Choice in the Presence of Experts: The Role of General Practitioners in Patients' Hospital Choice*

Walter Beckert†

June 7, 2018

Abstract

This paper considers the micro-econometric analysis of patients' hospital choice for elective medical procedures when their choice set is pre-selected by a general practitioner (GP). GPs have a dual role with regard to elective referrals in the English NHS, advising patients and at the same time taking account of the financial implications of referral decisions on local health budgets. The paper proposes a two-stage choice model that encompasses both patient and GP level optimization. It demonstrates that estimators that do not take account of strategic pre-selection of choice sets may be biased and inconsistent. We find that GPs as patients' agents select choice options based on quality, but as agents of health authorities also consider financial implications of referrals. When considering these choice options, patients focus on tangible hospital attributes, like amenities.

Keywords: Discrete choice, patient, principal, GP, agent, expert, endogenous

*I thank the editor, Luigi Siciliani, and two anonymous referees for very valuable comments and suggestions that greatly improved the paper. I am also grateful to Richard Blundell, Kate and Clare Collyer, Laura Cornelsen, Richard Disney, Jennifer Dixon, Pinelopi Goldberg, Rachel Griffith, Sandeep Kapur, Elaine Kelly, Alistair McGuire, Chris Pike, Carol Propper, Pasquale Schiraldi, Richard Smith, Ron Smith and Marcos Vera-Hernandez, Chris Walters and seminar audiences at the Nuffield Trust and NHS Improvement for very useful comments and discussions, and to the Health and Social Care Information Centre for providing access to the Hospital Episode Statistics under the Bespoke Data Re-Use Agreement CON-205762-B8S7B. This paper has been screened to ensure no confidential information is revealed.

†Birkbeck College, University of London

choice sets, competition, hospital choice, elective medical procedure.

JEL classification: D120, C510, I110, G110.

1 Introduction

In conventional discrete choice analysis, e.g. conditional logit (McFadden (1974)) and its variants, choice sets are assumed to be exogenous. In choice situations involving credence goods an "expert" agent with arguably superior information strategically presents a set of pre-selected choice alternatives to a principal decision maker. These pre-selected choice sets are endogenous. Choice of National Health Service (NHS) funded hospital services in England is an important case in point: legislation in the mid 2000s gave patients free choice of hospital for elective medical procedures, but choice is implemented by a referral from the patient's general practitioner (GP) who is mandated to offer patients a set of choice alternatives.[1] This paper discusses the design and estimation of a choice model for the patient / GP decision process and identifies biases in estimation when the potential endogeneity of choice sets is ignored in the econometric model that forms the basis of analysis.

UK legislation (Department of Health (2004)) mandated that, from 2006, patients be given choice among 5 hospital, and from 2008 patients' hospital choice was entirely unrestricted. For common elective procedures, like hip replacements, patients have several hundred choice alternatives. For most patients, in the role of the principal beneficiary of the choice outcome, such a choice problem is intractable. They typically exercise their choice following discussions with a GP who advises on their choice as a medical expert. Patients need a GP referral to access elective care, and medical expertise places the GP in the role of the gatekeeper who narrows the patient's choice problem down to a more manageable set of pre-selected choice alternatives.

GPs arguably possess superior information about salient attributes of the set of conceivable choice alternatives, notably with regard to the quality of medical

[1]See the National Health Service Commissioning Board and Clinical Commissioning Groups (Responsibilities and Standing Rules) Regulations 2012, available at http://www.legislation.gov.uk/uksi/2012/2996/part/8/made

treatment at a given hospital. Hospital quality is multi-dimensional and notoriously difficult to assess (Gowrisankaran and Town (1999), Gutacker et al. (2016)). In light of such information asymmetries, patients tend to defer to GPs' medical expertise, both when it comes to the need for treatment and the assessment of treatment quality at hospitals.[2] But GPs, to some extent, are also agents for hospitals and health authorities more generally. In 2011/12, the period of our study, local healthcare budgets were controlled by Primary Care Trusts (PCTs).[3] These budgets for the cost of care for the local population were fixed annually, and hospitals were paid a fixed price per referral. So the costs of referrals by GPs fall on the fixed budget of the PCT to which the GP belongs. This raises the question of whether GPs internalize these costs.[4]

Consequently, when pre-selecting sets of choice alternatives for patients, GPs may face a conflict of interest which induces a misalignment of their incentives with patients' incentives. This wedge driven between the GP's and patients' incentives renders choice sets endogenous.

Choice analysis with limited choice-sets has been considered by McFadden (1977) who offers two conditions - positive and uniform conditioning, characterizing an exogenous selection mechanism - that are sufficient to yield consistent estimates in the presence of exogenously limited choice sets; Santos et al. (2013) refer to this result as justification for the consistency of their maximum likelihood estimator with imposed choice sets that are subsets of the true choice sets.

The literature on general econometric choice models that allow for endogenous choice sets is still relatively sparse. The choice modelling literature refers to pre-

[2]For example, Monitor (2015), the then sector regulator for health services in England, found that "many [patients] were also thought to be happy to be guided by their GP" as regards their choice of health care provider. As of April 2016, Monitor is part of NHS Improvement, a government authority responsible for overseeing foundation trusts and NHS trusts, as well as independent providers that provide NHS-funded care.

[3]Primary Care Trusts (PCTs) are publicly funded local bodies that purchase hospital services for the local population on behalf of their associated GPs. Going forward, the Health and Social Care Act (2012) abolished PCTs and, from 2013/14, transferred budgetary management responsibilities to GP practices, now referred to as Clinical Commissioning Groups (CCGs). This system post-dates the data used in this study.

[4]See, for example, GPs referrals fall amid claims of rationed care in stretched NHS, available at `https://www.theguardian.com/society/2011/sep/09/gp-referrals-fall-stretched-nhs`

selected choice sets as consideration sets (Howard and Sheth (1969)). Mehta, Rajiv and Srinivasan (2003) estimate a dynamic structural model of consideration set formation and brand choice model in the context of price discovery for experience goods that are frequently purchased. Unlike in the context of the present paper where the pre-selected choice-set for a credence good is governed by a third-party agent, the consideration set formation process in Metha et al. is part of the sole decision maker's fixed-sample search strategy. Sovinsky Goeree (2008) proposes a model of consideration set formation that treats the inclusion decisions with respect to each choice alternative as endogenously driven by product advertisement, absent a constraint on the choice set size.

In the healthcare literature, the standard approach has been to treat the GP and the patient as a single decision maker (e.g Beckert et al. (2012)). Gaynor, Propper and Seiler (2016) are the first to model the GP led consideration set formation. In their model, consideration sets are generated subject to a constraint on the choice set size, by requiring that included choice alternatives be within a fixed distance of the alternative associated with maximal utility; the resulting choice sets are allowed to vary by GP and PCT. We offer a complementary approach. In our model the cost across experts (GPs) of acquiring and disseminating information is modelled as a determinant of choice set size and composition, and it is quantified explicitly. This approach has a particularly intuitive appeal in light of information asymmetries.

From an econometric perspective, the endogeneity of the set of choice alternatives constitutes a potential sample selection problem. It essentially arises from correlation between unobservables in the agent-level selection model and those in the principal-level final outcomes (choice) model. Such correlation may bias estimation results. This is similar to the well-known issue of incidental truncation (Heckman (1976)) whereby decision outcomes of interest are only observed for a selected subsample and where failure to properly model the sample selection mechanism induces the estimates of the outcome relationship to be biased and inconsistent. This has also been noted by Eizenberg (2014) and Jacobi and Sovinsky (2016). Similar issues also arise in the analysis of endogenous sample attrition (Hausman and Wise (1979)).

Methodological econometric issues aside, why is the distinction between principal and agent, when agents are imperfect, relevant for applied work? It is well

established that misalignment of incentives between principal and agent can give rise to market failures, resulting in suboptimal outcomes. In the present context, patients may be nudged into choosing a hospital that they would not have chosen had they been given different options. The distinction also matters for competition analysis. Demand estimation and merger simulation often feature in antitrust authorities' investigations of mergers. Beckert et al. (2012) discuss conventional hospital choice analysis, under the assumption of exogenous choice sets, and its use in hospital merger analysis. This sort of analysis does not distinguish between patients and GPs and their respective incentives. If hospitals compete for patients indirectly, via competing for GPs, then ignoring this distinction may lead to an incomplete competition assessment.

This paper proposes a micro-founded two-stage choice framework that links the pre-selection of a choice set of hospitals by the GP, as an "expert" agent on the first stage, with the choice of an alternative out of this set at the second stage by the patient, the principal and ultimate beneficiary of the choice outcome. It thereby provides a definition of "expert" agent, as opposed to "layman" principal. The model is applied to Health Episode Statistics (HES) data for hip replacement patients. This type of data is widely used in the empirical health care economics literature (Beckert et al. (2012), Beckert and Kelly (2017), Gaynor et al. (2016), Santos et al. (2013), Sivey (2012)), notably for the purpose of demand analysis. Importantly, HES data is also a primary source used by the UK competition authority, the Competition and Markets Authority (CMA).

The empirical analysis in this paper presents results that demonstrate the potential inconsistency of estimators when the endogeneity of choice sets is ignored. Estimates for the GP-level model proposed in this paper reveal that pre-selection by the GP is primarily driven by distance to the hospital, hospital quality and cost of treatment to the Clinical Commissioning Group that the GP is accountable to. The latter finding is consistent with GPs' conflict of interest at the intersection of their roles of agents of both, patients and health authorities. Once these drivers of GP-level pre-selection are accounted for by the pre-selected choice set, the results show that patients do not care about residual quality differences and that they focus instead on other tangible hospital attributes. In particular, it shows that waiting times - once their endogeneity is taken account of -, residual distance and hospital amenities are critical attributes to patients. In existing choice models, the effects of

these attributes sometimes appear with unexpected signs (e.g. Gaynor et al. (2016) who report positive waiting time effects for coronary artery bypass grafts[5]) or statistically insignificant.[6] At the same time, the residual distance effect that emerges is much more muted from the patient's perspective than has been found in other models (e.g. Beckert et al. (2012), Gaynor et al. (2016)).

The paper proceeds as follows. Section 2 provides an overview of the institutional background with regard to patient choice in the English NHS. Section 3 describes the data that forms the empirical basis of the study. Section 4 provides an overview of our econometric model for the patient / GP decision process and discusses identification, with details on specification and estimation relegated to a technical appendix. Section 5 presents results from the estimation of these models. And Section 6 concludes.

2 Institutional Background

The majority of primary and secondary health care in England is provided through the taxpayer funded National Health Service (NHS).[7] For patients, it is free at the point of use. Primary care is provided by General Practitioners (GPs). In the period studied in this paper, 2011/12, publicly funded local bodies, Primary Care Trusts (PCTs), make up the NHS commissioning system, i.e. they manage health care budgets and purchase secondary care, e.g. for elective medical procedures and other hospital services, for the local population. GPs thereby make referral decisions and so get to decide how some of the health care budgets is spent.[8] Patients can only

[5]They do point out that this finding can be rationalized in light of the severity of the underlying medical condition and the risk of the procedure; additional waiting time may leave the patient time to arrange necessary personal affairs.

[6]Goldman and Romley (2008), using stated preference data, do demonstrate significant amenity effects. Sivey (2012), using HES data for cataract surgery patients, finds negative waiting time effects. The systematic review by Aggarwal et al (2016) reports negative effects of waiting time on patients' choice outcomes.

[7]A private health care market exists in the UK, but it is excluded from the analysis of this paper.

[8]Patient choice of GP is relatively limited and typically restricted to GPs whose practices are local to the patient's area of residence; i.e. patients living in a given PCT are registered with a GP in the same PCT.

obtain access to secondary care through a referral from their GP. GPs therefore act as gatekeepers to secondary care, both with regard to in-patient and out-patient appointments.

Several waves of legislative reforms of the NHS over the past decade have increased the choice patients have over where they receive elective care. The first set of reforms gave patients a formal choice over where to attend a first outpatient appointment when referred by their GP (or consultant). From January 2006, GPs were required to offer patients a choice of (four to) five hospitals. They were also required to raise awareness of patients' right to choose. This replaced a system where patients could state preferences but GPs were under no obligation to offer patients a choice. In 2008, essentially all restrictions on the number of providers patients were able to choose from were removed. This established "free choice" of provider. These reforms were part of a set of market-related policies intended to improve hospital efficiency, make hospitals more responsive to patients needs, improve quality and improve equity by opening choice to all (Dixon et al. (2010)). A series of work has estimated the impact of patient choice on hospital quality by comparing areas with different degrees of potential competition, and finds that higher degrees of competition are associated with greater improvements in quality (Cooper et al. (2011), Gaynor et al. (2013)).

From a practical point of view, the choice architecture was implemented through an electronic booking system, "Choose and Book", which allows GPs to shortlist appropriate hospital services for their patients and, subsequently, enables patients to book their appointment, either at the GP practice, by phone or online. In this institutional setting, the GP is critical to the patient's exercise of choice.

NHS funded hospital care has historically been delivered by state owned and state run NHS Acute Trusts, or hospitals.[9] Under the Payments by Results NHS funding architecture,[10] commissioners (PCTs) pay health care providers, such as hospitals, a national tariff, i.e. a per patient payment based on the treatment they provide.[11] There is some variation in tariffs across hospitals captured by a Market

[9] A NHS Acute Trust may be comprised of a single hospital or multiple hospital sites within the same geographic area.

[10] `https://www.gov.uk/government/publications/simple-guide-to-payment-by-results`

[11] Hospital care is grouped into Healthcare Resource Groups (HRGs), which are similar to Diagnostic Resource Groups in the US. Prices or Tariffs are then set at a national level based on the

Forces Factor (MFF) which is an adjustment to the national tariff. This adjustment is unique to each provider and reflects that it is more expensive to provide health care services in certain areas, e.g. due to local estate costs or wage levels.

In 2011/12, such treatments were funded from fixed PCT budgets. These budgets for the cost of care for the local population were fixed annually, hospitals were paid a fixed price per referral, and hence the costs of GP referrals fall on the fixed budget of the PCT to which the GP belongs. This raises the question whether GPs internalize these costs and take account of the externalities of a given referral on the opportunities of other patients to receive treatment. Indeed, GPs may find themselves as a rationing agents on behalf of the PCT which pays for care.[12] GPs may therefore face conflicts of interest, balancing interests of a given patient with those of other patients and with the objectives of PCTs.[13]

A Report published in July 2011 by the Cooperation and Competition Panel into the operation of any willing provider for the provision of routine elective care (Cooperation and Competition Panel, 2011) shows the ways in which some PCTs sought to influence GPs' referrals for elective care. PCTs cited concerns regarding value for money as the primary reason for influencing GPs. The report notes that the MFF is one way in which these value-for-money considerations arise: "In relation to the market forces factor, one provider showed us correspondence relating to a patient referral decision where a GP had persuaded a patient not to go to their preferred provider of routine elective care because of the higher cost to the PCT of that provider arising from the market forces factor. We also saw correspondence between an SHA [Strategic Health Authority] and a provider concerning the higher cost of sending patients to that provider as a result of the market forces factor, and the actions that PCTs in the SHA were, as a result, taking to ensure that patients were treated by other providers. A PCT chief executive told us that "in London where you have providers in close proximity with different uplifts, it is clearly in the PCT's interest to encourage activity to go to those with a lower market forces

average cost of providing the associated care.

[12]For example, Blundell et al. (2010) report GPs' "preoccupation with overreferral" and interpret it as "a worrying consequence of perceptions of pressure on limited healthcare resources."

[13]GPs located in a given PCT can have significant portions of patients who live in adjacent PCTs. But it is the PCT that the practice belongs to that bears the the costs of all patients registered at the practive, irrespective of whether the patients live within the PCT's geographic boundaries.

factors and reinvest the savings in other health improving outcomes". The Report goes on to say not all PCTs necessarily respond to this incentive, however.[14]

Studies of earlier periods during which some GP practices participated in a fund-holding scheme (1991 - 1999) aimed at increasing GPs' budgetary responsibilities (Dusheiko et al. (2006), and Dusheiko et al. (2007)) show that the patient-physician agency relationship is weak, i.e. GPs do not act solely in the interest of patients.

To the extent that such incentives induced by PCTs exist, this is not unlike in the pre-reform period when PCTs contracted secondary care provision out to local NHS Trusts (bulk contracts). GPs were expected to refer their patients to contracted hospitals only and had to justify any referrals to non-contracted hospitals in light of the extra costs to the PCT caused by such off-contract referrals.

During the period of our study, some PCTs introduced formal referral management systems. Such referral management systems involve a spectrum of approaches and objectives. One of the dominant first stage objectives is to determine whether a referral is necessary, following clinical triage. This objective aims at demand management at the extensive margin, in order to reduce unnecessary referrals in the first place. This objective interacts with a subsequent, second stage objective of determining a suitable referral destination, in consultation with the patient and GP.

The analysis in this paper relates to the second stage in the referral management process, choice of hospital at the point of referral. While we do not model their role explicitly, we note that there are at least two ways how referral management systems might affect the model.

First, conditional on passing the first stage necessity test, referral management systems might impose further constraints and incentives on the GP level hospital evaluation and selection process - e.g. inducing diversions to out-of-hospital services, at the expense of longer distances - and thereby enhance the GP's role in the joint patient-GP decision process and, to some extent, curtail the patient's freedom to choose. As such, they may change the composition of the set of choice alternatives

[14]It quotes a PCT as saying that "when going to AWP [Any Willing Provider] for services we include a maximum tariff to ensure providing value for money. Whatever pricing agreements are in place, the choice of provider is still left to patients and the PCT does not direct patients to certain providers".

that a patient is being presented with.

Second, to the extent that such systems might de-motivate GPs and erect barriers between GPs and consultants, they may actually increase the cost of inclusion of choice alternatives in the pre-selected set and thereby constrain patients' choice options, i.e. they may reduce the number of choice alternatives and thus the size of the set of choice alternatives being presented to patients.

Such restrictions on the size and the composition of choice sets presented to patients are also supported by Rosen et al. (2007) who find that "the referral management centre opened by one primary care trust [at the time of their study] was seen to restrict choice".

Hence, referral management systems are likely to reinforce the relationship identified in our study.

3 Data

This study uses administrative data, Health Episode Statistics (HES), collected by the UK Department of Health for every NHS funded inpatient admission in England. HES data are widely used in academic research (Beckert et al. (2012), Beckert and Kelly (2017), Gaynor et al (2016), Santos et al. (2013), Sivey (2012)) and also constitute the primary empirical basis for any quantitative work in the area of health care demand carried out by UK competition authorities.

The study considers approximately 30,000 NHS funded hip replacement patients in 2011/12.[15] These patients were advised at 4721 GP practices; for ease of exposition, GP and GP practice are treated synonymously in the remainder of the paper.[16] Appendix B provides details on how the GP sample was selected. Patients in the

[15]The analysis uses selected orthopaedic treatments, so called Healthcare Resources Groups (HRGs) at spell level derived from the Secondary Uses Service (SUS) Payments by Results (PbR) data - HB11, HB12, HB13 and HB14 - and, within these, treatment specifications relating to general surgery and trauma and orthopaedics - Treatment Function Codes 100 and 110. HES data only record treatments, i.e. patients who actually had a hip replacement; patients contemplating a hip replacement, but ultimately choosing not to undergo surgery or to do so at a private clinic, are not recorded. Therefore, in this application there is no outside option.

[16]This is the approach also taken in Gaynor et al. (2016).

sample were treated at one of 168 NHS hospital sites that carried out at least 10 hip replacements in 2011/12 and for which a set of hospital attributes is available.[17] The analysis only considers GP practices that refer to between one and seven NHS hospitals.

For each patient episode in the sample, the data record the patient's age, local area of residence and the site of the hospital where the patient was treated. They record the dates of referral and treatment from which we compute the patient's waiting time, i.e. the time that elapsed between referral and treatment. From these waiting times, hospital level median waiting times can be constructed as a hospital attribute. Various hospital attributes can be merged in, from publicly accessible databases maintained by the Health and Social Care Information Centre (HSCIC). They include quality measures, such as Hospital Standardised Mortality Ratios (HSMR) which relate the actual number of deaths at the hospital to the expected number of deaths, given the characteristics of the patients treated at the hospital (case mix). They also include the Market Forces Factor (MFF) and hospital amenities, such as parking spaces at the hospital. Table 1 provides summary statistics on the hospital attributes included in this study.

HES records also record the GP practice that made the referral for treatment at a hospital site. Using the GP practice identifier, practice attributes can be included, some of them also from HSCIC sources. Practice attributes will be relevant to the extent that they act as drivers of practice level costs of pre-selecting choice alternatives.[18] They include the number of GPs at the practice: Larger practices

[17]We exclude NHS funded treatments at private providers in this study. For the choice model considered in this paper, we cannot include private providers - Independent Sector Treatment Centres (ISTCs) or Independent Sector Providers (ISPs) because there are no clinical quality measures comparable to HSMRs for our period. We would have to use a different measure, e.g. estimated readmissions rates or survey data, such as PROMs. We decided to use a quality measure that is officially published and, as such, is the same piece of (relatively) "hard intelligence" for all GPs. See the Identification section 4.2 below, for further discussion.

We acknowledge that excluding the private sector, where waiting times are considerably lower, may lead us to underestimate the waiting time effect. We note in this context that Dixon and Robertson (2009) find that the "independent sector was not perceived as much of a threat" and rather "acted as a partner for the NHS, providing extra capacity to help the NHS meeting waiting time targets". Beckert and Kelly (2017) study hip replacement patients' sorting between private and public providers.

[18]The following section provides a detailed exposition of the two-stage choice model that discusses

enjoy a richer pool of experience and information and hence are likely to more easily facilitate choice. The analysis also considers measures of the heterogeneity of the practice's patient pool. From HES records, we construct the coefficient of variation with respect to age at the practice level as a measure of dispersion. In our GP pre-selection model, we allow the cost of offering a set of alternatives to depend on the heterogeneity of patients, so that the net value of a given choice set is allowed to be higher the more diverse the pool of patients is to whom this set if offered.[19]

The first columns in Table 2 summarizes the distribution of the GP level coefficient of variation with respect to age. The age variation at the practice level is skewed to the left, i.e. towards practices with more homogeneous patient pools with respect to age. The median coefficient of variation with respect to age at the practice level is 19.2 percent, while the lower (upper) quartile is 13.3 (28.4) percent. Practices referring to a single hospital tend to have more homogenous patients, with a median coefficient of variation with respect to age of 16.6 percent, compared to 20.5 percent for practices referring to several hospitals. The former practices also tend to be smaller, with an average of 4.2 GPs pre practice, versus 5.4 for the latter.

The locational information regarding patients, GPs and hospitals sites permits calculating straight-line distances between hospitals and GPs, and residual distances between hospitals and patients, respectively.[20] We refer to the latter as residual distances to account for the fact that distance may already be taken into account by GPs pre-selecting hospitals choice sets for patients to choose from.

These GP-level referral data allow to construct hospitals' catchment areas with respect to hip replacements, i.e. the set of GP practices that refer hip replacement patients to them. The panel structure of the data, which associates multiple patients at the practice with potentially different treatment destinations, allows us to infer, or at least approximate, the set of hospital alternatives pre-selected by the GP as the set of hospitals that patients at a given practice were referred to and treated at. This is the same evidence base as in Gaynor et al. (2016). The approach

the role of costs at the first stage of GP-level pre-selection.

[19]For example, Harding et al. (2014) report that older patients, while valuing the freedom to choose, tend to shun exercising choice and to revert to their local hospital. This would suggest that the cost of promoting choice is higher at practices with predominantly older patients.

[20]The GP level distances are calculated using GP practice postcodes. The patient level distances are calculated using the patients LSOA.

taken in this paper implicitly assumes that hospitals that were chosen by at least one patient are part of the choice set and discussion between GP and patients.[21] It also assumes that the sample is informative enough to separate with reasonable reliability hospitals that were never chosen from those that were chosen by some patients. This leaves a risk of potential measurement error in the construction of the pre-selected choice sets at the GP practice level, which will be considered when assessing potential resulting biases in estimation.[22]

The approximation adopted in this paper, in our view, is the best possible approach given the available empirical basis for health care demand analysis. HES data are currently the most comprehensive data records for this kind of undertaking. Details of conversations between GPs and patients are confidential and not recorded. And additional data gathering exercises to date have proven unfruitful. For example, an alternative approach to identify the set of hospitals pre-selected by GPs would be to conduct a survey and use the results to explore the factors that these agents take into account when guiding patients choices. However, previous attempts to survey GPs have been frustrated by very low response rates. For example, in the Competition Commission's (CC) Royal Bournemouth and Christchurch Hospitals NHS Foundation Trust and Poole Hospital NHS Foundation Trust merger inquiry (2013), the important role of GPs in the referral process was recognized, but no strong conclusions could be drawn (para 6.98, Final Report), because out of 1099 GPs in the hospitals' catchment areas only 36 GPs (associated with 23 GP practices)

[21] In the context of our paper, if two patients have the same GP, then the distribution of their consideration sets satisfies the uniform conditioning property in McFadden (1977); if two patients have different GPs, then the distribution of their consideration sets satisfies McFadden's positive conditioning property.

[22] It may be worth mentioning that selection of information on outcomes is not uncommon as consideration sets are rarely observed. Gaynor et al. (2016) use the same data to model the choice options GPs offer to patients for their choice of hospitals when undergoing coronary artery bypass graft surgery. And Eizenberg (2012), in a study of the home PC market, also proceeds in a similar fashion: he infers the feasible set of Intel chips as those that PC manufacturers chose to offer in their products and that sold at least 10,000 units. Furthermore, our approach is essentially the same as the Inter-Personal Logit (IPL) in Crawford et al. (2016). Our setting meets their Condition 1 on sufficient sets which, together with the logit assumption, yields consistent estimators (Prop.1 in Crawford et al. (2016)); see section 3.4.5 of Crawford et al. (2016). We thank R. Griffith for pointing this out. In fact, their notion of sufficient sets is reminiscent of the sets of quasi regular states in the stochastic process theory literature.

provided complete survey responses (GfK presentation to CC, 2013). Furthermore, stated preference surveys risk to yield biased responses in this context. The use of revealed preference data allows the analyst to overcome these challenges.

Table 3 shows the distribution of the number of hospitals referred to, at the GP practice level. Even though patients choice had already been mandated for several years by 2011/12, a large fraction of GP practices (43.15 per cent in the sample used in the analysis) only referred to a single hospital (that meets the attribute data requirements); this is consistent with GP survey evidence (e.g. Monitor (2015)) that many GPs identify a "default provider". And over ninety percent refer to no more than three; also this is consistent with GP survey evidence (Monitor (2015), Dixon et al. (2010)) that most GPs discuss two or three, and at most five, hospital alternatives with their patients.[23]

The average age of hip replacement patients in our sample is 68.6, and the median is 68, with lower quartile 58 and upper quartile 76 and a standard deviation of 17.71. So the patient age distribution is almost symmetric about its mean. Similarly, at the practice level, the distribution of average patient age is fairly symmetric, with median average age at the practice level of 66.33, lower quartile 60.5 and upper quartile 70.97 (standard deviation 11.34). Table 2, second column, provides summary statistics on the GP level average patient age distribution.

The remaining columns of Table 2 shows some further GP practice level statistics. The mean number of GPs at the practice level is just below 4, skewed to the left, i.e. to practices with a small number of GPs (see also Kelly and Stoye (2014)). With regard to distances, Table 2 shows that pre-selected hospitals are closer than the average hospital in the GP's consideration set. The distribution of distance to hospitals in GPs' consideration sets is broadly similar to the distribution of distances

[23]Evidence provided by the King's Fund (Dixon et al. (2010)) shows that about 49 percent of patients say they were given two hospitals to choose from, 49 percent said they could choose between three and five, and only two percent reported having more than five hospitals to choose from. Monitor (2015), presenting GP survey evidence on referral practice, find: "This GP uses Choose and Book and gets a list of providers local to the patient. She then selects those NHS providers that are closest and discusses which the patient would prefer"; hospitals local to the patient are also local to the GP practice as patient overwhelmingly choose nearby GP practices; and GP survey respondents say they typically discuss no more than two or three, and at most five, hospitals options.

between patients and hospitals in their GPs' consideration sets.

4 Econometric Model

4.1 Model Overview

This section provides a high-level overview of the two-stage model for the GP and patient level choice process that we estimate in this paper. The model encompasses the GP's pre-selection of a choice set of hospital alternatives at the first stage, from which the patient makes a final choice at the second stage. Readers with interest in the details of the econometric model specification and estimation are referred to Appendix C.

The modelling approach is motivated by qualitative evidence on the merely partial overlap of the information sets that patients and GPs base their valuations and decisions on. We review this evidence in the following subsection 4.2 on identification, in order to justify our modelling assumptions.

GPs – as experts not only on medical diagnosis, but also on health care providers – can assess the quality of hospital alternatives. Patients – as laymen – either do not have access to this information, or they typically find it difficult to interpret. This makes GP level pre-selection of choice alternatives efficient. However, GPs as agents of health authorities may face incentives that are irrelevant to patients, e.g. with regard to financial implications of referral decisions captured by the MFF. We label such attributes of hospital j that are considered by the GP, i.e. that are in the GP's information set and that the GP acts upon, by $\mathbf{x}_j^a$, $j \in \mathcal{J}$, where $\mathcal{J}$ is the set of all hospitals considered by patient i's GP.

Patients, in turn, may pay attention to hospital amenities that do not matter to the GP, i.e. while they are in the GP's information set the GP does not act upon them when making a choice set pre-selection decision. We label such hospital attributes $\mathbf{x}_j^p$. GPs typically have incomplete information on patient preferences and all the hospital attributes that are salient to patients. Our model accounts for such information asymmetries. This information asymmetry is one justification for a mandate for GPs to offer choice to patients.

Hospital attributes considered by both, GPs and patients – i.e. attributes that are in the information sets of both and acted upon by both – include waiting times and distance. These are labelled $\mathbf{x}_{ij}^c$.

Table 4 summarizes the information asymmetries of the micro-theoretic GP and patient level choice models.[24]

Pre-selecting a choice set $\mathcal{J}_i^a \subset \mathcal{J}$ for patient i to choose from is costly for patient i's GP. We model this as a unit cost $C > 0$ for the GP to include a hospital into the set of choice alternatives. This cost may be specific to the GP. For example, in the context of hospital choice in the UK where a GP (practice) plays the role of the patient's agent, this cost might be expected to be a convex function $c(\mathbf{z})$ of practice level patient heterogeneity and the number of GPs in the practice. It imposes a constraint that can be thought of as the effort the GP needs to exert in order to explain the features, pros and cons of the alternative to the patient.

We model the GP's pre-selection decision as a concave constrained maximization problem in which the GP maximizes the value $I_{\mathcal{G}}(\mathbf{x}_i^c, \mathbf{x}^a)$ of candidate sets $\mathcal{G} \subset \mathcal{J}$ of hospitals, on the basis of $\mathbf{x}_j^a$ and $\mathbf{x}_{ij}^c$, $j \in \mathcal{G}$, subject to incremental costs $c(\mathbf{z})$, i.e.

$$\mathcal{J}_i^a = \arg\max_{\mathcal{G}\in\mathcal{P}} \left\{ I_{\mathcal{G}}(\mathbf{x}_i^c, \mathbf{x}^a) - c(\mathbf{z})\#\mathcal{G} \right\},$$

where $\mathcal{P}$ is the set of all subsets of $\mathcal{J}$ and $I_{\mathcal{G}}(\mathbf{x}_i^c, \mathbf{x}^a)$ is increasing in $\#\mathcal{G}$.

The GP thus reduces the complexity of the choice problem for the patient, by narrowing the set of conceivable choice alternatives $\mathcal{J}$ down to $\mathcal{J}_i^a$. At the same time, the information asymmetry and ensuing misalignment of assessment criteria between patient and GP results in a choice set $\mathcal{J}_i^a$ which may be suboptimal when evaluated on the basis of the attributes $\mathbf{x}_i^c$ and $\mathbf{x}^p$ relevant to patient i.

The pre-selected choice sets $\mathcal{J}_i^a$ vary across patients i, to the extent that the attributes considered by both, GP and patient, $\mathbf{x}_{ij}^c$, vary with i; e.g. distance between i and hospital j. In practice, the GP may pre-select a uniform choice set $\mathcal{J}^a$ at the outset on the basis of $\mathbf{x}^a$ and $\mathbf{x}^c$ as they relate to an "average" or "hybrid" patient and then offer this set to all patients at the practice. In light of the limitations on our data, we adopt this modelling assumption.

We model the second stage patient choice, given $\mathcal{J}^a$ and hospital attributes $\mathbf{x}_i^c$

[24] Table 5 summarises the econometric version of the model, as it is detailed in Appendix C.

and $\mathbf{x}^p$ salient to patient i's choice, as a multinomial logit model. In our econometric analysis, we allow for correlation between unobservables (to the econometrician) in the GP level pre-selection model and unobservables in the patient level choice model, in order to guard against possible non-random sample selection effects.

4.2 Identification

In this section, we discuss the identification of our econometric model.[25] The patient's indirect conditional utility δ_{ij}, conditional on the pre-selected $\mathcal{J}^a$, is identified through patients' choices from this set and variation in attributes across choice alternatives. Regarding the GP's pre-selection model, the GP's assessment of patient i's benefit of hospital j, α_{ij}, is identified through variation in attributes across alternatives and their inclusion in, respectively exclusion from, $\mathcal{J}^a$.[26]

Furthermore, the value $I_{\mathcal{J}^a}(\mathbf{x}^c, \mathbf{x}^a)$ of $\mathcal{J}^a$ is increasing in the imprecision of the GP's incomplete information about patients' valuation criteria, albeit less than linearly.[27] This imprecision is captured by a scale parameter σ on the extreme value errors of the patient level logit model. This scale parameter is identified through variation in set sizes across agents with the same levels of cost drivers. This feature of the constrained pre-selection model is an interesting departure from the usual lack of identification of scale on the selection stage in non-random selection (incidental truncation) models absent constraints.

Unless the coefficients θ_c on the attributes $\mathbf{x}^c_{ij}$ considered by both, GP and patient, are restricted to be identical across the patient and GP models, the log-likelihood of the two-stage model splits into a part that captures the GP's pre-selection and a part that captures the patient's choice, conditional on the pre-selected choice set. In this case, there are no parametric restrictions across the two parts, so

[25] The notation in this section therefore refers to the econometric model specification that is detailed in Appendix C.

[26] For example, if $\alpha_{ij} = v$ for all $j \in \mathcal{J}$, then the inequalities C-3 in Appendix C imply

$$\ln\left(\frac{J^a+1}{J^a}\right) \leq c(\mathbf{z}) \leq \ln\left(\frac{J^a}{J^a-1}\right).$$

Hence, the cardinality $J^a = \#\mathcal{J}^a$, i.e. the size of the pre-selected choice set, next to variation in cost drivers $\mathbf{z}$, identifies the agent's cost function $c(\mathbf{z})$.

[27] The GP's incomplete information is captured by ξ_{ij} in Table 4.

they can be estimated separately and consistently under the aforementioned identifying assumption. This is the approach taken below. The model by Gaynor et al (2016) shares this feature. The first-stage GP level pre-selection amounts to a nonlinear version of the classical incidental truncation model. The analogy to the classical linear incidental truncation model makes clear that for identification of the two-stage model, it is necessary that the true coefficients on $\mathbf{x}^a$ and $\mathbf{x}^c$, θ_a and θ_c, satisfy $\theta_a \neq \mathbf{0}$ and $\theta_p \neq \mathbf{0}$, i.e. exclusion restrictions must be in place that ensure independent exogenous variation at both, the GP and the patient stage. Therefore, absent any restriction on θ_c across the two stages of the model, the GP level pre-selection model can be estimated separately and inverted to retrieve imputations of unobservables (to the econometrician) in the GP pre-selection model μ_{ij}; these can be used to impute unobservables (to the econometrician) in the patient choice model ζ_{ij} which, in turn, can be used as embedded regressors in a second-step estimation of the patient's choice model. Appendix C.3 provides technical details.

The following approach is taken with regard to the exclusion restrictions. It is motivated by qualitative evidence in Rosen et al. (2007) who observe that "patients and GPs seek partially overlapping, but different characteristics when choosing a hospital." Their study finds that for GPs clinical aspects of care and waiting times are the most important hospital attributes, and that constraints on geography, i.e. distance, are an important equity issue GPs consider when offering choice.

Dixon and Robertson (2009) find that patients do not take hospital quality into consideration when choosing a hospital: "Patients made little use of available information on the performance of hospitals; just 4 percent consulted the NHS Choices website [$\cdots$] Instead, they relied heavily on [$\cdots$] the advice of their GP." In other words, patients obviously care about clinical outcomes and quality, but they defer to their GPs' assessment when making decisions.

They also report that "GPs we spoke to did not think patients were interested in information about comparative performance". And they report a similar view held by providers: "Providers were quite sceptical about the extent to which patients were acting as informed consumers. Any observed changes in referral patterns were largely seen to be a result of GP decisions".

With regard to patients, the importance of distance and waiting times is sup-

ported by these and other studies (Beckert et al. (2012), Gaynor et al. (2017)). With regard to distance, our analysis distinguishes distances between GPs and hospitals, and residual distances between patients and those hospitals that are pre-selected by the GP and presented to patients as choice alternatives.

The importance of amenities to patients is supported by Dixon and Robertson's finding that providers emphasise delivering a positive overall patient experience. They quote a Foundation Trust representative as saying "[patients] would make that choice on the basis of a whole range of indicators in terms of patient experience and what matters most, and that's from car parking to clinical outcomes."

The role of more broadly defined amenities in patients' hospital choice (including pleasant surroundings, attentive staff) has also been documented by Goldman and Romley (2008) on the basis of stated preference (survey) data. Patients' focus on amenities is also echoed by Coulter et al. (2005). Hence, hospital amenities (in the form of parking space) are attributes $\mathbf{x}^p$ that are assumed to solely matter to the patient, but not to the GP.

The analysis considers two hospital attributes that are assumed to be considered solely by the GP, $\mathbf{x}^a$. The first is the hospital's clinical quality. It is well recognized that hospital quality is difficult to assess, even for quantitative researchers, because many quality measures suffer from selection bias (Gowrisankaran and Town (1999)) and this is unlikely to be taken into account by the patient.

The GP level model considers the Hospital Standardised Mortality Ratio (HSMR) as measure of clinical quality. It puts the actual number of deaths at the hospital in relation to the expected number of deaths, given the characteristics of the patients treated at the hospital (case mix); the case mix adjustment is designed to guard against selection bias.

Alternative quality measures have been considered in the literature. Gutacker et al. (2016) and Skellern (2016) consider Patient Reported Outcome Measures (PROMs) as quality measure. PROMs data directly capture health gains as experienced by patients. The potential draw-back is that they are self-reported and thereby also risk being contaminated by selection bias. Valderas et al. (2011) point out that the link between healthcare outcomes and self-reported health status is unclear. Hutchings et al. (2012) report that reponse rates vary with patient charac-

teristics and that non-respondents have worse outcomes. Beckert and Kelly (2017) use their own estimates of hospital level readmission rates for orthopaedic surgeries, adjusted for case mix. This measure is not available to GPs and hence not in their information set, while HSMRs are formal published measures of quality and as such in GPs information set. Since our model revolves specifically around information asymmetries, the public availability and absence of selection bias justifies the choice of HSMR as clinical quality measure.

While hospital quality is clearly relevant to the patient, patients typically rely on expert advice to judge the quality of health care provision, so it seems reasonable to include HSMR in $\mathbf{x}^a$. This is in line with survey evidence collected by the King's Fund (Dixon and Robertson (2009)) that patients don't use quality measures when choosing a hospital. Nonetheless, the model allows patients' perceptions of hospital quality to affect their choice. Appendix C provides details on how the model allows patient level valuations to be correlated with GP level valuations. Such correlation would be expected to arise if the GP's quality assessments, unobserved by the econometrician, were factored into the patient's valuations. The model thus allows patients to respond to hospital quality as they perceive it through their GP. We return to this when discussing our results.

The second attribute in $\mathbf{x}^a$ is the hospital's Market Forces Factor (MFF), which is an adjustment to the national tariff NHS hospitals are compensated at for specific treatments such a hip replacements; this adjustment is unique to each provider and reflects that it is more expensive to provide health care services in certain areas, e.g. due to local estate costs or wage levels. Propper and Van Reenen (2010) argue that, because local wages do not adjust to the MFF, this causes lower hospital quality. Another hypothesis might be that referrals for treatment at hospitals with high MFF are more expensive and, in light of budgetary constraints, discouraged by the Primary Care Trust that the GP belongs to. Figure 1 shows that the MFF within and across GP practices exhibits considerable variation and hence is not merely a measure of the GP practice's geographic location. Hospitals attributes $\mathbf{x}_c$ that are assumed to be considered by both, patient and GP, include the respective distance to a hospital and the (median) waiting time until treatment at the hospital.

As alluded to earlier, the cost function $c(\mathbf{z})$ needs to be weakly convex in order to guarantee an interior solution, i.e. a pre-selected set $\mathcal{J}^a$ that is a (strict)

subset of $\mathcal{J}$. Costs in our model are in the same units as indirect utility. Hence, the average level of costs, which is not attributed to cost drivers, and the average level of indirect utility, which is not due to alternative specific attributes, cannot be identified separately. Metha et al. (2003) encounter an analogous lack of identification. Furthermore, this cost function must be specified at the GP (practice) level, i.e. it cannot vary with hospital alternative j; if it did, then for an included hospital alternative it would be indistinguishable from the utility contribution of that hospital to the inclusive value associated with $\mathcal{J}^a$. For GPs at the practice, including a hospital in the choice set $\mathcal{J}^a$ may be costly because its salient characteristics need to be researched and because its suitability for a patient with given characteristics needs to be assessed. For example, a report by the National Audit Office (NAO (2005)) documents that 90 percent of GPs believe their overall workload will increase as a result of the implementation of Choose and Book, and that only 3 percent feel very positive and 15 percent a little positive about the introduction of choice. The analysis considers two GP practice attributes $\mathbf{z}$ that may determine the cost $c(\mathbf{z})$ of inclusion of choice alternatives in the pre-selected choice set $\mathcal{J}^a$. First, the number of GPs at the practice, as a measure of collective experience with regard to referral success, may be hypothesised to lower the cost of inclusion. Second, relatively homogeneous patients are likely to benefit less from the inclusion of additional choice alternatives than patients with heterogeneous characteristics and needs. This makes the opportunity cost of not including more choice alternatives relatively low for practices with homogeneous patients, compared to practices with more heterogeneous patients. To control for this, the analysis considers as a second cost driver the coefficient of variation with respect to age of patients at the practice level.

Finally, the GP's consideration set needs to be defined in a practical manner. This problem is not new: Gaynor et al. (2016), using HES data as well for coronary artery bypass graft (CABG) patients, face essentially the same problem, except that there are only 29 hospitals performing CABGs, while the number of NHS hospitals preforming at least ten hip replacements in 2011/12 is 168 and as such renders the dimensionality of the GP level pre-selection problem impractically large. In fact, the set $\mathcal{J}$ that a GP (practice) considers is very likely much smaller. In our approach, all GPs have at least 7 hospitals to choose from; Figure 2 shows the distribution of consideration set sizes across GPs.

Our approach uses the following algorithm in order to construct the sets $\mathcal{J}$ considered by GPs from which the choice sets $\mathcal{J}^a$ are pre-selected. For each NHS hospital site, the hospital's catchment area in terms of GP practices is defined as the smallest set of GP practices that collectively refer at least 80 per cent of the hospital's hip replacement patients. The geographic size of the hospital's catchment area is then determined as the maximum distance between the hospital and any of the GP practices in this set; the median of the maximal distances is 66km. And the geographic catchment area of the hospital is given by the circular area about it, radially defined by that maximal distance. The hospital is included in a GP practice's consideration set $\mathcal{J}$ if the practice is in its geographic catchment area. For some GP practices, located in large metropolitan areas, the cardinality of $\mathcal{J}$ determined in this manner is rather large. To reduce the dimensionality of the pre-selection problem for such practices, $\mathcal{J}$ is defined as the intersection of these sets and the set of the k nearest hospitals. The sensitivity of this definition of GP level consideration sets with respect to k reveals that, for 86 per cent of GP practices, no more than one patient chooses to be treated at a hospital that is not among the $k = 15$ nearest hospitals, and for only one GP practice there are 5 patients who choose more distant hospitals. Such referrals are ignored by the present analysis and $k = 15$ is chosen as cut-off. Given that most patients report to have been given no more than 5 choice alternatives (Dixon et al. (2010)), this approach appears to err on the side that is generous towards GPs. Our approach may simply eliminate atypical choice situations, i.e. the choice outcome may well be due to reasons unidentifiable in the data, e.g. the patient has family living near such relatively distant hospitals.[28]

Also, to place this approach into the context of research practice, defining the consideration set via a limit on joint market share to manage the computational burden is not uncommon. For example, Eizenberg (2012) in his study of the home PC market restricts the number of product lines to those whose joint market share is 70 percent. Our approach conforms with the notion of sufficient sets and the Inter-Personal Logit (IPL) model in Crawford et al. (2016).

[28]The elimination of atypical choice outcomes is akin to the restricting the analysis to "quasi regular" (Oxtoby, 1952) or "nontrivial invariant" sets (Billingsley, 1995), i.e. to sets of states of the world that are aperiodic and recurrent, in the terminology of the theory of ergodic stochastic processes. States in this set are visited with positive probability in the limit. The implicit assumption here is that the initial "burn-in" period of the Markov chain that generates the sequence of pre-selected sets has been transcended.

5 Results

5.1 Estimation of Pre-Selection Model

Table 6 present estimation results for the model of GP level pre-selection. The table presents both, the estimates of the constrained choice model, with the cost function specified as $c(\mathbf{z}) = \exp(\mathbf{z}'\tau)$, and for comparison estimates of a linear probability model absent cost constraints. The former is estimated by Maximum Simulated Likelihood, with the unobservables (to the econometrician) in the GP pre-selection model $\{\mu_{ij}, j \in \mathcal{J}\}$ being i.i.d. draws from a standard normal distribution.[29]

The results of both models are qualitatively similar with regard to the hospital attributes included in $\mathbf{x}_c$ - distance and waiting time - and $\mathbf{x}^a$ - HSMR and MFF. They show that distance is the dominant hospital attribute in the GPs' pre-selection of hospitals into $\mathcal{J}^a$. GPs tend to pre-select closer hospitals. The coefficient on distance is about four times as large as the second most important attributes, the market forces factor (MFF).[30] The MFF also weighs negatively on the GP's inclusion decision, as does hospital quality, measured by the hospital's HSMR. If HSMR were regarded as fully controlling for hospital quality of care, then it could be argued that the negative effect of the MFF would suggest that GPs tend to refer to hospitals that are cheaper from the point of view of the local Primary Care Trust. This finding suggests that GPs to some extent internalize the costs of their referrals that fall on PCT budgets. This finding complements research on the implementation of GP fundholding reforms in the early 1990s that found that altruism is not an important motive of GPs and that direct financial incentives are required to induce GPs to take account of the externalities their referrals create (Croxson et al. (2001), Dusheiko et al. (2006)). This finding is also important in light of the recent changes to the institutional design of the NHS. With the formation of Clinical Commissioning Groups following the Health and Social Care Act (2012), GPs have greater responsibility for budgets. These changes have likely sharpened the incentives for

[29] See Appendix C.3 for details.

[30] If GPs took account of distances between the patient's LSOA (instead of GP practice) and hospital sites so that the distance variable would be measured with error, then results by Griliches and Ringstad (1970) about measurement error in nonlinear models imply that our distance coefficient estimate is biased towards zero, so the true distance effect is even stronger.

GPs to take account of financial implications of their referral decisions.

The results show that specifying the pre-selection model properly affects the relative importance of explanatory variables. The effect of quality relative to waiting times is bigger in the constrained model than in the unconstrained model. In the constrained model, hospitals are ranked relative to one another and included up to the marginal hospital, and our estimates show that the cutoff is driven more by quality than waiting time. In the unconstrained model, on the other hand, every hospital is evaluated only on its own merits and is potentially the marginal hospital. And when each hospital is considered in isolation, waiting time appears a more dominant attribute than quality.

The linear probability model does not constrain the cardinality of the pre-selected choice set. In contrast to that, the constrained pre-selection model does. Its estimates show that the cost of including choice alternatives in $\mathcal{J}^a$ is driven predominantly by the GP practice size in terms of number of GPS at the practice. The larger the practice, the lower the cost of including hospitals into the pre-selected choice sets. As discussed earlier, one may not be able to entirely rule out the presence of measurement error in the construction of consideration sets. If this measurement error were correlated with practice size, then the coefficient on the number of GPs at the practice level would be biased upward in absolute value. The homogeneity of the patient pool at the GP practice level in terms of age plays a role as well, albeit a more muted one. The estimates show that practices with a more homogeneous patient pool in terms of age, i.e. with a lower coefficient of variation for patient age, face higher costs of, or lower net benefits from, including hospitals into $\mathcal{J}^a$.

A key policy issue is how demand responds when hospital attributes change. While this can be easily and unambiguously assessed at the patient level - because patients make a binary choice -, responses at the GP level cannot be unambiguously assessed because GPs make choices on sets of hospital alternatives, and these choices involve both, the ranking of hospital alternatives and the marginal contribution of a hospital to the valuation of the set of highest ranking hospital alternatives. So the GP level response will depend on whether the hospital whose attributes changes is or is not a marginal hospital in the GP's choice of pre-selected set. Appendix C.4 provides details of the various cases that can arise.

From the estimation of the pre-selection model, we retain those draws μ_{ij} that are consistent with the bounds $(C-4)$ identified through the GP's pre-selection outcomes. These are estimates of residual hospital quality at the GP level.

5.2 Patient Level Choice

The patient level hospital choice model is specified as a multinomial logit model. Next to $\mathbf{x}_c$ - residual distance and waiting time -, the model includes, as $\mathbf{x}^p$, the number of parking spaces at the hospital as an amenity that is considered by the patient, but not the GP.

At the level of actual patient choice, waiting time is treated as potentially endogenous. Indeed, patients may face longer waiting times at higher quality hospitals that are popular with, and chosen by, many patients; a regression of waiting times on mortality rates (HSMR) and the exogenous hospital attributes yields a statistically significant negative coefficient. The analysis therefore employs the control function approach (Blundell and Powell (2003), Petrin and Train (2010)), including the residuals from the regression of waiting times on HSMR and exogenous hospital attributes (wait res) among the hospital attributes. This approach is valid because, absent parametric restrictions across the GP and patient parts of the model, the GP level pre-selection model – that uses HSMR among the hospital attributes – can be estimated separately,[31] while HSMR is excluded from the patient level hospital choice model.[32]

To control for the effect of pre-selection our estimates of residual hospital quality, the residuals backed out from the pre-selection model estimations, $\{\hat{\mu}_{ij}, j \in \mathcal{J}_i^a\}$, are also included. To the extent that GPs convey to patients any quality information about the pre-selected hospitals that does not only factor into the GPs' pre-selection, but also into patients' choice decisions, e.g. through patients' own quality assessments, these residuals would be expected to show up statistically significant in the patient level choice model.

[31] The model by Gaynor et al. (2016) exhibits the same separability property.

[32] We also considered the number of A&E admissions and the fraction of A&E admissions that exceeded the 4-hour waiting time target as possible instruments. However, this data is only available at the Trust level, and some Trusts have multiple hospital sites with an A&E department.

Table 7 presents coefficient and elasticity estimates of the patient level hospital choice model, conditional on the choice sets pre-selected by the patient's GP. Both sets of residuals, from the constrained pre-selection and the unconstrained linear probability model, are accounted for.

In line with the the existing hospital choice literature (e.g. Beckert et al. (2012), Beckert and Kelly (2017), Gaynor et al. (2016)), distance - albeit to be interpreted as residual distance - is the dominant hospital attribute from the patient's perspective.

Waiting times are also found to be substantively and statistically significant. The magnitude of the waiting time elasticity is smaller than the one reported by Sivey (2012). Our preferred waiting time coefficient estimate of -0.6189 implies that a change in waiting time by 3 months, i.e. about 3 standard deviations, results in a demand reduction on the order of 1.86 percent. Sivey (2012) reports an estimate of demand being reduced by 5 percent.

We note that our finding on waiting time is also shared with Beckert et al. (2012) and Beckert and Kelly (2017), but Gaynor et al., in their analysis of coronary artery bypass graft surgery, find no or positive waiting time effects. The result that the first-stage residuals from the regression of waiting times on HSMR enter as statistically significant into the model is novel and establishes the endogeneity of waiting times.

We now turn to the role of patients' perception of residual hospital quality in their choice from the pre-selected set. The residuals obtained from the constrained pre-selection model appear insignificant in the patient level model. This means that there is no correlation between μ_{ij} and ζ_{ij}, i.e. between patients' and GPs' quality assessment unobserved by the econometrician.

This is not what we would necessarily expect. For example, if both a distant and a nearby hospital are pre-selected by the GP, then the negative effect of distance on utility implies that the residual quality of the distant hospital is higher than that of the nearby hospital, and the GP conveys this to the patient via the inclusion of the distant hospital in the pre-selected set. Our model allows patients to take account of such differences in residual quality. Our estimates, however, show that such residual quality differences do not affect patients' choices.

One interpretation of this finding is that patients perceive all hospitals included

by the GP in the pre-selected set as of sufficiently high quality, and that any variation in residual quality above that threshold is irrelevant to them. This is consistent with qualitative evidence that patients themselves to not take quality in account (Dixon and Robertson (2009)). This finding supports the modelling strategy that patients defer to GPs with regard to hospital quality. And it supports the estimation strategy by which the GP level pre-selection and the patient level choice models can be estimated separately without bias provided the coefficients on $\mathbf{x}_c$ are allowed to differ between patient and GP. As discussed earlier, joint estimation is required if the model imposes a parametric restriction across the GP and patient parts of the model.

The residuals from the linear probability model do enter the model as statistically significant, with a positive coefficient. But the reason for this finding is that these residuals can be thought of as embedding a hospital fixed effect which is proportional to the fraction of GP practices that include a given hospital in the set $\mathcal{J}^a$ of pre-selected hospitals. Hence, the residuals from the linear probability model merely capture the frequency with which hospitals are offered, and more frequently offered hospitals are more likely to be chosen.[33] Beckert et al. (2012) report a similar result. [34] This also explains the slightly higher value of the log likelihood function in the model using this set of residuals.

Finally, Table 8 presents the same two multinomial logit specifications without conditioning on $\mathcal{J}^a$ and, instead, simply considering the set of the fifteen nearest hospitals as the patient's choice set. Comparing these with the results from the models that condition on $\mathcal{J}^a$, as in Table 7, it is seen that the residual distance effect and the residual distance elasticity are overestimated in absolute value. The reason is that distance was seen to be the dominant pre-selection criterion on the part of the GP. Therefore, non-selected hospitals, among the 15 nearest in $\mathcal{J} \setminus \mathcal{J}^a$, tend to be more distant on average, and in estimation the low choice incidence of distant hospitals among patients induces a large (in absolute value) estimate of the distance

[33]For example, consider hospitals A,B, and C in GP1's consideration set, and hospitals C,D and E in GP2's consideration set; suppose, GP1 selects B and C, and GP2 selects C and D. Then the FE for C is higher than for B and D, simply because it is in both GPs' consideration set, even if GP1 ranks B higher than C and GP2 ranks D higher than C. Everything else equal, the FE for C is twice the FE for B and D, respectively.

[34]See their Table 1, which reports a positive coefficient on GP referral frequency.

coefficient. At the same time, the waiting time effect is slightly underestimated compared to the model that conditions on $\mathcal{J}^a$. This may be explained by the fact that patients, when facing a set $\mathcal{J}^a$ of nearby, roughly equidistant hospitals of similar quality pre-selected by the GP, prefer hospitals with shorter waiting times. Finally, the effect of amenities, like parking, is not identified. While they are known to matter to patients (Dixon and Robertson (2009), Goldman and Romley (2008)), their effect risks being diluted when patient and GP are collapsed into a seemingly sole decision making entity. So, on the basis of the plausibility of the substantive findings, this model does worse than the comparator model that conditions on $\mathcal{J}^a$.

Comparing the purely statistical performance of the patient level choice models with the residuals from the constrained GP pre-selection model, the model conditioning on $\mathcal{J}^a$ exhibits a substantially higher log-likelihood function than the one conditioning on $\mathcal{J}$. In terms of its ability to predict, the proportion of correctly predicted choices, 0.74, is about as high as the one of the model conditioning on $\mathcal{J}$, 0.75. So the statistical metrics corroborate our preference for the model that conditions on the GP level pre-selected set $\mathcal{J}^a$ and the imputed residuals from the constrained GP pre-selection model.

In summary, our analysis may caution against ignoring, and simplistic modelling, of strategic pre-selection of choice sets, especially in the class of logit models popular with applied researchers.

6 Conclusions

This paper considers the microeconometric analysis of GP / patient choice processes in which the ultimate beneficiary of the choice outcome, the patient in the role of the principal, is advised by a GP, the principal's agent, through the GP's strategic pre-selection of a choice set for the patient. The paper presents a specific application to hospital choice for an elective procedure, hip replacements, in the setting of the English NHS. The empirical analysis illuminates the biases and inconsistencies that may result from ignoring the strategic pre-selection of choice sets on the part of the agent. Our analysis offers a refined perspective on the impact of distances and waiting times, distinguishing their contributions to GP pre-selection and patient

choice, and unlike many conventional models it identifies the effect of attributes that for many patients shape their perioperative experience, like amenities.

The results of the two-stage model show that GPs, in their role as agent of the patient, consider hospital quality when offering choice alternatives to patients, along with other attributes like distance and waiting times that patients are known to care about. However, the results also show that patients do not care about residual hospital quality differences and, instead, focus on attributes such as amenities that for them are tangible and relevant, but are unlikely to be considered by GPs.

Furthermore, the results provide evidence that GPs to some extent internalize costs of referrals that fall on PCT budgets. They respond to some incentives, like the MFF, that arise from their other role as agent of health authorities and the need to manage a budget for provision of care for the whole local population. The finding that GPs respond to financial incentives during the period when PCTs had the legal responsibility for NHS budgets is novel, points to potential conflicts of interest on the part of GPs, and as such is important for policy makers and potentially controversial.

It is worth noting in this regard that the NHS went broadly through three funding regime: (1) the fund holding period (1991-1999)[35], the PCT period which we study, and the subsequent period, from 2013, with CCGs replacing PCTs. In the CCG period, policy pressures on GP are likely the highest across the three regimes, particularly with regard to finance, so that the patient-physician agency relationship can be expected to be even weaker. So the fact that we find that GPs respond to financial incentives in the period when policy pressures were weakest implies that GPs conflict of interest and influence of choice set pre-selection in the other two periods can be expected to be even stronger.

There are at least five policy implications of this finding. First, in a system that promises equal access for equal need, heterogeneity in referral patterns across GPs as a consequence of the idiosyncratic incentives they face needs to be monitored. Indeed, it may raise the question whether there ought to be some national guidance

[35]During the fund holding period, GPs could personally benefit from referral decisions. Studies of the fund holding period into GPs budgetary responsibilities (Dusheiko et al. (2006), and Dusheiko et al. (2007)) show that the patient-physician agency relationship was weak, i.e. GPs do not act solely in the interest of patients.

with regard to referral advice, once a referral is deemed necessary (an assessment that may result from a referral management system).

Second, to the extent that patients may concentrate at hospitals that GPs favour to bolster up local trusts, irrespective of hospital quality, such patterns would tend to produce worse health outcomes. That would be problematic in its own right. And it would be an even bigger concern if it were to affect the most vulnerable patients.

Third, to the extent that GPs conflicts of interest are more acute in the current CCG regime, it would be worth monitoring whether GP led choice diversion is further accelerated.

Fourth, our results stress that GPs operate as businesses, and as such they face a range of incentives, from financial to non-financial ones. The latter may include their reputation with patients, with consultants, with health administrators; their day-to-day work load, etc.; and these non-financial incentives may be as strong as, if not stronger than, the financial ones we quantify in our study. Hence, the general finding that GPs are imperfect agents of their patients can be expected to hold irrespective of the particular funding arrangements that GPs operate under. Of course, the converse may also be true: In a competitive fee-for-service market, GPs may be too lenient in their gatekeeper role; see for example the study by Markussen and Røed (2017) of the Norwegian market for paid sick leave in this Journal, or Brekke et al. (2017). For this reason, we concur with Siciliani et al. (2017) in that policy design needs to take the role of the GP into account, and it needs to do so within the specific context of the respective health care funding architecture.

Fifth, the results could be of interest to policy makers because of their impact on competition in the health care sector. They show that GPs make some fairly complex trade-offs, which would suggest they shape competition in publicly funded health care services, equilibrating between excessive quality competition in a fixed-price system and excessive price competition at the expense of quality. In fact, this is in line with how hospitals appear to interact with GPs, as conduits to patients. Merger investigations by the UK competition authority, for example, have found evidence of hospitals focusing their marketing efforts on GPs. For example, in Royal Bournemouth and Christchurch Hospital NHS Foundation Trust / Poole Hospital NHS Foundation Trust merger inquiry (2013), the Competition Commis-

sion found that the merging parties had strategies to engage with GPs via a GP newsletter. Those examples are consistent with evidence from the Cooperation and Competition Panel of hospitals responding to competitive incentives in a variety of ways, including proactive GP engagement. Recognising the pivotal role of GPs in the competitive make-up of the NHS funded health care architecture in England, researchers have used qualitative methods to try to understand what drives GPs' choices. The analysis in this paper, to our knowledge, is among the first to formally model the role of GPs and quantify their incentives and their impact on patient choice outcomes.

We close by suggesting potential avenues of future research. Our analysis is limited by the data available to us: Neither the GPs' consideration sets, nor the patients' true choice sets are unambiguously observed; future research would benefit from data that encompass information on the set of alternatives that GPs consider and the actual sets of alternatives that are being discussed with each patient. Given our data limitations, we assume that pre-selected choice sets are tailored to a hybrid patient, i.e. they are uniform across patients at a GP practice; this is practicable in our situation, but with less limited data, our methodology could be adapted to allow for pre-selected sets that are tailored to each patient. Future research might also explore the role of GPs in the choice of NHS funded patients between public and private providers, notably in light of different capacity and performance metrics, in terms of breadth and depth, published by the two types of hospitals. It might also be useful to investigate whether GPs' financial incentives have strengthened since CCG are in control of budgets. And it might be interesting to see whether "disruptive entrants" into the health care sector, such as Artificial Intelligence harvesting big data to predict outcomes[36], fundamentally alter the patient – provider relationship.

References

[1] Aggarwal, A., Lewis, D., Mason, M., Sullivan, R. and J. van der Meulen (2017): "Patient Mobility for Elective Secondary Health Care Services in Repsonse to

[36]See, for example, JAMA, doi:10.1001/jama.2016.17216, a study into detection of diabetic retinopathy by means of "deep learning algorithms".

Patient Choice Policies: A systematic Review", *Medical Care Research and Review*, **74(4)**, 379-403

[2] Anderson, S.P. and A. de Palma (1992): "The Logit as a Model of product Differentiation", *Oxford Economic Paper*, **44**, 51-67

[3] Armstrong, M. and J. Zhou (2011): "Paying for Prominence", *The Economic Journal*, Vol. 121, Issue 556, F369-395

[4] Beckert, W., Christensen, M. and K. Collyer (2012): "Choice of NHS-Funded Hospital Services in England", *The Economic Journal*, Vol. 122, Issue 560, 400-417

[5] Beckert, W. and E. Kelly (2017): "Divided by Choice? Private Providers, Patient Choice and Hospital Sorting in the English National Health Service", Institute for Fiscal Studies working paper W17/15

[6] Besanko, D., Perry, M.K. and R.H. Spady (1990): "The Logit Model of Monopolistic Competition: Brand Diversity", *The Journal of Industrial Economics*, **38(4)**, 397-415

[7] Billingley, P. (1996): *Probability and Measure*, New York: Wiley

[8] Blundell, N., Clarke, A. and N. Mays (2010): "Interpretations of referral appropriateness by senior health managers in five PCT areas in England: a qualitative investigation", *Quality and Safety in Health Care*, **19(3)**, 182-186

[9] Blundell, R. and J.L. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models", in: *Econometric Society Monographs*, **36**, 312-357.

[10] Brekke, K.R., Holmas, T.H., Monstad, K. and O.R. Straume (2017): "Competition and Physician Behaviour: Does the Competitive Environment Affect the Propensity to Issue Sickness Certificates?", CESifo Working Paper No. 6672

[11] Cardell, N.S. (1991): "Variance Components Structures for the Extreme Value and Logistic Distributions", mimeo, Washington State University

[12] Chamley, C.P. (2004): *Rational Herds: Economic Models of Social Learning*, Cambridge: Cambridge University Press

[13] Cooper, Z., Gibbons, S., Jones, S. and and A. McGuire (2011): "Does Hospital Competition Save Lives? Evidence From The English NHS Patient Choice Reforms", *Economic Journal*, **121(554)**, 228-260

[14] Cooperation and Competition Panel (2011): "Review of the Operation of Any Willing Provider for the Provision of Routine Elective Care", available at `http://webarchive.nationalarchives.gov.uk`

[15] Crawford, G.S., Griffith, R. and A. Iaria (2016): "Demand estimation with unobserved choice set heterogeneity", unpublished manuscript

[16] Croxson, B., Propper, C. and A. Perkins (2001): "Do doctors respond to financial incentives? UK family doctors and the GP fundholder scheme", *Journal of Public Economics*, **79(2)**, 375-398

[17] Coulter, A., le Maistre, N. and L. Henderson (2005): "Patients' experience of choosing where to undergo surgical treatment - evaluation of London patient choice scheme", Ocford: Picker Institute Europe

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

[19] De Corniere, A. and G. Taylor (2014): "Integration and Search Engine Bias", *RAND Journal of Economics*, **45(3)**, 576-597

[20] Department of Health (2004): *The NHS Improvement Plan: Putting people at the heart of public services*. London: Department of Health

[21] Dinerstein, M., Einav, L., Levin, J. and N. Sunderesan (20014): "Consumer Price Search and Platform Design in Internet Commerce", mimeo, Stanford University

[22] Dixon, A. and R. Robsertson (2009): "Choice at the Point of Referral", available at `http://www.kingsfund.org.uk/publications/choice-point-referral`

[23] Dixon, A., Robertson, R., Appleby, J., Burge, P., Devlin, N. and H. Magee (2010): "Patient Choice", available at `http://www.kingsfund.org.uk`

[24] Dusheiko, M., Gravelle, H., Jacobs, R. and P. Smith (2006): "The effect of financial incentives on gatekeeping doctors: evidence from a natural experiment", *Journal of Health Economics*, **25(3)**, 449-478

[25] Dusheiko, M., Gravelle, H., Yu, N. and S. Campbell (2007): "The impact of budgets for gatekeeping physicians on patient satisfaction: evidence from fundholding", *Journal of Health Economics*, **26(4)**, 742-762

[26] Eliaz, K. and R. Spiegler (2011): "A Simple Model of Search Engine Pricing", *The Economic Journal*, Vol. 121, Issue 556, F329-339

[27] Eizenberg, A. (2014): "Upstream innovation and product variety in the US home PC market", *The Review of Economic Studies*, **81(3)**, 1003-1045

[28] Gaynor, M (2006): "What Do We Know About Competition and Quality in Health Care Markets?", *Foundations and Trends in Microeconomics*, **2(6)**, 441-508

[29] Gaynor, M., Moreno-Serra, R. 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)**, 134-166

[30] Gaynor, M., Ho, K. and R. Town (2014): "The Industrial Organization of Healthcare Markets", NBER working paper 19800

[31] Gaynor, M., Propper, C. and S. Seiler (2016): "Free to Choose? Reform and Demand Response in the English National Health Service", *The Amercian Economic Review*, **106(11)**, 3521-57

[32] Goldman, D. and J.A. Romley (2008): "Hospitals as Hotels: The Role of Patient Amenities in Hospital Demand", NBER working paper 14619

[33] Gowrisankaran, G. and R.J. Town (1999): "Estimating the quality of care in hospitals using instrumental variables", *Journal of Health Economics*, **18(6)**, 747-767

[34] Gutacker, N., Siciliani, L., Moscelli, G. and H. Gravelle (2016): "Choice of hospitals: Which type of quality matters?", *Journal of Health Economics*, **50**, 230-246

[35] Griliches, Z. and V. Ringstad (1970): "Error-in-theVariables Bias in Nonlinear Contexts", *Econometrica*, **38(2)**, 368–370

[36] Hagiu, A. and B. Jullien (2011): "Why Do Intermediaries Divert Search", *RAND Journal of Economics*, **42**, 337-362

[37] Harding, A. J., Sanders, F., Lara, A. M., van Teijlingen, E. R., Wood, C., Galpin, D., Barond, S., Crowe, S. and S. Sharma (2014): "Patient Choice for Older People in English NHS Primary Care: Theory and Practice", *ISRN family medicine*, 742676, published online 2014 Mar 4. doi: 10.1155/2014/742676

[38] Hausman, J.AS. and D.A. Wise (1979): "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment", *Econometrica*, **47(2)**, 455-473

[39] Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models", *Annals of Economic and Social Measurement*, **5**, 475-492

[40] Howard, J.A. and J.N. Sheth (1969): *The Theory of Buyer Behavior*, New York: Wiley

[41] Ishii, J. (2005): "Compatibility, Competition, and Investment in Network Industries: ATM Networks in the Banking Industry", mimeo, Yale University

[42] Jacobi, L. and M. Sovinsky (2016): "Marijuana on Main Street? Estimating Demand in Markets with Limited Access", *American Economic Review*, **106(8)**, 2009-2045

[43] Kelly, E. and G. Stoye (2014): "Does GP Practice Size Matter? GP Practice Size and the Quality of Primary Care", IFS Report R101

[44] Leary, T.B. (2001): "An Inside Look at the Heinz Case", available at: `https://www.ftc.gov/public-statements/2001/12/inside-look-heinz-case`

[45] Maddala, G.S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

[46] McFadden, D.L. (1974): "Conditional logit analysis of qualitative choice behaviour", in P. Zarembka, ed.: *Frontiers in Economics*, 105-142, New York: Academic Press

[47] McFadden, D.L. (1977): "Modelling the Choice of Residential Location", *Cowles Foundation Discussion Paper* No. 477

[48] McFadden, D.L. (1978): "Modelling the Choice of Residential Location", in: A. Karlqvist, L. Lundqvist, F. Snickars, J.W. Weibull (Eds.), *Spatial interaction theory and planning models*, North-Holland, Amsterdam, pp. 75 - 96

[49] Markussen, S. and K. Røed (2017): "The market for paid sick leave", *Journal of Health Economics*, **55**, 244–261

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

[51] Monitor (2015): "Choice in Adult Hearing Services: Exploring How Choice is Working for Patients", available from `https://www.gov.uk/government/`

[52] National Audit Office (2005): "Knowledge of the Choose and Book Programme Amongst GPs in England - A Survey of GPs' Opinions for the National Audit Office", available at `https://www.nao.org.uk/wp-content/uploads/2005/01/0405180_survey.pdf`

[53] Oxtoby, J.C. (1952): "Ergodic sets", *Bulletin of the American Mathematical Society*, **58(2)**, 116-136

[54] Pakes, A., Porter, J., Ho, K. and J. Ishii (2011): "Moment Inequalities and Their Application", mimeo, Harvard University

[55] Petrin, A. and K. Train (2010): "A control function approach to endogeneity in consumer choice models", *Journal of marketing research*, **47(1)**, 3-13

[56] Propper, C. and J. Van Reenen (2010): "Can Pay Regulation Kill? Panel Data Evidence on the Effect of Labor Markets on Hospital Performance", *Journal of Political Economy*, **118(2)**, 222-73

[57] Rosen, R., Florin, D. and R. Hutt (2007): "An Anatomy of GP Referral Decisions - A Qualitative Study of GPs' Views on Their Role in Supporting Patient Choice", available from `www.kingsfund.org.uk/publications`

[58] Santos, R., Gravelle, H. and C. Propper (2013): "Does quality affect patients' choice of doctor? Evidence from the UK", University of York, Centre for Health Economics, working paper No. 88

[59] Siciliani, L., Chalkley, M. and H. Gravelle (2017): "Policies Towards Hospital and GP Competition in Five European Countries, *Health Policy*, **121(2)**, 103-110

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

[61] Skellern, M. (2016): "The hospital as a multi-product firm: Measuring the effect of hospital competition using value-added, procedure-specific indicators of clinical quality", London School of Economics, unpublished manuscript

[62] Sovinsky Goeree, M. (2008): "Limited Information and Advertising in the US Personal Computer Industry", *Econometrica*, **76(5)**, 1017-74

[63] Weyl, E. G. and Fabinger, M. (2013): "Pass-through as an economic tool: Principles of incidence under imperfect competition", *Journal of Political Economy*, **121(3)**, 528-583

[64] White, H. (1994): *Estimation, Inference and Specification Analysis*, Econometric Society Monographs No.22, Cambridge: Cambridge University Press

Appendices

A Tables and Figures

Table 1: **Hospital Attributes**

	Waiting time[a]	Residual Distance [b]	HSMR[c]	MFF[d]	Parking[e]
Min	3	0.11	0.71	0.93	0
Lower Quartile	75	20.15	0.93	0.96	241
Median	93	37.20	1.02	0.97	376
Upper Quartile	108.5	64.23	1.06	1.06	545
Max	204	411.09	1.25	1.20	3215
Std. Dev	29.66	50.33	0.10	0.07	385.67

Source: HES and Health and Social Care Information Centre (HSCIC). 168 NHS hospital sites. [a] Median waiting time, calculated from HES inpatient records, at site level, in days. [b] Residual distances are straight line distances between the patient's LSOA and the postcode of the NHS hospital site in the patient's GP's pre-selected choice set. [c] Hospital standardized mortality ratio, at Trust level. [d] Market Forces Factor, at Trust level. [e] Number of parking spaces, at site level.

Table 2: **GP Practice Attributes**

	Coeff. of Var. w.r.t. Age[a]	Mean Patient Age[b]	Consideration Set Distance[c]	Pre-selected Set Distance[d]	Number of GPs
Min	5.16	.	0.11	0.11	1
Lower Quartile	13.27	60.5	16.56	5.15	3
Median	19.22	66.33	31.08	11.33	4
Upper Quartile	28.37	70.97	56.11	23.26	7
Max	50.13	96	411.09	317.04	21
Std. Dev.	.	11.34	44.62	23.42	3.06

Source: HES and Health and Social Care Information Centre (HSCIC). Based in 4,721 GP practices, selected as described in B. [a] The coefficient of variation w.r.t. age is the standard deviation, divided by the mean, of patient age at the GP practice level, in percent. [b] Mean patient age is calculated at practice level. [c] Distances are average straight line distances, in km, to NHS hospitals in GP consideration sets. [d] Distances are straight line distances, in km, between practice postcode and NHS hospitals in pre-selected sets.

Table 3: **Number of Hospitals Referred to, at GP Practice Level**

#	Freq.	Percent	Cum.
1	2,037	43.15	43.15
2	1,633	34.59	77.74
3	703	14.89	92.63
4	253	5.36	97.99
5	75	1.59	99.58
6	18	0.38	99.96
7	2	0.04	100.00
Total	4,721	100.00	

Source: HES.

Table 4: **Taxonomy of Choice Models: Who observes, resp. considers, what?**

symbols		variable	GP	patient
		indirect utility		
v_{ij}	$\mathbf{x}_j^a$	MFF	✓	not cons.
		HSMR	✓	not cons.
		⋮	✓	not cons.
	u_{ij} $\mathbf{x}_{ij}^c$	distance	✓	✓
		wait.time	✓	✓
	$\mathbf{x}_{ij}^p$	parking	not cons.	✓
		⋮	not cons.	✓
	ξ_{ij}		unobs.	✓
		cost		
	$\mathbf{z}$	GPs	✓	not rel.
		Coeff.Var.Age	✓	not rel.

Constrained Pre-Selection: v_{ij} – benefit of hospital j for patient i, as evaluated by i's GP; u_{ij} – patient i's indirect conditional utility of hospital j; ξ_{ij} – GP's incomplete information: hospital attributes relevant to patient i, but unobserved by patient i's GP. Variable classification. MFF: market forces factor; HSMR: Hospital standardised mortality ratio.

Table 5: **Taxonomy of Econometric Model: Who observes, resp. considers, what?**

symbols	variable	GP	patient	econometrician
	indirect utility			
α_{ij} $\mathbf{x}_j^a$	MFF	✓	not cons.	✓
α_{ij} $\mathbf{x}_j^a$	HSMR	✓	not cons.	✓
α_{ij} δ_{ij} $\mathbf{x}_{ij}^c$	distance	✓	✓	✓
α_{ij} δ_{ij}	wait.time	✓	✓	✓
α_{ij} δ_{ij} $\mathbf{x}_{ij}^p$	parking	not cons.	✓	✓
μ_{ij} μ_j^a		✓	not rel.	unobs.
μ_{ij} ζ_{ij} μ_{ij}^c		✓	✓	unobs.
μ_{ij} ζ_{ij} $\xi_{ij}^c = \xi_{ij}$		unobs.	✓	unobs.
μ_{ij}^p		not cons.	✓	unobs.
ξ_{ij}^p		not cons.	✓	unobs.
$\epsilon_{ij} = \mu_{ij}^p + \xi_{ij}^p$		not rel.	✓	unobs.
	cost			
$\mathbf{z}$	GPs	✓	not rel.	✓
	Coeff.Var.Age	✓	not rel.	✓

Constrained Pre-Selection: Notation as detailed in Appendix C.2. Variable classification. MFF: market forces factor; HSMR: Hospital standardised mortality ratio.

Table 6: **GP Pre-Selection**

	Constrained Choice[a]		Unconstr. Linear Prob. Model[b]	
	Coeff.	Std.Err.	Coeff.	Std.Err.
distance	-0.0666***	0.0007	-.0721***	.0012
mff	-0.0173***	0.0005	-.0255***	.0017
hsmr	-0.0150***	0.0033	-.0074***	.0017
waiting time	-0.01207***	0.0005	-.01468***	.0012
const			.1259***	.0012
σ	0.0876***	0.0008		
τ_0	-0.1939***	0.0003		
GPs	-0.3649***	0.0010		
Coeff. of Var. w.r.t. Age	-0.0299***	0.0009		

HES and Health and Social Care Information Centre (HSCIC).
*** significant at 1 percent level; ** significant at 5 percent level; * significant at 10 percent level. All regressors are standardized. mff: market forces factor, at Trust level; hsmr: hospital standardised mortality rate, at Trust level. [a] Constrained GP level pre-selection model, based on C-4; details in Appendix C.2.3 and C.3. [b] Unconstrained linear probability model; estimated by OLS.

Table 7: **Patient Hospital Choice, Conditional on $\mathcal{J}^a$**

	Residuals from Constrained Choice Model		Residuals from Unconstrained Linear Prob. Model	
	Coefficient Estimates			
	Coeff.	Std.Err.	Coeff.	Std.Err.
residual distance	-1.9992***	0.0409	-2.5540***	0.1431
parking	0.0218**	0.0118	0.0248**	0.0118
waiting time	-0.6189***	0.0831	-.5932***	0.0832
wait res[a]	0.0246***	0.0028	0.0208***	0.0029
constr res[b]	-0.0389	0.0343		
unconstr res[c]			6.423***	1.577
log lik	-16315.218		-16307.72	
MSPE[d]	0.74		0.79	
	Elasticity Estimates			
	Elasticity[e]	Std.Err.[f]	Elasticity[e]	Std.Err.[f]
residual distance	-1.879	1.912	-2.443	2.455
waiting time	-1.811	0.600	-1.740	0.674
parking	0.023	0.021	0.027	0.025

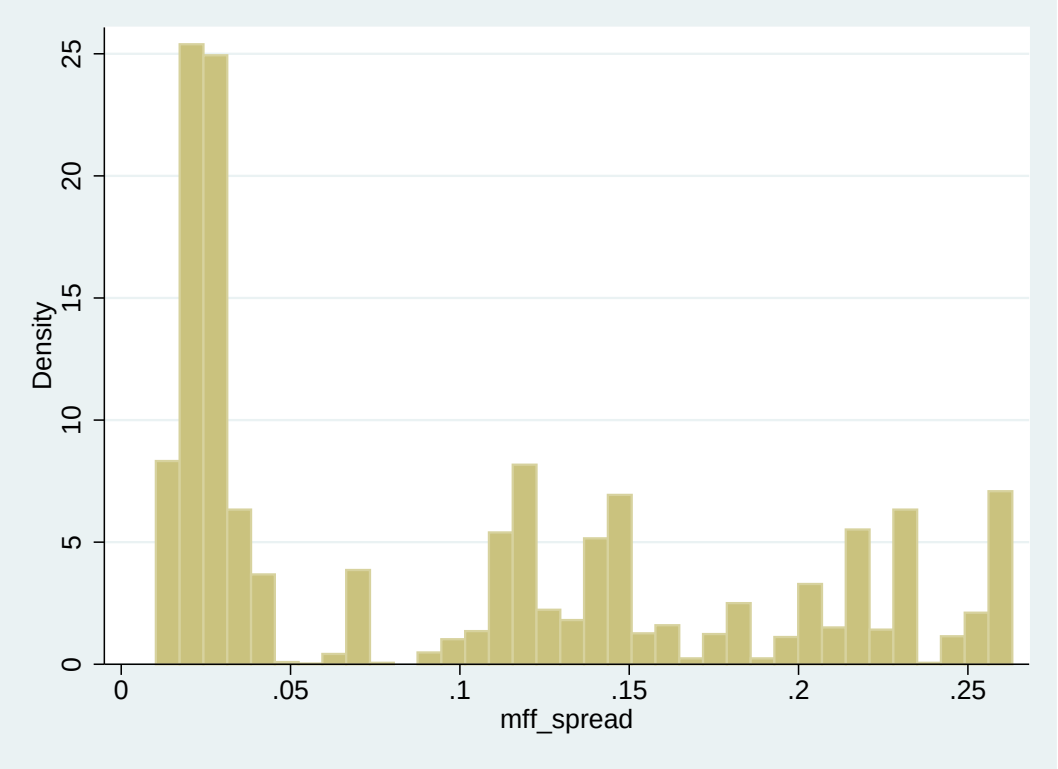
Notes: Conditional Logit model estimates, conditional on GP level pre-selected choice sets, $\mathcal{J}^a$. *** significant at 1 percent level; ** significant at 5 percent level; * significant at 10 percent level. The regressors dist, parking and waiting time are standardized; [a] wait res: residual from 1st stage regression of waiting times on hospital quality measures. [b] Constrained residuals imputed from GP pre-selection model. [d] Mean square prediction error, which corresponds to the fraction of correctly predicted choices. [c] Unconstrained residuals obtained from linear probability model for GP pre-selection. [e] Mean estimated elasticity. [f] Empirical standard error of estimated elasticities.

Table 8: **Patient Hospital Choice, Conditional on $\mathcal{J}$**

	Residuals from Constrained Choice Model		Residuals from Unconstrained Linear Prob. Model	
	Coefficient Estimates			
	Coeff.	Std.Err.	Coeff.	Std.Err.
residual distance	-6.6719***	0.0499	-2.9645***	0.0939
parking	0.0137	0.0093	0.0308	0.0123
waiting time	-0.2095***	0.0669	-0.5211***	0.0889
wait res[a]	0.0107***	0.0022	0.0175***	0.0030
constr res[b]	-0.0118	0.0314		
unconstr res[c]			11.0478***	0.9466
log lik	-27543.777		-24250.632	
MSPE[d]	0.75		0.71	
	Elasticity Estimates			
	Elasticity[e]	Std.Err.[f]	Elasticity[e]	Std.Err.[f]
residual distance	-6.383	6.414	-2.982	2.995
waiting time	-0.614	0.225	-1.531	0.599
parking	0.015	0.014	0.033	0.030

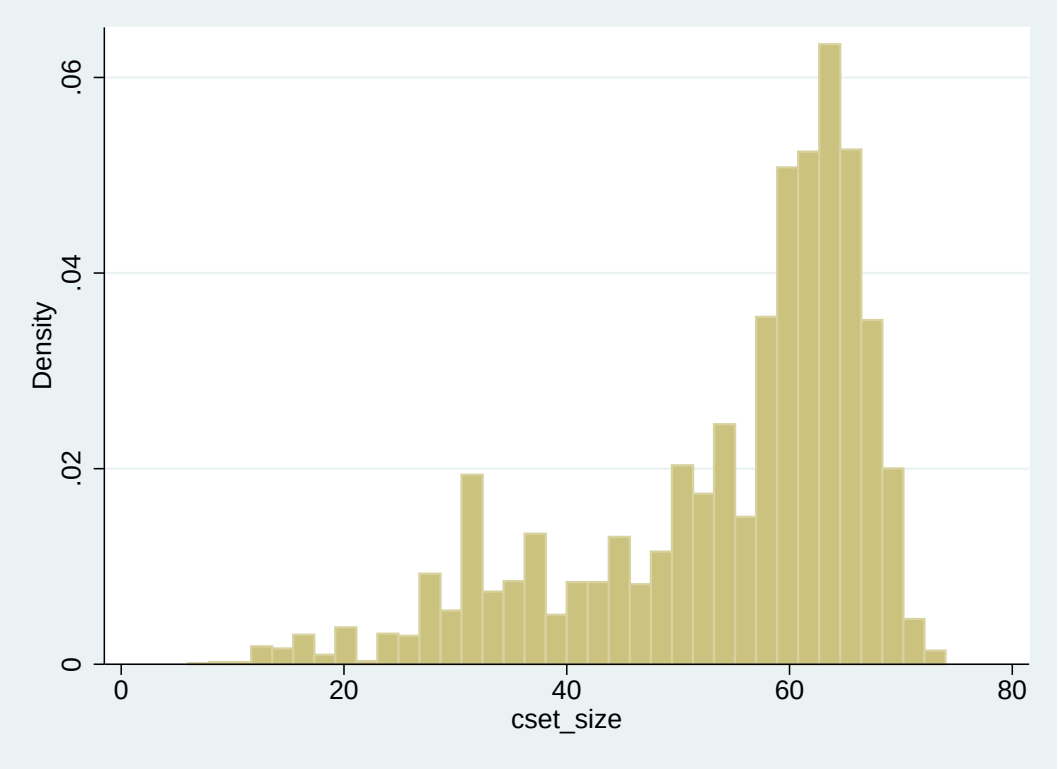
Notes: Conditional Logit model estimates, conditional on choice sets comprising 15 hospitals closest to patient, $calJ$. *** significant at 1 percent level; ** significant at 5 percent level; * significant at 10 percent level. The regressors dist, parking and waiting time are standardized; [a] wait res: residual from 1st stage regression of waiting times on hospital quality measures. [b] Constrained residuals imputed from GP pre-selection model. [d] Mean square prediction error, which corresponds to the fraction of correctly predicted choices. [c] Unconstrained residuals obtained from linear probability model for GP pre-selection. [e] Mean estimated elasticity. [f] Empirical standard error of estimated elasticities.

Figure 1: **MFF Spread at GP Practice Level**



Notes: The MFF spread is defined as the difference between maximum and minimum MFF among hospitals in the GP practice's consideration set. The minimum MFF across all GP practices is 0.929279, while the maximum MFF is 1.202005.

Figure 2: **Consideration Set Sizes at GP Practice Level**



Notes: The minimum number of hospitals considered at any GP practice is 7.

B Selection of GP Sample

This appendix provides details on how the sample of 4,721 GPs was selected.

There are 7,966 GP practices in our data that refer patients with relevant HRGs (healthcare resource groups). Of these, only 5,209 GP practices refer patients with the relevant TFCs (treatment function codes). Of these, 99.88% (i.e. all but 6) refer to no more than 7 hospitals. These are large practices with more than 30 GPs. They are neither typical of our sample of practices nor for the population of practices over the 2011/12 period we study (Kelly and Stoye (2014)) and therefore excluded.

Consider the consideration set definition according to the algorithm described in the Identification section of the paper. We start with 418 sites in our data. For 198 sites we have full attributes data (incl. amenities, e.g. parking etc.). For 189 sites we also have a provider MFF and HSMR. For 182 in addition we have (median) waiting time. Of these, 168 are in the GPs' consideration sets, i.e. the respective GP is in at least one of these hospitals' catchment area. There are 196 GP practices that are not in any of these hospitals' catchment areas. This leaves 5,007 GP practices that are in the catchment areas of the 168 hospitals.

These GPs have 50,497 hip replacement patients. Of these, some choose sites that are not in the GPs' consideration sets (e.g. sites that are dropped because there are no attributes, or distant site, etc.). And some GP practices have fewer than 15 hospitals in their consideration sets (see Figure 2).

267 GP practices refer only to more distant hospitals than the 15 closest in their consideration sets. 19 GPs have only patients who choose hospitals that are not in the GPs' consideration sets because the GPs are not in the respective hospitals' catchment areas. Eliminating these GP practices results in our sample of 4,721 GP practices.

C Econometric Model Specification

We start with a description of the micro-theoretic modelling approach to the GP's pre-selection problem. We then turns to the econometric specification of this GP pre-selection model and of the patient level choice model.

C.1 Micro-theoretic Modelling Approach to GP Pre-Selection

The model proposed in this section encompasses costs of information acquisition and dissemination. Such costs are low for "experts" such as GPs, but high for "laymen" such as patients. They thereby create a role for the former to pre-select choice sets out of the universe of choice alternatives for the benefit of the latter. The model shows how misalignment of incentives between GP and patients leads to different sets of hospitals to choose from. This can be interpreted as an inefficiency in the choice process, in that it induces a divergence between the distribution of choice outcomes under pre-selection and the distribution of choice outcomes in the absence of information costs. It also shows that, to the extent that the GP does not have complete information about the patients' evaluation criteria and does not tailor the pre-selected choice sets to the idiosyncratic evaluation outcomes of the patient, but instead offers a uniform choice sets to all patients, a further divergence is introduced, enhancing the level of inefficiency of the choice process.

Let u_{ij} denote patient i's indirect conditional utility of hospital alternative j, and v_{ij} the benefit assessment of hospital j for patient i from the perspective of patient i's agent (GP). The valuations u_{ij} and v_{ij} may differ for two reasons. First, patients and GP may evaluate different sets of attributes of hospital j. For example, GPs have access to hospital quality information that patients either do not have or find difficult to interpret. And they may reflect incentives the GP faces as agent of health authorities, e.g. with regard to financial implications captured by the MFF. We label such hospital attributes by $\mathbf{x}_j^a$. Patients may pay attention to hospital amenities that do not matter to GPs. Such hospital attributes are labelled by $\mathbf{x}_{ij}^p$. Hospital attributes considered by both, GP and patient, are labelled by $\mathbf{x}_{ij}^c$; they include, for example, distance and waiting time. The misalignment assumptions with

regard to GP and patient evaluation criteria imposed in this model are justified in Section 4.2.

Second, some of the attributes that matter to patients may not be observed by the GP, i.e. GPs may have incomplete information about the full set of evaluation criteria relevant to patients. We label such unobservables to the GP by ξ_{ij}. The term ξ_{in} could represent, for example, that someone among patient i's family and friends did, or did not, endorse hospital j; the patient's GP may not know about this, and indeed the patient may not want to divulge this, for fear of appearing prejudiced. This incomplete information assumption is necessary to motivate that GPs are imperfect agents for patients. As a consequence, they present a set of options, rather than simply making a choice on behalf of patients. It is for this reason that governments mandate choice.

As a reference for this subsection, the columns labelled "GP" and "patient" of Table 4 summarize the (mis-)alignment structure of the micro-theoretic GP and patient models and the GP's incomplete information. It encapsulates the micro-theoretic model of patient i's GP's valuation of hospital j,

$$v_{ij} = \alpha_{ij} + \xi_{ij} = \mathbf{x}_j^{a\prime}\theta_a + \mathbf{x}_{ij}^{c\prime}\theta_c + \xi_{ij},$$

and the micro-theoretic model of the patient's valuation,

$$u_{ij} = \delta_{ij} = \mathbf{x}_{ij}^{c\prime}\theta_c + \mathbf{x}_j^{p\prime}\theta_p,$$

where θ_a, θ_c and θ_p are parameter vectors that capture the sensitivity of valuations to the respective hospital attributes.

Turn now to the GP's problem of pre-selecting the composition of the set of hospitals $\mathcal{J}_i^a$ for patient i to choose from, out of the set $\mathcal{J}$ of all hospitals that i's GP considers. Suppose that, from the GP's perspective, there is a unit cost $C > 0$ of including a hospital alternative into $\mathcal{J}_i^a$. This cost may be specific to the GP. For example, in the context of hospital choice in the UK where a GP (practice) plays the role of the patient's agent, this cost might be expected to be a convex function $c(\mathbf{z})$ of practice level patient heterogeneity and the number of GPs in the practice. It imposes a constraint that can be thought of as the effort the GP needs to exert in order to explain the features, pros and cons of the alternative to the patient. This

perspective on GP decision making is supported by qualitative evidence (Rosen et al. (2007)).

Let $\mathcal{P}$ denote the set of all subsets of $\mathcal{J}$, i.e. $\mathcal{P} = \{\mathcal{G} \subset \mathcal{J} : \#\mathcal{G} \leq \#\mathcal{J}\}$. The value of any candidate set $\mathcal{G}$ of hospitals for patient i from the GP's perspective is $\mathbb{E}\left[\max_{j\in\mathcal{G}}(v_{ij} + \xi_{ij})\right]$, net of the cost of including the hospitals in the set, $c(\mathbf{z})\#\mathcal{G}$. If the unobservables ξ_{ij} are i.i.d. extreme value with location parameter zero and scale parameter σ, then the gross benefit of the set $\mathcal{G}$ is the inclusive value of $\mathcal{G}$ in the terminology of the discrete choice literature,

$$\begin{aligned} I_{\mathcal{G}}(\mathbf{x}_i^c, \mathbf{x}^a) &= \mathbb{E}\left[\max_{j\in\mathcal{G}}(v_{ij} + \xi_{ij})\right] \\ &= \ln\left(\sum_{j\in\mathcal{G}} \exp\left(\frac{v_{ij}}{\sigma}\right)\right), \end{aligned}$$

where $\mathbf{x}_i^c = [\mathbf{x}_{ij}^c]_{j\in\mathcal{G}}$, and

$$\mathcal{J}_i^a = \arg\max_{\mathcal{G}\in\mathcal{P}} \left\{I_{\mathcal{G}}(\mathbf{x}_i^c, \mathbf{x}^a) - c(\mathbf{z})\#\mathcal{G}\right\}.$$

The solution to this problem is to rank hospital alternatives in $\mathcal{J}$ in terms of v_{ij} and to choose the highest ranked alternatives, up to the point where the marginal contribution to $\mathbb{E}\left[\max_{j\in\mathcal{G}} v_{ij}\right]$ of the next ranked hospital is less than the marginal cost of its inclusion, $c(\mathbf{z})$, i.e.

$$\begin{aligned} I_{\mathcal{J}_i^a\cup\{j\}}(\mathbf{x}_i^c, \mathbf{x}^a) - I_{\mathcal{J}_i^a}(\mathbf{x}_i^c, \mathbf{x}^a) &< c(\mathbf{z}) \ \forall j \in \mathcal{J} \setminus \mathcal{J}_i^a \\ I_{\mathcal{J}_i^a}(\mathbf{x}_i^c, \mathbf{x}^a) - I_{\mathcal{J}_i^a\setminus\{j\}}(\mathbf{x}_i^c, \mathbf{x}^a) &\geq c(\mathbf{z}) \ \forall j \in \mathcal{J}_i^a. \end{aligned} \tag{C-1}$$

It is at this stage of pre-selection that the distinction between the GP as expert agent and the patient, as layman principal, emerges and can be defined: The GP (expert) has sufficient information and expertise to establish a ranking of the alternatives in $\mathcal{J}$, while the patient (layman) does not; for laymen, the cost of establishing such a ranking are likely to be prohibitive. This distinction is an implicit assumption in the present setup. The distinction creates a role for the GP, namely to pre-select, and thereby narrow down, the set of choice alternatives in order to render the patient's choice problem less complex and more tractable.

The set $\mathcal{J}_i^a$ resulting from the GP's pre-selection may differ, however, from the one that would be chosen if the assessment were based on u_{ij} (encompassing $\mathbf{x}_i^c$

and $\mathbf{x}_i^p$), instead of v_{ij} (encompassing $\mathbf{x}_i^c$ and $\mathbf{x}^a$), i.e. if the patient's and GP's assessment criteria were perfectly aligned, in the sense that they were to consider the same set of attributes of the choice alternatives as decision relevant. Denote the choice set that would have been pre-selected on the basis of $\{u_{ij}\}$ by $\mathcal{J}_i^p$. The efficiency loss due to pre-selection by the GP can then be cast as

$$\begin{aligned}\Delta_i &= I_{\mathcal{J}}(\mathbf{x}_i^c, \mathbf{x}_i^p) - I_{\mathcal{J}_i^a}(\mathbf{x}_i^c, \mathbf{x}_i^p) \\ &= I_{\mathcal{J}}(\mathbf{x}_i^c, \mathbf{x}_i^p) - I_{\mathcal{J}_i^p}(\mathbf{x}_i^c, \mathbf{x}_i^p) + I_{\mathcal{J}_i^p}(\mathbf{x}_i^c, \mathbf{x}_i^p) - I_{\mathcal{J}_i^a}(\mathbf{x}_i^c, \mathbf{x}_i^p).\end{aligned}$$

The first term captures the efficiency loss due to the reduction in complexity of the choice problem, while the second term captures the additional efficiency loss arising from a misalignment of assessment criteria between patient and GP which results in a choice set $\mathcal{J}_i^a$ which may be suboptimal when evaluated on the basis of the attributes $\mathbf{x}_i^c$ and $\mathbf{x}^p$ relevant to patient i.

The pre-selected choice sets $\mathcal{J}_i^a$ vary across patients i, to the extent that the attributes considered by both, GP and patient, $\mathbf{x}_{ij}^c$, vary with i; e.g. distance between i and hospital j. In practice, the GP may pre-select a uniform choice set $\mathcal{J}^a$ at the outset on the basis of $\mathbf{x}^a$ and $\mathbf{x}^c$ as they relate to the "average patient" and then offer this set to all patients at the practice. This wedge between the pre-selected choice set based on average attributes, rather than those specific to i, introduces yet another layer of potential inefficiency into the choice mechanism, so that the total inefficiency across all patients i at a GP practice can be measured by

$$\Delta = \sum_i \left[I_{\mathcal{J}^a}(\mathbf{x}_i^c, \mathbf{x}_i^p) - I_{\mathcal{J}_i^a}(\mathbf{x}_i^c, \mathbf{x}_i^p) + \Delta_i \right].$$

C.2 Econometric Choice Model

This section describes the econometric model specifications of the patient level choice model and the GP level pre-selection model. It details the econometrician's incomplete information about various model components and describes how this information structure can give rise to econometric selection biases.

C.2.1 Econometric Specification: The Patient's Choice Problem

As above and in Table 4, let $\mathbf{x}_{ij}^c$ denote hospital j's attributes that are taken into account by both, GP and patient; $\mathbf{x}_{ij}^p$ those that only matter to the patient; and $\mathbf{x}_j^a$ those that only matter to the GP, in the role of the patient's agent. For simplicity, suppose that patient and GP attach the same weights (coefficients) θ_c to $\mathbf{x}_{ij}^c$, and specify

$$\delta_{ij} \quad = \quad \mathbf{x}_{ij}^{c'}\theta_c + \mathbf{x}_{ij}^{p'}\theta_p,$$

where θ_c and θ_p are parameter vectors. The indirect utility of alternative j to patient i, latent to the econometrician, is then

$$u_{ij}^\star \quad = \quad \delta_{ij} + \zeta_{ij} + \epsilon_{ij} = \mathbf{x}_{ij}^{c'}\theta_c + \mathbf{x}_{ij}^{p'}\theta_p + \zeta_{ij} + \epsilon_{ij},$$

where ζ_{ij} and ϵ_{ij} are errors unobserved by the econometrician. The decomposition of the econometric error into ζ_{ij} and ϵ_{ij} and how it relates to the econometric errors in the GP model is explained below.

Let Y_{ij} be a binary indicator taking value one if i chooses alternative j, i.e. if $u_{ij}^\star = \max\{u_{ik}^\star, k \in \mathcal{J}_i^a\}$, and zero otherwise. Condition on the set of hospital alternatives $\mathcal{J}_i^a$ pre-selected by the GP.[37] Under the assumption that the errors ϵ_{ij}^p are i.i.d. type 1 extreme value and assuming that patient i takes the pre-selected choice set $\mathcal{J}_i^a$ as given[38], conditional on $\zeta_i' = [\zeta_{ij}]_{j\in\mathcal{J}}$,

$$\begin{aligned} \Pr(Y_{ij} = 1|\mathcal{J}_i^a, \zeta_i) \quad &= \quad \frac{\exp(\delta_{ij} + \zeta_{ij})}{\sum_{k\in\mathcal{J}_i^a}\exp(\delta_{ik} + \zeta_{ik})}, \qquad j \in \mathcal{J}_i^a \\ &= \quad 0 \qquad\qquad\qquad\qquad\qquad\quad j \notin \mathcal{J}_i^a. \end{aligned}$$

Except for the unobserved ζ_{ij}s, this is a conventional logit model with choice set $\mathcal{J}_i^a$.

C.2.2 Econometric Specification: The GP's Selection Problem

Recall that the GP's assessment of i's valuation of alternative j, is $v_{ij}^\star = \alpha_{ij} + \xi_{ij}$, where $\alpha_{ij} = \mathbf{x}_{ij}^{c'}\theta_c + \mathbf{x}_j^a\theta_a$ is a linear function of $\mathbf{x}_{ij}^c$ and $\mathbf{x}_j^a$, θ_a is a vector of parameters,

[37] In the general setting of this subsection, $\mathcal{J}_i^a$ may depend on i, to the extent that the agent wholly espouses the attributes that principal i values and that these vary with i, e.g. distance.

[38] This amounts to assuming that the patient behaves non-strategically and does not question how the GP arrived at the pre-selection outcome $\mathcal{J}_i^a$.

and ξ_{ij} is an error term. It relates to the error term in the patient's model as follows. Suppose that the error term ζ_{ij} in the patient's valuation model $u^{\star}_{ij}$ can be decomposed into uncertainty $\mu^{c}_{ij} + \xi^{c}_{ij}$ with regard to the attributes taken into account by both, patient and GP,

$$\zeta_{ij} = \mu^{c}_{ij} + \xi^{c}_{ij},$$

while the remaining uncertainty with regard to attributes that only matter to the patient is captured by $\mu^{p}_{ij} + \xi^{p}_{ij} = \epsilon_{ij}$. Here, μ^{c}_{ij} and μ^{p}_{ij} are those parts of the econometrician's uncertainty about the two parts of δ_{ij} that are known to the GP, while ξ^{c}_{ij} and ξ^{p}_{ij} are unknown to both, GP and econometrician. From the perspective of the GP who cares only about the utility contribution related to $\mathbf{x}^{c}_{i}$, only the former matters. So, $\xi_{ij} = \xi^{c}_{ij}$. Consequently, from the perspective of the econometrician, in the model for the GP, μ^{c}_{ij} matters in addition to $\xi_{ij} = \xi^{c}_{ij}$.

To facilitate an overview of the information and consideration structure of this model as it relates to the GP, patient and econometrician, Table 5 provides an taxonomy of the components of the econometric model. It encapsulates the econometric version of the information structure of the mirco-theoretic model presented in Table 4 above.

Assuming, as above, the ξ_{ij} are i.i.d. extreme value with location parameter zero and scale parameter σ, the distribution of choice outcomes from the GP's perspective is given by logit choice probabilities based on attributes $\mathbf{x}^{c}$ and $\mathbf{x}^{a}$. Denote the econometrician's incomplete information about the GP (agent) specific relevant attributes $\mathbf{x}^{a}$ by μ^{a}_{j}. Once the $\{\xi_{ij}\}_{i\in\mathcal{J}}$ are integrated out, the econometrician's remaining uncertainty with regard to the agent's assessment of alternative j is therefore $\mu_{ij} = \mu^{c}_{ij} + \mu^{a}_{j}$. The solution to the GP's optimization problem is defined by

$$\mathcal{J}^{a}_{i} = \arg\max_{\mathcal{G}\in\mathcal{P}} \left\{ I_{\mathcal{G}}(\mathbf{x}^{c}_{i}, \mathbf{x}^{a}, \mu_{i}) - c(\mathbf{z})\#\mathcal{G} \right\}. \tag{C-2}$$

Analogous to C-1, the solution yields bounds on cost,

$$\underline{U}(\alpha_{i}, \mu_{i}) \leq c(\mathbf{z}) \leq \overline{U}(\alpha_{i}, \mu_{i}) \tag{C-3}$$

where $\alpha_{i} = (\alpha_{ij}, j \in \mathcal{J})$ and $\mu_{i} = (\mu_{ij}, j \in \mathcal{J})$. Appendix C.2.3 provides formal derivations.

Note that, considering just the GP level pre-selection of choice sets as the first part of the entire, two-stage choice model, the inequalities above allow moment based estimation of the set of values of $C = c(\mathbf{z})$ consistent with the above inequalities, next to the parameters in α_{ij}, using the methodology proposed in Pakes et al. (2011) and applied in Ishii (2005). In the present instance, moments are obtained by integrating out $\{\mu_{im}, m \in \mathcal{J}_i^a\}$ in the upper bounds, and in addition $\{\mu_{ij}, j \notin \mathcal{J}_i^a\}$ in the lower bounds. The setting differs from the one in Ishii (2005) in that in her work only the cardinality of the optimal set is chosen, while here in addition the specific elements of the optimal set are determined.[39]

Notice also that this model of GP pre-selection is reminiscent of the one proposed by Mehta et al. (2003). While these authors directly motivate their selection model in terms of the (inclusive) value of sets of alternatives, the model presented here motivates the way in which these inclusive values determine the pre-selected sets in terms of cost constrained optimization. This model can also be seen as an alternative to the selection model of Gaynor et al. (2016). In their model, the distance metric that defines the size of the pre-selected set is specified as a fixed distance from the alternative with maximal utility. The model of this paper can be interpreted instead as distance measured in terms of Kullback-Leibler divergence.[40] In the context of incomplete and asymmetric information, this information theoretic measure has particular intuitive appeal.

The econometrician cannot observe the ranking of the alternatives included in $\mathcal{J}_i^a$. From the inequalities C-5 provided in the Appendix, the set $\{\mu_{ij}\}_{j\in\mathcal{J}_i^a}$ must satisfy the necessary condition for inclusion of the jth alternative, so that

$$\begin{aligned} G(\mathcal{J}_i^a; \alpha_i, C) &= \left\{\{\mu_{ij}\}_{j\in\mathcal{J}_i^a} : \underline{U}(\alpha_i, \mu_i) \leq c(\mathbf{z}) \leq \overline{U}(\alpha_i, \mu_i)\right\} \qquad \text{(C-4)} \\ \Pr(\mathcal{J}_i^a; C) &= \Pr\left(G(\mathcal{J}_i^a; \alpha_i, C)\right). \end{aligned}$$

To the extent that $\mu_{ij} = \mu_{ij}^c + \mu_{ij}^a$ is correlated with ζ_{ij} through μ_{ij}^c, i.e. to the extent that μ_{ij}^c is non-zero with positive probability, observing $\mathcal{J}_i^a$ is informative

[39]Mapping the present setting onto the framework in Pakes et al. (2011), the agent level unobservable $\xi_{ij} = \xi_{ij}^c$ corresponds to their ν_1 terms, while the econometrician level unobservable $\mu_{ij} = \mu_{ij}^c + \mu_{ij}^a$ corresponds to their ν_2 terms.

[40]For example, Δ_i is the KL divergence between the distributions induced by $\{\Pr(Y_{ij}|\mathcal{J}_i^a), j \in \mathcal{J}_i^a\}$ and $\{\Pr(Y_{ij}|\mathcal{J}), j \in \mathcal{J})\}$, respectively.

about ζ_{ij}, so that $\Phi(\alpha_i, C) = \mathbb{E}[\zeta_{ij}|G(\mathcal{J}_i^a; \alpha_i, C)]$ accounts for pre-selection in this model. Just as in linear models with sample selection (Heckman (1976)), omission of such selection terms will yield biased and inconsistent estimates. The selection term here does not permit a closed-form solution and needs to be simulated. Details on simulation assisted estimation are provided in appendix C.3.

The contribution of patient i to the likelihood function is then given by

$$\Pr\left(Y_{ij}^p = 1|\mathcal{J}_i^a\right) \Pr(\mathcal{J}_i^a; C),$$

where

$$\Pr\left(Y_{ij}^p = 1|\mathcal{J}_i^a\right) = \frac{\exp\left(\delta_{ij} + \Phi(\alpha_i, C)\right)}{\sum_{k \in \mathcal{J}_i^a} \exp\left(\delta_{ik} + \Phi(\alpha_i, C)\right)}.$$

C.2.3 Details on Bounds in GP Pre-Selection Model

Recall that the solution to the GP's optimization problem C-2 is to order the alternatives in $\mathcal{J}$ according to their indirect utilities,

$$\begin{aligned} \exp\left(\frac{\alpha_{i(1:J)} + \mu_{i(1:J)}}{\sigma}\right) &= \exp\left(\frac{\mathbf{x}_{i(1:J)}^{c'}\theta_c + \mathbf{x}_{(1:J)}^{a'}\theta_a + \mu_{i(1:J)}}{\sigma}\right) \\ &\geq \cdots \\ &\geq \exp\left(\frac{\alpha_{i(J:J)} + \mu_{i(J:J)}}{\sigma}\right) \\ &= \exp\left(\frac{\mathbf{x}_{i(J:J)}^{c'}\theta_c + \mathbf{x}_{(J:J)}^{a'}\theta_a + \mu_{i(J:J)}}{\sigma}\right) \qquad \text{(C-5)} \end{aligned}$$

and to include the ones up to the point that

$$\begin{aligned} J_i^a &= \arg\max_{h \in \{1,\cdots,J\}} \left\{ \ln\left(\sum_{k=1}^{h} \exp\left(\frac{\alpha_{i(k:J)} + \mu_{i(k:J)}}{\sigma}\right)\right) - \ln\left(\sum_{m=1}^{h-1} \exp\left(\frac{\alpha_{i(m:J)} + \mu_{i(m:J)}}{\sigma}\right)\right) \geq c(\mathbf{z}) \right\} \\ &= \arg\max_{h} \left\{ -\ln\left(1 - \frac{\exp\left(\frac{\alpha_{i(h:J)} + \mu_{i(h:J)}}{\sigma}\right)}{\sum_{m=1}^{h} \exp\left(\frac{\alpha_{i(m:J)} + \mu_{i(m:J)}}{\sigma}\right)}\right) \geq c(\mathbf{z}) \right\} \end{aligned}$$

This yields an upper bound $\overline{U}(\alpha_i, \mu_i)$ on cost which is given by the increment to the inclusive value of $\mathcal{J}_i^a$ by the marginal included hospital,

$$\overline{U}(\alpha_i, \mu_i) = \ln\left(\sum_{k=1}^{J_i^a} \exp\left(\frac{\alpha_{i(k:J)} + \mu_{i(k:J)}}{\sigma}\right)\right) - \ln\left(\sum_{m=1}^{J_i^a - 1} \exp\left(\frac{\alpha_{i(m:J)} + \mu_{i(m:J)}}{\sigma}\right)\right).$$

This also implies that

$$-\ln\left(1-\frac{\exp\left(\frac{\alpha_{i(k:J)}+\mu_{i(k:J)}}{\sigma}\right)}{\sum_{m=1}^{J_i^a}\exp\left(\frac{\alpha_{i(m:J)}+\mu_{i(m:J)}}{\sigma}\right)}\right)\geq c(\mathbf{z})\quad\text{for }\ k=1,\cdots,J_i^a.$$

Similarly,

$$J_i^a+1=\arg\min_h\left\{-\ln\left(1-\frac{\exp\left(\frac{\alpha_{i(h:J)}+\mu_{i(h:J)}}{\sigma}\right)}{\sum_{m=1}^{h+1}\exp\left(\frac{\alpha_{i(m:J)}+\mu_{i(m:J)}}{\sigma}\right)}\right)\leq c(\mathbf{z})\right\}$$

implies a lower bound $\underline{U}(\alpha_i,\mu_i)$ on cost which is given by the increment to the inclusive value of $\mathcal{J}_i^a$ by the marginal excluded hospital,

$$\underline{U}(\alpha_i,\mu_i)=\ln\left(\sum_{k=1}^{J_i^a+1}\exp\left(\frac{\alpha_{i(k:J)}+\mu_{i(k:J)}}{\sigma}\right)\right)-\ln\left(\sum_{m=1}^{J_i^a}\exp\left(\frac{\alpha_{i(m:J)}+\mu_{i(m:J)}}{\sigma}\right)\right).$$

Also, for any $j\notin\mathcal{J}_i^a$,

$$-\ln\left(1-\frac{\exp\left(\frac{\alpha_{ij}+\mu_{ij}}{\sigma}\right)}{\exp\left(\frac{\alpha_{ij}+\mu_{ij}}{\sigma}\right)+\sum_{m\in\mathcal{J}_i^a}\exp\left(\frac{\alpha_{i(m:J)}+\mu_{i(m:J)}}{\sigma}\right)}\right)\leq c(\mathbf{z}).$$

C.3 Details on Estimation

The lower and upper bounds $\underline{U}(\alpha_i,\mu_i)$ and $\overline{U}(\alpha_i,\mu_i)$ of cost implied by the GP level pre-selection model depend on estimable parameters (θ_c,θ^a) - coefficients on distance, waiting time, MFF and HSMR - via α_i, and on errors $\mu_i=\{\mu_{ij},j\in\mathcal{J}\}$ which are unobserved by the econometrician. This is summarized in the columns labelled "GP" and "econometrician" in Table 5.

The unobservables μ_i induce the probability of observing the pre-selected set $\mathcal{J}_i^a$ out of the GP's consideration set $\mathcal{J}$. And ensemble of such probabilities constitutes the likelihood function of (θ_c,θ^a), given the collection of observed pre-selected sets.

The probabilities associated with the observed preselected sets are analytically intractable, i.e. the model does not permit a closed form expression for them. Therefore, they need to be simulated, i.e. replaced by simulated approximations. And

consequently, instead of estimating the parameters (θ_c, θ^a) and the parameters of the cost function (i.e. the coefficients on $\mathbf{z}$ in $c(\mathbf{z})$) by maximizing the log-likelihood function, they must be estimated by maximizing the simulated log-likelihood function (maximum simulated likelihood, MSL, estimation).

In order to simulate the probabilities of the observed pre-selected sets, 100 standard normal draws for each hospital j in a GP's consideration set $\mathcal{J}$ are obtained and retained throughout the estimation. For each candidate value of the parameters, the values $\alpha_{ij} + \mu_{ij}$ are computed and ranked in descending order.

Refer to a consistent simulation sample draw as one for which the simulated values $\alpha_{ij} + \mu_{ij}$ for $j \in \mathcal{J}_i^a$ are the highest in the ranking of the J alternatives and the implied bounds hold. The fraction of consistent simulation sample draws, out of the 100 draws, is a simulation estimator of the probability of observing the set $\mathcal{J}_i^a$. The SML estimation methodology uses these to build up the simulated log-likelihood function and maximizes it over the parameter space.

The consistent simulation sample draws are retained and constitute estimates of residual hospital quality at the GP level. They averages, at the hospital level for each GP, are used as regressors in the second stage choice model for that GP's patients.

C.4 Effect of Attribute Changes on Pre-Selected Sets

Let j^m denote the marginal included hospital alternative, and let j^e denote the marginal excluded hospital alternative.

The GP pre-selection problem, as set out in section C.1, implies that

$$\ln\left(\sum_{j\in\mathcal{J}_i^a}\exp\left(\frac{v_{ij}}{\sigma}\right)\right) - \ln\left(\sum_{j\in\mathcal{J}_i^a\setminus\{j^m\}}\exp\left(\frac{v_{ij}}{\sigma}\right)\right) \;\geq\; c(\mathbf{z}) \qquad \text{(C-6)}$$

$$\ln\left(\sum_{j\in\mathcal{J}_i^a\cup\{j^e\}}\exp\left(\frac{v_{ij}}{\sigma}\right)\right) - \ln\left(\sum_{j\in\mathcal{J}_i^a}\exp\left(\frac{v_{ij}}{\sigma}\right)\right) \;<\; c(\mathbf{z}). \qquad \text{(C-7)}$$

The interpretation of C-6 is that the contribution of the marginally included hospital

j^m to the inclusive value of the the pre-selected set $\mathcal{J}^a$ exceeds the cost of including an alternative.[41] And C-7 means that the contribution of the marginally excluded hospital j^e is less than that cost.

Note that

$$\ln\left(\sum_{j\in\mathcal{J}_i^a}\exp\left(\frac{v_{ij}}{\sigma}\right)\right)-\ln\left(\sum_{j\in\mathcal{J}_i^a\setminus\{j^m\}}\exp\left(\frac{v_{ij}}{\sigma}\right)\right)=-\ln\left(1-\frac{\exp\left(\frac{v_{ij^m}}{\sigma}\right)}{\sum_{j\in\mathcal{J}_i^a}\exp\left(\frac{v_{ij}}{\sigma}\right)}\right), \tag{C-8}$$

and similarly,

$$\ln\left(\sum_{j\in\mathcal{J}_i^a\cup\{j^e\}}\exp\left(\frac{v_{ij}}{\sigma}\right)\right)-\ln\left(\sum_{j\in\mathcal{J}_i^a}\exp\left(\frac{v_{ij}}{\sigma}\right)\right)=-\ln\left(1-\frac{\exp\left(\frac{v_{ij^e}}{\sigma}\right)}{\sum_{j\in\mathcal{J}_i^a\cup\{j^e\}}\exp\left(\frac{v_{ij}}{\sigma}\right)}\right), \tag{C-9}$$

Suppose a hospital attribute of hospital j changes, e.g. its waiting time increases. Then, three cases can arise.

First, hospital j can be an infra-marginal hospital. In that case, if the change is small enough in order not to affect the hospital ranking, then it follows from C-8 that the contribution of the marginal hospital to the inclusive value of the pre-selected set increases, so that this tends to enlarge the pre-selected set.

Second, if hospital j is (or becomes) the marginal hospital, then again C-8 implies that this change diminishes the contribution of the marginal hospital to the inclusive value of the pre-selected set. Hence, in this case the change tends to decrease the size of the pre-selected set.

Third, if hospital j is the excluded hospital, then this change is irrelevant for the pre-selected set because it further diminishes the value of the excluded alternative.

If, on the other hand, hospital j's attribute improves, i.e. its waiting time decreases, for example, then this change tends to decrease the size of the pre-selected set if hospital j is infra-marginal - the contribution of the marginal hospital is smaller -, or increase it if hospital j is the marginal hospital or initially excluded.

[41] Recall that the inclusive value of the pre-selected set $\mathcal{J}^a$ is given by $\ln\left(\sum_{j\in\mathcal{J}^a}\exp\left(\frac{v_{ij}}{\sigma}\right)\right)$.