on Case-Mix Adjusted Mortality Rate	0.0042	-0.1652**	0.0488	-0.0667	0.0416	-0.0692*	-0.0373	-0.0127
	(0.0552)	(0.0622)	(0.0531)	(0.0744)	(0.0267)	(0.0321)	(0.0379)	(0.0466)
Hospital Fixed Effects	No	No	No	No	Yes	Yes	Yes	Yes
Number of Observations	142	143	142	143	142	143	142	143
Hospitals	29	29	29	29	29	29	29	29
Quarters	5	5	5	5	5	5	5	5

Table 7: Reduced-Form Evidence: Changes in the Expected Mortality Rate

	Sample	Mean Pre	Mean Post	Difference in Means
Mortality Rate (Raw Rate)	All Patients	1.344	0.948	-0.396
	Patients Visiting the Nearest Hospital	1.287	1.022	-0.265
	Patients Not Visiting the Nearest Hospital	1.462	0.779	-0.683
Mortality Rate (Case-Mix Adjusted)	All Patients	1.471	0.748	-0.723
	Patients Visiting the Nearest Hospital	1.352	0.809	-0.543
	Patients Not Visiting the Nearest Hospital	1.716	0.606	-1.110

Table 8: Regression Results from the First Step of the Estimation[1]

			Coeff.	S.E.	
Income	Waiting	Pre	0.01	0.68	
Deprivation	Times	Post	-3.85	0.65	**
Index	Mortality	Pre	0.12	0.86	
	Rate	Post	-0.02	1.13	
Co-	Waiting	Pre	5.67	0.43	**
Morbidity	Times	Post	4.10	0.58	**
Count	Mortality	Pre	-10.11	0.56	**
	Rate	Post	-13.18	0.94	**
Patient	Waiting	Pre	1.25	0.66	*
Informedness	Times	Post	-4.41	0.65	**
	Mortality	Pre	5.17	0.83	**
	Rate	Post	0.01	1.06	
Unobserved	Waiting	Pre	-0.22	75.36	
Preference	Times	Post	-0.26	76.70	
Heterogeneity	Mortality	Pre	35.02	0.87	**
	Rate	Post	39.04	1.81	**
Distance	Linear		-14.86	0.21	**
	Square		4.91	0.11	**
	Cube		-0.57	0.02	**
	Closest Dummy		1.07	0.02	**
	Closest "Plus 10" Dummy		-0.01	0.00	*
	Closest "Plus 20" Dummy		0.01	0.05	

[1]All the patient characteristics used in the regression are standardized in order to make the magnitudes of the coefficients comparable.

Table 9: Regression Results from the Second Step of the Estimation

			Baseline Specification			Sensitivity Check		
			Coeff.	S.E.		Coeff.	S.E.	
Average Effect	Waiting Times	Pre	-4.24	3.15		0.43	4.09	
		Post	6.25	4.74		13.45	7.53	
	Quality	Pre	-4.85	3.70		-1.63	3.81	
		Post	-12.40	4.00	**	-11.39	3.96	**
Hospital Fixed Effects			Constant Fixed Effects Pre- and Post-Reform			Separate Fixed Effects Pre- and Post-Reform		

[1]The table reports the elasticity of demand at the patient-level with respect to the case-mix adjusted mortality rate. The values reported in the first column represent the percentage change in market-share when the hospital increases the mortality rate by one standard deviation. The second column reports the elasticity of demand. Specifically, we compute the first column as $[100 * \beta_x * Std_x * (1 - Pr)]$. The elasticity is computed as $[\beta_x * x * (1 - Pr)]$.

Table 10: Patient-Level Elasticities of Demand with Respect to Mortality[1]

	Impact on Patient's Purchase Probability From 1 S.D. Shift in Adjusted Mortality	Elasticity
Pre-Reform		
Average Patient	-2.69	-0.021
Lower Income	-2.63	-0.021
Higher Comorb	-8.30	-0.066
More Informed	0.18	0.001
Post-Reform		
Average Patient	-7.08	-0.056
Lower Income	-7.09	-0.056
Higher Comorb	-14.40	-0.114
More Informed	-7.08	-0.056

Table 11: Hospital-Level Responsiveness of Demand with Respect to Mortality[1]

1-S.D. Shift	Mean	S.D.	25th Perc.	Median	75th Perc.
Pre-Reform	-0.36	5.11	-1.73	0.11	0.56
Post-Reform	-4.83	4.73	-5.66	-3.14	-2.34
Change	-5.38	5.81	-6.25	-2.88	-2.28
Elasticities	Mean	S.D.	25th Perc.	Median	75th Perc.
Pre-Reform	0.02	0.16	-0.03	0.00	0.01
Post-Reform	-0.12	0.07	-0.16	-0.10	-0.05
Change	-0.14	0.19	-0.15	-0.07	-0.05

[1]The table reports the elasticity of aggregate demand at the hospital-level with respect to the case-mix adjusted mortality rate. The values reported represent the percentage change in market-share when the hospital increases the mortality rate by one standard deviation. The impact on market-shares is computed for each hospital individually. The table reports the distribution of changes within the choice-set of hospitals faced by patients. The lower panel reports elasticities across the hospitals in the choice-set. See text for more details on how the elasticities are computed.

Table 12: Changes in Survival Probability due to the Reform[1]

Change in Survival when Choices Post-Reform are Made with Pre-Reform Parameters		-12.17
Post-Reform (5 Quarters)	Admissions	20338
	Deaths	393
	Mortality Rate	1.93
	Recomputed Mortality Rate	1.87

[1]The table reports the change in the number of survivals for a counterfactual scenarios. The lower panel reports what change in the mortality rate is entailed by the changes in absolute numbers. For the post-reform period the number of admissions, deaths and the mortality rate are reported for the 5 month period used in the estimation (see text for details on the sample period used for estimation).

Table 13: Impact on Market Shares from an Earlier Adoption of the Reform[1]

Mean	S.D.	Min	25th Perc.	Median	75th Perc.	Max
0.25	4.56	-8.68	-1.25	-0.30	0.44	14.66

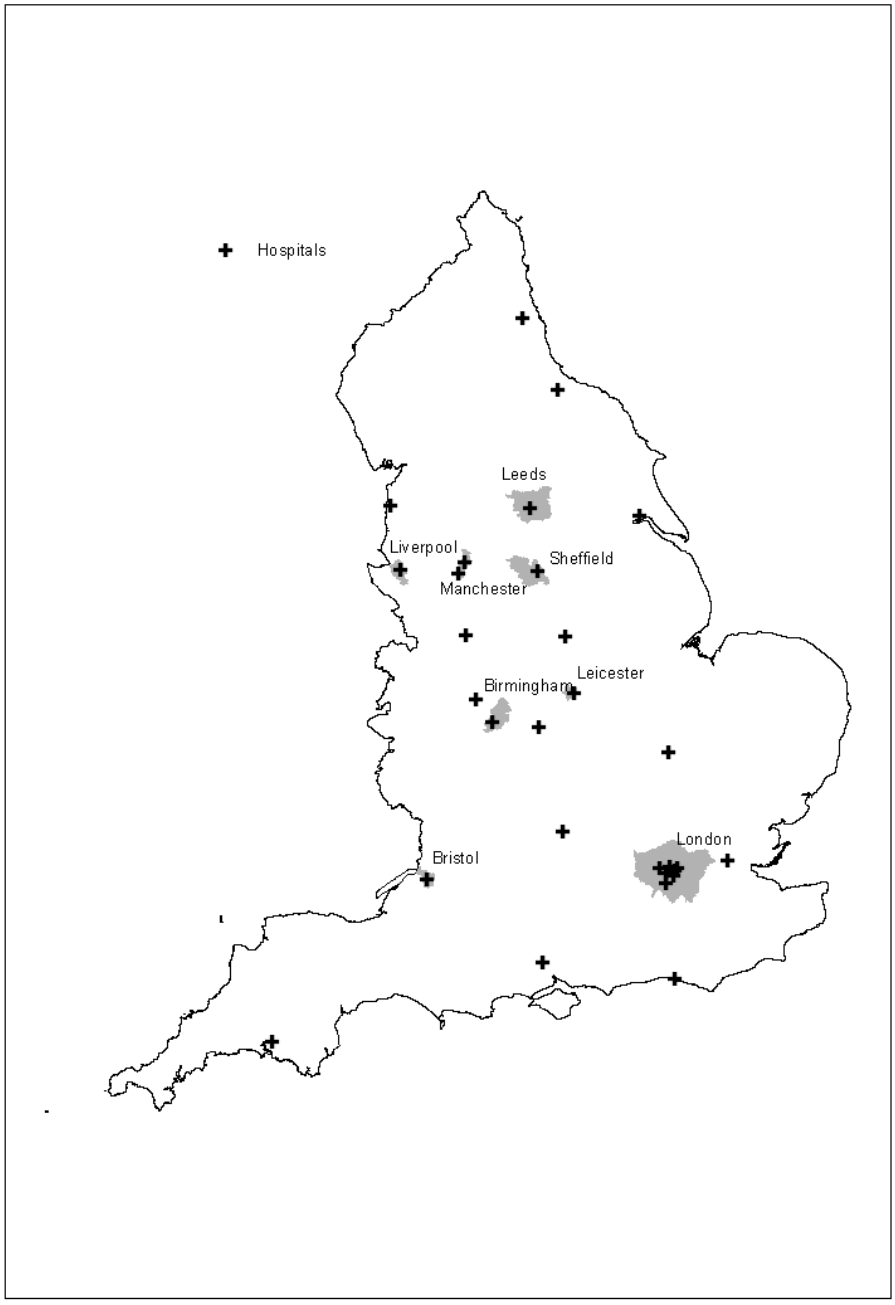
[1]The table shows the changes in market-shares across hospitals for the counterfactual scenario of an earlier adoption of the reform. This counterfactual is conducted by simulating choice in the pre-reform environment using post-reform preferences. This entails a zero-sum game of market-share "reshuffling" between hospitals. The distribution of changes over all hospitals is reported.

Table 14: Supply-Side Response[1]

Dependent Variable	Change in Case-Mix Adjusted Mortality Rate
Change in the Elasticity of Demand with Respect to the Mortality Rate	-0.1296*** (0.0209)
Observations	27

[1]The table reports results from a difference-in-difference OLS regression. A unit of observation is a hospital that existed both pre- and post-reform.

Figure 1: **Map of Hospital Locations**



Appendix

A Implementation of the Mortality Rate Adjustment

A.1 Instrumenting Hospital Fixed Effects

In order to estimate the mortality equation (**??**) we need to instrument the hospital fixed effects, which we do using distance to the hospital (D_{ij}). This regression provides us with fitted values of the hospital-quarter dummies, $\widehat{\psi}_{jt}$, which are then used as the quality of care measure in the utility function, i.e. we set $Z_{jt} = \widehat{\psi}_{jt}$ in equation (1).

$$M = HD\psi + H^{obs}\gamma_{obs} + H^{unobs}\gamma_{unobs} + \eta$$

Since we allow the hospital fixed effects to vary over time, we need to instrument $(J_t - 1)$ variables in each time period (a set of hospital dummies minus a constant). In order to do this we need at least as many instruments. We choose to use the distance to each hospital and a set of dummies equal to one for the closest hospital. This yields a total of $(2 \cdot J_t)$ instruments for each time period (quarter). As indicated previously, the identifying assumption that allows us to obtain a causal effect on patient survival is the exogeneity of patients' locations with respect to their unobserved health status.

In terms of observable health status (H^{obs}) we include the patient's age, sex and co-morbidities, which are widely considered important determinants of patient severity of illness.[2] Since there is no reason to constrain the time period over which we can estimate the impact of hospitals on health outcomes, we use 20 quarters of data from 2003 to 2007.[3] We estimate the production function relationship jointly for all quarters. This is done by arranging the data in a block-diagonal fashion such that each quarter constitutes a block in the matrix. Both the hospital dummies and the instruments are arranged in this way. Therefore distance and the closest-hospital dummy are effectively operating as quarter-specific instruments for the hospital dummies of the particular quarter. This yields twice as many instruments $(2 \times J_t)$ as the number of dummies in each quarter. γ_{obs} is restricted to be the same across all time periods, i.e. the effects of case-mix variables cannot change over time. This regression provides us with fitted values of the hospital-quarter dummies, $\widehat{\psi}_{jt}$, which are then used as the quality of care measure in the utility function, i.e. we set $Z_{jt} = \widehat{\psi}_{jt}$ in equation (1).

[2] Specifically, we use a female indicator, age of the patient, the count of co-morbidities (capped at 6), the female indicator interacted with age, and the interaction of the female indicator with the co-morbidities count.

[3] This is in contrast to the demand estimation for which the institutional setup constrains us to use data only from a selected period of time (see earlier discussion in Section (3.2).

A.2 Implementation

In order to obtain adjusted mortality rates we run a linear probability model, regressing a dummy for death on a set of quarter-specific hospital dummies. The hospital dummies are stacked in a block-diagonal matrix, each block representing one quarter out of 20 quarter for the time period 2001 to 2005. The case-mix is restricted to enter in the same way in all quarters.

Specifically, the data are arranged as follows:

$$X = \begin{bmatrix} X_1 & & & CM_1 \\ & X_2 & & CM_2 \\ & & \ddots & \vdots \\ & & X_{20} & CM_{20} \end{bmatrix}$$

Where CM_t denotes a matrix with various variables capturing the health status of patients within a particular quarter t. All elements in the matrix other then the matrices X_1 to X_{20} and CM_1 to CM_{20} are equal to zero. The block-diagonal elements are given by:

$$X_t = \begin{bmatrix} x^t_{11} & \cdots & x^t_{1k_t} \\ \vdots & \ddots & \vdots \\ x^t_{n_t1} & \cdots & x^t_{n_tk_t} \\ & & \end{bmatrix}$$

Where n_t denotes the number of patients in a particular quarter t, i.e. the number of observations in the data. k_t denotes the number of hospital dummies in each quarter. This number varies across quarters because of hospital entry, exit and mergers.

The matrix of instruments is arranged in a similar fashion:

$$Z = \begin{bmatrix} Z_1 & & & CM_1 \\ & Z_2 & & CM_2 \\ & & \ddots & \vdots \\ & & Z_{20} & CM_{20} \end{bmatrix}$$

with

$$Z_t = \begin{bmatrix} z^t_{11} & \cdots & z^t_{1l_t} \\ \vdots & \ddots & \vdots \\ z^t_{n_t 1} & \cdots & z^t_{n_t l_t} \end{bmatrix}$$

Where n_t denotes the number of patients in a particular quarter t (as in the X-matrix above). l_t denotes the number of quarter-specific instruments. In general we need the condition $l_t > k_t - 1$ to be fulfilled in all quarters (We do not need an instrument for the constant in each quarter, i.e. the average quarterly death rate over all hospitals). In practice we use the distance to each hospital available in the quarter and a dummy for whether this is the closest hospital for the individual patient as instruments. This yields $l_t = 2 * k_t$ instruments for each quarter.

A.3 Production Function Estimates

When estimating the production function we obtain a large set of 275 hospital-quarter fixed effects as an outcome of the regression. It is therefore not convenient to present the parameter estimates here. Instead we give some simple intuition for how the selection mechanism affects our results and provide several formal specification checks.

We find that our estimated quality measure has a correlation of 0.584 with the raw mortality rate in the data. Figure A1 displays a plot of the two measures of quality against each other. One can see that our preferred quality measure, the case-mix adjusted mortality rate, has a larger variance than the unadjusted mortality rate. Specifically, when comparing the scatterplot to the 45-degree line, one can see that the adjusted mortality rate makes good hospitals look even better, and bad hospitals worse. This is precisely what one would expect from a procedure that adjusts for case-mix selection. In the raw data better hospitals look worse than they actually are because they attract relatively sicker patients, while the opposite is true for bad hospitals. Our adjustment leads to a larger spread in the quality distribution by removing the selection effect.

We also provide a set of formal specification tests in Table (A1). We first test for the validity of our overidentifying restrictions using the Sargan-Hansen overidentification test (Hansen 1982). We fail to reject the null – that the overidentifying restrictions are valid. We then test for weak instruments using Anderson's canonical correlation likelihood-ratio test (Anderson 1984) and find that we can strongly reject the null of weak instruments. Both tests provide statistical evidence in favor of our IV specification. However, we fail to reject the null hypothesis of the Durbin-Wu-Hausman test, that the hospital dummies are exogenous (more precisely that the OLS and IV estimates are not statistically different). Despite this we still favor using the case-mix adjusted hospital mortality rates in order to obtain an appropriate measure of the quality of care.

Finally, we examine our identifying assumption that patient location is uncorrelated with health status

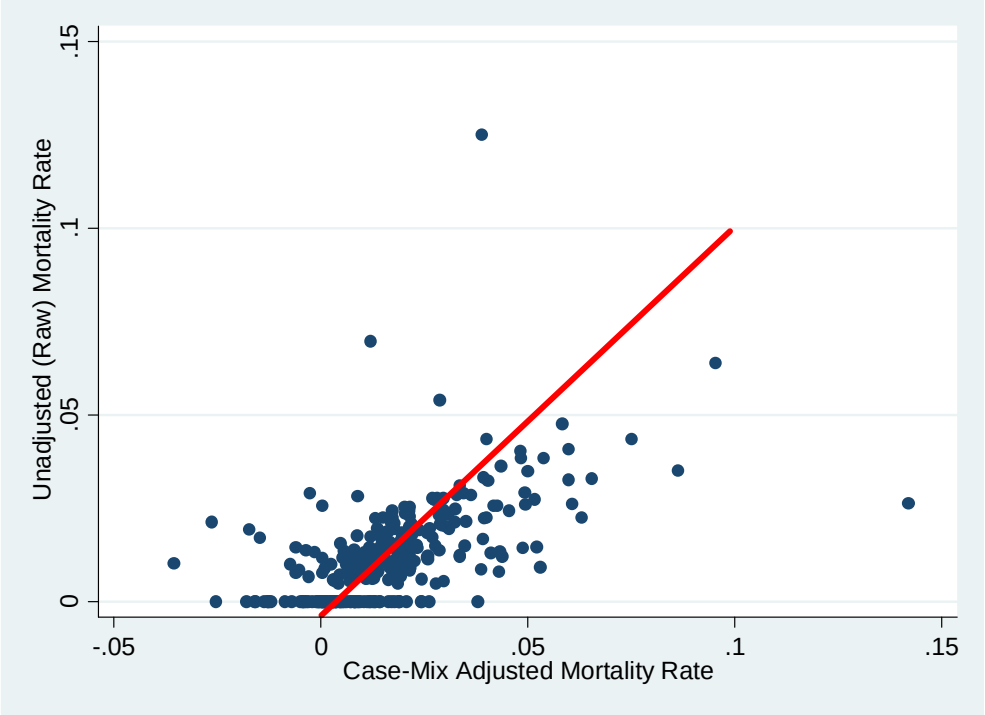
by analyzing whether there is any geographic variation left in unobserved health status after controlling for observable patient characteristics. To this end we run our mortality regression without instruments and include (in addition to the hospital dummies and patient characteristics) a set of 6-digit postcode dummies.[4] We find that the postcode dummies are not jointly significant (while the set of hospital dummies is jointly significant) suggesting that after controlling for observed patient characteristics there is no significant amount of geographic variation in health status left that might invalidate our IV approach.

[4]This is roughly equivalent to controlling for zipcode dummies in the case of the US. Geweke, Gowrisankaran, and Town (2003) use a similar test with zipcodes for US data. There are 7,727 6-digit postcode dummies in our sample of 64,082 patients.

Table A1: Specification Test Statistics for the Mortality IV Regression

Sargan-Hansen	χ^2	585.74
Overidentification Test	P-value	0.20
Anderson Canonical	χ^2	2,557.67
Correlations Test	P-value	0.00
Wu-Hausman Test	χ^2	562.05
	P-value	0.57
Number of Hospitals	28 / 29	(Varies Across Quarters)
Number of Quarters	20	(2003-2007)
Number of Patients	64,082	

Figure A1: **Relationship Between Raw and Adjusted Mortality Rate**



B Data Sources

Patient Choice Data	Hospital Episodes Statistics (HES) dataset. Administrative discharge dataset that covers all patients the underwent treatment in an NHS hospital.
Index of Multiple Deprivation	UK Census (http://www.communities.gov.uk/communities/research/indicesdeprivation/deprivation10/). The measure is defined at the Middle Layer Super Output Area (MSOA). There are about 6,800 MSOAs in England with an average population of 7,200.
Patient Informedness Data	NHS Patient Choice Survey (http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsStatistics/DH_094013) This measure is defined at the Primary Care Trust Level (PCT). There are about 150 PCTs in England