Institute for Fiscal Studies

Choice in the presence of experts: the role of general practitioners in patients' hospital choice

IFS Working Paper W16/21

Walter Beckert
Kate Collyer

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

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

Walter Beckert† and Kate Collyer‡

October 27, 2016

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). It proposes a two-stage choice model that encompasses both, patient and GP level optimization, and it discusses identification. The empirical analysis demonstrates biases and inconsistencies that arise when strategic pre-selection is not properly taken into account. We find that patients defer to GPs when assessing hospital quality and focus on tangible attributes, like hospital amenities; and that GPs, in turn, as patients' agents present choice options based on quality, but as agents of health authorities also consider their financial implications.

Keywords: Discrete choice, patient, principal, GP, agent, expert, endogenous choice sets, competition, hospital choice, elective medical procedure.

JEL classification: D120, C510, I110, G110.

*With thanks to Penelope Goldberg, Rachel Griffith, Sandeep Kapur, Elaine Kelly, Chris Pike, Carol Propper, Ron Smith and Marcos Vera-Hernandez 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 NIC-211948-F5J9K. This paper has been screened to ensure no confidential information is revealed.

†Birkbeck College, University of London

‡Competition and Markets Authority

1 Introduction

In choice situations involving credence goods in which an "expert" agent with arguably superior information strategically presents a set of pre-selected choice alternatives to a principal decision maker, 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 General Practitioner (GP) who advises on their choice as a medical expert. Next to legal requirements, 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 treatment at a given hospital. 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,

[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

[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

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. As a result, GPs had to take account of the financial implications of their referral decisions.[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.

In conventional discrete choice analysis, e.g. conditional logit (McFadden (1974)) and its variants, choice sets are assumed to be exogenous. 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 parse. The choice modelling literature refers to pre-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 the third-party agent, the consideration set formation process in Metha et al. is part

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 postdates 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`

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 independent and exogenously driven by product advertisement, absent a constraint on the choice set size. Gaynor, Propper and Seiler (2016) model the GP led consideration set formation 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. Their model can be regarded as an alternative to the one proposed in this paper where distance is given an information theoretic interpretation and where heterogeneity in cost associated with utilitarian distance across experts (GPs) is modelled and 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 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 a principal and an 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 inaccurate 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 the 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 (2016), Gaynor et al. (2016), Santos et al. (2013)), 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 consider the hospital alternatives in this set as being of comparable quality and that they focus on other tangible hospital attributes. In particular, it shows that waiting times, once their endogeneity is taken account of, and hospital amenities are critical attributes to patients. In competing choice models, the effects of these attributes either appear implausible (e.g. Gaynor et al. (2016) who report positive waiting time effects for coronary artery bypass grafts[5]) or statistically in-

[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.

significant. 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, where it has conventionally been found to be the dominant driver of choice (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 lays out econometric models for the patient / GP decision process and discusses pertinent identification and estimation issues. Section 4 presents results from the estimation of these models. And Section 5 concludes, with a view to adaptations of the empirical strategy of this paper to similar principal-agent choice settings.

2 Institutional Background

The majority of primary and secondary health care in England is provided through the taxpayer funded National Health Service (NHS).[6] 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.[7] Patients obtain access to secondary care through a referral from their GP. GPs therefore act ask 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

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

[7] 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.

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 motivated by both, the belief that patients valued the choice over their care, and evidence that health care competition when prices were fixed could improve quality (Gaynor (2006)). 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, under the moniker "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 a pivot 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.[8] Under the Payments by Results NHS funding architecture,[9] commissioners (PCTs) pay health care providers, such as hospitals, a national tariff, i.e. a per patient payment based on the treatment they provide.[10] There is some variation in tariffs across hospitals captured by a Market 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. Since such treatments are funded from fixed PCT budgets, GP referral decisions have financial

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

[9] `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213150/PbR-Simple-Guid`

[10] 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 average cost of providing the associated care.

implications. Therefore, when making referrals, an important part of the GP's role is to act as a rationing agent on behalf of the PCT which pays for care (Blundell et al. (2010)).

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.

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 (2016), Gaynor et al (2016), Santos et al. (2013)) 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.[11] 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. Patients in the sample were treated at one of 168 hospitals that carried out at least 10 hip replacements in 2011/12 and for which a set of hospital attributes is available. The analysis only considers GP practices that refer to between one and seven hospitals.[12]

[11]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.

[12]There is a small tail of very large GP practices that refer to several dozen hospitals. These

For each patient 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 put 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). They also include the aforementioned Market Forces Factor (MFF) and hospital amenities, such as parking spaces at the hospital.

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.[13] They include the number of GPs at the practice: Larger practices enjoy a richer pool of experience and information and hence are likely to more easily facilitate choice. The analysis also considers measures of the homogeneity 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. This is motivated by evidence (Harding et al. (2014)) 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 patients of older ages.

The locational information regarding patients, GPs and hospitals sites permits calculating distances between hospitals and patients, and GPs respectively.

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

practices and their patients are excluded.

[13]The following section provides a detailed exposition of the two-stage choice model that discusses the role of costs at the first stage of GP-level pre-selection.

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 taken in this paper implicitly assumes that hospitals that were never chosen are not part of the choice set and discussion between GP and patient; and that even if they had featured in discussions, yet were never chosen, they would be eventually dropped, being irrelevant alternatives. 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.[14]

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)

[14]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.

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 giving patients choice was mandated already 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.[15]

The average age of hip replacement patients is 68.6, but the variation at the practice level is skewed to the left, i.e. towards practices with more homogeneous patient pools with respect to age. The mean number of GPs at the practice level is just below 4, equally skewed to the left, i.e. to practices with a small number of GPs. Table 4 summarizes these practice characteristics.

4 Econometric Model

This section describes a two-stage model for the GP and patient level choice process. It captures 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. In order to bring out the sample-selection issues arising in this context, a simple GP-level model is sketched first, absent any constraints on the size of the pre-selected set. This serves as a backdrop to the main model of GP-level cost-constrained pre-selection. The section also offers a discussion of salient identification issues.

[15]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.

4.1 Unconstrained Pre-Selection

This section presents a simple econometric two-stage decision model in which the first-stage pre-selection mechanism is unconstrained. It shows how choice set pre-selection at the first-stage induces features of incidental truncation into discrete choice analysis at the second stage that parallel the ones identified by Heckman (1976) in linear models. It thereby provides a reference point for a more comprehensive, strategic model of cost-constrained pre-selection in the subsequent section.

Consider patient i, the principal beneficiary of the choice outcome. The patient is to make a discrete choice out of a set of hospital alternatives $\mathcal{J}^a$ that is pre-selected by the GP, who acts as the patient's agent. The GP, in his capacity of medical expert, arguably possesses superior information, say on the hard-to-assess quality of all possible choice alternatives, collected in the set $\mathcal{J}$ of all conceivable hospital alternatives. Viewed as a two-stage decision mechanism, the role of the GP is, at a first stage, to pre-select $\mathcal{J}^a \subseteq \mathcal{J}$ for the benefit of the patient who selects an alternative out of the set $\mathcal{J}^a$ at the second stage.

Consider the stage of the GP's pre-selection. Denote the GP's latent assessment of alternative j's net benefit by $v_j^\star$; this could incorporate anticipated benefits accruing to patients, any benefits accruing to the expert as a result of incentivization schemes put in place by the hospital j or the PCT both it and the GP are located in; or any benefits accruing to the expert's reputation from promoting hospital j. Suppose that the GP includes j in $\mathcal{J}^a$ if, and only if, $v_j^\star > 0$:

$$\begin{aligned} v_j^\star &= \alpha_j + \xi_j, \\ v_j &= 1_{\{v_j^\star > 0\}}, \quad j \in \mathcal{J}, \end{aligned}$$

where α_j denotes the measurable component of $v_j^\star$, ξ_j is unobserved by the econometrician, and v_j is a binary inclusion indicator, taking value one when the agent's net benefit assessment is positive so that j is included in $\mathcal{J}^a$, and zero otherwise. Here, ξ_j might capture, in particular, the unquantifiable quality assessment of alternative j by the agent, e.g. to the extent that it affects the agent's prospective reputation or other subjective or "soft" attributes of alternative j. In this preliminary and simple framework, the GP has all the information relevant to him, each choice alter-

native is assessed by the GP individually and independently on its own merits, and $\mathcal{J}^a = \{j \in \mathcal{J} : v_j = 1\}$. An alternative and more realistic pre-selection mechanism is outlined below.

Now consider patient i, the ultimate beneficiary of the choice outcome. Suppose with any conceivable choice alternative i associates an indirect conditional utility $u^\star_{ij}$,

$$u^\star_{ij} = \delta_{ij} + \zeta_{ij} + \epsilon_{ij},$$

that comprises a measurable component δ_{ij}, next to unobserved components ζ_{ij} and ϵ_{ij}. Here, δ_{ij} might capture observable attributes of j that relate directly to i, e.g. geographic distance, coverage of specific idiosyncratic risks, etc. The (to the econometrician) unobservable ζ_{ij} might reflect quality aspects of alternative j that are unobserved by the econometrician, and it may or may not vary with i; a precise structure for ζ_{ij} is given in the following subsection. Patient i's idiosyncratic taste or preference for j, modelled by ϵ_{ij}, is also unobserved by the econometrician. The indirect utility that patient i associates with alternative j is latent, but inference about δ_{ij} is possible to the extent that j is included in $\mathcal{J}^a$, in that it can be observed whether or not j is chosen by i. Consider the case when ξ_j and ζ_{ij} are allowed to be correlated. This may arise when unobserved quality aspects of alternative j are at least partly relevant to both, the patient and the GP. This is plausibly so when the GP's reputation hinges on matching up decision makers, like patient i, with beneficial choice outcomes, like j. It can also arise from subjective assessments of "soft" (i.e. not easily quantifiable or measurable) attributes of the choice alternative.[16] Then, given $j \in \mathcal{J}^a$,

$$\begin{aligned} \tilde{u}^\star_{ij} &:= \mathbb{E}\left[u^\star_{ij} | j \in \mathcal{J}^a\right] \\ &= \delta_{ij} + \mathbb{E}[\zeta_{ij} | -\xi_j < \alpha_j] + \epsilon_{ij} \\ &= \delta_{ij} + \phi(\alpha_j) + \epsilon_{ij}, \end{aligned}$$

where $\phi(\alpha_j) = \mathbb{E}[\zeta_{ij} | -\xi_{ij} < \alpha_j]$ accounts for the effect of the GP's inclusion of

[16] In the medical context, for example, the patient and GP may differ in terms of what they consider relevant aspects of the perioperative care and environment: The GP may focus on strictly medical aspects (e.g. availability of specialist expertise for treating any comorbidities), while a patient may focus also on psychosocial aspects (e.g. psychological support to mitigate anxiety) which may affect somatic recovery.

j in $\mathcal{J}^a$; for example, if ξ_j and ζ_{ij} are bivariate standard normal with correlation $\rho \in (-1, 1)$, then this terms is the well-known Mills ratio, evaluated at α_j and pre-multiplied by ρ. The observed choice outcome is an indicator Y_{ij} taking value one when i choose j out of $\mathcal{J}^a$, i.e.

$$Y_{ij} = 1_{\left\{\tilde{u}^{\star}_{ij} = \max\{\tilde{u}^{\star}_{ik}, k \in \mathcal{J}^a\}\right\}},$$

and the probability distribution of Y_{ij} is induced by distributional assumptions on ϵ_{ij}.

Under extreme value type 1 assumptions on the ϵ_{ij}s, this yields choice probabilities of the logit form (McFadden (1978), Cardell (1991)),

$$\begin{aligned} \Pr\left(Y_{ij} = 1 | \, \mathcal{J}^a\right) &= \mathbb{E}[Y_{ij} | \mathcal{J}^a] \\ &= \frac{\exp((\delta_{ij} + \phi(\alpha_j)))}{\exp(I_{i\mathcal{J}^a})}, \end{aligned}$$

where $I_{i\mathcal{J}^a} = \ln\left(\sum_{k \in \mathcal{J}^a} \exp((\delta_{ik} + \phi(\alpha_k)))\right)$ is the inclusive value of the set comprising the pre-selected hospitals. The expression for $\Pr(Y_{ij} = 1 | \mathcal{J}^a)$ demonstrates that the selection terms $\phi(\alpha_j)$, $j \in \mathcal{J}^a$, constitute regressors that are omitted in analyses that ignore strategic choice set pre-selection by the GP, provided correlation between ξ_j and ζ_{ij} cannot be ruled out and the selection terms vary across $j \in \mathcal{J}^a$. Such omission will yield inconsistent maximum likelihood estimates, as a consequence of model mis-specification.

Sovinsky Goeree (2008) presents a related model of random choice or consideration sets at the level of the decision maker in which the probability of the decision maker being informed about a choice alternative j takes the place of the inclusion probability $\Pr(v_j = 1)$. In her model of the US personal computer industry, these probabilities are exogenously driven by product level advertising and consumer level media exposure.[17] Her model can be viewed as a special case of the present model in which ζ_{ij} and ξ_j are independent, conditional on observed attributes. Eizenberg (2014) and Jacobi and Sovinsky (2016) also estimate similar models and discuss Heckman style corrections for selection.

[17] See also Dinerstein et al. (2014) for an application to consumer search in internet commerce. Gaynor et al. (2014) emphasize the promise this approach holds in health care industrial organization research.

This type of model is less compelling in situations when information about choice alternatives is asymmetrically distributed and costly to acquire and disseminate. On the one hand, information acquisition costs render decision making complex for the uninformed layman principal. And on the other hand, they create a role for informed experts as agents, namely to reduce the complexity of the decision process for the layman. The following subsection describes an alternative model that captures these ideas.

4.2 Constrained Pre-Selection

4.2.1 Modelling Approach

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 merely partial alignment of relevant evaluation criteria between GPs and patients (experts and laymen, or the agent and the principal) introduces an inefficiency into 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 possess 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.

As a reference for this subsection, the columns labelled "GP" and "patient" of Table 1 summarize the (mis-)alignment structure of the GP and patient models and the GP's incomplete information. Details on the econometric specification and the econometrician's information will be provided in Sections 4.2.2-4.2.4.

The model distinguishes attributes of hospital j that matter to patient i, summarized in indirect utility u_{ij}, that are not perfectly aligned with those that the

Table 1: **Taxonomy of Choice Models: Who observes, resp. considers, what?**

symbols		var.	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: Variable classification. MFF: market forces factor; HSMR: Hospital standardised mortality ratio.

GP considers, summarized in v_{ij}. Attributes solely relevant to the GP, denoted $\mathbf{x}_j^a$, reflect incentives the GP faces as agent of health authorities, e.g. with regard to financial implications captured by the MFF. Attributes solely relevant to the patient, denoted $\mathbf{x}_{ij}^p$, reflect hospital amenities, e.g. parking. Attributes $\mathbf{x}_j^c$, such as distance, are considered by both. The misalignment assumptions imposed in this model are justified in Section 4.2.4. Finally, and importantly, ξ_{ij} captures attributes relevant to the patients that the GP does not observe. 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.

The patient's choice model is simply to select the hospital that is associated with maximal indirect utility u_{ij} out of the set J_i^a of hospital options pre-selected by the GP, i.e.

$$Y_{ij}^p = 1_{\{u_{ij}=\max\{u_{ik},k\in\mathcal{J}_i^a\}\}}.$$

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. 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 list size, practice level patient heterogeneity, the number of GPs in the practice, their work experience and whether they obtained their qualification in the UK or abroad. 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 partitions of $\mathcal{J}$, i.e. $\mathcal{P} = \{\mathcal{G} \subset \mathcal{J} : \#\mathcal{G} \leq \#\mathcal{J}\}$. Suppose the GP's objective in selecting $\mathcal{J}_i^a$ is to minimize the divergence of the distribution of patient level choice outcomes under pre-selection relative to their distribution absent pre-selection. The distribution of choice outcomes from the GP's perspective is induced by the GP's evaluation criteria v_{ij} which only partially overlap with those of the patient, and the GP's uncertainty about the patient's other evaluation criteria, ξ_{ij}. Hence, the GP's model of the patient's choice is

$$Y_{ij}^a = 1_{\{v_{ij}=\max\{v_{ik}+\xi_{ik},k\in\mathcal{J}\}\}}.$$

An information theoretic measure for the divergence between the two distributions

of outcomes with and without pre-selection is the Kullback-Leibler measure,[18]

$$D(\mathcal{J}_i^a||\mathcal{J}) = \sum_{j\in\mathcal{J}_i^a} \Pr(Y_{ij}=1|\mathcal{J}_i^a)\ln\left(\frac{\Pr(Y_{ij}^a=1|\mathcal{J}_i^a)}{\Pr(Y_{ij}^a=1|\mathcal{J})}\right),$$

where the probabilities are induced by $\{\xi_{ij}\}_{j\in\mathcal{J}}$, capturing the GP's incomplete information about the patient's salient evaluation criteria. Assuming that the ξ_{ij}s are i.i.d. extreme value with location parameter zero and scale parameter σ,

$$\begin{aligned} D(\mathcal{J}_i^a||\mathcal{J};\mathbf{x}_{c_i},\mathbf{x}^a) &= \ln\left(\sum_{k\in\mathcal{J}}\exp\left(\frac{v_{ik}}{\sigma}\right)\right) - \ln\left(\sum_{m\in\mathcal{J}_i^a}\exp\left(\frac{v_{im}}{\sigma}\right)\right) \\ &= I_{\mathcal{J}}(\mathbf{x}_{c_i},\mathbf{x}^a) - I_{\mathcal{J}_i^a}(\mathbf{x}_{c_i},\mathbf{x}^a), \end{aligned}$$

where $I_{\mathcal{J}}(\mathbf{x}_{c_i},\mathbf{x}^a)$ is the inclusive value of the choice alternatives in set $\mathcal{J}$, and similarly for $I_{\mathcal{J}_i^a}(\mathbf{x}_{c_i},\mathbf{x}^a)$. This divergence can be viewed as a loss in efficiency that arises from reducing the complexity of the choice problem, limiting it to evaluating $J_i^a = \#\mathcal{J}_i^a$ alternatives, instead of $J = \#\mathcal{J} \geq J_i^a$. The smaller this efficiency loss, the greater the benefit to the patient arising from the GP level pre-selection. The GP's optimization problem then is to select

$$\mathcal{J}_i^a = \arg\min_{\mathcal{G}\in\mathcal{P}} D(\mathcal{G}||\mathcal{J};\mathbf{x}_{c_i},\mathbf{x}^a) + c(\mathbf{z})\#\mathcal{G}.$$

The solution to this problem is to rank the hospital alternatives in terms of indirect utility v_{ij} so that the contribution of the marginal hospital to the inclusive value of the set of the highest value options $I_{\mathcal{J}_i^a}(\mathbf{x}_{c_i},\mathbf{x}^a)$ just exceeds the costs $c(\mathbf{z})$.[19] 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 $\mathcal{J}$ without cost, while the patient (layman) does not; for layman, the cost of

[18] The Kullback - Leibler divergence for two measures P and Q is $D(P||Q) = \mathbb{E}_P[\ln(P/Q)] = \sum_j P(j)\ln(P(j)/Q(j))$ and not symmetric. It requires that $Q(j)=0$ implies $P(j)=0$, i.e. that $\Pr(Y_{ij}=1|\mathcal{J})=0$ implies $\Pr(Y_{ij}=1|\mathcal{J}_i^a)=0$. In the present model, this is plausible. White (1994) offers an interpretation that, adapted to this model, implies that the divergence measures the "surprise" from learning that decision outcomes are in fact governed by $\{\Pr(Y_{ij}=1|\mathcal{J}_i^a), j\in\mathcal{J}\}$, rather than by $\{\Pr(Y_{ij}=1|\mathcal{J}), j\in\mathcal{J}\}$; his Assumption 3.4 is satisfied because $\{\Pr(Y_{ij}=1|\mathcal{J}), j\in\mathcal{J}\}$ is a probability distribution.

[19] Details of this solution are provided in Section 4.2.3.

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 $\mathbf{x}_{c_i}$ and $\mathbf{x}_i^p$), instead of v_{ij} (encompassing $\mathbf{x}_{c_i}$ 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

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

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 $\mathbf{x}_c$ and $\mathbf{x}^p$ relevant to the patient.

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, $\mathbf{x}_{c_{ij}}$, 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 measured by the KL divergence is

$$\begin{aligned}
\Delta &= \sum_i \left[D(\mathcal{J}^a||\mathcal{J}_i^a;\mathbf{x}_{c_i},\mathbf{x}_i^p) + D(\mathcal{J}_i^p||\mathcal{J};\mathbf{x}_{c_i},\mathbf{x}_i^p) + D(\mathcal{J}_i^a||\mathcal{J}_i^p;\mathbf{x}_{c_i},\mathbf{x}_i^p)\right] \\
&= \sum_i \left[D(\mathcal{J}^a||\mathcal{J}_i^a;\mathbf{x}_{c_i},\mathbf{x}_i^p) + \Delta_i\right].
\end{aligned}$$

Uniformity of the pre-selected choice set across i adds, for each patient i, an additional potential efficiency loss.

4.2.2 Econometric Specification: The Patient's Choice Problem

As above and in Table 1, let $\mathbf{x}_{c_{ij}}$ 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 $\mathbf{x}_{c_{ij}}$, and specify

$$\begin{aligned} \delta_{ij} &= \mathbf{x}'_{c_{ij}}\theta_c + \mathbf{x}_{ij}^{p'}\theta^p, \\ \alpha_{ij} &= \mathbf{x}'_{c_{ij}}\theta_c + \mathbf{x}_j^{a'}\theta^a \end{aligned}$$

where θ^a and θ^p are parameter vectors and α_{ij}, taking the role of α_j above, reflects the possible variation of $\mathbf{x}$ across i, in addition to j. The indirect utility of alternative j to patient i, latent to the econometrician, is then

$$u_{ij}^\star = \mathbf{x}'_{c_{ij}}\theta_c + \mathbf{x}_{ij}^{p'}\theta^p + \zeta_{ij} + \epsilon_{ij},$$

where, as above, ζ_{ij} and ϵ_{ij} are unobserved by the econometrician.

Condition on the set of hospital alternatives $\mathcal{J}_i^a$ pre-selected by the GP.[20] 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[21], conditional on $\zeta_i' = [\zeta_{ij}]_{j\in\mathcal{J}}$,

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

while, absent the pre-selection,

$$\Pr(Y_{ij}=1|\mathcal{J},\zeta_i) = \frac{\exp(\delta_{ij}+\zeta_{ij})}{\sum_{k\in\mathcal{J}}\exp(\delta_{ik}+\zeta_{im})} \quad j\in\mathcal{J}.$$

[20] In the 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.

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

This implies that the divergence of the distribution of patient level choice outcomes under pre-selection relative to their distribution absent pre-selection, in terms of the Kullback-Leibler measure, is

$$\begin{aligned} & D(\mathcal{J}_i^a || \mathcal{J}; \mathbf{x}_{c_i}, \mathbf{x}_i^p, \zeta_i) \\ = & \ln\left(\sum_{k\in\mathcal{J}} \exp(\mathbf{x}_{c_{ik}}\theta_c + \mathbf{x}_{ij}^{p'}\theta^p + \zeta_{ij})\right) - \ln\left(\sum_{m\in\mathcal{J}_i^a} \exp(\mathbf{x}'_{c_{im}}\theta_c + \mathbf{x}_{im}^{p'}\theta^p + \zeta_{im})\right) \\ = & I_{\mathcal{J}}(\mathbf{x}_{c_i}, \mathbf{x}_i^p, \zeta_i) - I_{\mathcal{J}_i^a}(\mathbf{x}_{c_i}, \mathbf{x}_i^p, \zeta_i). \end{aligned}$$

4.2.3 Econometric Specification: The GP's Selection Problem

Let the GP's assessment of i's valuation of alternative j, latent to the econometrician, be $v_{ij}^\star = \alpha_{ij} + \xi_{ij}$, where 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 $u_{ij}^\star$ can be decomposed into uncertainty $\mu_{ij}^c + \xi_{ij}^c$ with regard to the attributes taken into account by both, patient and GP,

$$\zeta_{ij} = \mu_{ij}^c + \xi_{ij}^c,$$

while the remaining uncertainty with regard to attributes that only matter to the patient is captured by $\mu_{ij}^p + \xi_{ij}^p = \epsilon_{ij}$. Here, μ_{ij}^c and μ_{ij}^p are those parts of the econometrician's uncertainty about the two parts of δ_{ij} that are known to the GP, while ξ_{ij}^c and ξ_{ij}^p are unknown to both, GP and econometrician. From the perspective of the GP who cares only about the utility contribution related to $\mathbf{x}_c$, only the former matters. So, $\xi_{ij} = \xi_{ij}^c$. Consequently, from the perspective of the econometrician, in the model for the GP, μ_{ij}^c matters in addition to $\xi_{ij} = \xi_{ij}^c$. To facilitate an overview of the information and consideration structure of this model as it relates to the GP, patient and econometrician, Table 2 provides an taxonomy of the components of the econometric model.

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

Table 2: **Taxonomy of Econometric Model: Who observes, resp. considers, what?**

symbols	var.	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		✓	NR	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.
$\epsilon_{ij} = \mu_{ij}^p + \xi_{ij}^p$		NR	✓	unobs.
	cost			
$\mathbf{z}$	GPs	✓	NR.	✓
	Coeff.Var.Age	✓	NR	✓

Constrained Pre-Selection: Variable classification. MFF: market forces factor; HSMR: Hospital standardised mortality ratio; NR: not relevant in model for respective column.

relevant attributes $\mathbf{x}^a$ by μ_j^a. 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 $\mu_{ij} = \mu_{ij}^c + \mu_j^a$. The solution to the GP's optimization problem

$$\mathcal{J}_i^a = \arg\min_{\mathcal{G}\in\mathcal{P}} D(\mathcal{G}||\mathcal{J}; \mathbf{x}_{c_i}, \mathbf{x}_i^p, \mu_i) + c(\mathbf{z})\#\mathcal{G}.$$

is to order the alternatives in $\mathcal{J}$ according to their indirect utilities,

$$\begin{aligned}
\exp\left(\frac{\alpha_{i(1:J)}+\mu_{i(1:J)}}{\sigma}\right) &= \exp\left(\frac{\mathbf{x}'_{c_{i(1:J)}}\theta_c+\mathbf{x}^{a'}_{(1:J)}\theta^a+\mu_{i(1:J)}}{\sigma}\right)\\
&\geq \cdots\\
&\geq \exp\left(\frac{\alpha_{i(JU:J)}+\mu_{i(J:J)}}{\sigma}\right)\\
&= \exp\left(\frac{\mathbf{x}'_{c_{i(J:J)}}\theta_c+\mathbf{x}^{a'}_{(J:J)}\theta^a+\mu_{i(J:J)}}{\sigma}\right) \qquad (4\text{-}1)
\end{aligned}$$

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 also implies that

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

Since $C=c(\mathbf{z})$ is unknown to the econometrician, this identifies an upper bound on C. 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, i.e. for any $j\notin\mathcal{J}_i^a$,

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

For example, suppose $\alpha_{i(m:J)}+\mu_{i(m:J)}=v$ for all $m=1,\cdots,J$. Then, the inequalities above imply

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

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.[22]

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 an information theoretic efficiency minimization problem subject to a cost constraint. 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 proposes instead the Kullback Leibler divergence as a distance measure. 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 4-1 above, the set $\{\mu_{ij}\}_{j\in\mathcal{J}_i^a}$ must satisfy the necessary condition for inclusion of the jth alternative, so that

$$\begin{aligned} G(\mathcal{J}_i^a;\alpha_i,C) &= \left\{\{\mu_{ij}\}_{j\in\mathcal{J}_i^a} : -\ln\left(1-\frac{\exp\left(\frac{\alpha_{ij}+\mu_{ij}}{\sigma}\right)}{\sum_{m\in\mathcal{J}_i^a}\exp\left(\frac{\alpha_{im}+\mu_{im}}{\sigma}\right)}\right)\geq c(\mathbf{z})\right\} \\ \Pr(\mathcal{J}_i^a;C) &= \Pr\left(G(\mathcal{J}_i^a;\alpha_i,C)\right). \end{aligned}$$

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 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, analogous to $\phi(\alpha_i)$ in the model with unconstrained pre-selection. Unlike in

[22]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.

the model of unconstrained pre-selection, the selection term here does not permit a closed-form solution and needs to be simulated.

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

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

4.2.4 Identification

The patient's choice model, i.e. δ_{ij} conditional on the pre-selected $\mathcal{J}_i^a$, is identified through patients' choices from this set and variation in attributes across choice alternatives. Regarding the GP's pre-selection model, α_{ij} is identified through variation in attributes across alternatives and their inclusion in, respectively exclusion from, $\mathcal{J}_i^a$. As shown through the bounds on cost above,

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

the cardinality of $\mathcal{J}_i^a$, i.e. the size of the pre-selected choice set, next to variation in cost drivers $\mathbf{z}$, identifies the agent's cost function $c(\mathbf{z})$. Furthermore, since the inclusive value of $\mathcal{J}_i^a$ is increasing in σ, albeit less than linearly, 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 $\mathbf{x}_{ij}^c$ 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 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 $\theta^a \neq \mathbf{0}$ 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 μ_{ij}; these can be used to impute ζ_{ij} which, in turn, can be used as embedded regressors in a second-step estimation of the patient's choice model.

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 attributes when choosing a hospital. 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 medical quality, measured by the Hospital Standardised Mortality Ratio (HSMR) which 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). 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. 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 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 this model are in the same units as is 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. The following algorithm is used in order to construct the sets $\mathcal{J}$ considered by GPs from which the choice sets $\mathcal{J}^a$ are pre-selected. For each hospital, 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 (Dixen 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. The approach is also consistent with GP survey evidence collected by Monitor (2015) about their referral practice: "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. 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.

5 Results

5.1 Estimation of Pre-Selection Model

Table 5 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 $\{\mu_{ij}, j \in \mathcal{J}\}$ being i.i.d. draws from a standard normal distribution.

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). 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 is consistent with research on the implementation of GP fundholding reforms in the early 1990s. That research found that health care providers did respond to financial incentives offered by the scheme (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 GPs to take account of financial implications of their referral decisions.

A notable difference between the constrained pre-selection and the unconstrained linear probability model is that the effect of waiting time dominates the quality effect in the latter, while the reverse is the case in the former.

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 $\mathcal{J}^a$.

5.2 Patient Level Choice

The patient level hospital choice model is specified as a multinomial logit model. Next to $\mathbf{x}_c$ - 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) yields a statistically significant negative coefficient. The analysis therefore employs the control function approach (Blundell and Powell (2003)), including the residuals from the regression of waiting times on

HSMR (wait res) among the hospital attributes. To control for the effect of pre-selection, the residuals backed out from the pre-selection model estimations 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.

Table 6 presents the 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 (2016), Gaynor et al. (2016)), distance is the dominant hospital attribute from the patient's perspective. Waiting times are also found to be substantively and statistically significant. This finding is shared with the former two studies, 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.

The residuals obtained from the constrained pre-selection model appear insignificant in the patient level model. This is what one should expect. The pre-selection outcome is a set of selected hospitals $\mathcal{J}^a$ that can only be ranked collectively vis-à-vis hospitals that are not selected, $\mathcal{J} \setminus \mathcal{J}^a$. The constrained pre-selection does not convey any information to the econometrician that would allow to rank them individually. From a substantive point of view, the interpretation of this finding is that patients defer to GP when it comes to the assessment of hospital quality. This is consistent with qualitative evidence that patients themselves to not take quality in account (Dixon and Robertson (2009)). On the basis of this finding, 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, i.e. provided $\kappa \neq 0$. 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.[23] Beckert et al. (2012) report a similar result. [24] This also explains the slightly higher value of the log likelihood function in the model using this set of residuals.

Finally, Table 7 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 6, it is seen that the distance effect is 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 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 matter to patients, their effect risks being diluted when patient and GP are collapsed into a seemingly sole decision making entity.

[23]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.

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

Taken together, these comparisons 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. Apart from overestimating the importance to patients of distance and underestimating that of waiting time, conventional models struggle to identify the effect of attributes that for many patients shape their perioperative experience, like amenities. The results of the proposed two-stage model also show that patients defer to GPs when it comes to hospital quality and, instead, focus on attributes such as amenities that for them are tangible and relevant, but are unlikely to be considered by GPs. GPs, on the other hand, are found to consider hospital quality when offering choice alternatives to patients, next to other attributes like distance and waiting times that patients are known to care about. But the results also reveal that these are not the only attribute dimensions that GPs respond to, and that 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 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 of particular interest in light of GPs' enhanced budgetary responsibilities as part of Clinical Commissioning Groups following the Health and Social Care Act (2012).

The results could be of interest to policy makers because 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 Commission found that the merging parties had strategies to engage with GPs via a GP newsletter. Those examples are consistent with evidence from the Co-operation 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.

Advised choice situations are common, providing scope for suitable adaptations of the empirical strategy proposed in this paper. For example, endogeneity of choice sets is an issue in the area of financial decision making. Here, a financial advisor or broker may offer sets of financial contracts to a retail client (e.g. different investment funds or assets, out of all traded assets; or different insurance products). This is also an area of regulatory interest. The then Financial Services Authority[25], for example, in its recent Retail Distribution Review (RDR) proposed various changes to the remuneration, capital and independence requirements for financial advisors, with the ultimate objective to bring financial advice in line with retail investor needs and preferences. Some real estate decisions have similar characteristics, as do certain types of art purchases.[26]

Expert agents may be more broadly understood. They may be social media plat-

[25]Now, Financial Conduct Authority.

[26]Chamley (2004) summarizes the growing theoretical microeconomic literature on the role of experts in consumer and investor choice decisions.

forms or retailers, rather than traditional experts. Strategic composition of choice sets emerges, for example, as a feature of online markets. Social media platforms are at the point of becoming gateways to online service providers. For example, Facebook in the future may host contents of selected online news media[27] and already now acts as platform for app-install ads[28]. Furthermore, antitrust authorities have focussed on Googles competition with so-called "vertical", or specialised, search services, such as comparison shopping sites, travel search engines and search sites aimed at local services, out of concern that rivals are disadvantaged because Google's search platform allegedly gives preferential treatment to results from its own services; this concern has culminated in the launch of a formal inquiry by the European Commission's Directorate for Competition into Google's shopping searches[29]. In these instances, the design of the online platform, acting as a gateway to services relevant to their ultimate users, is likely governed by revenue considerations of the platform operator - such as revenue from advertisement or proprietary services - that are not aligned with those relevant to the service users. Similar issues of misalignment of incentives faced by platform operators versus consumers have been considered by Armstrong and Zhou (2011), De Corniere and Taylor (2014), Eliaz and Spiegler (2011) and Hagiu and Jullien (2011).

Proper modelling of choice in the present of third-party agents is important for the design of effective consumer policy and competition analysis. It is well established that misalignment of incentives between a principal and an agent can give rise to market failures. Traditional analyses of patient choice (e.g. Beckert et al. (2012)) ignore this distinction. This analysis has identified conflicts of interest that the agent may face. There are other examples that share such conflicts of interest of third-party advisers. For instance, in its 2013 investigation of the market for audit services, the Competition Commission found that competition between audit firms was focused towards satisfying demands from executive management, including

[27]See New York Times, 24 March 2015;

http://www.nytimes.com/2015/03/24/business/media/facebook-may-host-news-sites-content.html

[28]See New York Times, 26 March 2015;

http://www.nytimes.com/2015/03/26/technology/debunking-the-latest-predictions-of-facebooks-demise.h

[29]See, for example, Financial Times, 02 and 15 April 2015;

http://www.ft.com/cms/s/0/97a4dc62-e360-11e4-9a82-00144feab7de.html?siteedition=uk#axzz3XIZ3NHfN

http://www.ft.com/cms/s/0/0c2b2840-d8d3-11e4-8a23-00144feab7de.html?siteedition=intl#axzz3W8LdSMDi

instances where such demands are not fully aligned with the interest of shareholders and investors as those with a direct interest in the outputs of the audit.[30] Similarly, in merger analysis in consumer retail markets, improper modelling of the critical role that retailers play in the pre-selection of consumer choice sets is an acknowledged limitation of currently prevailing approaches and is an as of yet empirically largely unresolved consideration.[31]

References

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

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

[3] 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

[4] Beckert, W. and E. Kelly (2016): "Divided by Choice? Private Providers, Patient Choice and Hospital Sorting in the English National Health Service", Institute for Fiscal Studies, unpublsihed manuscript

[5] 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

[6] 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

[30] Statutory audit services for large companies market investigation, Competition Commission, 15 October 2013. `https://www.gov.uk/cma-cases/statutory-audit-services-market-investigation`

[31] See, for example, Leary (2001) on market definition in the Heinz / Beech-Nut merger, and OFT (2011) in its decision on the anticipated acquisition of Alberto Culver by Unilever.

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

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

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

[10] 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

[11] 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

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

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

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

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

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

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

[18] 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

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

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

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

[22] 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

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

[24] Gaynor, M., Propper, C. and S. Seiler (2016): "Free to Choose? Reform and Demand Response in the English National Health Service", unpublished manuscript

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

[26] 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

[27] 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

[28] 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

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

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

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

[32] 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`

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

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

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

[36] 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

[37] 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

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

[39] 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

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

[41] 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

[42] 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

[43] 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

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

[45] 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

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

A Tables and Figures

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: **Percentiles of GP Practice Attributes**

Percentile	Coeff. of Var. w.r.t. Age	Number of GPs
5	.0516129	1
10	.0822310	1
25	.1326908	3
50	.1922468	4
75	.2837015	7
90	.4085385	9
95	.5013276	10

Source: HES and Health and Social Care Information Centre (HSCIC).

Table 5: **GP Pre-Selection**

	Constrained Choice		Unconstr. Linear Prob. Model	
	Coeff.	Std.Err.	Coeff.	Std.Err.
dist	-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 Var, Age	-0.0299	0.0009		

HES and Health and Social Care Information Centre (HSCIC).

All regressors are standardized. mff: market forces factor; hsmr: hospital standardised mortality rate.

Table 6: **Patient Hospital Choice, Conditional on $\mathcal{J}^a$**

	Res Constr Choice		Res Unconstr. Lin. Prob. Model	
	Coeff.	Std.Err.	Coeff.	Std.Err.
dist	-1.9992	.0409	-2.5540	.1431
parking	.0218	.0118	.0248	.0118
waiting time	-.6189	.0831	-.5932	.0832
wait res	.0246	.0028	.0208	.0029
constr res	-.0389	.0343		
unconstr res			6.423	1.577
log lik	-16315.218		-16307.72	

HES and Health and Social Care Information Centre (HSCIC).

The regressors dist, parking and waiting time are standardized.

Notes: wait res: residual from 1st stage regression of waiting times on hospital quality measures.

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

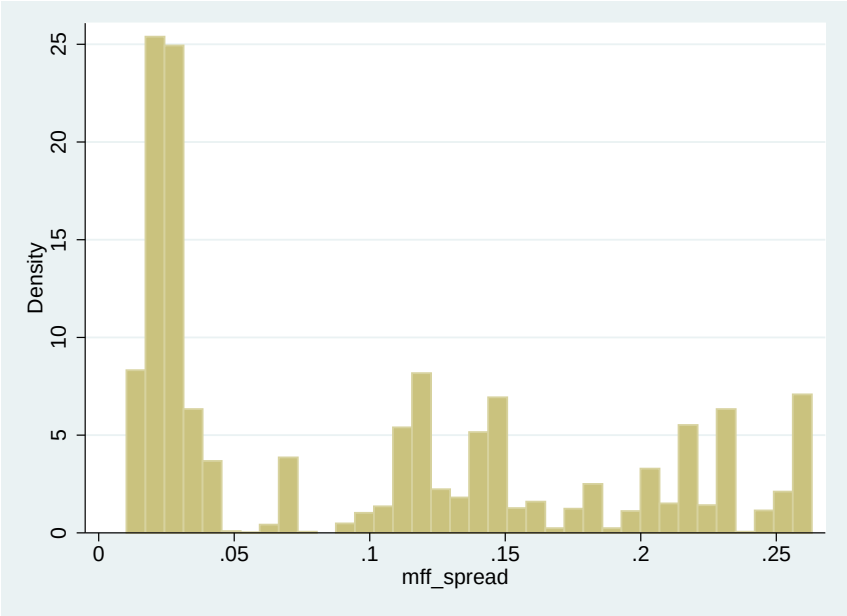
	Res Constr Choice		Res Unconstr. Lin. Prob. Model	
	Coeff.	Std.Err.	Coeff.	Std.Err.
dist	-6.6719	.0499	-2.9645	.0939
parking	.0137	.0093	.0308	.0123
waiting time	-.2095	.0669	-.5211	.0889
wait res	.0107	.0022	.0175	.0030
constr res	-.0118	.0314		
unconstr res			11.0478	.9466
log lik	-27543.777		-24250.632	

HES and Health and Social Care Information Centre (HSCIC).

The regressors dist, parking and waiting time are standardized.

Notes: wait res: residual from 1st stage regression; constr res: imputed residuals from GP pre-selection model.

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.