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

This draft: February 22, 2016

*We are grateful to the Editor and three anonymous referees, whose comments substantially improved the paper. Helpful comments were provided by participants at seminars at a number of universities and conferences. We are grateful for financial support to Propper from the ESRC under grant ES/J023108/1. 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 that explicitly captures the referral constraints imposed on patients to evaluate whether removing constraints on choice increased the demand elasticity faced by hospitals. Using data for an important surgical procedure we find that patients became more responsive to clinical quality. The increased demand responsiveness led to a modest reduction in mortality by re-allocating patients and a substantial increase in patient welfare. The elasticity of demand faced by hospitals increased substantially post-reform, giving hospitals stronger incentives to improve their quality of care. Finally, we find evidence that hospitals responded to the enhanced incentives by improving quality. 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. However, whether enhanced patient choice will make hospital choice more responsive to quality is not well established, although 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 issue we address here.

To do this we exploit a reform which introduced patient choice and tie this to the estimation of a structural demand model that explicitly incorporates the institutional features of the reform. This enables us to identify the effect of increasing choice on patient behavior. We use the model to quantify the gains from the reform in terms of patient welfare and survival and to analyze how the changes in patients' choices translate into changes in the competitive environment faced by hospitals.

The reform we exploit is from the English National Health Service (NHS). In 2006, the UK government mandated that patients in the English NHS had to be offered a choice of 5 hospitals when referred by their physician to a hospital for treatment. Prior to this reform, there was no requirement that patients be offered choice. The reform provides 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, allows us to cleanly identify the effect of greater choice while holding the underlying market structure fixed. We use this reform to estimate a structural model of demand under both pre-reform constrained choice and post-reform liberated choice.

In the post-reform period we assume that, as patients were mandated free choice, the choices made by the referring physician fully reflect patient preferences over hospital characteristics (quality of care, waiting time, travel distance). Pre-reform, patient choice was constrained. We do not observe the choice sets available to patients. Therefore, to model hospital choice pre-reform we adopt an approach which draws from the consideration set literature (see, e.g., Goeree 2008, Mehta, Rajiv, and Srinivasan 2003). We model the patient's choice set as containing a subset of the full set of available options. This subset is determined by the physician's preferences, which were shaped by the institutional structure of the pre-reform referral system. Specifically, in the pre-reform period referrals were paid for by a selective contracting system that covered referrals to a only a subset of all potential hospitals. This system, discussed in detail in Section 2.2 below, made it easier for a physician to refer patients to hospitals located within local administrative boundaries. Exploiting these institutional features, we estimate the extent to which the referring physician cares about patient welfare by allowing him to offer a (potentially limited) set of hospitals from which the patient chooses the one with the highest utility

according to her preferences.[1] This allows us to identify the extent to which patients were constrained pre-reform. To estimate the model we take the case of coronary artery bypass graft (CABG) surgery. This is well suited to our purpose as most CABGs are elective and scheduled well in advance, giving patients an opportunity to exercise choice if allowed to do so.[2]

We begin by providing descriptive evidence to show that, while on average distance travelled changed little post-reform, there is clear evidence of improved sorting of patients to higher quality hospitals after the reform. Furthermore, we see these patterns only for elective cases but not for emergencies, suggesting that choice is the key driver. We then estimate the structural model and use the estimates to quantify the impact of the reform. First, we analyze the direct impact of the removal of constraints on the ability of patients to choose a hospital according to their preferences. This leads to a reallocation of patients to higher quality hospitals and thus an improvement in patient welfare and expected survival. We find a modest decrease in patient mortality of around 3.5 more patients per year (approximately a 3 percent decrease) had patients had free choice in the pre-reform period. The utility increase for the average patient as a result of the reform is is equal to \$300,900.[3] We also find that the quality elasticity of demand increased for all types of patients. The demand elasticity increased relatively more for sicker and lower income patients. The latter effect is particularly interesting because there have been concerns expressed about the impacts of pro-choice policies on the poor (see, e.g., Cookson, Laudicella, and Donni 2013). We do not find these to be substantiated here.

Second, we analyze the change in the competitive environment in the hospital market. To quantify the effect on competition, we aggregate the patient-level elasticities to the hospital level and find that the competitive environment changed substantially. For the average hospital an increase in mortality leads to a five-times larger drop in market share post-reform relative to pre-reform. This lends support to the notion that hospitals had stronger incentives to improve quality due to the introduction of patient choice. Finally, we analyze the supply-side response to the reform and find that hospitals which experienced the largest increase in elasticity also had the biggest reduction in their mortality rates.

Our paper makes a contribution to two distinct literatures. The first is the impact of pro-competitive and choice based reforms on the quality of health care provision. Kessler and McClellan (2000) provided initial evidence on the effect of competition on quality in the US Medicare program. The literature on health care competition and quality in the US has grown greatly since then. The overwhelming majority of these papers take non-structural approaches. Gaynor and Town (2012) and Gaynor, Ho, and Town (2015) survey those papers and their results. Most (but not all) papers find that competition leads to enhanced quality. In the UK context, the primary analyses of the impact of the 2006 reforms on patient outcomes are Gaynor, Moreno-Serra, and Propper (2013) and Cooper,

[1] The model nests the extreme cases of unconstrained and fully constrained choice, where in the latter case the physician offers only a singleton choice set, his preferred hospital, to the patient.

[2] A relatively small proportion of CABG surgeries are performed on an emergency basis. We exclude those from our main analysis and use them as a placebo test for the robustness of our results. We provide factual details on CABGs in Section 2.1.

[3] Due to the absence of price in this market, we monetize the welfare gain using patient preferences over mortality and the value of a statistical life.

Gibbons, Jones, and McGuire (2011). They estimate reduced form models of changes in hospital mortality outcomes regressed on measures of market concentration. Both papers find that patient outcomes at the hospital level improved and conclude that competition and choice improved quality.

Our paper advances the prior literature by explicitly modeling the choice process before and after the reform, allowing us to estimate the extent of constraints pre-reform. The nature of our structural model allows us to carefully decompose the effect of the reform along various dimensions. As mentioned above, we first quantify the changes in patient mortality and utility from the reallocation of patients to better hospitals following the liberation of choice. Second, we assess the effect on supply-side incentives. We compute the change in hospital elasticities with respect to quality after the reform and assess whether hospitals reacted to the change in the competitive environment by improving their quality (by lowering mortality rates). The small number of earlier papers primarily focuses on just the last part of this assessment: hospitals' reactions to the reform in terms of quality improvements. We provide a much more comprehensive picture of the impact of increasing choice. Furthermore, while there is a large literature estimating demand models in health care, either directly or as part of a larger model (e.g., of hospital competition), these papers are not typically able to separately identify patient versus physician preferences due to a lack of sources of identifying variation.[4,5] The setting we study provides us with a unique opportunity to identify patient preferences separately from physician preferences because of the change in the choice process due to the removal of constraints.

Secondly, our paper contributes to the literature on consideration set formation, for example Roberts and Lattin (1991), Andrews and Srinivasan (1995), Bronnenberg and Vanhonacker (1996), Mehta, Rajiv, and Srinivasan (2003), Goeree (2008), and Seiler (2013). We add to this literature by exploiting a unique aspect of our data: the fact that we observe a change in the process by which consideration sets are formed. In particular, observing unconstrained choice post-reform allows us to estimate preferences in a setting where constraints had no impact on choice. Given those preference estimates, we then use choices in the pre-reform period to estimate the process driving the consideration set formation. Within the health care literature, Ho (2006) and Dafny, Ho, and Varela (2013) also analyze the effect of removing choice constraints. However, in their settings the constrained choice sets are observed in the data and the papers evaluate the welfare effect of their removal in a counterfactual. We observe a change in the way consideration sets are formed due to the reform, but as the precise choice sets being offered are not observed we have to infer the constraining process from the choice data.

The paper is structured as follows. Section 2 describes the institutional setting. Section 3 outlines the modeling framework. Section 4 describes the data and Section 5 presents econometric issues and estimation methods. Section 6 presents reduced-form results followed by Section 7, which presents the results from the structural estimation. Section 8 quantifies the impact of the reform along various

[4]Luft, Garnick, Mark, Peltzman, Phibbs, Lichtenberg, and McPhee (1990), Tay (2003), and Howard (2005) are examples of demand estimation for the US. Sivey (2008), Beckert, Christensen, and Collyer (2012), Varkevisser, van der Geest, and Schut (2012), and Moscone, Tosetti, and Vittadini (2012) are examples for Europe. Capps, Dranove, and Satterthwaite (2003), Gaynor and Vogt (2003), and Gowrisankaran, Nevo, and Town (2015) are examples of choice estimation within the context of models of hospital competition.

[5]Beckert (2015) proposes a methodological approach to identifying patient vs physician preferences.

dimensions using our model estimates. The final section contains concluding remarks.

2 Institutional Details

2.1 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.[6] 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.[7] 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.[8] 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 in a hospital who conducts tests and may then perform a non-surgical procedure to unblock the artery (called angioplasty or percutaneous coronary intervention, PCI). If this fails, the patient will then be referred for a CABG to be performed by a cardiac surgeon and put on an elective waiting list for this treatment. Cardiologists operate in almost all short term general NHS hospitals but CABGs are performed only at a limited number of hospitals. 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).[9]

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

[7] 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).

[8] 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.

[9] Payment for the treatment is discussed in the next section.

2.2 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 (NHS). Primary care is provided in the community by publicly funded physicians known as General Practitioners (GPs). GPs are self employed and work in practices which on average contain 4 to 5 GPs. They earn their income by providing services to the NHS. Patients have a very limited choice of GP.[10] GPs also act as gatekeepers for hospital-based (known as secondary) care, sending patients who need treatment to a specific hospital. Secondary care (including cardiology and 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 are generally employed only in one NHS Trust. Publicly funded bodies covering specific geographic areas, called Primary Care Trusts (PCTs), have the task of buying hospital-based health care for their population on behalf of the GP practices in their area. In the period we examine, PCTs also oversaw the GP practices in their area, monitoring their prescribing, inspecting their premises, providing financial assistance for practice computing, financing community nurses to complement GP services and providing information to practices to allow them to compare their performance with other practices in the PCT. On average each PCT had around 20 or so GP practices in their area (Santos, Gravelle, and Propper 2015).

In the pre-reform period, when purchasing hospital based care, buyers (PCTs) and sellers (the NHS Trusts) negotiated over price, service quality (mainly waiting times rather than 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 contracts negotiated by the PCTs on behalf of all the GPs in their area and covering all the patients registered with these GPs.[11] PCTs had to make additional, separate, payments for any referrals physicians made to non-contracted hospitals. Thus GPs whose patients were referred off-contract were likely to be subject to scrutiny by, and discussions with, the PCT about their behavior. In the choice of hospitals with which to contract, PCTs had a strong tendency to support local providers, i.e., ones that fell within their geographical boundaries. This was for both historic reasons (PCTs and hospitals had, pre-1991, been joint administrative entities) and to assure local supply of care. So for these contractual reasons physicians were more likely to refer patients to the nearby hospital(s) with which their PCT had negotiated a contract.

From late 2002 the government started to develop the components of a reform package intended to bring about hospital competition from 2006 onwards.[12] There were several elements to this policy. First, under law, after January 2006 patients had to be offered a choice of five providers for where they

[10] Patients almost always have to choose a GP located near to where they live. There are currently (as of 2016) proposals to increase choice.

[11] Almost all patients in the PCT will be registered with a GP in the PCT.

[12] A previous hospital competition policy had operated 1991-1997 but was dropped when the Labour government came to power in 1997 (Propper 2012).

had their hospital care (Farrar, Sussex, Yi, Sutton, Chalkley, Scott, and Ma 2007). GPs were required (and provided with software) to ensure that patients were made aware of, and offered, choice.

Second, 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 arrangements (http://www.chooseandbook.nhs.uk/).

Third, from 2006 onwards the NHS adopted a payment system in which hospitals were paid fixed, regulated, prices for treating patients (a regulated price system similar to the Medicare hospital payment system in the US). This fixed price system covered around 70% of hospital services, including CABG. This change in the remuneration system meant that GPs (and hospital specialists making referrals for treatment to a hospital other than their own) were no longer restricted in these referral decisions by their PCT's contractual arrangements with individual hospitals.[13]

In the particular case of CABG, which is a specialized treatment provided at only a few hospitals, the Choose and Book information system was probably less important than for more routine treatments. What was more important was the removal of selective contracting and the right to choose. These allowed the patient to choose, with the aid of their GP, the initial hospital in which to see a cardiologist, and gave the cardiologist freedom in where they sent patients for a CABG. In these choices the patient, and the physicians involved, were no longer restricted in their decisions by selective contracting.

It is important to note that the reforms did not change financial incentives for the patient or financial payments to referring physicians. Patients did not pay for medical care either before or after the reforms. Neither GPs nor hospital based specialists (including cardiologists) received payments that were linked to their referral advice to the patient, either before or after the reform. However, the reform did have a financial impact on PCTs, which were overseeing GPs, and pre-reform had an interest in referring patients to hospitals that were covered by contractual arrangements. Therefore, by ending selective contracting, what the reforms did was to remove legal and (indirect) financial restrictions on both physicians and patients which enabled referral decisions to be more flexible and tailored to individual patients.

It is possible that patient preferences influenced referral decisions even before the formal introduc-

[13]The reforms also promoted the 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 analyze. 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.

tion of patient choice. For instance, a well-informed patient with a strong preference might have been able to convince the physician to refer her to a specific hospital. The effect of the reforms was to make such choice available and far more explicit for all patients. Similarly, physician agency might have led to referrals pre-reform which were in line with patient preferences although not directly selected by the patient. However, the nature of selective contracting suggests that physician incentives were not fully aligned with patients' interests. In particular, contractual arrangements made by PCTs with nearby hospitals made referrals outside that set of hospitals substantially more difficult pre-reform. Rather than assuming that patients were constrained pre-reform, we let the data tell us the extent to which patient preferences influenced choice in the pre-reform period. In the econometric model we explicitly estimate the degree to which patient preferences were constrained before the reform, thus allowing for less than fully constrained choices.

3 Modeling Approach

Our framework for analyzing the impact of the reforms on hospital demand is comprised of two components: (1) a model of patient choice under the post-reform regime where choice was liberated and (2) a representation of the process by which choice was constrained before the reform.

3.1 Patient Preferences

In the post-reform period, when choice is liberated, we assume that referrals reflect choice from the full set of hospitals that perform CABG surgery based on patients' utility from this choice.[14] We note that in the case of a complex procedure such as CABG, patients will typically seek and follow advice from their referring physician. We make the key assumption that after the reform, the physician acts as a perfect agent for his patient by providing her with information and advice based on his best understanding of the patient's preferences. We make this assumption based on the institutional features described above. Specifically, neither the GP nor the cardiologist have any incentive post-reform not to act in the patient's best interest. They do not receive payments linked to where they refer and they are no longer bound by selective contracting. Furthermore, physicians were mandated to explicitly involve patients in the decision-making process by offering them choice from a set of hospitals. Both of these aspects of the reform together are the reason for assuming that post-reform referrals are based solely on patient preferences.

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 affect hospital choice are the quality of care, the

[14]There are only 29 (both pre- and post-reform) hospitals in England which offer CABGs and all treat a fairly large number of cases on a regular basis. We confirmed, from independent sources, that we cover the exhaustive set of possible CABG providers. So it is not the case that some options are excluded from the analysis because they were never chosen during the sample period. Further, while physicians were only required to offer patients the choice of 5 hospitals post-reform, they were able to refer to any hospital that performs CABGs. A non-negligible fraction of patients receive treatment at a hospital other than one of the 5 closest hospitals. We therefore assume in estimation that all 29 hospitals are available to all patients.

amount of time a patient has to wait for surgery, and distance from the hospital. We also allow for preference heterogeneity across different patient characteristics. Finally, 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.[15]

Let a patient i obtain the following utility from choosing hospital j:

$$U_{ij} = \beta_{w,i} W_{jt} + \beta_{z,i} Z_{jt} + f(D_{ij}) + \xi_j + \varepsilon_{ij} \tag{1}$$

where W_{jt} denotes the average waiting time for a CABG at hospital j in time period t, Z_{jt} denotes the quality of clinical care at the hospital in that time period, and D_{ij} is the distance from patient i's location to the location of hospital j.[16] The function $f(D_{ij})$ is a transformation of D_{ij} that reflects the (non-linear) preference for distance to the hospital. ξ_j denotes unobserved hospital quality and ε_{ij} is an idiosyncratic taste shock that is iid extreme value. We incorporate preference heterogeneity by allowing the coefficients on waiting times and mortality ($\beta_{w,i}$ and $\beta_{z,i}$) to vary with patient characteristics.

3.2 Constraints on Patient Choice

To model the fact that choice before the reform was limited, we assume that pre-reform, physicians offered their patients a limited set of options. This referral behavior was determined by their own utility, which we model based on the institutional features of the pre-reform contractual arrangement and referral process described in Section 2.2. As explained, for contractual reasons physicians were more likely to refer patients to the nearby hospital(s) with which their PCT had negotiated a contract and discouraged from referring elsewhere. In addition, physicians were likely to have preferences for referrals to certain hospitals based on past interactions with the hospital or particular surgeons at the hospital. We capture these referral incentives by making physician utility a function of hospital fixed effects and distance related variables as well as an indicator for whether the hospital was located in the PCT of the referring GP.

More formally, we define the utility the physician receives from referring patient i to hospital j as

$$V_{ij} = g(D_{ij}) + \zeta_j + \nu_{ij} \tag{2}$$

where $g(D_{ij})$ is a transformation of D_{ij} that reflects the (non-linear) preference for distance to the hospital which facilitates referrals.[17] ζ_j denotes unobserved (to the econometrician) physicians' assessment of hospital quality and ν_{ij} is an idiosyncratic shock that is iid extreme value. Note that, for simplicity of exposition, we index the physician's utility by the index i of the patient he is referring.

[15] This assumption is common in the healthcare literature. Capps, Dranove, and Satterthwaite (2003) and Ho (2006) (among others) make the same assumption when estimating demand models for hospital choice.

[16] Note that strictly speaking utility has a t subscript because hospital characteristics vary over time. However, we observe each patient only once and every patient i has a unique time-period t associated with his referral. We therefore denote the utility function above as patient/hospital but not time-period-specific (U_{ij}).

[17] The indicator for whether the hospital is located in the PCT of the referring GP is included in $g(D_{ij})$.

This has a close mapping to our data: the physician's influence via offering a limited set of hospitals constitutes a latent process and is not directly observable. In the data we only see the referral outcome for patient i. We maintain this notation for the remainder of the paper.

Details of the specific variables included in physician and patient utility are provided in the estimation section (Section 5). In brief, to implement the approach discussed here, we include only aspects of the hospital that the physician directly cares about in the physician's utility function. The preferences physicians are likely to have for referrals to certain hospitals based on contractual relationships or past interactions are captured through the set of hospital fixed effects (ζ_j) as well as a dummy for whether the hospital is located in the physician's PCT. Importantly, waiting times and quality are not included in physician utility, as we think of those variables as influencing the physician's decision only indirectly via their influence on patient utility.

Contrary to a standard choice model, we assume that the physician is not ultimately making the decision of which hospital to visit, but offers a set of options from which the patient chooses the highest utility one according to his preferences. More specifically, we assume that the physician includes hospital k in the consideration set, that is the (potentially limited) set of hospitals patients can choose from, if

$$V_{ik} \geq max_{j \in J}(V_{ij}) - \lambda_i \tag{3}$$

where $\lambda_i \geq 0$ and J denotes the full set of hospitals. In other words, every hospital that is within a distance of λ_i (in utility space) relative to the highest utility option is included in the consideration set. The highest utility hospital is always included, and the number of options included in the set increases the higher is the value of λ_i.

The particular value of the constraining parameter λ_i captures the extent to which the physician cares about patient utility and allows the patient's preferences to influence choice. In the case of $\lambda_i = 0$, the physician picks the highest utility option according to his preferences and the patient's preferences have no bearing on the choice. In this case, the consideration set formation process collapses to a discrete choice model based on physician preferences. For $\lambda_i > 0$ the physician might include multiple hospitals in the choice set. Whether and how many are included depends on the specific value of λ_i as well as how similar the utilities of all hospitals are to each other. λ_i is a parameter to be estimated and the main driver of consideration set size. In order to allow for the reform to differentially affect different groups of patients we allow λ_i to vary across patient characteristics.

A few comments on this way of modeling the consideration set process are in order. We believe that our approach reflects actual decision-making well and at the same time is parsimonious, in the sense that we are able to model the degree to which patient preferences are constrained through one parameter: λ. Our framework is different from other approaches in the literature due largely to the nature of the forces constraining patient choice. In the consideration set literature in consumer goods markets the limited nature of choice sets usually originates from limited information acquisition.[18]

[18]Mehta, Rajiv, and Srinivasan (2003), Kim, Albuquerque, and Bronnenberg (2010) and Honka (2014) model the

Instead, in our case, the constraining force is the joint-decision making process, which involves the patient and physician and in which the physician had a dominant role before the liberation of choice.[19]

The paper closest to ours is Goeree (2008), who models the effect of advertising on consumer's consideration sets. However, our modeling approach is different in a few key aspects. In contrast to our paper, Goeree (2008) models the probability of inclusion into the choice set separately for each product. This implies that the inclusion probabilities are independent across products. This is not the case in our setting, where a change in physician utility for one hospital can influence the inclusion probabilities of other hospitals. A second difference is that her setup allows for a strictly positive probability of an empty consideration set, which is not permitted in our model (patients have to be offered the choice of at least one hospital). Both features are appropriate in the context of advertising and personal computer purchases, but less attractive for our setting. First, our model assures that a patient needing a CABG will be offered the choice of at least one hospital. Second, it seems reasonable that the characteristics of all available hospitals shape the consideration set size and composition. In our model, the presence of a particularly attractive option can lead to a smaller choice set by "pushing" other hospitals out of the consideration set. For instance, a hospital at 10 kilometers distance is less likely to be included if another hospital exists that is located at a distance of only 5 kilometers.[20]

Finally, an alternative modeling approach we could take would be to have expected patient utility directly enter the physician's utility function. However, expected patient utility is defined over all possible consideration set permutations, of which there are $2^{29} - 1$ $(= 536,870,911)$ for 29 available options. This approach would therefore require us to write down physician utility over all the possible one billion compositions of the consideration set. Instead, our framework allows us to compute utility for each option separately and then derive the consideration set based on this set of utilities and the value of λ. This approach considerably decreases the computational burden and thereby makes the estimation of the model feasible.

4 Data and Descriptive Statistics

We use data from the UK Department of Health's Hospital Episode Statistics (HES) dataset, which is an administrative dataset containing information on every English NHS hospital inpatient admission. The data contain details of the medical procedures which the patient received (classified according to OPCS codes[21]) and up to 14 diagnoses, classified according to the ICD-10 classification.[22] We have

consumer's decision to gather information in a model of consumer search. Our case is less comparable with models of consumer search and has more in common with the case of external information provision, such as in Goeree (2008), in the sense that patients do not actively influence the formation of consideration sets.

[19] For other papers which allow for an agent to be involved in the choice process see Baker, Bundorf, and Kessler (2015) for an empirical examination and Beckert (2015) for an econometric model.

[20] More generally, increasing the utility of the highest utility hospital (holding everything else constant) will increase the maximum utility level. As a consequence fewer hospitals will tend to be within λ distance in utility-space from the highest utility hospital. This mechanism will lead to a smaller consideration set.

[21] 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.

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

data on the universe of inpatient discharges receiving CABG surgery from every hospital in the NHS in England from April 2003 to March 2008, corresponding to 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. 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 due to the fact that the reform was phased in gradually over 2005 and 2006. We provide more detail on the specific timing of the introduction of the reform in Appendix A.

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.[23] 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 hospital level CABG-specific waiting times and mortality rates by aggregating the data at the hospital level over the relevant time period. Finally, from the hospital location and the patient's postcode, we compute the distance to the hospital. A list of sources for the data is in Appendix E.

4.1 Measuring Clinical Quality of Care

We need to define an appropriate measure for the quality of clinical service. Due to the relatively high risk of death following a CABG procedure, we assume the survival probability at a specific hospital is the primary quality metric patients care about. We thus use mortality rates as a quality measure.

One might be concerned that mortality rates do not correctly reflect differences in survival probabilities due to differences in case-mix across hospitals. We therefore implement an empirical test to assess the role of case-mix differences across hospitals, which we describe in detail in Appendix B. In summary, we regress mortality (at the individual level) on a set of hospital dummies which we instrument with distance to each hospital, following similar approaches by Gowrisankaran and Town (1999) and Geweke, Gowrisankaran, and Town (2003).[24] Importantly, this IV estimation approach allows us to test whether case-mix varies significantly across hospitals by comparing the coefficients on the hospital dummies from an OLS regression (which are equal to the unadjusted mortality rate) with the estimates from the IV regression via a Hausman test.[25] Doing so, we find that we cannot reject the null hypothesis that the OLS and IV estimates are the same. We hence conclude that case-mix differences across hospitals are small enough not to affect the mortality rates significantly and therefore use the unadjusted mortality rate as a measure of the clinical quality of the hospital. This is simpler

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

[24] This assumes that people do not choose where they live relative to CABG hospitals based on their unobservable health status. 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). We provide additional evidence for this assumption in the appendix.

[25] We assess the strength of the instruments via F-tests for each of the 284 first stage regressions (one for each hospital/quarter pair). The instruments are strong: the mean of the F-statistic across all the regressions is 160.9. See Appendix B.2 for more details.

and avoids introducing another source of error into the estimation from case-mix adjustment.[26]

4.2 Hospital Characteristics

In contrast to many other procedures, CABGs are only offered by a small set of hospitals. Of around 170 short term general (acute) public hospitals within the NHS, only 29 hospitals offer bypass operations. There was almost no change in market structure around the time of the policy reform.[27] The choice set faced by patients is nearly identical before and after the reform, which allows us to separate the impact of greater choice from a possible change in market structure.[28] Figure 1 provides a map of the locations of NHS CABG-performing hospitals. Also, while in principle patients could choose privately funded treatment, in practice they did not.[29]

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.[30] 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).[31] 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 years with a slight decline towards the end of the sample period. There is substantial variation in mortality rates across hospitals in all years.

When using the 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 D1.

[26] We also note that the point estimates when using an adjusted mortality rate in the demand estimation are very similar to the ones we obtain when using the unadjusted rate (see Table D2 in the Appendix). This is consistent with the fact that we fail to reject the equality of the two mortality rates.

[27] The only changes are: (i) the merger of Hammersmith Hospital and St. Mary's Hospital, which 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. There are therefore 27 hospitals present in every period of the data. Table D1 lists the number of hospitals by quarter.

[28] 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.

[29] 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.

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

[31] This target policy ran most strongly from 2001-2005 i.e. before the choice reform. It has been shown that the fall in waiting times was primarily due to efficiency improvements and did not have any detrimental effect on health outcomes (see Propper, Sutton, Whitnall, and Windmeijer 2008).

4.3 Patient and Area Characteristics

We measure patient socio-economic status using the Index of Multiple Deprivation (IMD) in the small area (the Middle Super Output Area, MSOA) in which the patient lives. The IMD is a measure of income deprivation of the patient's neighborhood and is the best available metric on patient income in our data. This variable ranks a patient's local neighborhood from richest to poorest. The range is 0 to 1, with higher values implying higher deprivation.[32] In the estimation we employ an indicator for whether the IMD in a patient's neighborhood is below the median IMD (0.1), i.e., whether their neighborhood is above the median in income (since IMD decreases in income). Going forward, we simply refer to this variable as "income."

HES provides a list of co-morbidities at the patient-level. We use these to compute the Charlson index, which weights co-morbidities by their impact on mortality risk (Charlson, Pompei, Ales, and MacKenzie 1987). The higher the value the greater the patient's risk of mortality (the index for a patient with no co-morbidities has a value of zero).[33] We use an indicator for whether a patient has a Charlson Index above the median (≥ 1) in the estimation. We refer to this variable as "severity."

Table 2 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 considerable variation in patients' general health status, with a large fraction of patients for whom several co-morbidities are reported. About 40 percent of patients have a positive value of the Charlson index. Both income and the Charlson index are used in the demand estimation to analyze how the reform differentially affected different groups of patients.

The bottom two rows of Table 2 contains descriptive statistics on the distances patients travel for their CABG treatments.[34] We see that the average patient traveled a substantial distance (over 30 kilometers) and that there is a great deal of variation in how far patients traveled for care. It is also notable that there is very little difference in distance traveled between the pre- and post-reform time periods. This could occur if patients sorted themselves to better hospitals post-reform within approximately the same distance. In Section 6 we provide some reduced-form evidence that this is the most likely explanation for the lack of a change in distance traveled.

5 Structural Estimation

As outlined in the model exposition in Section 3, there are two parts of the model we need to specify: patients' preferences and the constraining process in the pre-reform time period. Post-reform, patients' choices are unconstrained. As a consequence, utility alone drives the choice of hospital. We first

[32] The IMD is computed by the government for geographical areas that comprise roughly 7,000 individuals. 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/.

[33] For patients in our sample the index takes on values 0, 1 and (rarely) 2. 60% of patients have a Charlson Index value of 0.

[34] We use the patient's 4 digit postcode available in HES to calculate straight line travel distances.

describe unconstrained post-reform choice, then move on to pre-reform choice under constraint.

5.1 Post-Reform (Unconstrained) Choice

Post-reform utility is (as described earlier in equation (1))

$$U_{ij} = \overline{U}_{ij} + \varepsilon_{ij} = \beta_{w,i} W_{jt} + \beta_{z,i} Z_{jt} + f(D_{ij}) + \xi_j + \varepsilon_{ij} \tag{4}$$

We define $f(\cdot)$ by allowing distance to enter linearly as well as with an indicator for whether the hospital was the closest one in the choice set:

$$f(D_{ij}) \quad = \quad \alpha_{d1} D_{ij} + \alpha_{d2} Closest_{ij},$$

where $Closest_{ij}$ is a dummy equal to one if hospital j is the closest one in the choice set of patient i. In what follows, when we write out the utility function we continue to use $f(D_{ij})$ to economize on notation. We estimate unobserved hospital quality (ξ_j) by including a set of hospital fixed effects. We also allow for observable heterogeneity in preferences for both waiting times and quality of service in the standard way:

$$\begin{aligned} \beta_{z,i} &= \overline{\beta}_z + \beta_z X_i \\ \beta_{w,i} &= \overline{\beta}_w + \beta_w X_i \end{aligned}$$

where X_i is comprised of observable patient characteristics on income and illness severity, as described previously in Section 4.3.

Based on this utility function, the probability of patient i choosing hospital k in the post-reform time period is given by

$$Pr_{ik}^{UNCON}(\Omega_{patient}) = \frac{\exp[\overline{U}_{ik}(\Omega_{patient})]}{\sum_{j \in J} \exp[\overline{U}_{ij}(\Omega_{patient})]}$$

where $(\Omega_{patient} = \beta_{wi}, \beta_{zi}, \alpha_d, \xi)$ is the vector of coefficients to be estimated pertaining to patient utility and includes the coefficients on waiting times and mortality as well as the interaction terms with observable patient characteristics, the distance coefficients, and the set of hospital fixed effects. J denotes the unconstrained set of all CABG performing hospitals in the UK.[35] We denote the probability

[35] We do not limit choice sets based on the location of the patient and hence J contains all CABG performing hospitals and does not vary across patients.

as Pr^{UNCON} to distinguish it clearly from the pre-reform choice probability under constraint.

5.2 Pre-Reform (Constrained) Choice

We now describe the process by which patient preferences are constrained prior to the reform. In particular, physician utility is distinct from patient utility (as described above, we use the patient index i to describe the physician's utility with regard to the referral of patient i to hospital j):

$$V_{ij} = \overline{V}_{ij} + \nu_{ij} = g(D_{ij}) + \zeta_j + \nu_{ij} \tag{5}$$

where $g(D_{ij})$ is a transformation of D_{ij} that reflects the physician's preference for distance to the hospital, ζ_j is a fixed hospital effect denoting (the physician's perception of) hospital quality and ν_{ij} is an idiosyncratic shock that is iid extreme value. We operationalize $g(D_{ij})$ in a similar way as in the patient utility function by including a linear distance term and a dummy for whether the hospital is the closest one, while allowing the parameters for doctors to differ from those for patients. Furthermore, we also include a dummy which is equal to one if the hospital is located within the PCT of the referring physician to capture the fact that it is more likely a contractual arrangement exists for any hospital within the PCT of the GP, thus making a referral to such a hospital easier for the physician prior to the reform.[36]

$$g(D_{ij}) \quad = \quad \gamma_{d1} D_{ij} + \gamma_{d2} Closest_{ij} + \gamma_{d3} WithinPCT_{ij},$$

The physician's utility function differs from patient utility in that it does not depend on waiting time or quality. Nonetheless, hospital quality and waiting times can affect pre-reform referrals if the physician allows some degree of choice and therefore the patient is not fully constrained.

We assume that the physician offers multiple hospitals to the patient to choose from. The physician will include hospital k in the consideration set, that is the (potentially limited) set of hospitals patients can choose from, if

$$V_{ik} \geq max_{j \in J}(V_{ij}) - \lambda_i \tag{6}$$

where $\lambda_i \geq 0$ and J denotes the full set of hospitals available. λ_i is a parameter to be estimated and the primary driver of consideration set size, i.e., the degree to which choice was constrained before the reform. To allow for the reform to differentially affect different groups of patients we allow for heterogeneity in λ_i. Similar to the way we modeled preference heterogeneity we assume

[36] As noted above, after the reform such incentives did not exist because contracts with specific hospitals were replaced with prospective payments.

$$\lambda_i = \overline{\lambda} + \lambda X_i \tag{7}$$

where X_i is the same set of variables used to capture patient preference heterogeneity, namely the severity of the case and the patient's income level.[37] λ_i determines the weight of patient (vs. physician) preferences in the decision-making process. For the case of $\lambda_i = 0$ only the physician's highest utility hospital is included in the consideration set, i.e., choice is driven entirely by physician preferences. The consideration set formation process collapses to a discrete choice model based solely on physician preferences. At the other extreme, as λ_i grows larger and $\to \infty$, more hospitals (eventually all) are included in the consideration set and ultimately patients' preferences are decisive, i.e., patient choices are not constrained by the physician. The consideration set formation process therefore nests both extreme cases of fully constrained and completely unconstrained choice. We denote the constrained set of hospitals offered by the physician by CS. The constrained set is a subset of the full choice set ($CS \subseteq J$) and contains at least one option ($CS \neq \emptyset$).

The probability that patient i is referred to hospital k in time-period t is given by the product of the probability that hospital k is included in the consideration set by the physician and the probability that the patient picks it from the consideration set:

$$Pr_{ik}^{CON}(\Omega_{patient}, \Omega_{physician}) = \sum_{CS_k} Pr(CS_k|\Omega_{physician}) Pr(k|CS_k, \Omega_{patient})$$

where CS_k denotes all consideration sets that contain hospital k. The second probability in the equation above is a function of the patient utility parameters ($\Omega_{patient}$) and is similar to the post-reform choice process without constraints. The only difference is that patient preferences determine the choice of hospital from the subset CS_k rather than the full choice set J. This yields the following conditional choice probability

$$Pr(k|CS_k, \Omega_{patient}) = \frac{\exp[\overline{U}_{ik}(\Omega_{patient})]}{\sum_{j \in CS_k} \exp[\overline{U}_{ij}(\Omega_{patient})]} \tag{8}$$

The consideration sets in our context are unobserved and the probability for each of the possible sets is driven by the parameters in the physician's utility ($\Omega_{physician} = \gamma_d, \lambda_i, \zeta_j$):

$$Pr(CS|\Omega_{physician}) = Pr(V_{ij \in CS} \geq [max_{j \in J}(V_{ij}) - \lambda_i], V_{ij \notin CS} < [max_{j \in J}(V_{ij}) - \lambda_i]|\Omega_{physician})$$

[37] As outlined above, we allow patients' preferences over quality of service and waiting times to vary with severity and income. Because λ_i plays the role of constraining patients' ability to react on their preferences regarding both waiting times and mortality (see the discussion of identification in Section 5.3 below), we allow the strength of the constraining effect to also vary with the same patient characteristics.

While $Pr(k|CS_k, \Omega_{patient})$ has an analytical expression, the probability of a particular consideration set occurring has no closed-form solution. For this reason, we take draws from the distribution of physician taste shocks ν_{ij} and simulate the resulting choice sets. We denote a set of taste shock draws for consumer i as $(\nu_{i1,s_i}, \nu_{i2,s_i}, ...\nu_{ij,s_i}, ...)$, where s_i is an index that denotes one set of simulation draws for consumer i. Hospital k is contained in the simulated consideration set $\widetilde{CS}_{s_i}$ of patient i (for the set of simulation draws s_i) if

$$\overline{V}_{ik} + \nu_{ik,s_i} \geq max_j(\overline{V}_{ij} + \nu_{ij,s_i}) - \lambda_i$$

For every set of draws, we obtain a different simulated consideration set. Conditional on the simulated consideration set $\widetilde{CS}_{s_i}$, we then use $Pr(k|\widetilde{CS}_{s_i}, \Omega_{patient})$ to determine the choice probability of each hospital in the set. This yields the simulated choice probability for hospital k of

$$\widetilde{Pr^{CON}}_{ik}(\Omega_{patient}, \Omega_{physician}) \quad = \quad \frac{1}{S_i} \sum_{s_i} 1(k \in \widetilde{CS}_{s_i}) Pr(k|\widetilde{CS}_{s_i}, \Omega_{patient})$$

where S_i is the number of draws and $1(\cdot)$ is an indicator function for $k \in \widetilde{CS}_{s_i}$.

Note that $\Omega_{physician}$ influences the choice probability by affecting the probability of a specific consideration set $\widetilde{CS}_{s_i}$ occurring. In the case that hospital k is included in the simulated consideration set, the predicted choice probability is given by $Pr(k|\widetilde{CS}_{s_i}, \Omega_{patient})$ and depends on the identities of the other hospitals included in the simulated set. If hospital k is not contained in the simulated set, the choice probability is equal to zero, which is captured by the indicator function. Averaging across simulation draws within each patient i yields the simulated choice probability.[38]

The model is estimated by using a simulated method of moments estimator where we set

$$\sum_i \sum_j [d_{ij} - Pr_{ij}(\Omega_{patient}, \Omega_{physician})] z_{ij} = 0.$$

In the equation above d_{ij} denotes a dummy variable which is equal to one for the hospital j patient i was referred to and zero otherwise. Pr_{ij} denotes the predicted choice probability ,where $Pr_{ij} = \widetilde{Pr^{CON}}_{ij}$ if patient i was referred pre-reform and $Pr_{ij} = Pr_{ij}^{UNCON}$ if the referral occurred post-reform. z_{ij} is a vector of instruments. In our case, the set of instruments is simply equal to a vector of hospital dummies and characteristics as well as interactions of hospital and patient characteristics. Specifically, z_{ij} contains hospital dummies, distance to the hospital, a dummy for the closest hospital, a within-PCT dummy and the two hospital characteristics: quality of service (mortality rate) and waiting times. The latter two are also included interacted with the severity of the case and income.

[38] In order to avoid discontinuities which occur for the simple frequency estimator described here, we implement a kernel-smoothed frequency estimator. More details are provided in Appendix C.

Finally, all instruments are included twice, interacted with both a pre-reform and a post-reform dummy variable.[39]

5.3 Identification

In this section we cover the sources of identification for the model. We first discuss in detail how we separately identify patient and physician preferences. We then turn to the identification of the specific patient preference parameters on waiting time and quality of service.

Separate Identification of Patient and Physician Preferences

The key source of identification that allows us to separately identify physician and patient preferences is the fact that we observe a change in the process by which consideration sets are formed due to the reform.

It is easiest to think about the logic underpinning our identification strategy by first considering the identification of patient preferences, which is relatively standard because we observe a time-period (post-reform) where choice is liberated, and hence only patient preferences drive choice.[40] This allows us to identify patient preferences from post-reform choice data.[41] Now consider hospital choice in the pre-reform period. If patient choice pre-reform was unconstrained (and if patient preferences are stable over time) then the post-reform estimates should predict hospital choice before the reform. If instead post-reform preferences do not predict referral patterns in the pre-reform time period, it has to be the case that the way in which referrals were made changed over time. Based on the institutional features of the market (see Section 2.2), we capture such differences in referral patterns by assuming that physicians offer patients a limited set of hospitals to choose from prior to the reform.[42]

We assume that these constrained choice sets are formed based on two factors. One is physicians' preferences over hospital characteristics that directly affect the convenience of referrals, and hence their utility. Due to the nature of the contractual arrangements (see Section 2.2 for details) between referring physicians and hospitals in the pre-reform period, we assume that distance to the hospital as well as hospital fixed effects and whether the hospital is in the physician's PCT enter physician utility directly. Distance captures the convenience of referrals and the hospital fixed effects and whether the hospital is in the physician's PCT capture the physician's past experience with a hospital. The preference weights of these characteristics are identified by the extent to which pre-reform referrals are responsive to the respective hospital characteristics.

[39] The only instrument which appears only for the pre-reform period is the PCT dummy, which only enters physician utility.

[40] As discussed above, (Section 2.2) we think of this as physicians having an important role providing information and guiding choice, but acting as patients' agents post-reform, so that choice reflects patients' preferences.

[41] We note that one could estimate the post-reform choice process on its own in the fashion of a standard random coefficient discrete choice model.

[42] We note that this identification strategy is different from other models of consideration set formation, for example Goeree (2008) or Mehta, Rajiv, and Srinivasan (2003), where typically the nature of the constraining process does not change over time.

The second is the constraining parameter λ, which is the extent to which the physician allows patient preferences to influence choice, and is the main driver of consideration set size. This is identified through the influence of the variables unique to patient utility on pre-reform choices. The key aspect that helps us identify λ is therefore an exclusion restriction on waiting times and mortality rates, which only enter patient utility. Thus, if waiting times and mortality rates have any impact on choice pre-reform, this can only be rationalized by patients being less than fully constrained and hence patient preferences affecting choices even before the reform. The extent to which the sensitivity of referrals to both of these hospital characteristics is lower pre-reform relative to post-reform determines the strength of the constraint.

The two key elements that are crucial for identification are the exclusion restriction on waiting times and mortality, which are assumed not to enter physician utility, and the assumption that patient preferences are stable over time. We now discuss the validity of both assumptions.

Exclusion Restrictions

Technically, a necessary requirement for our model to be identified is that at least one variable that enters patient utility is excluded from physician utility. If instead we were to include the identical set of variables in both patient and physician utility this would leave the constraining parameter λ unidentified. To see why this is so, note that the role of λ is to rationalize the change in the sensitivity of referrals to waiting times and mortality. Hence, if we included waiting times and mortality directly in physician utility, then the sensitivity change with respect to these characteristics could be rationalized either by their weights in physician preferences or by a different value of λ. Therefore, λ would not be separately identified.

An alternative way to think about the identification of physician preferences and constraints is to consider the set of instruments used to identify the parameters of the constraining process in the pre-reform period (conditional on having identified patient preferences). As instruments for the pre-reform choice process, we use the same set of hospital characteristics as those that identify preferences in the post-reform period. Given our assumption about the behavior of the physician, hospital dummies and distance respectively serve as instruments for the hospital fixed effects and distance coefficients that enter physician utility directly. Waiting time and quality of service (and their interactions with income and severity) serve as instruments to identify the constraining parameter λ. In the absence of waiting times and quality entering the physician's utility function, the only way to rationalize the observed sensitivity of referrals to those characteristics is through λ.

Economically, we exclude waiting times and mortality rate from physician utility because physician utility should only contain variables that influence the physician directly, i.e., those that affect the convenience of referrals. Any variables that affect the physician because he cares about patients' health outcomes reflect physician agency and affect the physician indirectly. We capture these impacts on physician agency by treating these variables as influencing referrals via a loosening of constraints. In other words, a physician who cares about his patients is modeled as allowing a greater degree of

choice for the patient and hence allowing patient preferences to influence choice.

We note that our setting is considerably more flexible than extant consideration set models, which are estimated for situations where the constraining regime does not change. Without a change in the decision-making process one needs to have a non-overlapping set of variables enter the preferences and constraints respectively. Any variable that enters both stages is only identified by functional form, due to the fact that choice is always driven by parameters in both stages.[43] In our setting instead, we can estimate preferences from post-reform data and then estimate the constraining process separately using the pre-reform data. This enables us to allow for some (but not complete) overlap in the variables that affect consumers' preferences and constraints. Such an overlap is desirable in our setting where we believe that some characteristics (such as distance) matter to both patients and (pre-reform) to physicians and hence should enter the utility functions of both parties. Our approach allow us to estimate such a model without having to rely on functional form for identification of variables entering patient and physician utility.

Finally, we note that we include the "within-PCT" dummy variable only in physician utility, but not in patient utility, based on the institutional features described earlier. While this helps with identification, it is not necessary to have a variable that only influences physician utility.

Stability of Patient Preferences

It is crucial for our identification strategy for patient preferences to be stable over time, because this allows us to recover patient preferences from post-reform data and then attribute any change in referral patterns over time to the constraining process. If instead, patient preferences pre-reform are different from post-reform preferences, then we would have to identify both pre-reform patient preferences and constraints from pre-reform data. We believe that the assumption of stable patient preferences is reasonable in our context but provide additional discussion here.

First, we assume that patient preferences regarding waiting times and mortality do not change over time. This assumption might be violated if the mix of patients seeking treatment changes over time. We do see a modest increase in patient severity over time. However, we allow preferences to be a function of severity. Therefore, as long as the change over time in patient severity is reflected in the observed severity measure, this does not pose a problem.

Second, we assume that patient preferences over unobserved hospital quality (ξ_j) are stable over time. One could imagine that as a consequence of the choice reform, hospitals attempted to attract patients by improving quality along dimensions other than waiting times and mortality, and hence this would lead to a change in patient preferences over unobserved hospital quality. We believe that the scope for such behavior in the UK market was very limited in the period we study. Hospitals did not

[43] For example, Goeree (2008) assumes that advertising (and interactions of advertising with demographics) only enters the constraining process, but not utility. Other variables such as price and physical product characteristics only affect utility but not the consideration set formation. Similar, Mehta, Rajiv, and Srinivasan (2003) model consideration set formation to be a function of whether the product is displayed or featured and familiarity with the store, whereas utility is a function of product dummies and price.

market themselves directly to patients. Published data on performance from the regulatory bodies focused on waiting times and clinical quality. Any changes in unobserved (to the econometrician) quality would have had to be observable to patients or referring physicians in the absence of active marketing activities by hospitals. By contrast, mortality and waiting times, which we do observe, are highly salient indicators. And while mortality may not be known by the patient, it is relatively easy to observe and interpret by the referring physician, who can then communicate this information to the patient.

Endogeneity of Waiting Time and Quality of Service

Finally, the endogeneity of waiting times and mortality in the utility function (4) is a potential concern. First, it is possible that unobservably better hospitals may have longer waiting times because they attract more patients. By increasing aggregate demand, higher unobserved quality from the patient's (ξ_j) or physician's perspective (ζ_j) will lead to longer waiting times, so $Corr(W_{jt}, \xi_j) \neq 0$ (or $Corr(W_{jt}, \zeta_j) \neq 0$) implies that we will be unable to obtain a consistent estimate of the effect of waiting times on hospital choice ($\beta_{w,i}$) without addressing this issue. The issue is very similar to the endogeneity of the price coefficient commonly encountered in the empirical literature in industrial organization. In that context, 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.

Second, a related concern also applies to our measure for quality of service, because hospitals which treat a larger number of cases might also exhibit higher quality. Such a relationship between volume and quality is well established and is likely to also apply to our setting.[44] The volume-quality channel is problematic in our context, because hospitals with higher unobserved quality will attract more patients, which in turn will lead to higher quality and hence a correlation of quality of service (mortality rate) with unobserved quality ξ_j (ζ_j).

In principle these endogeneity problems can be addressed either by using instrumental variables or by controlling for unobserved heterogeneity via fixed effects to absorb the variation in unobserved quality. As there are no obvious good instruments for waiting times and quality of service, we employ a fixed effects approach.[45] Specifically, we estimate a separate hospital fixed effect as part of both

[44] There is a very large literature on the "volume-outcome" relationship in health care. Some papers from that literature are Birkmeyer, Siewers, Finlayson, Stukel, Lucas, Batista, Welch, and Wennberg (2002), Silber, Rosenbaum, Brachet, Bressler, Even-Shoshan, Lorch, and Volpp (2010), and Halm, Lee, and Chassin (2002).

[45] One could consider employing the commonly used strategy of using values of the endogenous variable(s) from another (product or geographic) market. For example, we could consider using waiting times for other procedures at the same hospital as instruments. As a robustness check we implemented such an approach and find that our results are robust to instrumenting CABG waiting times (on top of including hospital fixed effects) with waiting times for other procedures (due to the non-linearity of the demand model, we implement this regression via a control function approach). However, these instruments are not without problems, as unobserved quality may be correlated across procedures. For instance, general hospital reputation might affect demand similarly across procedures. Furthermore, such an IV strategy is harder to implement for mortality rates, since quality measures are more difficult to compare across different procedures. See

patient and physician utility. This allows for unobserved hospital quality to differentially affect the pre- and post-reform periods, since quality effects on physician utility affect choice only in the pre-reform period.[46]

6 Reduced-Form Evidence

Before proceeding to the structural analysis, we look at patterns in the data to provide some simple empirical evidence on whether patients became more responsive to hospital quality after the reform. We start by running a simple linear regression of aggregate market shares on mortality rates to examine the impact of the introduction of choice on the responsiveness of market shares to the mortality rate. This allows us to illustrate some of the main patterns in the data in a simple way. We aggregate the patient-level data to the hospital-quarter level. The mortality rate is also defined at this level. We estimate separate OLS regressions with hospital fixed effects for the pre- and post-reform time periods.

The results are reported in Table 3, columns (1) and (2). These show that pre-reform higher quality hospitals did not have significantly larger market shares. Post-reform, however, a lower mortality rate is significantly associated with a higher market share. This provides 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 regressions of columns (1) and (2) using 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 3, columns (3) and (4), we see that hospital mortality rates have no statistically significant impact on emergency CABG market shares either pre- or post-reform.

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. In the first row of Table 4 we report the average mortality rate a patient faces pre- and post-reform. We find a substantial fall of about 30 percent (0.4 percentage points) in the mortality rate post-reform. 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, or 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

for example Gravelle, Santos, Siciliani, and Goudie (2012). 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 is not a good fit for our situation.

[46] We note that ideally we would want to control for unobserved quality even more rigorously by including a separate fixed effect for each hospital/quarter combination. However, other hospital characteristics, namely waiting times and the mortality rate, are defined at this level and there impact on choice would hence not be identified if hospital/quarter fixed effects were included.

patients who visited the nearest hospital and patients who 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 who visited the nearest hospital. Examining the patterns in Table 4 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 who visit the nearest hospital. In other words, we observe larger declines in mortality for patients who decide not to use their local hospital. Consistent with the results in Table 3, this supports the idea that these patients were not simply lucky that the local hospital improved its quality but, rather, that they sought 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.

7 Estimation Results: Structural Model

We first report the estimation results from our model of choice and constraints, then report how they translate into patient and hospital level elasticities of demand.

7.1 Parameter Estimates

The results from the estimation are reported in Table 5. For economy of exposition, the (large number of) fixed effect estimates are not reported. We find that patients care about distance to the hospital and both of the distance coefficients are highly significant in the patient utility function. The results also show that patients dislike higher mortality rates, i.e., lower quality. The effect is stronger for more severely ill patients as well as for lower income groups. The latter effect, however, is only significant at the 10 percent level. In terms of sensitivity to waiting times, we find an insignificant effect for low severity and low income patients. There is some evidence that higher income households care less about waiting times.[47]

In terms of physician preferences, we find that both distance terms as well as the within-PCT dummy are highly significant. We also find that most patients are severely, in fact fully, constrained before the reform took place. For low severity cases (regardless of income) λ is estimated to be equal

[47]The lack of an effect of waiting times is probably due to the large fall in waiting times that occurred just prior to the full roll out of the choice program in 2006. In 2001 the government instituted an aggressive national policy (dubbed "targets and terror") to reduce waiting times which had considerable success in lowering waiting times before the choice reforms (Propper, Sutton, Whitnall, and Windmeijer 2010) and may have meant that waiting times became less salient to patients. See also Gutacker, Siciliani, Moscelli, and Gravelle (2015) who find no effect of waiting times on demand for hip replacement surgery in England between 2010 and 2013. Our estimated net effect of waiting times for high income patients (i.e. adding the waiting time coefficient and the interaction of waiting times and high income) is positive, but only marginally significant (p-value 0.055). This either means that high income patients prefer longer waits (perhaps in order to arrange their affairs before entering hospital), or may indicate some residual endogeneity.

to zero, which implies that no choice was offered by the physician and patient preferences did not influence referrals. Only for high severity cases do we find a positive coefficient on λ.

7.2 Elasticities of Demand

As 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 post-reform, we focus on the elasticity of demand with respect to the mortality rate. We start by computing elasticities for individual patients with respect to the 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. We then compute hospital-level demand elasticities (by aggregating up changes in individual choice probabilities) to assess the impact on the demand faced by hospitals.

Patient-level Elasticities and Consideration Set Size

The top part of Table 6 reports the sensitivity of choice probabilities to changes in the mortality rate for different groups of patients pre- and post-reform. Bootstrapped standard errors are in parentheses. To simulate patient-level elasticities, we compute a one standard deviation shift in the mortality rate for each hospital and compute the change in choice probabilities entailed by this change for each patient in the relevant time period. We then average the changes across patients and hospitals. We also compute the average consideration set size pre-reform to give a sense of how constrained patients were in their choices. To examine the impact of patient characteristics, we simulate the reactions of all patients pre- and post-reform to a quality change for all four possible permutations of severity and income. (For each simulation we set all patients to have the same characteristics.) This allows us to isolate elasticity differences that are due to patient characteristics from any other factors, such as geographic location, that might be correlated with patient characteristics.

The first column shows that, pre-reform, choice was fully constrained for low severity patients regardless of income. This leads to a complete lack of responsiveness of referrals to quality for this group of patients, as shown in the second column. For high severity cases, irrespective of income, choices were not fully constrained. High severity patients in both income groups were offered an average of 1.61 hospitals. This leads to a non-zero, but small, sensitivity to quality pre-reform. The third column shows that the liberation of choice has a substantial impact for all four demographic groups. Sensitivity changes by 1.2 for low severity and low income cases and by 1.5 for high severity and low income cases. The respective increases for high income cases are of the order of 0.6 and 0.9. Post-reform, higher income households have a lower sensitivity to quality than low income households. Interestingly, in contrast to fears that choice-based reforms harm individuals from lower income groups (see Cookson, Laudicella, and Donni 2013), our analysis suggests that households from more deprived areas benefitted slightly more from the reform.

Hospital-level Elasticities

To assess of the impact of the choice reform on the quality provision by hospitals, the hospital-level elasticities are the most crucial factor. If the demand that hospitals face becomes more elastic with regard to quality, then relaxing the constraints on choice was successful in increasing hospitals' incentives to provide higher quality. We examine hospital-level demand sensitivity by simulating a one standard deviation change in mortality for each hospital in the choice set and computing the percentage change in the hospital's market share. The responsiveness to a change in mortality differs across hospitals as a function of the density of patients in the local area, the demographic composition of the local population, and the locations of other hospitals. The lower panel of Table 6 reports the distribution of elasticities across all hospitals.

The first column shows that when constraints on choice are relaxed post-reform a one standard deviation increase in the mortality rate leads to a 4.46 percent drop in market share for the average hospital. This compares to a much smaller decrease of 0.82 percent before the reform. The standard errors indicate this change is statistically significant. The highest quartile of the elasticity distribution pre-reform lies below the lowest post-reform. The distribution of hospital elasticities also shows substantial heterogeneity in the impact of the reform across hospitals. The additional market share loss after the reform is 4.37 at the 25th percentile and 2.04 at the 75th percentile of the distribution of elasticity changes.

Overall, the results suggest that relaxing the constraints on choice substantially increased hospitals' incentives to improve quality. In percentage terms, demand at the average hospital became over five times more responsive to quality. While the magnitude of the elasticity is not especially large in absolute terms, the reform led to a large increase in demand responsiveness from what had been a very low level. Further, there is large heterogeneity in the effect: many hospitals experienced substantial changes in the demand elasticities they faced.

8 Policy Evaluation

We provide an evaluation of the impact 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 and analyze the consumer welfare gains due to the relaxation of choice constraints. These calculations evaluate effects of the reform under the assumption that hospitals did not react to the change in demand conditions, so the survival and welfare gains are purely due to sorting of patients across hospitals. We then proceed to an analysis of how much the competitive environment changed with the introduction of the reform. Finally, we provide 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. Contrary to many other applications in the empirical industrial organization literature, we do not simulate changes caused by a hypothetical policy, but instead simulate behavior for the post-reform

population under the assumption that the reform had not taken place. This allows us to 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 the policy is the impact on the probability of survival following a CABG. 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 of any particular patient i in time-period t is given by

$$\begin{aligned} E(Mortality_i^{UNCON}) &= \sum_j Pr_{ij}^{UNCON}(\Omega_{patient}) \cdot E(Mortality_i|Choice = j) \\ E(Mortality_i^{CON}) &= \sum_j Pr_{ij}^{CON}(\Omega_{patient}, \Omega_{physician}) \cdot E(Mortality_i|Choice = j) \end{aligned}$$

where the two rows define the mortality probability under unconstrained and constrained choice respectively.[48] Note that the two expressions differ only in the choice probability. The first term in both equations denotes the probability of visiting hospital j, which can be computed from the demand model estimates. For the case of unconstrained choice this is determined by patient preferences. In the constrained case physician preferences together with patient preferences influence choice. $Mortality_i$ denotes an indicator variable which is equal to one if the patient dies during the surgery. The second term in both lines denotes the conditional mean of this variable, which is equal to the hospital-specific mortality rate.[49]

To obtain the expected difference in mortality across all patients we compute

$$E(\Delta Mortality_{total}) = \sum_{i \in PostReform} \left[E(Mortality_i^{CON}) - E(Mortality_i^{UNCON})\right]$$

In other words, we sum the changes in mortality probability for each patient in the post-reform period when choice is constrained relative to when there are no constraints. The latter constitutes the actual state of the world post-reform while the former is a counterfactual scenario in which the reform never happened and constraints are still in place.

[48] 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 yet know the patient-specific shock to the mortality outcome.

[49] The hospital-specific mortality rate is identical the mortality variable used as quality indicator Z_{jt} in the demand model.

The results are reported in Table 7. We estimate that 4.2 fewer patients would have survived had the reform not been implemented in 2005. This number is calculated over the entire five post-reform quarters used in the estimation and corresponds to 3.3 lives saved on a annual basis. The changes amount to about 0.02 percentage points or a 3 percent decrease in the mortality rate in the relevant time period. 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 $5.6 million per year in terms of value of life-years saved.[50]

8.2 Changes in Patient Welfare

Next we compute the welfare changes due to the removal of restrictions on choice. We simulate a post-reform scenario where the constraints are still in place as the counterfactual. The comparison with the unconstrained choice scenario allows us to quantify the welfare effect of the reform on the post-reform pool of patients. Expected consumer surplus (in utils) for consumer i when choice is unconstrained can be expressed using the standard formula:

$$E(Surplus_i^{UNCON}) = E[max_{j \in J}(\overline{U}_{ij} + \varepsilon_{ij})]$$

For the case of constrained choice utility is still given by the same patient utility function but choices are now determined by both patient preferences and the constraining forces of the physician's influence on the referral. Expected utility is equal to:

$$E(Surplus_i^{CON}) = E[max_{j \in CS_i}(\overline{U}_{ij} + \varepsilon_{ij})]$$

This expression differs from the unconstrained case only in the choice rule: the chosen hospital is the utility maximizing hospital from the constraint set CS_i rather than the full set J.[51]

Due to the iid extreme value assumption on the error term (see Small and Rosen 1981, Train 2003) we can re-write the surplus expressions above (up to an arbitrary constant) as a logit-inclusive value:

$$\begin{aligned} E(Surplus_i^{UNCON}) &= \ln \sum_{j \in J} \exp(\overline{U}_{ij}) \\ E(Surplus_i^{CON}) &= \ln \sum_{j \in CS_i} \exp(\overline{U}_{ij}) \end{aligned}$$

The welfare calculation for the unconstrained case is standard and the expression above can be

[50] $3.3 \times 17 \times 100,000 = 5,610,000$

[51] As before, J denotes the set of all CABG performing hospitals in the UK and is therefore not patient-specific. The constraint set CS_i instead is specific to patient i.

computed directly from the patient utility function conditional on the estimated model parameters (see, e.g., Nevo 2003, Ho 2006). The constrained case is more difficult to compute because the consideration set CS_i is unobserved. We therefore simulate consideration sets conditional on our parameter estimates (in a similar manner as we simulated consideration sets in the estimation). Specifically, we simulate the physician taste shocks (ν_{ij}) that determine which hospitals are included in CS_i. For a given set of draws, we then compute the associated consideration set and the expected surplus derived from choice from this set. Averaging over draws allows us to compute the expected surplus for each patient i.

The average change in surplus per patient is simply equal to the difference between the two expressions above, averaged across patients:

$$E(\Delta Surplus_i) = \frac{1}{\#Patients_{PostReform}} \sum_{i \in PostReform} E(Surplus_i^{UNCON}) - E(Surplus_i^{CON})$$

where $\#Patients_{PostReform}$ denotes the total number of referrals in the post-reform time-period.

We find that the freeing of choice led to an average increase of 1.04 units in expected utility.[52] Since there is no price mechanism in this market (and therefore no price coefficient in the demand model) we cannot directly translate the welfare change from utils into a dollar value. However, we can express the gain in terms of the hospital characteristics in the utility function. We therefore translate the welfare gain into units of mortality and then monetize the implied mortality rate reduction using the value of a statistical life (using the same value as previously). This calculation yields a welfare effect of $300,900 per person.[53]

Relative to the change in survival reported above, the welfare calculation provides a much broader assessment of the impact of the reform on patients. All components of consumer preferences including quality, waiting times, distance, as well as hospital fixed effects and idiosyncratic patient taste shocks are suppressed by the pre-reform constraints. Hence when constraints are removed, the patient's ability to choose based on her preferences leads to an increase in welfare. To further understand the source of the welfare change, we proceed to decompose the impacts of the different utility components that are suppressed by the choice constraints. We implement the decomposition by imposing counterfactual constraints to patient choice that are aligned or misaligned with patient preferences along various dimensions in order to isolate the impact of freeing up choice on the different components of patient preferences.

[52] Our calculations are for impacts on consumer welfare alone, not social welfare. Nonetheless, we note that these calculations assume the same amount of spending for CABGs under the counterfactual as the actual post-reform patterns. This is likely to be the situation. Since there is no difference between the actual post-reform case and the counterfactual in the number of CABG cases, and since the fixed payment per CABG case is the same across hospitals, there is no difference in total spending between the actual post-reform case and the counterfactual.

[53] Due to heterogeneity in preferences over mortality, we calculate the utility gain from the reform for each patient separately and then divide by (the negative of) the relevant mortality coefficient. This yields an average per patient gain in units of mortality (measured between 0 and 1) of 0.177, averaged over all individuals. Multiplying by the value of a statistical life ($100,000 per life year $\times$ 17 additional years of life) gives us $0.177 \times 100,000 \times 17 = 300,900$.

In a first step we take the estimated physician utility function, but replace the hospital fixed effects in his utility function with the patient's set of hospital fixed effects. We then estimate the welfare impact of post-reform patients being constrained by a physician with these preferences. Relative to the previous welfare comparison, the welfare decrease from imposing the constraints has to be smaller now, because the constraints are less "hurtful" to the patient due to the fact that patient and physician utility are more aligned and hence the physician is more likely to offer a hospital that the patient would have chosen herself. Comparing the welfare change due to the removal of constraints in this case with our baseline welfare change calculation, we find that the change is 17 percent smaller. Hence 17 percent of the welfare change is due to the patient being able to react to her preference over hospitals' unobserved quality.

Next, we further align physician and patient preferences by endowing the physician with the same preferences over distance, waiting time, mortality and unobserved quality as the patient. Thus the only remaining difference between the two parties are their idiosyncratic taste shocks. Aligning the deterministic part of utility in this way further reduces the welfare change from removing the constraints by 11 percent. The remaining (residual) difference in welfare with and without constraints constitutes 72 percent of the total welfare change and is due to the patient's ability to respond to her idiosyncratic random taste shocks. Idiosyncratic taste shocks could be, for example, whether a hospital is located near to a patient's family members, or if a patient particularly values being treated by a surgeon at a particular hospital.

8.3 Change in the Competitive Environment

The welfare analysis and the change in survival calculate only the changes that are achieved by reallocating patients, i.e. they do not take into account any supply side adjustments by hospitals to the new demand conditions. We now examine the further improvements that could be achieved if the reform also provided incentives for hospitals to improve quality.

First, we undertake a counterfactual calculation 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 without constraints, i.e., using the estimated post-reform parameters. This allows us to compute how much re-shuffling of market shares would have happened had patients had free choice earlier. When implementing this 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 actual pre-reform choice situation. We do not allow hospitals to adjust to the changes in demand caused by the removal of constraints. Holding hospital quality fixed is helpful because any movement we see in market shares in actual post-reform referrals will be due to both demand changes and hospitals' responses to these. Our counterfactual allows us to isolate the former effect to assess the pressure on hospitals from the reform, for a given quality level. The magnitude of the re-shuffling of market shares is therefore a valuable metric of how much incentives to improve quality changed for hospitals. It should also be noted that due to the absence of an outside option the simulated changes in market shares have to cancel out across hospitals. We are

therefore quantifying a re-allocation of a given set of patients.

To describe this, we report the change in market share at various percentiles of the distribution of changes in the middle panel of Table 7. We find that the introduction of choice had a significant impact on many hospitals. At the 25th percentile of the distribution hospitals would have experienced a roughly 15 percent decrease while at the 75th percentile hospitals experienced around a 15 percent increase in market-share. At the mean and median, the change is close to zero due to the fact that the changes in market-shares across hospital sum to zero.

8.4 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. We expect that hospitals in areas where demand became more elastic to improve their quality more than other hospitals. We test this hypothesis by regressing the change in the mortality rate on the change in hospitals' elasticity of demand with respect to quality.

This approach mirrors the difference-in-difference estimation conducted in Gaynor, Moreno-Serra, and Propper (2013) 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.[54] 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, constitutes a reduced-form way of capturing that the elasticity of demand is expected to change relatively more in high concentration areas. Here we are instead able to compute demand responsiveness directly from the model estimates, rather than having to use hospital concentration as a proxy. We are hence able to use a more direct measure of the change in competitive environment than the previous literature.

We use the observations on the 27 hospitals that are present in all periods of the data and use the change in demand responsiveness reported in the lower panel of Table 6 as the regressor.[55] We estimate the following OLS regression:

$$\Delta Mortality_j = \phi_0 + \phi_1 \Delta Elasticity_{j,Mortality} + e_j$$

where $Elasticity_{j,Mortality}$ denotes the percentage change in market share for hospital j when the mortality rate is increased by one standard deviation. For ease of interpretation, we use the absolute value of the elasticity in the regression. We note that, due to the differenced nature of the regression, any time-invariant factor that might differ across hospitals does not pose a threat to a causal interpretation. However, any time-varying factor that is correlated with the change in the demand elasticity could lead to a bias in the estimation. While there is no other obvious change over time that might correlate

[54] For the most part different versions of an HHI index are used in these papers. However both papers show robustness to a host of definitions of the measure of market structure.

[55] There were two new entrants and one merger over our period. See footnote 27 for details.

with the change in the competitive environment (see also the discussion in Gaynor, Moreno-Serra, and Propper (2013) and Cooper, Gibbons, Jones, and McGuire (2011)) we cannot fully rule out such confounds.

The results are reported in the bottom panel of Table 7. We find a negative and significant impact of the change in the demand elasticity on the change in the mortality rate. In other words, hospitals whose demand became more responsive to quality improved quality disproportionately more than other hospitals (by lowering the mortality rate). To get a sense of the magnitude of the coefficient, consider the change in the demand elasticity for the median hospital, 3.09, as reported in Table 6. This shift implies a drop of 1.01 in the mortality rate. This estimation of the effect of the reform on mortality at the median hospital is slightly larger than improving quality by one standard deviation in the across-hospital distribution and suggests that freeing up patient choice elicited a supply side response by hospitals that improved patient survival. In terms of magnitude, our point estimate implies that competition could have played a significant role in the overall drop in the CABG mortality rate from 2003 to 2007.[56]

We also undertake an exercise to illustrate what the welfare effect of quality improvement due to supply side response could be. We do this by simulating choice under the scenario that mortality rates had not changed due to the reform. In order to obtain counterfactual mortality rates in the absence of the reform, we assume that the change in mortality for each hospital is equal to the intercept in the regression above, ϕ_0. We hence assume that without the reform, no change in elasticity would have taken place and therefore $\Delta Elasticity_{j,Mortality} = 0$, which implies that the expected change in mortality does not depend on the slope parameter ϕ_1. We then simulate the change in welfare when applying constraints to the post-reform patients (as we did in Section 8.2) *and* setting the mortality rates to the higher counterfactual levels. We find a decrease in welfare that is 8 percent larger than the welfare change computed earlier from only the removal of constraints.[57] This is provides evidence that there is a further component of welfare improvements due to the supply-side reaction to the reform.

While these results are interesting, they do not come from a formal model of supply-side behavior and are based on only 27 observations. We therefore regard them as suggestive. To fully explore the supply response we would need to estimate a fully specified structural model including the supply side: we leave this for future research.

9 Summary and Conclusions

This paper takes advantage of a "natural experiment" in the English National Health Service that introduced patient choice among hospitals to examine the effect on patient behavior and supplier responses to that behavioral change. We evaluate whether increased choice resulted in increased

[56] We note that a volume-outcome effect (as discussed earlier) would lead the estimates to go the other way, since high mortality hospitals are those with the largest changes in their elasticities and the biggest improvement in quality. Therefore to the extent there is a volume-outcome effect, we may underestimate the supply response to the reform.

[57] The welfare calculation in Section 8.2 was based on a simulated removal of constraints, holding mortality rates constant.

elasticity of demand faced by hospitals with regard to two central dimensions of hospital service: clinical quality of care and waiting times. Using detailed patient-level data, we estimate a structural model of patient choices and constraints. On the methodological side we show how to explicitly model the choice constraints imposed by the pre-reform referral system.

We find substantial impacts of the removal of restrictions on patient choice. Patients are more responsive to the clinical quality of care at hospitals. Most patient groups are not more responsive to waiting times. There is, however, heterogeneity in these impacts. The more severely ill and those from low income areas benefit more from the removal of constraints. This increased demand responsiveness alone led to a reduction in mortality and an increase in patient welfare. The elasticity of demand faced by hospitals also increased post-reform. This gave hospitals incentives to improve their quality of care and we find 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. And, in contrast to fears that these pro-choice reforms would only benefit the better off, we find no evidence of this. This suggests that there is potential for choice based reforms to succeed and for competition in health care to enhance quality.

References

Andrews, R. L., and T. C. Srinivasan (1995): "Studying Consideration Effects in Empirical Choice Models Using Scanner Panel Data," *Journal of Marketing Research*, 32(1), 30–41.

Baker, L. C., M. K. Bundorf, and D. P. Kessler (2015): "The Effect of Hospital/Physician Integration on Hospital Choice," Working Paper 21497, National Bureau of Economic Research.

Beckert, W. (2015): "Choice in the Presence of Experts," Birkbeck Working Papers in Economics and Finance 1503, Birkbeck College, University of London.

Beckert, W., M. Christensen, and K. Collyer (2012): "Choice of NHS-funded Hospital Services in England," *The Economic Journal*, 122(560), 400–417.

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.

Birkmeyer, J. D., A. E. Siewers, E. V. Finlayson, T. A. Stukel, F. L. Lucas, I. Batista, H. G. Welch, and D. Wennberg (2002): "Hospital Volume and Surgical Mortality in the United States," *New England Journal of Medicine*, 346, 1128–1137.

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

Bronnenberg, B. J., and W. R. Vanhonacker (1996): "Limited Choice Sets, Local Price Response and Implied Measures of Price Competition," *Journal of Marketing Research*, 33(2), 163–173.

Capps, C., D. Dranove, and M. Satterthwaite (2003): "Competition and Market Power in Option Demand Markets," *RAND Journal of Economics*, 34(4), 737–763.

Charlson, M. E., P. Pompei, K. L. Ales, and C. MacKenzie (1987): "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," *Journal of Chronic Diseases*, 40(5), 373 – 383.

Cookson, R., M. Laudicella, and P. L. Donni (2013): "Does Hospital Competition Harm Equity? Evidence from the English National Health Service," *Journal of Health Economics*, 32(2), 410–422.

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

Dafny, L., K. Ho, and M. Varela (2013): "Let Them Have Choice: Gains from Shifting Away from Employer-Sponsored Health Insurance and toward an Individual Exchange," *American Economic Journal: Economic Policy*, 5(1), 32–58.

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., K. Ho, and R. J. Town (2015): "The Industrial Organization of Health-Care Markets," *Journal of Economic Literature*, 53(2), 235–84.

Gaynor, M., R. Moreno-Serra, and C. Propper (2013): "Death by Market Power: Reform, Competition, and Patient Outcomes in the National Health Service," *American Economic Journal: Economic Policy*, 5(4), 134–66.

Gaynor, M., and R. J. Town (2012): "Competition in Health Care Markets," in *Handbook of Health Economics*, ed. by T. G. McGuire, M. V. Pauly, and P. Pita Barros, vol. 2, chap. 9, pp. 499–637. Elsevier North-Holland, Amsterdam and London, 2012.

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.

Goeree, M. S. (2008): "Limited Information and Advertising in the US Personal Computer Industry," *Econometrica*, 76(5), 1017–1074.

Gowrisankaran, G., A. Nevo, and R. Town (2015): "Mergers When Prices Are Negotiated: Evidence from the Hospital Industry," *American Economic Review*, 105(1), 172–203.

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

Gravelle, H., R. Santos, L. Siciliani, and R. Goudie (2012): "Hospital Quality Competition under Fixed Prices," Research Paper 80, University of York Center for Health Economics.

Gutacker, N., L. Siciliani, G. Moscelli, and H. Gravelle (2015): "Do Patients Choose Hospitals That Improve Their Health?," Research Paper 111, University of York Center for Health Economics.

Halm, E. A., C. Lee, and M. R. Chassin (2002): "Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature," *Annals of Internal Medicine*, 137(6), 511–520.

Ho, K. (2006): "The Welfare Effects of Restricted Hospital Choice in the US Medical Care Market," *Journal of Applied Econometrics*, 21(7), 1039–1079.

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

Honka, E. (2014): "Quantifying Search and Switching Costs in the U.S. Auto Insurance Industry," *RAND Journal of Economics*, forthcoming.

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.

Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2010): "Online Demand under Limited Consumer Search," *Marketing Science*, 29(6), 1001–1023.

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.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*, 57, 995–1026.

Mehta, N., S. Rajiv, and K. Srinivasan (2003): "Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation," *Marketing Science*, 22(1), 58–84.

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.

Nevo, A. (2003): "New Products, Quality Changes, and Welfare Measures Computed from Estimated Demand Systems," *Review of Economics and Statistics*, 85(2), 266–275.

Propper, C. (2012): "Competition, Incentives and the English NHS," *Health Economics*, 21(1).

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.

——— (2010): "Incentives and Targets in Hospital Care: Evidence from a Natural Experiment," *Journal of Public Economics*, 94(3-4), 318–335.

Roberts, J. H., and J. M. Lattin (1991): "Development and Testing of a Model of Consideration Set Composition," *Journal of Marketing Research*, 28(4), 429–440.

Santos, R., H. Gravelle, and C. Propper (2015): "Does quality affect patients' choice of Doctor? Evidence from England," *The Economic Journal*, forthcoming.

Seiler, S. (2013): "The Impact of Search Costs on Consumer Behavior: A Dynamic Approach," *Quantitative Marketing and Economics*, 11(2), 155–203.

Silber, J. H., P. R. Rosenbaum, R. N. Brachet, T. J.and Ross, L. J. Bressler, O. Even-Shoshan, S. A. Lorch, and K. G. Volpp (2010): "The Hospital Compare Mortality Model and the Volume-Outcome Relationship," *Health Services Research*, 45, 1148–1167.

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

Small, K. A., and H. S. Rosen (1981): "Applied Welfare Economics with Discrete Choice Models," *Econometrica*, 49(1), 105–130.

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

Train, K. E. (2003): *Discrete Choice Methods with Simulation*. Cambridge University Press, New York.

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[1]

	Total Admissions		Waiting Times (Days)		Mortality Rate	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
2003	502.9	189.4	109.1	32.1	1.88	0.82
2004	507.5	200.0	100.5	20.7	1.93	0.78
2005	449.1	170.8	67.8	15.2	1.90	0.56
2006	425.4	172.7	65.6	17.3	1.95	0.79
2007	459.9	169.9	64.9	21.4	1.51	0.69

[1]The table reports descriptive statistics for all hospitals performing CABGs from 2003 to 2007. To compute the columns in the table, the hospital-year level values of the variables are calculated. The means and standard deviations are based purely on between-hospital variation within each year.

Table 2: Descriptive Statistics — Patient Characteristics

	Mean	Median	Standard Deviation	10th Percentile	90th Percentile
Age	65.76	66	55.04	53	76
Fraction Male	81.18%				
Index of Multiple Deprivation	0.14	0.11	0.12	0.04	0.31
Comorbidity Count	5.42	5	2.81	2	9
Charlson Index	0.55	0	0.71	0	2
Distance Pre-reform	34.93	22.34	44.97	4.77	71.40
Distance Post-reform	32.24	22.91	32.94	4.93	70.58

Table 3: Reduced-Form Evidence: Regressions using Aggregate Market-shares[1]

	(1)	(2)	(3)	(4)
Dependent Variable	Elective CABGs Market-share		Emergency CABG Market-share	
Time Period	Pre-reform	Post-reform	Pre-reform	Post-reform
Mortality Rate Coefficient	0.005 (0.046)	-0.188*** (0.033)	0.049 (0.065)	-0.027 (0.053)
Hospital Fixed Effects	Yes	Yes	Yes	Yes
Observations	142	143	142	143
Hospitals	29	29	29	29
Quarters	5	5	5	5

[1]*** denotes statistical significance at the 1 percent level, ** denotes statistical significance at the 5 percent level, * denotes statistical significance at the 10 percent level.

Table 4: Reduced-Form Evidence: Changes in the Expected Mortality Rate

Sample	Mean Mortality Rate Pre-reform	Mean Mortality Rate Post-reform	Difference in Means
All Patients	1.330	0.935	-0.395
Patients Visiting the Nearest Hospital	1.276	1.027	-0.249
Patients Not Visiting the Nearest Hospital	1.445	0.735	-0.711

Table 5: Structural Parameter Estimates[1]

		Coefficient	Standard Error
Patient Preferences	Distance	-6.983***	0.211
	Closest Hospital Dummy	1.341***	0.052
	Mortality Rate	-7.883***	2.229
	Mortality Rate * High Severity	-5.419**	2.467
	Mortality Rate * High Income	3.832*	2.320
	Waiting Times	-1.528	1.887
	Waiting Times * High Severity	-1.584	1.140
	Waiting Times * High Income	6.262***	1.196
Physician Preferences	Distance	-4.985***	0.207
	Closest Hospital Dummy	1.734***	0.110
	Within-PCT Dummy	1.309***	0.308
Choice Constraint Parameters	Constant	0.000	0.119
	High Severity	1.011***	0.178
	High Income	0.000	0.113

[1]*** denotes statistical significance at the 1 percent level, ** denotes statistical significance at the 5 percent level, * denotes statistical significance at the 10 percent level.

Table 6: Sensitivity of Demand with Respect to Quality[1]

Patient-level Sensitivity (by characteristics)	Consideration Set Size (Pre-reform)	Sensitivity to Quality Pre-reform	Sensitivity to Quality Post-reform
Low Severity, Low Income	1	0	-1.209
	(0.037)	(0.041)	(0.317)
Low Severity, High Income	1	0	-0.637
	(0.056)	(0.035)	(0.272)
High Severity, Low Income	1.611	-0.486	-1.972
	(0.110)	(0.090)	(0.354)
High Severity, High Income	1.611	-0.354	-1.438
	(0.108)	(0.083)	(0.323)

Hospital-level Sensitivity	Mean	S.D.	25th Perc.	Median	75th Perc.
Pre-reform	-0.82	0.65	-1.33	-0.56	-0.30
	(0.17)				
Post-reform	-4.46	2.57	-6.53	-3.69	-2.38
	(0.70)				
Change	-3.50	1.97	-4.37	-3.09	-2.04
	(0.60)				

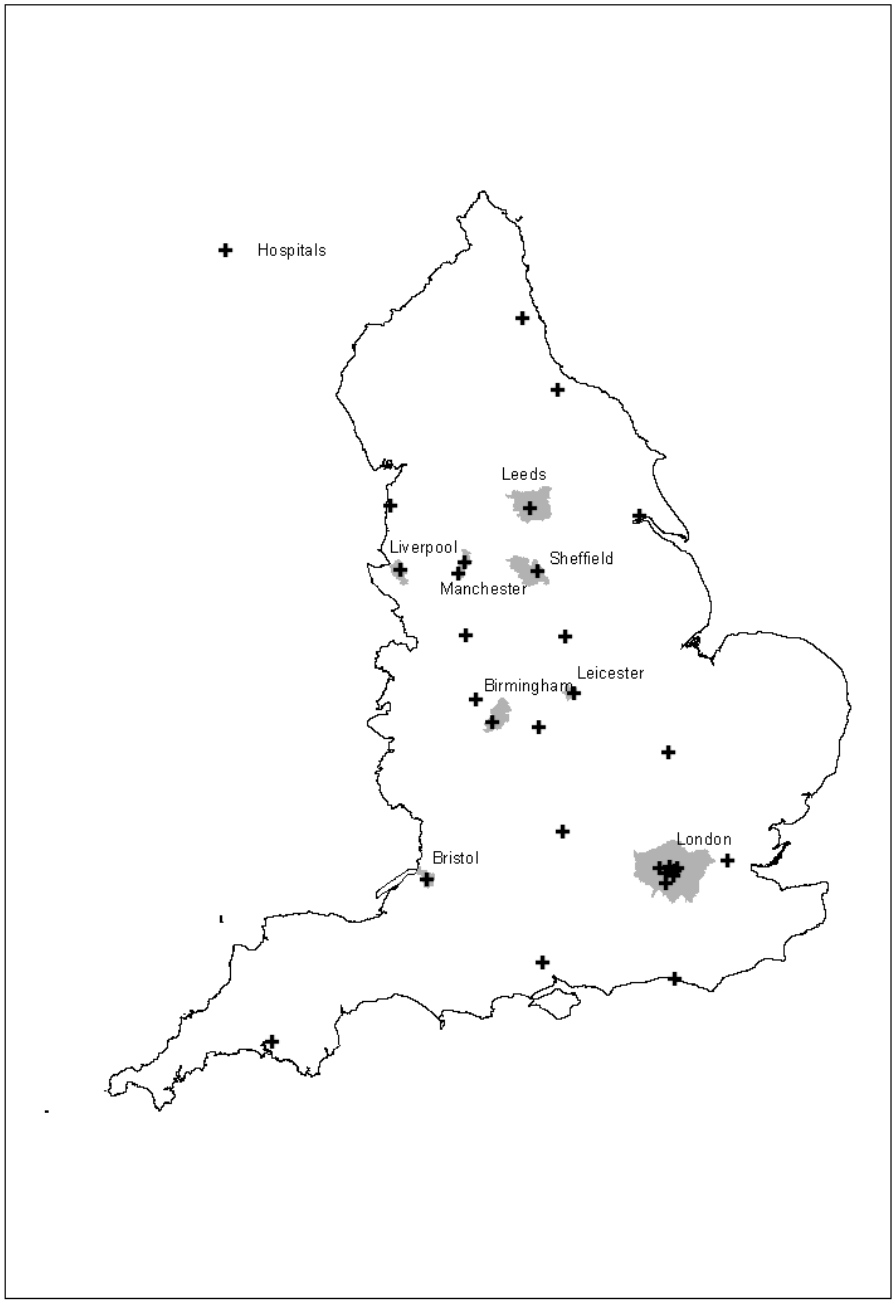
[1]The top panel reports the pre-reform consideration set size and the responsiveness of demand at the patient-level with respect to the mortality rate. The values reported in the second and third column represent the average percentage change in the choice probability when a hospital increases the mortality rate by one standard deviation. The bottom panel reports the distribution of percentage changes (across all hospitals) in market share when a hospital increases the mortality rate by one standard deviation. Bootstrapped standard errors are reported in parentheses.

Table 7: Policy Evaluation[1]

Impact on Patient Survival	Change in Survivals when Post-reform Choices are Constrained				-4.17
	Post-Reform (5 quarters)	Admissions			14,968
		Deaths			140
		Mortality Rate			0.94
		Recomputed Mortality Rate Under Constraints			0.96
Percentage Change in Market Shares Due to the Reform	Mean	S.D.	25th Perc.	Median	75th Perc.
	-3.77	22.83	-15.92	2.14	13.49
Supply-side Response	Dependent Variable				Change in Mortality Rate
	Change in the Elasticity of Demand with Respect to the Mortality Rate				-0.328** (0.128)
	Observations				27

[1]The top panel reports the change in the number of survivals when constraints are removed. The middle panel shows the changes in market-shares across hospitals for the counterfactual scenario of an earlier removal of constraints. This entails a zero-sum game of market-share reshuffling between hospitals. The distribution of changes across hospitals is reported. The lower panel reports results from an OLS regression of a change in mortality on the change in the elasticity of demand (derived from the demand model).

Figure 1: Map of Hospital Locations



Appendix

A Timing of the Choice Reform

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 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.

B Mortality Rate Adjustment

B.1 Regression Framework

In this section we describe a framework for risk adjusting hospital mortality rates, and assess whether risk adjustment has an impact on our estimation. To do this, we specify a linear probability model of patient mortality in which we regress an indicator for whether the patient died after the surgery on a set of hospital-time period fixed effects. Let the mortality of patient i in period t at hospital j be determined as follows,

$$M = JT\psi + (\gamma H + \eta) \tag{9}$$

where the last two terms in parentheses denote two components of the econometric error term. 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 ψ is a vector of coefficients. We define a time period to be a quarter. H represents the patient's health status. η is a vector of iid error terms.

We use the regression above 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 necessarily allow us to uncover the true causal relationship. Hospital choice will likely be correlated with patient health status, which will be subsumed in the empirical error term. If sicker patients choose systematically different hospitals, health status is predictive of which hospital dummy is "switched on." Therefore, any arbitrary column of the hospital dummy matrix, JT will correlated with the error term. This endogeneity problem is very closely related to the fact that the hospital's mortality rate is "contaminated" by differences in patient case-mix. In fact, when running the above regression by OLS, the fitted hospital fixed effects $\widehat{\psi}$ will be equal to the hospital/quarter-specific mortality rates.[2] The linear probability regression therefore recasts the issue of case-mix affecting the mortality rate as an endogeneity problem.

In order to uncover the causal effect on the mortality from visiting a specific hospital, we need to instrument the hospital dummies JT, which we do using distance to the hospital (D_{ij}). 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). The identifying assumption that allows us to obtain a causal effect on patient survival is the exogeneity of patients' locations with respect to their health status. Under this assumption, we can use distance as an instrument for choice. It is a relevant instrument in the sense that distance is highly predictive of choice, as our demand model estimates show. It is a valid instrument under the assumption that patients' health status is not correlated with their locational choice and in particular with the distance of where they live to different hospitals that offer CABGs. Formally, in the actual IV regression this corresponds to the assumption that distance (to each hospital) is uncorrelated with the error term of the mortality regression (i.e. patient health status or the patient-specific probability of survival independent of hospital choice). The estimated hospital fixed effects in the IV regression can be interpreted as a case-mix adjusted mortality rate due to the fact that the instruments remove any impact of patients selecting according to severity.

One might also wonder whether the recovered LATE effect of the IV regression constitutes the relevant quality measure. This might be a concern if the LATE was very different from the average treatment effect. In our case, the LATE might differ from the average treatment effect if hospitals' ability to perform a CABG differs across patient types. For instance, it is conceivable that visiting a high quality hospitals makes more of a difference to the survival probability for a difficult case relative to a more standard one. The question then is whether distance has a differential effect on choice for different parts of the severity distribution, which is something we can test using the hospital choice data used to estimate the demand model. In order to explore whether the LATE is likely to differ from the average treatment effect we run a simple multinomial-logit regression using as covariates distance as well as an interaction of distance with the severity of the case. We run several different specifications

[2] 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 deaths divided by the total number of admissions, i.e. the mortality rate.

using either linear distance or a closest hospital dummy or both as measures of travel distance. In all cases we find that severity does not alter the coefficient on distance in the utility function that predicts hospital choice. The results are consistently insignificant and of small magnitude. Based on these regressions we conclude that the recovered LATE effect is likely to be very similar to the average treatment effect.

B.2 Test for Exogeneity

Our primary focus of the IV regression is to establish whether case-mix differences did play an important role and hence generated an endogeneity problem in our linear regression model. The IV regression provides us with a straightforward test for this. Specifically, when running the regression as OLS without instrumenting, the estimated values of hospital/quarter fixed effects $\widehat{\psi}$ will be equal to the hospital/quarter-specific (unadjusted) mortality rates. Instead, the IV allows us to isolate the contribution of the hospital to the probability of patient survival independent of case-mix. Therefore, if case-mix did differ across hospitals, the estimated coefficient on the hospital dummies will be significantly different between the OLS and IV regression.

Before testing for differences between the OLS and IV estimates, we first establish the strength of our first stage. As a simple measure for the strength of the instruments, we compute the F-statistic for each of the individual first stage regressions (each hospital/quarter dummy is an endogenous regressors, hence there are 284 first stages for each of the 285 hospital/quarter dummies minus a constant). Doing so, we find an average F-statistic across all first stage regressions of 160.9 and a median F-statistic of 100.7 (the lowest F-statistic still has a relatively large value of 15.6). This is unsurprising, as our demand model clearly shows that distance is a strong predictor of hospital choice and this is reflected in a strong first stage in the case-mix adjustment regression.

Finally and most importantly, we run a Durbin-Wu-Hausman test to test for endogeneity of the hospital dummies. We fail to reject the null hypothesis of the test, that the hospital dummies are exogenous. We find an F-stat of 0.9735 and corresponding p-value of 0.6138. In other words, the OLS and IV estimates are not statistically different from each other and hence the case-mix adjusted mortality rate (obtained from the IV-regression) and the unadjusted rate (obtained from the OLS) are not significantly different from each other. We also note that distance is a strong predictor of hospital choice and it is hence not the case that we fail to reject the null due to a lack of statistical power.

As a final robustness check, we estimate our demand model with the adjusted mortality rate rather than the unadjusted one. The results from this regression are reported in Table (D2) of the appendix. The point estimates on mortality as well as other variables are very similar to our baseline specification using the unadjusted mortality rate. This is perhaps unsurprising and simply another way to confirm that the two mortality rates are not very different from each other.[3]

[3]Note that the coefficient estimate on mortality is slightly smaller when using the adjusted mortality rate. However, the case-mix adjusted mortality rate has a higher standard deviation (1.46 times larger standard deviation than the raw mortality rate) and in terms standard deviations of the underlying variable, the point estimates are therefore more similar.

C Kernel-smoothed Frequency Estimator

The simple frequency estimator described in the main text suffers from discontinuities in the likelihood. Specifically, in our context, a small movement in a parameter which influences the consideration set process might not alter the simulated consideration set for any consumer / simulated draw. In this case the choice probabilities and the corresponding likelihood are unchanged. In order to smooth the choice probabilities we employ a kernel-smoothed frequency estimator (McFadden (1989)). Note that the smoothing only happens at the "upper-level" of the consideration set process. Patient choices (conditional on a given consideration set) do not need to be simulated and therefore do not suffer from discontinuities.

In order to implement the smoothing, we compute for each draw the simulated consideration set $\widetilde{CS}_{s_i}$ (as defined in the main text) as well as a secondary consideration set $\widetilde{CS^+}_{s_i}$ that contains one additional option. Specifically, we add the highest utility option among the hospitals *not included* in the consideration set to form this secondary set. We then compute the choice probabilities conditional on both the main and the secondary simulated consideration set and take a weighted average between the two. The choice probability for any given set of draws s_i (which enter physician utility and therefore the consideration set process) is thus given by

$$\begin{aligned}\widetilde{Pr^{CON}}_{s_ik}(\Omega_{patient},\Omega_{physician}) &= 1(k\in\widetilde{CS^+}_{s_i})[w(\Omega_{physician})Pr(k|\widetilde{CS}_{s_i},\Omega_{patient}) \\ &\quad +(1-w(\Omega_{physician}))Pr(k|\widetilde{CS^+}_{s_i},\Omega_{patient})],\end{aligned}$$

where $w(\Omega_{physician})$ is the weight associated to the primary consideration set. Note that if hospital k is contained in $\widetilde{CS}_{s_i}$, it is also included in $\widetilde{CS^+}_{s_i}$ (because by construction $\widetilde{CS}_{s_i}\subset\widetilde{CS^+}_{s_i}$).[4] The simple frequency simulator corresponds to the expression above with $w(\Omega_{physician})=1$. Summing over draws yields the simulated choice probability for individual i:

$$\begin{aligned}\widetilde{Pr^{CON}}_{ik}(\Omega_{patient},\Omega_{physician}) &= \frac{1}{S_i}\sum_{s_i}\widetilde{Pr^{CON}}_{s_ik}(\Omega_{patient},\Omega_{physician}) \\ &= \frac{1}{S_i}\sum_{s_i}1(k\in\widetilde{CS^+}_{s_i})[w(\Omega_{physician})Pr(k|\widetilde{CS}_{s_i},\Omega_{patient}) \\ &\quad +(1-w(\Omega_{physician}))Pr(k|\widetilde{CS^+}_{s_i},\Omega_{patient})].\end{aligned}$$

We define the weight as follows

[4] For the one option that is only contained in $\widetilde{CS^+}_{s_i}$, but not $\widetilde{CS}_{s_i}$, the weight $w(\Omega_{physician})$ is equal to zero (because the primary consideration set $\widetilde{CS}_{s_i}$ does not contain this option).

$$w(\Omega_{physician}) = \frac{1}{1 + \kappa \exp(V_{ij^+t} - [max_{j \in J}(V_{ijt}) - \lambda_i])}$$

where V_{ij^+t} denotes patient utility for hospital j^+ which is the highest utility hospital not included in the primary simulated consideration set (but which is included in the secondary consideration set). $(V_{ij^+t} - [max_j(V_{ijt}) - \lambda_i])$ denotes the distance in utility-space of this hospital to the threshold. Note that $V_{ij^+t} < max_j(V_{ijt}) - \lambda_i$, i.e. the distance is measured as a negative number, because otherwise j^+ would be included in the primary consideration set $\widetilde{CS}_{s_i}$. The further away V_{ij^+t} is from the inclusion threshold $(max_j(V_{ijt}) - \lambda_i)$, the closer the weight is going to be to one. When the distance is smaller, the weight decreases and the influence of the secondary consideration set becomes larger. κ is a scaling parameter that governs the degree of smoothing. For small κ the estimator is closer to the simple frequency estimator. Due to a large number of draws (20 draws per individual across 32,714 patients) we are able to set $\kappa = 5$ which implies only a small amount of smoothing.

D Additional Tables

Table D1: Descriptive Statistics — Mortality Rate and Waiting Times at the Quarter Level

		Waiting Times (Days)		Mortality Rate	
	Number of Hospitals	Mean	S.D.	Mean	S.D.
2004q1	28	113.7	36.3	1.14	0.94
2004q2	28	106.1	26.8	1.60	0.89
2004q3	28	102.5	26.6	1.39	1.10
2004q4	29	100.5	23.6	1.60	1.70
2005q1	29	93.4	21.6	1.12	0.90
2007q1	28	66.7	19.0	1.62	1.45
2007q2	28	66.2	18.5	0.67	0.93
2007q3	29	65.3	22.7	0.95	1.03
2007q4	29	63.9	23.9	1.19	2.33
2008q1	29	66.1	23.3	1.12	1.45

Table D2: Robustness Check: Demand Estimation with Unadjusted and Case-Mix Adjusted Mortality Rates[1]

		(1) Unadjusted Mortality Rate		(2) Case-Mix Adjusted Mortality Rate	
		Coefficient	Standard Error	Coefficient	Standard Error
Patient Preferences	Distance	-6.983	0.211	-7.026	0.218
	Closest Hospital Dummy	1.341	0.052	1.311	0.053
	Mortality Rate	-7.883	2.229	-3.828	1.490
	Mortality Rate * High Severity	-5.419	2.467	-3.661	1.660
	Mortality Rate * High Income	3.832	2.320	1.446	1.672
	Waiting Times	-1.528	1.887	-2.623	1.657
	Waiting Times * High Severity	-1.584	1.140	-1.580	1.062
	Waiting Times * High Income	6.262	1.196	5.972	1.297
Physician Preferences	Distance	-4.985	0.207	-5.046	0.561
	Closest Hospital Dummy	1.734	0.110	1.721	0.231
	Within-PCT Dummy	1.309	0.308	1.285	0.268
Choice Constraint Parameters	Constant	0.000	0.119	0.021	0.492
	High Severity	1.011	0.178	0.963	0.809
	High Income	0.000	0.113	0.018	0.653

[1]The first column corresponds to the baseline specification of the paper. The second columns presents results from the same model, using the case-mix adjusted mortality rate rather than the unadjusted one. All other aspects of the estimation are identical.

E Data Sources

Patient Choice Data	Hospital Episodes Statistics (HES) dataset. (http://www.hscic.gov.uk/hes) Administrative discharge dataset that covers all patients that 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.