Demand Analysis under Latent Choice Constraints*†

Nikhil Agarwal and Paulo Somaini‡

June 25, 2024

Abstract

Consumer choices are constrained in many markets due to either supply-side rationing or information frictions. Examples include matching markets for schools and colleges; entry-level labor markets; limited brand awareness and inattention in consumer markets; and selective admissions to healthcare services. Accounting for these choice constraints is essential for estimating consumer demand. We use a general random utility model for consumer preferences that allows for endogenous characteristics and a reduced-form choice-set formation rule that can be derived from models of the examples described above. The choice-sets can be arbitrarily correlated with preferences. We study non-parametric identification of this model, propose an estimator, and apply these methods to study admissions in the market for kidney dialysis in California. Our results establish identification of the model using two sets of instruments, one that only affects consumer preferences and the other that only affects choice sets. Moreover, these instruments are necessary for identification – our model is not identified without further restrictions if either set of instruments does not vary. These results also suggest tests of choice-set constraints, which we apply to the dialysis market. We find that dialysis facilities are less likely to admit new patients when they have higher than normal caseload and that patients are more likely to travel further when nearby facilities have high caseloads. Finally, we estimate consumers' preferences and facilities' rationing rules using a Gibbs sampler.

*We are grateful to USRDS for facilitating access to the data. The authors acknowledge support from the NSF (SES-1254768), the NIH (P30AG012810), the Sloan Foundation (FG-2019-11484), and the MIT SHASS Dean's Research Fund.

†The data reported here have been supplied by the United States Renal Data System (USRDS). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy or interpretation of the U.S. government.

‡Agarwal: Department of Economics, MIT and NBER, email: agarwaln@mit.edu. Somaini: Stanford Graduate School of Business and NBER, email: soma@stanford.edu. We thank Mert Demirer, Liran Einav, Alessandro Iaria, Francesca Molinari, Aviv Nevo, Peter Reiss and participants at several seminars and conferences for helpful feedback, and to Chris Conlon and Alex McKay for constructive discussions of our paper. We also thank Felipe Barbieri, Sichang Chen, Alden Cheng, Idaliya Grigoryeva, Lia Petrose, Ricardo Ruiz, and Yucheng Shang for their excellent research assistance.

1 Introduction

Textbook discrete choice models assume that consumers pick their most preferred option from a known choice set at posted prices. Further, the models assume that there is no excess demand or supply at these prices, which makes prices the sole instrument that clears the market.[1] In many instances, demand is rationed by information frictions or by supply-side policies other than prices: students must be admitted by a school or a college, healthcare providers may be selective about their patients or be fully booked, and information frictions may result in consumers being unaware of certain products. The final allocation, in these cases, depends on the constraints on the choice sets in addition to preferences and prices.

With the few exceptions that are discussed below, existing approaches for estimating preferences with latent choice constraints are based on assuming specific models of latent choice set formation. In two-sided matching models – school or college admissions (e.g. Agarwal and Somaini, 2018; Fack et al., 2019), and certain labor markets (e.g. Boyd et al., 2013; Agarwal, 2015) – choice sets are determined by supply-side preferences and screening (Roth and Sotomayor, 1990), whereas search costs and incomplete information are the source of limited choice sets in models of consumer search (Hortaçsu et al., 2017; Heiss et al., 2021) or consideration sets (e.g. Manski, 1977; Swait and Ben-Akiva, 1987; Alba et al., 1991; Roberts and Lattin, 1991; Goeree, 2008; Abaluck and Adams-Prassl, 2021; Barseghyan et al., 2021a,b). Perhaps the only apparent similarity between these models is that consumers are not unconstrained to choose from the full set of options in the market.

This paper presents a unified analysis of a large class of empirical models of consumer choice in the presence of latent choice-set constraints. Our model combines a general random utility model for consumer preferences (Block and Marshak, 1960; Matzkin, 1993) with a reduced-form model for choice set formation. We show, by way of examples, that many commonly used models of latent choice sets discussed above are consistent with this general reduced-form. Our primary contribution is to show conditions under which this general model is non-parametrically identified using data on final allocations in the presence of preference and choice set shifters. We also propose a tractable estimation procedure. Finally, we apply our methods to data from the market for kidney dialysis to test for supply-side rationing and to describe the potential biases from ignoring constraints in choice sets.

Our model has two components. The first is a random utility model for consumer preferences, which allows for rich observed and unobserved heterogeneity in consumer preferences.

[1] These prices may either be set to maximize profits or set competitively. In both cases, firms produce exactly enough quantity to satisfy the demand at these prices.

We allow for product unobserved attributes that may be correlated with observed product characteristics as in Berry (1994) and Berry et al. (1995). The model accommodates most random utility models with single-unit demand, including product space and characteristic space models. The second component is a reduced-form which can capture various models that yield latent constraints on choice sets. We show that our reduced form is consistent with models of two-sided matching (e.g. matching of students to schools or colleges); dynamic models in which profit motives induce firms to be selective in their admission policies; as well as certain models of consideration sets, consumer search and informational advertising.

The empirical challenge is that the observed allocations depend both on the preferences of agents and the choice set formation process, making it hard to disentangle the two. In particular, standard methods for estimating the distribution of preferences based on inverting market shares to estimate key demand parameters (e.g. Berry, 1994; Berry et al., 2013) are inapplicable in the presence of product-specific unobservables that influence choice sets. Intuitively, the product chosen most often need not be the one most preferred by the largest proportion of customers.[2] We show that our model is non-parametrically identified in the presence of two sources of variation. The first is an observable that affects choice-set constraints, but is excluded from consumer preferences. The second is an observable that influences consumer choices but is excluded from the choice-set constraints. We show how to combine these two observables to trace-out the joint distribution of consumer preferences and latent choice sets. Moreover, we formally show that our model is not identified if either set of shifters is not available.

At the cost of requiring shifters on both sides, our results place minimal functional form and statistical restrictions on preferences and latent choice sets. We allow for the preference shifter to enter non-linearly in our utility specification; functional form restrictions on the choice-set shifters are similarly weak. Moreover, we allow unobservables that affect the choice set to be arbitrarily correlated with unobservable determinants of preferences. Specific models of choice set formation and other approaches typically require stronger restrictions, either on functional forms and/or the joint distribution of unobservables. The non-identification result in the absence of the shifters indicates that these restrictions are necessary, and substitute for exogeneous variation in the data.

[2] A salient example is colleges – the largest colleges need not be the most desirable. Consider that Stanford University has an undergraduate enrollment higher than that of MIT. Tuition at Stanford is also higher. One of the authors of this study claims that MIT has a lower enrollment only because it has a lower capacity and is therefore more selective. Even when confronted with Stanford's lower overall acceptance rates, the author rebuts by suggesting that acceptance rates are a biased measure of selectivity because the applicant pools are endogenous and different.

We also incorporate unobserved product characteristics that influence preferences or choice sets into our identification analysis. These unobservables may be correlated with observable characteristics, creating an endogeneity concern akin to those in the analysis of demand (see Berry et al., 1995, for example). We show that across-market variation in instruments can be used to solve this endogeneity concern, by adapting index and invertibility restrictions from Berry and Haile (2014) to the case with constrained choice sets.

As an illustrative application, we apply our methods to the kidney dialysis market in California. Patients with low enough kidney function need to undergo regular dialysis, typically thrice weekly for several hours at a time. The procedure requires the use of expensive machines, nursing care and physical space to accommodate a patient. These resource constraints can limit the number of patients a facility can serve. Most of the costs of dialysis are borne by the taxpayer since Medicare provides near universal coverage for costs related to kidney failure, irrespective of age. With approximately 750,000 patients on dialysis currently in the US, these costs approach 1% of the national healthcare expenditure (Chapter 10, U. S. Renal Data System, 2020).

The choice-set shifter in this application is a measure of the facility's capacity constraints when patient i begins dialysis. This measure is constructed as the difference between the number of patients being treated at the facility when patient i begins dialysis and an estimated target. We exclude this short-term variation in facility utilization from patient preferences while we allow for long-term facility fixed effects. Thus, the argument is that patient preferences do not depend on short-term variation in a facility's caseload, but that this variation can result in supply-side rationing. A similar instrument is used in Gandhi (2021) to estimate preferences in nursing homes. As a test of the model, we show that this variable predicts whether or not a new patient is admitted into a facility even after controlling for facility-quarter fixed effects. This is the first piece of evidence that suggests that supply-side rationing due to capacity constraints can constrain a consumer's choice set.

The shifter of consumer preferences is the distance between the facility and the patient's residence. This variable is excluded from the choice set formation process but included in consumer preferences because dialysis involves several weekly visits and long post-dialysis trips can be particularly demanding on patients.

Consistent with the hypothesized effects of these shifters, we document that distance to the facility chosen by a patient is higher if nearby facilities have higher than usual caseloads. These results provide evidence that supply-side rationing affects allocations substantially.

The main challenge in estimating our model is that the number of potential choice sets is large, even if a patient has relatively few facilities to consider. This curse of dimensionality

creates a computational burden for approaches that integrate over all possible choice sets when computing the likelihood. Indeed, applications based on these cases have sometimes been limited to a small number of choices (e.g. Abaluck and Compiani, 2020). We solve this problem by estimating a parametric version of our model using a Gibbs sampler (see also Logan et al., 2008; Menzel and Salz, 2013; He et al., 2024). This procedure uses data augmentation in order to condition either on choice sets or utilities when drawing the parameters governing the other component. Doing so avoids the curse of dimensionality and reduces each component to a standard problem. The Bernstein-von Mises Theorem implies that the posterior mean of the sampling chain we generate is asymptotically equivalent to a maximum likelihood estimator (van der Vaart, 2000, Theorem 10.1).

The empirical model we take to the data allows for preferences to be correlated with choice sets due to unobserved factors. Our estimates indicate that selective admissions practices are common in the dialysis market. The probability that a patient is accepted at her first choice facility is only 59.1%, and this probability varies by facility. Because selective admissions push patients to less desirable facilities, models that do not account for choice set constraints yield biased estimates. These models misestimate the desirability of various facilities as abstracting away from selective admissions would yield estimates in which the largest facilities are also the most desirable.

We compare our approach to alternatives that naively correct for capacity constraints by including our measure of occupancy in the utility function. A stark prediction of these models is that a patient lost by a facility because of variation in this variable is diverted to the same set of facilities as a patient lost by a facility due to reductions in quality. In both cases, the facility loses a patient who is close to indifferent between two facilities. In contrast, in models with selective admissions, the patients that a facility loses is marginal for the facility, but the patient strictly prefers this facility to others. The facilities that a patient is diverted to are different, depending on which margin a patient is pulled from. We show that not capturing this difference yields quantitatively different estimates of diversion ratios.

Related Literature

A large literature – dating back to Block and Marshak (1960) and Manski (1977) – presents several specific models with latent constraints on choice sets. A much more recent literature has attempted to understand the identification of these models. This body of work has studied models of consideration sets (Abaluck and Adams-Prassl, 2021; Barseghyan et al., 2021b,a); two-sided matching models (Menzel, 2015; Diamond and Agarwal, 2017; He et al., 2024); and models of consumer search (Abaluck and Compiani, 2020). Our approach covers

models in each of these three groups.[3] And, at the cost of requiring both shifters of choice sets and shifters of preferences, our results are able to achieve point identification using fewer functional-form restrictions on preferences (c.f. Diamond and Agarwal, 2017; Abaluck and Adams-Prassl, 2021; Barseghyan et al., 2021b,a; Abaluck and Compiani, 2020; He et al., 2024; Barseghyan, 2022) or on the dependence between preferences and choice-sets (c.f. Menzel, 2015; Abaluck and Compiani, 2020; Abaluck and Adams-Prassl, 2021). It is worth reiterating in the context of this discussion that our non-identification results show that either two sets of shifters or these additional restrictions are necessary to achieve identification. In requiring two sets of shifters, our results hold under more general conditions than those in He et al. (2024), which requires additional shifters and non-primitive rank restrictions. We compare and contrast these papers with ours in greater detail as we develop our results.

In addition to these differences, we also address common endogeneity concerns when estimating demand models, extending results in Berry and Haile (2014) by allowing for constrained choice sets. This solution can be useful for a number of applications. For example, existing work on estimating school demand to study equilibrium effects (Neilson, 2020; Dinerstein and Smith, 2021; Allende, 2019) typically abstracts away from selective admission due to capacity constraints. Similar issues are likely important in other settings where prices are not the sole market clearing mechanism.

A small recent literature studies the industrial organization of the dialysis industry. Many of these studies are based on quasi-experimental research designs (e.g. Dafny et al., 2018; Wollmann, 2022), or focus on longer-run supply side issues such as the quality/quantity trade-off or investment/entry decisions (Grieco and McDevitt, 2017; Eliason, 2019; Eliason et al., 2020; Kepler et al., 2021). In contrast, our focus is on estimating demand and the supply-side rationing policies in response to shorter-term capacity constraints while keeping investment and quality decisions fixed. Previous approaches to estimating demand in this setting have abstracted away from supply-side rationing.

Our empirical model is closest to those of selective admission practices in nursing homes (Ching et al., 2015; Gandhi, 2021), although these papers do not formally consider the identification. Our identification results also cover models of two-sided matching in education or entry-level labor markets with fixed prices (e.g. Dagsvik, 2000; Agarwal, 2015; Azevedo and Leshno, 2016); models of consumer choice with incomplete consideration sets (e.g. Manski, 1977; Swait and Ben-Akiva, 1987; Alba et al., 1991; Roberts and Lattin, 1991; Goeree, 2008);

[3] The set of models covered by any one of these papers may not be nested with the models that we consider. For example, Abaluck and Compiani (2020) consider consumer search with hidden attributes. While it is possible to cast fixed-sample search in either our framework or the one in Abaluck and Compiani (2020), their paper allows for other models of consumer search (e.g. sequential search) that does not fit our framework.

models with strict capacity constraints (de Palma et al., 2007); and models of consumer stock-outs (Conlon and Mortimer, 2013; Hickman and Mortimer, 2016). Our reduced-form approach to supply-side rationing accommodates several of the reasons for incomplete choice sets discussed above. Estimating a more primitive model of the supply side requires additional assumptions on the structural model that we avoid because they are, by nature, application specific. We discuss the interpretation of our model in these specific applications in further detail in Section 2.2.

Overview

The paper proceeds as follows. Section 2 presents our model. It includes a discussion of the models of supply-side rationing that yield the reduced-form of interest in our paper. Section 3 presents the identification results and the estimator. Section 4 describes the dialysis industry and presents descriptive evidence on supply-side rationing. Section 5 presents results from our estimates. Section 6 concludes. All proofs not included in the main text are in the appendix.

2 Model

We will consider markets, indexed by t, in which agents can be divided into two sets, I_t and J_t. We will refer to the set I_t as *consumers* and the set J_t as *products*. Consumers, indexed by $i \in I_t$, have unit demand. We will say that consumer i is *matched* with product j if it is in the consumer's choice set and the consumer chooses it. A product can match with many consumers. The outside option, denoted with 0, is always in the consumer's choice set.[4] Each consumer i participates in only one market, thus the index i implies the associated market index t. Section 2.2 describes several models that fit this formulation.

2.1 Preferences and Choices

We adopt a random utility model for consumer preferences. The indirect utility of consumer i for matching with product j is given by

$$v_{ij} = u_{jt}\left(d_i, \omega_i\right) - g_{jt}\left(d_i, y_{ij}\right), \tag{1}$$

[4]That the consumer can always choose the outside option avoids empty choice sets. There is limited loss of generality in this assumption because the outside option can be defined as a composite of alternatives outside the market considered.

where d_i is a vector of observed consumer attributes; y_{ij} is a scalar observed attribute that varies at the consumer-product level; and ω_i is a random vector of arbitrary dimension that introduces unobserved consumer-specific preference heterogeneity. We impose the following normalizations, which are without loss of generality (Matzkin, 2007): we normalize the utility of the outside option v_{i0} to zero for each i; for some known value y_0 and a fixed j in each t, we set $\left|\frac{\partial g_{jt}}{\partial y}(d_i, y_0)\right| = 1$ for all d_i; and we set $g_{jt}(d_i, y_0) = 0$ for every j, t and d_i. The restrictions on v_{i0t} and the partial derivative of $g_{jt}(\cdot)$ are familiar location and scale normalizations. The restriction that $g_{jt}(d_i, y_0) = 0$ is without loss because a constant shift in $g_{jt}(\cdot)$ can be subsumed in $u_{jt}(\cdot)$.

This model places minimal restrictions on the representation of preferences. The term ω_i allows for multi-dimensional unobserved heterogeneity, including idiosyncratic product-specific preference shocks. The functions $u_{jt}(\cdot)$ and $g_{jt}(\cdot)$ are indexed by product and market, indicating that they can vary arbitrarily along these dimensions. Thus, these functions can vary due to both observed and unobserved market-product specific attributes. The term d_i may include attributes that vary at the consumer-product level in addition to those that only vary at the consumer level. The main distinction between y_{ij} and other consumer-product observables included in d_i is that y_{ij} only affects the indirect utility of product j and is separable from ω_i.

Unlike standard consumer choice models, consumers cannot simply choose their most preferred product. In education markets, students must be accepted by the school; in healthcare markets, patients need appointments; in labor markets, applicants need job offers; in models of consumer search or consideration sets, choice sets are incomplete. Although our model is intended for any of these settings, for the sake of uniformity of nomenclature, we personify products and say that they must accept the consumer. Let

$$\sigma_{ijt} = \sigma_{jt}(d_i, \omega_i, z_{ij}) \in \{0, 1\} \tag{2}$$

denote this latent acceptance decision, where $\sigma_{jt}(d_i, \omega_i, z_{ij}) = 1$ denotes that consumer i was accepted by product j in market t. We refer to the function $\sigma_{jt}(\cdot)$ as the *acceptance policy function*. It is indexed by product and market, allowing it to depend on market-product specific observables and unobservables. The product's decision to accept the consumer depends arbitrarily on ω_i as well. Therefore, utilities and acceptance decisions may be correlated due to unobservables.[5]

[5]This formulation and the results below do not impose restrictions on the dependence between indirect utilities and choice sets. To see this, one can write $\omega_i = (\omega_i^u, \omega_i^\sigma)$ each of arbitrary dimension, and $u_{jt}(\cdot)$ and $\sigma_{jt}(\cdot)$ only depend on ω_i^u and ω_i^σ. At one end, ω_i^u and ω_i^σ can be independent, whereas at the other end, the

The term z_{ij} is a consumer-product specific observable scalar characteristic that affects the decision of the product to accept the consumer that is excluded from the consumer's utility. As opposed to d_i, the scalar characteristic z_{ij} can only affect acceptances by product j, not product k. This rules out strategic interactions between products on this dimension, but it does allow for strategic interactions on the basis of aggregate conditions of the market and on consumer i's characteristics via the dependence of $\sigma_{jt}(\cdot)$ on t and on d_i.

We assume that each consumer is matched with one of her most preferred products that accepts her. Formally, each consumer i's (latent) choice set in their market t is given by the set of products that accept the consumer:

$$O_i = \{j \in J_t : \sigma_{ij} = 1\} \cup \{0\}.$$

She picks a product with the highest indirect utility within this set. Let $c_{ij} \in \{0, 1\}$ be an indicator for consumer i matching with $j \in O_i$. We assume that $\sum_{j \in O_i} c_{ij} = 1$ and $c_{ij} = 1$ only if $j \in \arg\max_{k \in O_i} v_{ik}$. Moreover, if $\arg\max_{k \in O_i} v_{ik}$ is not a singleton, then the tie between the products with the highest indirect utilities is broken independently of $y_i = (y_{ij})_{j \in J_t}$ where t is the market to which i belongs.[6] Thus, we assume that the only source of friction in the economy is through the choice set formation process. We will see that this formulation accommodates many forms of consumer search frictions.

We will make the following assumption throughout the paper:

Assumption 1. *In each market t, the unobserved term ω_i is conditionally independent of the vector (y_i, z_i) given d_i.*

This assumption places two substantive restrictions. First, the conditional independence of ω_i from y_i implies that each component y_{ij} is shifts preferences without interacting with consumer-specific unobservables that affect either preferences or choice sets. The effect of a marginal change in y_{ij} can depend on d_i as well as on product-market specific characteristics through the function $g_{jt}(\cdot)$. Second, it implies that unobserved determinants of preferences and choice sets are independent of z_i given d_i. Thus, z_i is an instrument that shifts choice sets without affecting the distribution of preferences. The assumption does not rule out correlation between y_i and z_i conditional on d_i. The plausibility of these restrictions is specific to the empirical application and the available data. For now, we defer the discussion of these issues in the context of our specific empirical application.

---

two are perfectly correlated.

[6] Formally, for all t and (O_i, d_i, z_i, j), $P(c_{ij} = 1 | O_i, t, d_i, y_i = y, z_i) = P(c_{ij} = 1 | O_i, t, d_i, y_i = y', z_i)$ if $g(y) = g(y')$.

We assume that the data are generated by sampling the random vector ω_i independent and identically across consumers. Therefore, for each market t, the choice set and preferences of consumer i are independent from those of other consumers in market t conditional on the observables (d_i, y_i, z_i), where $y_i = (y_{ij})_{j \in J_t}$ and $z_i = (z_{ij})_{j \in J_t}$. However, consumer preferences and choice sets may be correlated within a market via the functions $u_{jt}(\cdot)$, $g_{jt}(\cdot)$ and $\sigma_{jt}(\cdot)$. As we discuss in the examples below, this assumption is satisfied in standard consumer choice and consumer search models; and also in two-sided matching markets in which there are many consumers relative to the number of products.

The assumptions on the data generating process imply that the share of consumers with observables (d_i, y_i, z_i) that are matched with product j in market t is given by

$$s_{jt}(d_i, y_i, z_i) = \sum_{O \in \mathcal{O}} P(O_i = O, c_{ij} = 1 | t, d_i, y_i, z_i),$$

where $\mathcal{O}$is the set of all possible choice sets. The information in the data consists only of these market shares for each value of (d_i, y_i, z_i) in its support. Assumption 1 implies that the shares $s_{jt}(\cdot)$ can be re-written as

$$s_{jt}(d_i, y_i, z_i) = \sum_{O \in \mathcal{O}} P(c_{ij} = 1 | O_i = O, t, d_i, y_i, z_i) P(O_i = O | t, d_i, z_i). \tag{3}$$

The first term in the summand is the probability that a consumer with attributes (d_i, y_i, z_i) is matched with product j when faced with the choice set O, whereas the second term is the probability of choice set O given (d_i, z_i). Assumption 1 allows us to omit the conditioning on y_i when writing the second term. However, we cannot omit z_i from the first term because we allow for dependence between preferences and choice sets conditional on observed characteristics through ω_i, which is often not allowed in the prior literature. Since the distribution of ω_i conditional on $O_i = O$ depends on z_i, the distribution of c_{ij} conditional on $O_i = O$ also depends on z_i.

We assume that the shifter z_{ij} has a monotonic effect on choice sets:

Assumption 2. *The function $\sigma_{jt}(d_i, \omega_i, z_{ij})$ is non-increasing in z_{ij}.*

Monotonicity requires that product j is more likely to be in consumer i's choice set if the value of z_{ij} is lower. This assumption is natural in the examples discussed in section 2.2.

Define the cut-off quantity, $\pi_{jt}(d_i, \omega_i) = \sup\{z : \sigma_{jt}(d_i, \omega_i, z) = 1\}$ where we adopt the convention that $\pi_{jt}(d_i, \omega_i) = \infty$ if $\sigma_{jt}(d_i, \omega_i, z) = 0$ for all z and $\pi_{jt}(d_i, \omega_i) = -\infty$ if

$\sigma_{jt}(d_i, \omega_i, z) = 1$ for all z.[7] Under assumption 2, the function $\pi_{jt}(\cdot)$ determines product j's acceptance policy for almost every z since $z < \pi_{jt}(d_i, \omega_i)$ implies $\sigma_{jt}(d_i, \omega_i, z) = 1$, and $z > \pi_{jt}(d_i, \omega_i)$ implies $\sigma_{jt}(d_i, \omega_i, z) = 0$. However, the acceptance policy function can take any value when $z = \pi_{jt}(d_i, \omega_i)$.

Under this assumption, a key primitive for each market t is the joint distribution of the random vector $(u_{it}, \pi_{it}) = (u_{1t}(d_i, \omega_i), \ldots, u_{J_t t}(d_i, \omega_i), \pi_{1t}(d_i, \omega_i), \ldots, \pi_{J_t,t}(d_i, \omega_i))$ conditional on d_i and t, and the function $g_{jt}(\cdot)$. To see this, although we do not impose this restriction, consider the case when (u_{it}, π_{it}) admits a density, which implies that ties in utility and acceptance cutoffs are zero-probability events. In this case, the terms in summand in equation (3) can be re-written to yield

$$s_{jt}(d_i, y_i, z_i) = \sum_{O \in \{O \in \mathcal{O} : j \in O\}} \int\int 1\{u_{ijt} - g_{jt}(d_i, y_{ij}) \geq u_{ikt} - g_{kt}(d_i, y_{ik}) \ \forall k \in O\} \times \left[\prod_{k \notin O} 1\{\pi_{ikt} < z_{ik}\} \prod_{k \in O} 1\{\pi_{ikt} > z_{ik}\}\right] f_{U,\Pi|d_i,t}(u_{it}, \pi_{it}) \, \mathrm{d}u_{it} \mathrm{d}\pi_{it}. \quad (4)$$

Hence, the vector of market shares in t is determined by $F_{U,\Pi|d_i,t}(u_{it}, \pi_{it})$ and the functions $g_{jt}(\cdot)$. These features of the model also determine the effects of changes in y_{ij} and z_{ij} on consumer and producer surplus.

This equation also shows that the share of consumers that are matched with product j in market t depends both on the preferences of the consumers and the acceptance policies of all the products in the market. Unlike in standard models of consumer demand, the market share of product j does not directly reveal the fraction of consumers who prefer j to all other products. Therefore, commonly used demand-inversion methods would yield invalid mean utility measures whenever relevant latent choice set constraints are ignored (c.f. Berry, 1994; Berry et al., 1995, 2013).

2.2 Examples

We start by showing that our preference model is general enough to accommodate commonly used random utility models in the analysis of discrete choice demand functions. Then, we work out several different examples that yield constrained consumer choice sets that are compatible with the acceptance policy function described above.

Example 1. (Preference Model) Our formulation encompasses the widely used discrete choice models with random coefficients and a linearly separable index for product-specific

[7] If $\sigma_{jt}(d_i, \omega_i, z) = 0$ for all z or $\sigma_{jt}(d_i, \omega_i, z) = 1$ for all z, the vector $\pi_{it} = (\pi_{1t}(d_i, \omega_i), \ldots, \pi_{J_t,t}(d_i, \omega_i))$ is an extended random variable with components that take on values in $\mathbb{R} \cup \{-\infty, \infty\}$.

unobservables ξ_{jt} (e.g. Berry et al., 1995; Petrin, 2002):

$$v_{ij} = d_i \Gamma x_{jt} + x_{jt}\beta_i + y_{ij} + \xi_{jt} + \varepsilon_{ij},$$

where each individual i belongs to only one market t. We can nest this specification by setting $u_{jt}(d_i, \omega_i) = d_i \Gamma x_{jt} + x_{jt}\beta_i + \xi_{jt} + \varepsilon_{ij}$, $\omega_i = (\beta_i, \varepsilon_{i1}, \ldots, \varepsilon_{iJ})$ and $g_{jt}(d_i, y_{ij}) = -y_{ij}$. Thus, the unobserved term ξ_{jt} together with the observed characteristics x_{jt} is subsumed into the function $u_{jt}(\cdot)$. The price of good j in market t can be included as an observed characteristic in x_{jt}. The random coefficients β_i induce preference heterogeneity that results in rich substitution patterns between the different goods j. The matrix Γ captures interactions between consumer characteristics d_i and observable product characteristics x_{jt}. Our identification results will accommodate most commonly used distributional assumptions on ε_{ij}, including those that yield the familiar logit or nested-logit models (Train, 2009). We can also accommodate other random utility models of preferences such as the pure characteristics model of Berry and Pakes (2007).

Example 2. (Selective Acceptance in Healthcare) Our acceptance policy function accommodate the model of supply-side rationing in skilled nursing facilities in Gandhi (2021). Facility j accepts a new patient if the patient's profitability exceeds a threshold which is a function of the facility's current caseload. In our notation:

$$\sigma_{ijt}(d_i, \omega_i, z_{ij}) = 1\left\{NPV_{jt}(\omega_i, d_i) + V_j(z_{ij}+1) - V_j(z_{ij}) > 0\right\},$$

where $NPV_{jt}(\omega_i, d_i)$ denotes the present value of variable profits from patient i at facility j, and $V(z_{ij}+1) - V(z_{ij})$ is the change in the continuation value given an caseload of z_{ij} at the time of arrival of patient i. The terms d_i and ω_i denote observable and unobservable characteristics of patient i. The term $V_j(z_{ij}) - V_j(z_{ij}+1)$ is a threshold equal to the opportunity cost of accepting a new patient. Gandhi (2021) shows that this threshold is increasing in z_{ij}. In principle, d_i can include aggregate market conditions at the time of i's arrival which could also enter in the continuation value $V_j(\cdot)$.

Example 3. (Two-Sided Matching) Our framework encompasses models used in the empirical analysis of two-sided matching markets with non-transferable utility under pairwise stability. Examples include the matching of students to schools or colleges, and entry-level labor markets with fixed pay scales (e.g. Dagsvik, 2000; Agarwal, 2015). Let $e_{jt}(d_i, \omega_i, z_{ij})$ be an unknown rule that school or college j employs in market t to evaluate candidates. This rule depends on observable and unobservable characteristics. For example, in the case of

college acceptances, d_i may contain demographic information and observable exam scores, ω_i includes unobservable essay quality or other hard to codify aspects of an application, and z_{ij} is an observed characteristic that varies at the student-school level. Azevedo and Leshno (2016) showed that a pairwise stable allocation in a many-to-one two-sided matching models can be described by a set of cutoffs p_{jt} for each school $j \in J_t$ and market t. These cutoffs are such that each agent i is assigned to her most preferred facility in the set $O_i = \{j \in J_t : e_{jt}(d_i, \omega_i, z_{ij}) \geq p_{jt}\} \cup \{0\}$. Thus, in our notation:

$$\sigma_{ijt} = 1\left\{e_{jt}(d_i, \omega_i, z_{ij}) - p_{jt} > 0\right\}.$$

The identification of a similar model of two-sided matching was recently studied in He et al. (2024). Our results will require fewer exogenous shifters and place fewer restrictions on primitives, a comparison that we further flesh out when discussing our theoretical results in section 3.

Example 4. (Consideration Sets) Several models in marketing and economics assume that consumers choose among the subset of products in the market (see Manski, 1977; Swait and Ben-Akiva, 1987; Alba et al., 1991; Roberts and Lattin, 1991; Goeree, 2008). In our framework, product j belongs to the latent consideration set O_i if $\sigma_{jt}(d_i, \omega_i, z_{ij}) = 1$. Since d_i and ω_i are arguments in $u_{jt}(\cdot)$, consideration sets can be correlated with utilities. The main requirement of our model is that there are consumer-product specific characteristics z_{ij} that affect the probability that product j belongs to i's consideration set. This requirement is satisfied by a number of microfoundations. We discuss a few below:

Brand Awareness: Butters (1977) and Eliaz and Spiegler (2011) model advertising as affecting the probability with which a consumer is informed about a product. Goeree (2008) estimates an empirical model that uses the interaction between a product's advertising expenditure and a consumer's exposure to advertising to construct a variable analogous to z_{ij}. Another example is Gaynor et al. (2016), which models a physician who decides whether a patient should have hospital j in their consideration set. It is natural to expect the consideration sets to be correlated with preferences in this setting, as is allowed in our framework.

Inattention and Defaults: Consumers in some models are inattentive and choose a default unless sprung into action (e.g. Heiss et al., 2021; Ho et al., 2017; Hortaçsu et al., 2017). These models often feature strong defaults where only the characteristics or utility of the default option influences attention.[8] In some of these models, attention is binary where a consumer either pays attention to all products or none. Our framework allows for a version

[8] See the default specific consideration and hybrid cases in Abaluck and Adams-Prassl (2021).

with certain products being much more likely a part of the consideration set than others, but it will require that characteristics of any of the products that are excluded from preferences, z_{ij}, influence product-specific consideration.

Fixed Sample Search: A number of papers model fixed sample search by modeling choice over a latent subset of heterogeneous products (see Honka, 2014; Honka et al., 2017, for example). Assume that consumers know their preferences for the products except that they do not know the price that they will be quoted for a product. The consumer decides the portfolio of products for which to obtain a the price quote based on its ex-ante distribution (Chade and Smith, 2006). The consumer incurs a search cost for each quote. Thus, the decision to search for a product is given by the search policy function $\sigma_{jt}(\cdot)$.

In our framework, let y_{ij} be the price that is unobserved by the consumer prior to search. The realized values of σ_{ijt} can depend on the ex-ante price distribution, the other components of indirect utility, and search costs. Thus, σ_{ijt} can be correlated with v_{ijt}, but it is not a deterministic function of v_{ijt}.[9] We also require an observable z_{ij} that is excluded from preferences, but shifts the probability that consumer i searches for product j. For example, informative advertising or distance to the product may affect search probabilities – the former through awareness and the latter through search costs – while being independent of preferences.

Stock-outs: Consider a case in which a product may not be available on the shelves when a consumer arrives. Hickman and Mortimer (2016) distinguish two data environments depending on whether stock-out events are observed and recorded in the data. When stock-out events are observed, they provide an opportunity to estimate demand cross-elasticities as in Conlon and Mortimer (2013). Alternatively, a product may or may not be available for all consumers within a market in which case, the aggregate market share of the product will be zero in that market (see Dubé et al. (2021)). In this case, choice sets are effectively observed.[10] However, when the dataset does not record or allow the researcher to infer specific stock-out events, consumer choice sets are latent and cross-elasticities are generally not identified. We model latent choice sets by letting σ_{ijt} denote whether product j was available at the time agent i arrived at store t. The choice set shifter z_{ij} may be the time-lag between when product j was last restocked and when consumer i checked out. We show that variation in z_{ij} can restore identification of demand.

[9] Models of sequential search do not naturally fit our framework because the decision to continue searching depends on the highest utility amongst the goods already searched (Weitzman, 1979). In this case, y_{ij} cannot be excluded from $\sigma_{jt}(\cdot)$.

[10] Dubé et al. (2021) also require an observable that shifts choice sets that is excluded from demand, but are able to make progress with a shifter that is product-specific because choice sets are both observed/inferred and common to all consumers within a market.

3 Identification

The first goal we pursue (in subsection 3.1) is to identify the joint distribution $F_{U,\Pi|d_i,t}$ and the function $g_{jt}(\cdot)$. We will show that choice-set shifters are necessary for this goal (subsection 3.2). For example, identifying the distribution of u_{it} in a neighborhood and the derivative of the function $g_{jt}(\cdot)$ are sufficient for identifying changes in demand in response to changes in y_i and to perform welfare analysis if $g_j(d_i, y_i)$ is an appropriate numeriare. Identifying π_{it} allows us to obtain $\sigma_{jt}(\cdot)$, which are product-specific acceptance policy functions.[11]

This analysis will condition on d_i and t, focusing on within market variation in (y_i, z_i). The conditioning on t fixes product-level observables and unobservables for all products in a market. In subsection 3.3, we will micro-found the dependence of $u_{jt}(\cdot)$, $\pi_{jt}(\cdot)$ and $g_{jt}(\cdot)$ on observed product attributes x_{jt} and a vector of unobservables ξ_{jt}, allowing for the case when observed product attributes are endogeneous. We will then show how instruments can be used to identify the joint distribution of (u_{it}, π_{it}) and $g_{jt}(\cdot)$ as a function of (x_{jt}, ξ_{jt}). Solving this endogeneity problem allows us to identify the effects of changing x_{jt} while holding ξ_{jt} fixed on market shares, as well as consumer surplus.

3.1 Identification within a market

We will build the main result of this subsection (theorem 1) in two steps. First, lemma 1 shows identification given that the functions $g_j(\cdot)$ are known (section 3.1.1). Second, lemma 2 shows that the functions $g_j(\cdot)$ are identified under slightly stronger assumptions (section 3.1.2). These two results together will imply our main theorem (section 3.1.3). Throughout this subsection, we omit the market subscript t because we condition on it.

3.1.1 Identification with known $g(\cdot)$

Let $u_i = (u_j(d_i, \omega_i))_{j\in J}$, $\pi_i = (\pi_j(d_i, \omega_i))_{j\in J}$ and $\sigma_i = (\sigma_j(d_i, \omega_i, z))_{j\in J}$. In the case when $g(\cdot) = (g_j(\cdot))_{j\in J}$ is known, identification of the joint distribution of (u_i, π_i) given d_i, which implies identification of the joint distribution of (v_i, σ_i) given (d_i, y_i, z_i), can be achieved without any further assumptions.

Lemma 1. *Fix d_i. Suppose that assumptions 1 – 2 are satisfied, and $g(\cdot)$ is known. Let χ be the interior of the support of (g, z) given d_i. The joint distribution of (u_i, π_i) conditional*

[11] The random variable π_{it} directly yields preferences in the case of static two-sided matching models. In the case of a dynamic acceptance policy $\sigma_{jt}(\cdot)$, we may use Hotz and Miller (1993) inversion to recover payoffs.

on $(u_i, \pi_i) \in \chi$ *and* d_i *is identified.*

Proof. See appendix A.1 □

The idea of the proof is best described with the aid of two figures. Assume for this illustration that (u, π) admits a density, a requirement that our formal results dispense with but is useful for exposition to avoid carefully tracking zero-measure sets with mass points in the distribution of (u, π). Consider the probability that a consumer is not matched to any of the products in the market. This probability, which is observed, is equal to the probability that for every product j either $u_j < g_j$ or $\pi_j < z_j$, where ties are zero probability events. The cross-hashed region in figure 1 shows this set projected on the $u_1 - \pi_1-$hyperplane. That is, the random variables $u_2, \ldots, u_J$ and $\pi_2, \ldots, \pi_J$ are marginalized conditional on $u_j < g_j$ or $\pi_j < z_j$ for $j > 1$. The point $(\bar{g}_1, \bar{z}_1)$ collects the first components of any vector $(\bar{g}, \bar{z}) \in \chi$ that we fix in the remainder of the argument. Now, consider a small $\Delta > 0$ such that all points that are at most Δ away from each component of $(\bar{g}, \bar{z})$ belong to the interior of the support of $g(\cdot)$ and z. Perturb $\bar{z}_1$ by Δ to obtain the region between $\bar{z}_1$ and $\bar{z}_1 + \Delta$ that lies above $\bar{g}_1$. The probability that (u_1, π_1) falls within this region is equal to the increase in the probability from a consumer remaining unmatched at $(\bar{g}, \bar{z})$ to remaining unmatched when $\bar{z}_1$ is increased by Δ. This is because the change from $\bar{z}_1$ to $\bar{z}_1 + \Delta$ only affects consumers who would like to match with product 1 but the product drops out of the choice set due to this change. Since this increase in probability is observed, we can determine the probability that (u_1, π_1) belongs to the set $[\bar{g}_1, \infty) \times [\bar{z}_1, \bar{z}_1 + \Delta]$. Using a similar argument and subtracting observed probabilities, we can determine the probability that (u_1, π_1) belongs to the yellow square, with $u_2, \ldots, u_J$ and $\pi_2, \ldots, \pi_J$ marginalized as before. We can determine the density at the point $(\bar{g}_1, \bar{z}_1)$, marginalized over the other components, by considering an arbitrary small Δ.

In the special case when $J = 1$ so that there is only one inside option, the perturbations above have intuitive interpretations. Specifically, variation in $\bar{g}_1$ only affects the match of consumers on the margin between choosing the sole inside option and the outside option, and variation in $\bar{z}_1$ affects the match of consumers that are on the margin of being acceptable for product 1. Thus, the two perturbations together yield the density at the point $(\bar{g}_1, \bar{z}_1)$.

The argument outlined above only provides us with only the marginal density of (u_1, π_1). This is because the shaded yellow box from figure 1 is the projection on the $u_1 - \pi_1-$hyperplane. When projected on the $u_2 - \pi_2$ hyperplane, the set has still the L-shape implied by the conditions $u_2 < \bar{g}_2$ or $\pi_2 < \bar{z}_2$. The yellow region in figure 2 illustrates this set projected on the $u_1 - u_2 - \pi_2$ hyperplane for a particular value of $(\bar{g}, \bar{z})$. Observe that this region conditions

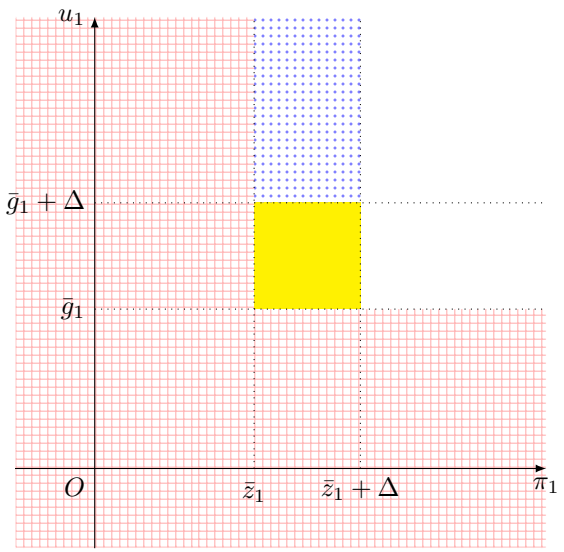


Figure 1: Two Dimensions

on the event that $u_2 < \bar{g}_2$ or $\pi_2 < \bar{z}_2$ in order to focus on the set of consumers that would not be matched with product 2 if $\bar{g}_1$ or $\bar{z}_1$ were perturbed.

Our approach uses mathematical induction to extend this argument to higher dimensions, ultimately recovering the joint distribution of (u, π). The inductive step is also illustrated in figure 2. We can perturb $\bar{z}_2$ to $\bar{z}_2 + \Delta$ and repeat the steps of perturbing $\bar{z}_1$ and $\bar{g}_1$ at the value $\bar{z}_2 + \Delta$ to obtain the probability that $u_2 < \bar{g}_2$ or $\pi_2 < \bar{z}_2 + \Delta$, while focusing on consumers such that $(u_1, \pi_1) \in [\bar{g}_1, \bar{g}_1 + \Delta] \times [\bar{z}_1, \bar{z}_1 + \Delta]$. Similarly, we can perturb $\bar{g}_2$ to $\bar{g}_2 + \Delta$ to obtain the analogous quantity at $\bar{g}_2 + \Delta$. Subtracting these two quantities yields the probability that $(u_2, \pi_2) \in [\bar{g}_2, \bar{g}_2 + \Delta] \times [\bar{z}_2, \bar{z}_2 + \Delta]$, $(u_1, \pi_1) \in [\bar{g}_1, \bar{g}_1 + \Delta] \times [\bar{z}_1, \bar{z}_1 + \Delta]$ and for $j > 3$, $u_j < \bar{g}_j$ or $\pi_j < \bar{z}_j$. This set is the cross-hashed cube in Figure 2.

Although an illustration in higher dimensions is challenging, this process can be used to determine the probability that (u, π) belongs to the set $\prod_{j=1}^{J} [\bar{g}_j, \bar{g}_j + \Delta] \times [\bar{z}_j, \bar{z}_j + \Delta]$. This probability, for arbitrarily small Δ, yields the density of (u, π) if it exists. The proof formalizes this intuition without requiring that (u, π) admits a density by identifying the mass accumulated in sets that generate the Borel sigma algebra.

The message of the result is intuitive. When two sets of instruments are present, one that shifts choice sets and one that shifts preferences, they can be used together to identify the distribution of utilities and acceptance decisions. The argument uses the variation in match probabilities with respect to the shifters g and z for preferences and acceptance decisions respectively. Assumption 1 implies that each shift leaves the joint distribution of (u, π)

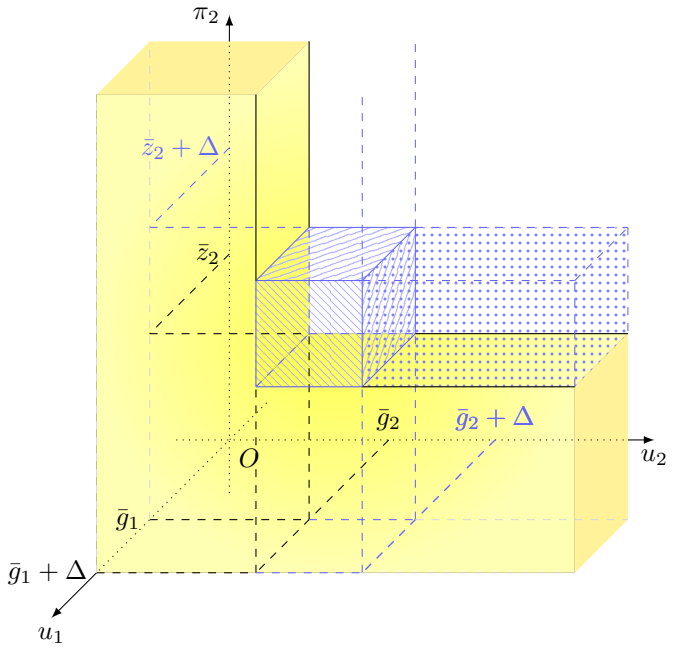


Figure 2: Three Dimensions

unchanged. And, since π implies the vector of acceptance decisions σ for a given z, the result implies the identification of acceptance decisions jointly with the distribution of indirect utilities, u.

This part of our argument is closely related to prior work in He et al. (2024), which shows identification in models of two-sided matching markets while relaxing previous restrictions on preference heterogeneity (e.g. Diamond and Agarwal, 2017) or on tail behavior on unobservables (e.g. Menzel, 2015). Although He et al. (2024) (henceforth HSS) show a result similar to lemma 1, there are three ways in which the results in HSS require stronger assumptions.

First, HSS requires exogenous continuous variation in d_i, y_i and z_i (see Condition 3.3 and Proposition 3.4 in HSS), while we dispense away with any requirement of variation (continuous or not) in d_i. Second, HSS identifies the functions $u_j(d_i)$, $g_j(y_{ij})$ and $\pi_j(d_i)$ in a first-step using a non-primitive rank condition.[12,13] While we take $g(\cdot)$ to be given for now,

[12]In our notation HSS assume $v_{ijt} = u_{jt}(d_i) - g_{jt}(y_{ij}) + \omega_{ijt}$, $\sigma_{ijt} = 1\{\pi_{jt}(d_i) - h_{jt}(z_{ij}) + \eta_{ijt} > 0\}$. Their results assume a rank condition on the matrix of derivatives of market shares with respect to each of the observable characteristics (Condition 3.3, HSS). One interpretation of lemma 1 is that it provides a primitive condition for their results in a more general model. In appendix B, HSS also show identification of certain derivatives of indirect utility functions with non-separable unobserved heterogeneity. However, these results are not sufficient for identification of the distribution of preferences.

[13]We also require knowledge of $g_j(\cdot)$ in Lemma 1, but will provide weak sufficient conditions for identification in the next subsection.

our assumptions in section 3.1.2 for identifying this function can be verified from model primitives. Third, HSS require that the unobservable ω_i is separable from both d_i and y_{ij} so that $v_{ij} = u_j(d_i) + \omega_{ij}^u - g_j(y_{ij})$ and $\pi_{ij} = \pi_j(d_i) + \omega_{ij}^\pi$. This restriction rules out models with non-separable unobserved heterogeneity, which include important cases such as random utility models in characteristic space (e.g. Berry and Pakes, 2007) and certain random coefficient models.[14] Our approach allows d_i and ω_i to be non-separable. Finally, section 3.3 considers a model with endogenous product characteristics.

3.1.2 Identification of $g(\cdot)$

The results above assume that the functions $g_{jt}(\cdot)$ are known. A commonly-studied special case is that of a special regressor, $g_j(d_i, y_{ij}) = y_{ij}$ (Lewbel, 2007). This functional form restriction does not always yield from desirable primitive economic assumptions. Thus, we will now show that $g_j(\cdot)$ is also non-parametrically identified under weak assumptions. In order to develop this result, we need to introduce further notation and assumptions.

Definition 1. Goods j and k are strict substitutes in y at (d_i, y_i, z_i) if $\frac{\partial}{\partial y_{ik}} s_j(d_i, y_i, z_i)$ and $\frac{\partial}{\partial y_{ij}} s_k(d_i, y_i, z_i)$ exist and are non-zero.

This notion is a mild strengthening of requirements imposed in equation (1) and assumption 1, which together imply that the market share of each good k is weakly increasing (decreasing) in y_{ij} if $g_j(d_i, y_{ij})$ is weakly increasing (decreasing) in y_{ij}. It further requires the existence of cross-partials and assumes that they are non-zero.[15]

Define $\Sigma_{j,k}(d_i, y_{ij}, y_{ik}) = 1$ if there is $\bar{y}_i$ and z_i in their respective supports such that goods j and k are strict substitutes in y at $(d_i, \bar{y}_i, z_i)$, $\bar{y}_{ij} = y_{ij}$, and $\bar{y}_{ik} = y_{ik}$. For two goods j and k, we say that there is a path connecting two values y_j and y_k, respectively, if there is a sequence of goods m_l and values of y_l, $(j, y_j) = (m_1, y_1), (m_2, y_2), ..., (m_n, y_n) = (k, y_k)$, such that such that for all $l = 2, ..., n$, $\Sigma_{m_{l-1}, m_l}(d_i, y_{l-1}, y_l) = 1$.

Assumption 3. *For every d_i, every good k, and almost all values of y_{ik} in its support, there exists a path connecting good k and value y_{ik}, (k, y_{ik}), to the reference good j and the reference value y_0, (j, y_0), for which we have normalized $\left|\frac{\partial g_j(d_i, y_0)}{\partial y}\right| = 1$.*

[14] In appendix A, He et al. (2024) also show identification of certain derivatives of indirect utility functions with a particular form of non-separable unobserved heterogeneity that is not nested in our model. The results in that appendix are not sufficient for identification of the distribution of preferences.

[15] Berry et al. (2013) show that invertibility of demand does not require smoothness. The purpose of the assumption in our exercise is different from invertibility.

This assumption requires some degree of substitution between the goods considered but is weaker than requiring strict substitution between every pair of goods at all values of y_i and z_i. Moreover, the condition is testable. In models of unit demand with latent choice sets, there are at least two important reasons why a given pair of goods j and k may not be substitutes. First, preferences for goods may restrict substitution patterns between goods that are considered. Salient examples include models with vertical preferences where consumers only substitute to goods that are adjacent in quality ranking or the pure characteristics model of Berry and Pakes (2007). Nonetheless, these models often admit a path connecting any pair of goods, thereby satisfying assumption 3 (see Berry et al., 2013, for related ideas). Second, choice sets may restrict substitution in demand. For example, if latent choice sets are of the form that goods j and k never appear in the choice set together then the relevant cross-partials the shares of these goods would be zero. However, there will still be a path connecting two values of their shifters y_j and y_k, if there is a third good, l and a shifter value y_l, such that the pairs $\{j, l\}$ and $\{l, k\}$ are strict substitutes. Furthermore, the shifters y_i and z_i at which $\{j, l\}$ are strict substitutes can be different than those at which $\{l, k\}$ are strict substitutes. However, the assumptions rules out particular cases where the share of a good k is constant with respect to y_k for an interval range of values at every value of z or cases in which good k only substitutes with the outside good.

Although assumption 3 is non-primitive, proposition 2 in the appendix shows weak primitive conditions under which goods j and k are strict substitutes. It shows that goods j and k are strict substitutes in y_i at (d_i, y_i, z_i) if the pair of goods $\{j, k\}$ belong to the choice set O_i with non-zero probability, the derivatives of $g_j(d_i, \cdot)$ and $g_k(d_i, \cdot)$ are non-zero, and the joint distribution of indirect utilities implies substitution between the goods in demand. This requirement is satisfied for well-behaved pure characteristics models, including versions with vertical preferences. Hence, a researcher may justify assumption 3 either by evaluating the assumption directly in the data or by arguing for the sufficient conditions based on either proposition 2 or corollary 3.

While assumptions 1 and 2 have allowed for atoms in the joint distribution of (u_i, π_i), assumption 3 requires that some regions admit a density between pairs of components of u_i. If the distribution of u_i (conditional on d_i) has an atom at $g(d_i, y_i)$, then $s(d_i, y_i, z_i)$ may not be differentiable with respect to y_i at that value even if the function $g(d_i, y_i)$ is differentiable. In this case, goods j and k may not be strict substitutes in y_i at (d_i, y_i, z_i). We view this restriction as mild.

Finally, we require the following weak support and regularity conditions to identify $g(\cdot)$:

Assumption 4. *(i) The support of the random vector y_i, denoted Y, is rectangular with*

non-empty interior.

(ii) For each d_i and j, the function $g_j(d_i, y_j)$ is continuously differentiable in y_j

Part (i) places a weak requirement on the support of Y that is used mostly for tractability and allow us to write $Y = \prod_j Y_j$ where Y_j is a non-empty closed interval. Part (ii) implies that the functions $g_j(d_i, y_j)$ are smooth with respect to the second argument. These assumptions together imply that $g_j(\cdot)$ is identified on its support:

Lemma 2. *Suppose that assumptions 1, 3 and 4 hold and $|J| > 1$. Then, for every $j \in J$, the function $g_j(d_i, \cdot)$ is identified for all $y_j \in Y_j$.*

Proof. See appendix A.3. □

The argument first identifies the ratio of $g_k'(d_i, y_{ik})$ and $g_j'(d_i, y_{ij})$ for goods j and k that are strict substitutes in y at (d_i, y_i, z_i). Consider the inclusive value of a consumer conditional on (d_i, z_i, y_i). Dropping the conditioning on d_i and z_i, this inclusive value is given by

$$V^*(g(y_i)) = \sum_{O \in \mathcal{O}} E\left(\max_{j \in O} u_j(\omega_i) - g_j(y_{ij}) \middle| O, g(y_i)\right) P(O),$$

where $g(y_i) = \left(g_1(y_{i1}), \ldots, g_J\left(y_{i|J|}\right)\right)$ and assumption 1 implies that the probability that $P(O)$ does not depend on y_i. The envelope theorem implies that

$$\frac{\partial V^*(g(y_i))}{\partial g_j} = -s_j(y_i).$$

This result is a version of Roy's identity for stochastic choice models (see McFadden, 1981), but for models with latent choice set constraints.[16] Assume that $V^*(g)$ is twice-continuously differentiable, a requirement that our proof dispenses with but is useful to understand the core of the argment. Then, the partial derivative of this equation with respect to y_{ik} yields that

$$\frac{\partial s_j(y_i)}{\partial y_{ik}} = -\frac{\partial^2 V^*(g(y_i))}{\partial g_j \partial g_k} g_k'(y_{ik}).$$

[16] This proof technique is also related to methods used in Allen and Rehbeck (2019) to consider latent utility models with additive heterogeneity. There are three differences worth noting. First, we avoid the representative agent's problem that is central to the arguments in Allen and Rehbeck (2019), resulting in a more direct approach to results on identification. Second, our model involves a two-sided problem with latent consumer-specific choice sets whereas choice sets are observed in Allen and Rehbeck (2019) and McFadden (1981). Third, we provide testable or primitive conditions – assumption 3 and proposition 2 – that imply the required sufficient conditions on the cross-partials of $V^*(g(y_i))$.

Taking the ratio of the partial derivatives of $s_j(\cdot)$ with respect to y_{ik} and of $s_k(\cdot)$ with respect to y_{ij}, and applying Young's theorem, we get that the ratio

$$\frac{g_k'(y_{ik})}{g_j'(y_{ij})} = \frac{\partial s_j(y_i)}{\partial y_{ik}} \Big/ \frac{\partial s_k(y_i)}{\partial y_{ij}} \tag{5}$$

is identified. Our model has not imposed enough smoothness assumptions to ensure continuity of second partial derivatives of V^*, a key requirement for Young's Theorem. Instead, our model and assumptions 1 and 3 are sufficient to show symmetry of the cross partial derivatives.

If all pairs of goods are strict substitutes at all values of (y_i, z_i) (for each d_i), we could directly use the normalizations that $g_j(y_0) = 0$, $\left|\frac{\partial g_j(y_0)}{\partial y}\right| = 1$ and assumption 4 to solve for $g_k(\cdot)$ and $g_j(\cdot)$. While not all pairs of good are strict substitutes, assumption 3 guarantees that there is a path connecting good k for almost all values of y_{ik} to the reference good j at the reference value y_0. Thus, the ratio of derivatives $\frac{g_k'(y_{ik})}{g_j'(y_{ij})} = \prod_{l=1}^{n} \frac{g_{j_l}'(y_{ij_l})}{g_{j_{l-1}}'(y_{ij_{l-1}})}$ is identified. The normalizations that $g_j(y_0) = 0$, $\left|g_j'(y)\right| = 1$ and assumption 4 can again be used to solve for $g_k(\cdot)$ and $g_j(\cdot)$.

As argued above, each function $g_{jt}(d_i, \cdot)$ can be identified when $J > 1$ under the assumptions outlined earlier. In the case when $|J| = 1$, we can assume without loss that $g_j(d_i, \cdot)$ is known as long as it is monotonic since the outside option is normalized to zero.[17]

This result, which shows the identification of $g(\cdot)$, allows us to achieve identification wihout relying on quasi-linear special regressors. It differentiates our approach from that of Abaluck and Adams-Prassl (2021), which uses the assumption that true choice probabilities exhibit Slutsky symmetry (e.g. in y) to identify three specific models of consideration set formation.[18] Instead, we use a reduced-form model of latent choice sets and allow for asymmetries to arise from non-linearity of indirect utilities in y_{ij}. As before, the cost of this greater generality is the need for choice-set shifters.

[17] To prove this claim, assume that $g_1(\cdot)$ is increasing and note that the market share of good 1 conditional on (y, z) is $s(y, z) = \int 1\{u_1(\omega) > g_1(y_1), \pi(\omega) > z_1\} \, \mathrm{d}F_\omega = \int 1\left\{g_1^{-1}(u_1(\omega)) > y_1, \pi(\omega) > z_1\right\} \mathrm{d}F_\omega$ where the equality follows because $g_1(\cdot)$ is monotonically increasing. Thus, the model is observationally equivalent to one in which $u_1(\cdot)$ is replaced with $g_1^{-1}(u_1(\cdot))$ and y_1 enters linearly. If $g_1(\cdot)$ is decreasing, then the market share will increase in y_1 and the argument follows by inverting $-g_1(y_1)$ instead.

[18] Slutsky symmetry requires that if all options are in the choice set, then $\partial s_j(y_i)/\partial y_{ik} = \partial s_k(y_i)/\partial y_{ij}$. A necessary and sufficient condition for the above in our model is that $g_{jt}(\cdot)$ is linear in y_{ij} (see equation 5). Abaluck and Adams-Prassl (2021) uses departures from Slutsky symmetry combined with specific models of consideration to identify incomplete consideration sets.

3.1.3 Main Result

Lemmas 1 and 2 above yield the main identification result of the paper:

Theorem 1. *If assumptions 1 – 4 hold and* $|J| > 1$*, then for every* w*, (i) the function* $g_j(d, \cdot)$ *is identified for every* $j \in J$ *and* $y_j \in Y_j$*, and (ii) the joint distribution of* u_i *and* π_i *is identified for every value* (u, π) *in the interior of* $g(d, Y) \times Z = \prod_{j=1}^{J} g_j(d, Y_j) \times Z$*, where* $g_j(d, Y_j)$ *is the image of the set* Y_j *under* $g_j(d, \cdot)$ *and* Z *is the support of the random vector* z_i*.*

Proof. Condition on d_i and drop it from the notation. Lemma 2 directly implies part (i). For part (ii), take any $\bar{g} \in int\, g(d_i, Y)$ and $\bar{z} \in int\, Z$. By lemma 1, the distribution of (u, π) conditional on d and (u, π) in the interior of $g(d, Y) \times Z$ is identified. □

The techniques used in this section rely only on local variation in the shifters y_i and z_i. The benefit of this approach is that it does not lean on "identification at infinity" arguments (see He et al., 2024, for example). For example, an alternative method for identifying the distribution of indirect utilities would be to focus on extreme values of z_i under which consumers can choose any product in the market and then rely on previous results. Such an argument would extrapolate the preferences of all consumers from a subset. The identification results do not rely on such extreme values belonging to the support of the shifters y_i and z_i.

Of course, we can learn about the distributions of u_i and π_i in only the regions that correspond to the support of the observables. When the observables have full support, we can identify the joint distribution of (π_i, u_i) everywhere. We formalize this point in following corollary to theorem 1:

Corollary 1. *Suppose the hypotheses of theorem 1 hold. If the support of* (u_i, π_i) *is a subset of* $int\,(g(d_i, Y) \times Z)$*, the joint distribution of* u_i *and* π_i *conditional on* d_i *is identified.*

This joint distribution of u_i and π_i contains information about a host of economic phenomena based on unobservable factors. For example, correlation between u_{ij} and $u_{ij'}$ implies that products j and j' are close substitutes., i.e. consumers who like one tend to also like the other one. Correlation between π_{ji} and $\pi_{j'i}$ suggests that products j and j' tend to prefer the same set of consumers. Moreover, correlation between u_{ij} and π_{ji} suggests that consumers tend to prefer products that are likely to admit them.

The results above also imply that local variation in (y_i, z_i) can be used to identify the distribution of (u_i, π_i). However, the effects of marginal changes in the shifters on the probability that j is chosen from the set $O \supseteq \{j\}$ or on the probability that O is the choice set for a fixed

value of y and z requires full or large support assumptions on $(g(d,Y),Z)$. An alternative approach to requiring large support on the choice-set and preference shifters is to further restrict the model (see Barseghyan, 2022, for example). The trade-off between these strategies depends on the empirical setting, available data, and the questions of interest.

3.2 Necessity of Choice Set Shifters for Identification

An advantage of the results above is that they do not need to rely on strong assumptions on the latent choice set formation, but they come at the cost of greater demands on the data in terms of the choice set shifters. Thus, a natural question is whether we can achieve identification without these shifters.

One might conjecture that choice set shifters may not be necessary because a model with full choice sets is testable as long as an additively separable shifter of preferences is available. To see this, suppose that assumption 1 is satisfied, the joint distribution of (π_i, u_i) admits a continuous density function, and that $P(O=J)=1$. The density of indirect utilities at a point $g \in \mathbb{R}^J$ can be recovered either by using only local variation in g in the market share of the outside good or the market share in any good j.[19] Since the densities recovered in these two alternative ways must be equal to each other, the model is over-identified. This observation suggests that it may be possible for the restrictions implicit in the model to help discriminating between preferences and latent choice sets.

Our next result shows that this conjecture is false. That is, without further restrictions, it is not possible to identify both the distribution of latent choice sets and indirect utilities unless both sets of shifters are available.

Proposition 1. *Suppose assumption 1 is satisfied, and the joint distribution of u_i admits a density function. Further assume that the support of z_i is a singleton $\{\bar{z}\}$ and $g(d_i, y_i)$ is observed and has full support on $\mathbb{R}^{|J|}$. If there exists an open set $B \subset \mathbb{R}^{|J|}$ and a choice set $O \subsetneq J$ such that for all $u \in B$, $f_U(u) > 0$ and $P(O|u) > \kappa > 0$, then $f_U(u)$ is not identified.*

Proof. See appendix A.4. □

The result shows that if variation from a shifter of choice sets is not available, then we cannot recover the distribution of utilities if we allow for incomplete latent choice sets. Therefore,

[19] Observe that $s_0(g) = \int 1\{u \leq g\} f_U(u)\,du$ and $s_j(g) = \int 1\{u_j - g_j > 0\} \prod_{k \neq j} 1\{u_k \leq u_j + \tilde{g}_k\} f_U(u)\,du$ where $\tilde{g}_k = g_k - g_j$. Using these expressions, it is easy to see that

$$\frac{\partial^{|J|} s_0}{\partial g_1 \dots \partial g_{|J|}}(g) = \frac{\partial^{|J|} s_0}{\partial \tilde{g}_1 \dots \partial \tilde{g}_{j-1} \partial g_j \partial \tilde{g}_k \dots \partial \tilde{g}_{|J|}}(g) = f_U(g).$$

the conclusions of lemma 1 and theorem 1 do not hold. Our proof explicitly constructs an alternative distribution of indirect utilities and latent choice set probabilities that result in an identical market share function. Intuitively, we can explain the probability that a product is chosen either using preferences conditional on a choice set or using the probability that a product is in the choice set.

The distribution of preferences is not identified even though we allow for the shifter of preferences to have full support on its domain. Of course, the under-identification issue would be more severe if the support of $g(d_i, y_i)$ is more limited or if $g(\cdot)$ were unknown. The main requirement is that choice sets cannot be complete for all u. As discussed above, if latent choice sets are complete, the distribution of preferences is over-identified under the remaining assumptions. We demonstrate that preferences are not identified once incomplete choice sets have non-zero probability because multiple preference distributions can rationalize observed market shares by altering the probabilities of various choice sets.

This result implies that, without variation in the choice-set shifter, the special case of our model with complete choice sets is essentially the only one when identification can be achieved. Since simply allowing for incomplete latent choice sets results in under-identification, the results indicate that the conditions in theorem 1 are sharp. Any alternative to using shifters of choice sets would therefore require further restrictions on the model. There are two existing approaches that we are aware of. The first, proposed in Abaluck and Adams-Prassl (2021) uses specific models of choice set formation and assumes that the functions $g_j(\cdot)$ are known. The models of choice set formation include those in which the probability that an alternative is in the choice set is independent across alternatives and independent of preferences, or models in which the consumer is either inattentive or picks from the full set of available alternatives. The second approach, proposed in Barseghyan et al. (2021a) and Barseghyan (2022), uses a characteristic space model for the distribution of preferences. In this approach the distribution of indirect utilities lies in a lower-dimensional manifold of $\mathbb{R}^{|J|}$. An example is the pure characteristic model of Berry and Pakes (2007), which cannot allow for idiosyncratic product-specific preferences. Our approach does not require these *a priori* restriction.

3.3 Introducing Endogeneity

A challenge in estimating discrete choice demand systems is that certain characteristics may be unobserved to the econometrician. Correlation between these observables and observed characteristics may bias estimates of demand (Berry, 1994; Berry et al., 1995). For example, product prices may be set strategically in oligopolistic markets. This type of endogeneity

is usually analyzed using models in which indirect utilities depend on both observable and unobservable product characteristics (see Berry and Haile, 2014, for example).

We now assume that indirect utilities and profits can be written, with a slight abuse of notation, as follows:

$$u_{ijt} = \tilde{u}\left(x_{jt}, \xi^u_{jt}, \omega_i\right)$$
$$\pi_{ijt} = \tilde{\pi}\left(x_{jt}, \xi^\pi_{jt}, \omega_i\right),$$

where ξ^u_{jt} and ξ^π_{jt} are scalar unobservables, x_{jt} denotes a vector of observable product characteristics that are potentially correlated with the unobservables $\xi_{jt} = \left(\xi^u_{jt}, \xi^\pi_{jt}\right)$, and $u(\cdot)$ and $\pi(\cdot)$ are unknown functions. We have dropped d_i from the notation because our arguments will condition on it. Thus, the unobservable ξ_{jt} is implicitly d_i-specific. The combination of the assumptions that (i) the unobservables are scalars and (ii) the unknown functions are not indexed by j and t, makes this specification more restrictive than the ones analyzed in the prior subsections.[20]

We assume that the data generating process starts by sampling markets with characteristics of all the products in market t, namely $(x_t, \xi_t) = \{x_{jt}, \xi_{jt}\}_j$, drawn i.i.d. from a joint distribution that is common across markets. Then, ω_i and (y_i, z_i) are drawn i.i.d. across consumers, with $\omega_i \perp (y_i, z_i)$ as before. Each consumer belongs to only one market. Whereas the results in the prior subsections did not require across-market variation, the results in this subsection will exploit both cross-product and cross-market variation. We will therefore include the market index t on certain random variables for clarity.

Our goal is to identify the joint distribution

$$u_{it}, \pi_{it} | \, x_t, \xi_t$$

on the relevant support. The joint distribution that we previously identified conditioned on the market's identity t and implicitly on all the products in the market as well, but could not separate the effects of observables and unobservables. Now we want to condition on specific values of x_{jt} and ξ_{jt} for each of the products in the market. Knowledge of these distributions is sufficient for identification of several quantities of interest. These include identification of

[20] We can also allow for observables in $g_{jt}(\cdot)$ that may be correlated with unobservables ξ^g_{jt}. Theorem 1 and corollary 1 imply that $g_{jt}(y_{ij})$ is identified on the support of y_{ij} in each market t. When $g_{jt}(y_{ij})$ takes the form $g\left(x_{jt}, \xi^g_{jt}, y_{ij}\right)$ and x_{jt} is potentially correlated with ξ^g_{jt}, then identification of the function $g(\cdot, \cdot, \bar{y})$ for a fixed value of $\bar{y}$ follows from existent results for non-linear IV models. Chernozhukov and Hansen (2005) show identification of this model assuming the availability of instruments $r_{jt} \perp \xi^g_{jt}$, and additional regularity and support conditions.

choice probabilities under any choice set as well as identification of counterfactual choices with exogenous changes in x_t. The argument will solve the endogeneity problem and recover ξ_{jt}. Once we recover these unobservables, we will obtain the joint distribution of (u_{it}, π_{it}) conditional on the full vector of observed and unobserved characteristics.

We made the following restriction on $u(\cdot)$ and $\pi(\cdot)$:

Assumption 5. *Index restrictions.* x_{jt} *can be partitioned into* $\left(x^*_{jt}, \left(x^\delta_{jt}, x^\gamma_{jt}\right)\right)$ *such that indirect utility and profits take the form* $u_{ijt} = u\left(x^*_{jt}, \delta_{jt}, \omega_i\right)$ *and* $\pi_{ijt} = \pi\left(x^*_{jt}, \gamma_{jt}, \omega_i\right)$, *where* $\delta_{jt} = x^\delta_{jt} + \xi^u_{jt}$ *and* $\gamma_{jt} = x^\gamma_{jt} + \xi^\pi_{jt}$.

The index restrictions above are similar to those imposed in Berry and Haile (2014) to identify demand without choice set constraints. Although the observable components x^δ_{jt} and x^δ_{jt} are one-dimensional, this restriction is inessential in a linear model as long as one of the components is known to have a non-zero coefficient as the model can be re-normalized. In other words, the observables x^δ_{jt} and x^γ_{jt} set the units for ξ_{jt}. Linearity can also be relaxed, but we do not develop this extension for simplicity of notation and exposition. [21] Finally, the observable component of the indices, namely x^δ_{jt} and x^γ_{jt}, may be the same although this is not required as long as x^*_{jt} does not contain $\left(x^\delta_{jt}, x^\gamma_{jt}\right)$.

We now turn to the key assumption that forms the basis of our solution:

Assumption 6. *Invertibility. There exists a function* $\psi(\cdot, \cdot; x^*)$ *such that for any two markets* t *and* t' *with* $x^*_t = x^*_{t'} = x^*$, $\psi(\delta_t, \gamma_t; x^*_t) = \psi(\delta_{t'}, \gamma_{t'}; x^*_{t'})$ *implies* $(\delta_t, \gamma_t) = (\delta_{t'}, \gamma_{t'})$. *Moreover, for each market* t, $\phi_t = \psi(\delta_t, \gamma_t; x^*_t)$ *is known.*

This invertibility restriction requires that the model places sufficient restrictions such that (δ_t, γ_t) is invertible in the observable quantity ϕ_t. It is worth emphasizing that the analyst need not know the function $\psi(\cdot)$, only the realized value of ϕ_t for any market. This assumption parallels the literature on the identification of demand. Specifically, Berry and Haile (2014) assume that the index of demand – δ_t in our case – is invertible in the vector of market shares – ϕ_t in our notation – and the unknown function maps δ_t to market shares – $\psi(\cdot)$ in our notation. Primitive conditions for invertibility in the case of demand (without constraints) are studied in Berry et al. (2013).

[21] The case when $\delta_{jt} = \tilde{\delta}(x_{jt}) + \xi^u_{jt}$ and likewise for γ_{jt} follows from an extension based on results in Matzkin (2007). This case requires additional normalizations on the function $\tilde{\delta}(\cdot)$. The non-separable case follows from Chernozhukov and Hansen (2005), which requires strengthening the mean-independence restriction in assumption 1(i) below. Berry and Haile (2014) provide additional details regarding these extensions in an appendix, but focus their analysis on a base case that is similar to ours.

Our approach is similar. Recall that theorem 1 and corollary 1 show that the joint distribution of (u_{it}, π_{it}) is identified on the support of $(g(Y), Z)$. To solve the endogeneity problem, we will require the analyst to place sufficient primitive restrictions on the model to guarantee that these features are sufficient to identify ϕ_t.

We present two examples that satisfy assumption 6 below:

Example 5. Suppose that $(\delta_{jt}, \gamma_{jt})$ and x^*_{jt} are additively separable in both utility and profitability,

$$u\left(x^*_{jt}, \delta_{jt}, \omega_i\right) = u_0\left(\delta_{jt}, \omega_i\right) + u_1\left(x^2_{jt}, \omega_i\right)$$
$$\pi\left(x^*_{jt}, \gamma_{jt}, \omega_i\right) = \pi_0\left(\gamma_{jt}, \omega_i\right) + \pi_1\left(x^2_{jt}, \omega_i\right),$$

and that $E\left[u_0\left(\delta_{jt}, \omega_i\right) | \delta_{jt}\right]$ and and $E\left[\pi_0\left(\gamma_{jt}, \omega_i\right) | \gamma_{jt}\right]$ are strictly monotonic in δ_{jt} and γ_{jt}, respectively. The linear random coefficients preference model (example 2.2) satisfies these assumptions. Taking expectations conditional on market observed characteristics and indices, we get that

$$E\left[u_{ijt} | \delta_t, x^*_t\right] = E\left[u_0\left(\delta_{jt}, \omega_i\right) | \delta_{jt}\right] + E\left[u_1\left(x^*_{jt}, \omega_i\right) \middle| x^*_{jt}\right]$$
$$E\left[\pi_{ijt} | \gamma_t, x^*_t\right] = E\left[\pi_0\left(\gamma_{jt}, \omega_i\right) | \gamma_{jt}\right] + E\left[\pi_1\left(x^*_{jt}, \omega_i\right) \middle| x^*_{jt}\right],$$

where the equality follows because ω_i is independent of (δ, γ, x^*). This model satisfies assumption 6 with $\psi\left(\delta_t, \gamma_t; x^*_t\right) = \left\{E\left[u_{ijt} | \delta_t, x^*_t\right], E\left[\pi_{ijt} | \gamma_t, x^*_t\right]\right\}$. With a sufficiently large support of (Y, Z) is sufficient to identify these expectations (see corollary 1).[22]

Example 6. Assumption 6 also holds under weaker requirements on support but stronger functional form assumptions. Consider the following vertical model:

$$u\left(x^*_{jt}, \delta_{jt}, \omega_i\right) = \alpha_i u_0\left(\delta_{jt}, x^*_{jt}\right)$$
$$\pi\left(x^*_{jt}, \gamma_{jt}, \omega_i\right) = \beta_i \pi_0\left(\gamma_{jt}, x^*_{jt}\right),$$

for positive valued functions u_0 and π_0 that are strictly monotone if their first argument. Assume that $\omega_i = (\alpha_i, \beta_i)$ has support on $\mathbb{R}^2_+$. In our empirical context δ_{jt} in this model can be interpreted as the unobserved determinant of quality of facility j in market t and α_i can be interpreted as the preference for quality by patient i. Analogously, β_i may represent patient

[22] The large support assumption on (Y, Z) can be relaxed if ω_i is excluded from $u_0\left(\delta_{jt}, \omega_i\right)$ and $\pi_0\left(\gamma_{jt}, \omega_i\right)$ and $u_1\left(x^2_{jt}, \omega_i\right)$ and $\pi_1\left(x^2_{jt}, \omega_i\right)$ are unimodal. In this case, ϕ_t is the $2J-$dimensional vector with the mode of the joint distribution of u_{ijt} and π_{ijt} in the j and $j + J$ positions.

i's profitability and γ_{jt} denotes an unobserved determinant of preference for profitability by facility j.

If the joint distribution of (α_i, β_i) is unimodal and the support of (Y, Z) in market each t identifies the mode of $\left(u\left(x_{jt}^*, \delta_{jt}, \omega_i\right), \pi\left(x_{jt}^*, \gamma_{jt}, \omega_i\right)\right)$ (via corollary 1), then assumption 6 follows with $\psi\left(\delta_t, \gamma_t; x_t^*\right)$ equal to the $2J$ vector with the mode of the joint distribution of u_{ijt} and π_{ijt} in the j and $j+J$ positions.[23] Note that the support condition on (Y, Z) is weaker than those needed to identify expectations and that the assumption that ω_i is unimodal is testable.

Finally, we require the availability of instruments for x_t, which may be endogenous. Specifically, we impose:

Assumption 7. *(i) Availability of instruments.* $E\left[\xi_t | r_t\right] = 0$ *for all* r_t.[24]

(ii) Completeness. For any function $B\left(\phi_t, x_t\right)$ *with finite expectation,* $E\left[B\left(\phi_t, x_t\right) | r_t\right] = 0$ *a.e. in* r_t *implies that* $B\left(\phi_t, x_t\right) = 0$ *a.e. in* $\left(\phi_t, x_t\right)$.

This assumption is standard in the analysis of non-parametric instrumental variable models (see Newey and Powell, 2003) and is also required by Berry and Haile (2014). The completeness condition in part (ii) is the non-parametric analog to a rank condition in linear instrumental variable models.

We are now ready to prove our main result for the section:

Theorem 2. *If assumptions 5-6 are satisfied, then* $\delta\left(\cdot\right)$, $\gamma\left(\cdot\right)$ *and* $\left(\xi_t, \gamma_t\right)$ *are identified. In particular, the conditional distribution of* $u_{it}, \pi_{it} | x_t, \xi_t$ *is identified on the interior of the support of* $\left(g\left(Y_t\right), Z_t\right)$.

Proof. Assumptions 5 and 6 imply that there exists a function $\psi^{-1}\left(\cdot; x^*\right)$ such that $\left(\delta_t, \gamma_t\right) = \psi^{-1}\left(\phi_t; x_t^*\right)$. Assumption 7(i) implies that $E\left[\psi^{-1}\left(\phi_t; x_t^*\right) - \left(x_t^\delta, x_t^\gamma\right) \middle| r_t\right] = E\left[\xi_t | r_t\right] = 0$. Let $\tilde{\psi}^{-1}\left(\cdot; x^*\right)$ be an alternative function such that $E\left[\psi^{-1}\left(\phi_t; x_t^*\right) - \tilde{\psi}^{-1}\left(\phi_t; x^*\right) \middle| r_t\right] = 0$ almost everywhere. Assumption 7(ii) implies that $\psi^{-1}\left(\phi_t; x^*\right) = \tilde{\psi}^{-1}\left(\phi_t; x^*\right)$ almost everywhere. Therefore, $\psi^{-1}\left(\cdot; x^*\right)$ is identified. Since ϕ_t is known, we know that $\left(\delta_t, \gamma_t\right) = \psi^{-1}\left(\phi_t; x_t^*\right)$ is identified and so is $\xi_t = \left(\delta_t, \gamma_t\right) - \left(x_t^\delta, x_t^\gamma\right)$. □

The key hypothesis of Theorem 2 is Assumption 6. This invertibility assumption represents the main difference relative to those used for identifying and estimating models of demand

[23] A similar result holds under the assumption that the joint distribution of $(\log \alpha_i, \log \beta_i)$ is unimodal.

[24] We sample r_t jointly with (x_t, ξ_t).

without choice set constraints (e.g. Berry and Haile, 2014; Berry et al., 1995). In these analyses, identification arguments are based on a $J-$dimensional vector of indices containing product-level unobservables to be invertible in the $J-$dimensional vector of market shares. However, the existence of such an inverse – proved in Berry et al. (2013) for the case of demand – is not available in our case because we need to invert a $2J-$dimensional vector, (δ_t, γ_t), whereas market shares only have dimension J.[25] This creates the new challenge we solve in our analysis.

Our solution exploits the features of the model that are identified using the within-market variation in preference and choice-set shifters, Y and Z. Corollary 1 shows that this variation allows us to identify the distribution of the $2J$-dimensional random variable (u_{it}, π_{it}) on the relevant support (corollary 1). We therefore base our inversion on a $2J$-dimensional vector ϕ_t. To justify this approach, the researcher needs to place sufficient restrictions on the model so that ϕ_t is a known function of this joint distribution and is identified for each market.[26]

Our results therefore allow for endogenous characteristics in the presence of constrained choices, can be relevant for a number of applications. For example, a growing literature uses estimates of school demand to study the effects of school investment (Dinerstein and Smith, 2021) or the effects of competition between schools in prices and quality (Neilson, 2020; Allende, 2019). An important goal is to estimate the elasticity of school demand with respect to these observables in order to predict equilibrium effects of various policy reforms. While this work incorporates unobserved factors that affect school demand, it abstracts away from the possibility that schools select students by assuming that each student is matched with their most preferred school in equilibrium, an assumption that may not be reasonable in markets with selective school admissions. Our framework, to our knowledge, is the first to accommodate both these features.

[25] Of course, one possibility would be to omit product-level unobserved characteristics in the model of choice sets. This approach may allow us to obtain an inversion from the vector of market shares to δ_t. While this approach will allow for counterfactuals varying x_t while fixing ξ_t^u, the within-market variation utilized in section 3.1 will still be necessary to disentangle preferences from choice-set constraints.

[26] The use of within-market variation in the shifters Y and Z to identify demand resembles the work of Berry and Haile (2020). However, we require cross-market or cross-product variation in an instrument r_{jt} to identify the model instead of relying solely on within-market "micro data."

4 Data and Descriptive Analysis

4.1 Background

Dialysis is the predominant form of treatment for patients with End Stage Renal Disease (ESRD). It is a procedure that removes toxins that are otherwise filtered by a functioning kidney. Even with dialysis, median survival for ESRD patients is about five years (Figure 5.7, U. S. Renal Data System, 2020). Although kidney transplantation has much better outcomes, organs for transplantation are scarce, making dialysis the only feasible option for the majority of patients.

There are two ways in which dialysis can be performed. The first and most commonly used method in the US is hemodialysis, accounting for about 90% of dialysis patients (Figure 1.2, U. S. Renal Data System, 2020). This method circulates the patient's blood through an extracorporeal artificial kidney. Hemodialysis is usually performed in an outpatient facility that focuses exclusively on dialysis treatments. It lasts between three to four hours and must be performed two to three times a week depending on the patient's residual kidney function. The second method, peritoneal dialysis, requires a catheter to be surgically inserted into the patient's body which is then used to administer a cleansing fluid and to collect waste. A patient's choice between the two dialysis modalities depends on numerous factors, including medical conditions, lifestyle and preferences (Lee et al., 2008). Our study focuses on facility-based hemodialysis patients, considering the choice of alternative treatment modalities as part of the outside option.

Facilities performing hemodialysis are regulated – they are required to employ skilled staff, use highly specific capital and adhere to health and safety requirements (Department of Health and Human Services, 2008). The most binding constraint in the medium-term is the number of kidney dialysis stations in the facility. Dialysis machines are large, dedicated to a single patient at a time, and must be placed adjacent to a chair or a bed where a patient can be stationed for several hours. Short-term inputs influencing capacity include nursing staff and technicians that can operate the machines. The staff monitors patients, provides medications, administers injections, and cleans and services the machines prior to use by every patient. These staffing, capital and space requirements make capacity adjustments to demand fluctuations a slow response (Eliason, 2019; Grieco and McDevitt, 2017).

Medicare provides insurance for costs related to ESRD for all US patients, irrespective of age. This coverage is secondary for patients with a private or employer health insurance plan during first 30 months after diagnosis of ESRD, called the coordination period. Each patient-

year on hemodialysis costs approximately $90,000 at Medicare rates, and higher at private rates (Chapter 10, U. S. Renal Data System, 2020). With approximately 750,000 patients suffering from ESRD in the US, Medicare costs of patients with kidney failure totaled to $49.2 billion in 2018 (Chapter 1 and 10, U. S. Renal Data System, 2020). This figure is more than 7% of all Medicare claims and more than 1% of national health care spending (Chapter 10, U. S. Renal Data System, 2020).

4.2 Data

The data for this study are taken from the US Renal Data System (U. S. Renal Data System, 2020). These data are assembled from various sources, including Medicare claims, facility reports and data on patient outcomes collected as part of the regulatory process. There are two important pieces of information that we will use for our study. First, we observe the residential zip-code, demographics, employment status and co-morbidities of each patient, as well as the facility where each patient is being treated. These data include patients who are initially covered by a private or employer health insurance plan because the start of dialysis determines the date at which a patient becomes eligible for full coverage by Medicare.

Second, the role of Medicare as the near-universal insurer in this market allows us to track the number of patients that are being treated in each facility on any given day. Further, we can determine whether an ESRD patient was cared for using hemodialysis or peritoneal dialysis.

Our analysis sample focuses on patients whose first treatment commenced at a facility in California between 2015 and 2018. There are two main restrictions imposed by this choice. First, the restriction to a single state is for tractability. Although we chose California since it is the largest state in terms of population, it also happens to be the case that the vast majority of its population does not live close to a neighboring state. Given the role of Medicare in this part of the healthcare sector, idiosyncrasies regarding California's healthcare sector are less relevant for our study. Our sample selection procedure is further described in appendix B.

Second, we focus on the first facility where a patient begins dialysis to abstract away from considerations that are unique to switching facilities, which include interference with continuity of care and administrative or financial barriers.[27] In our sample, 74.2% of patients are treated at only one facility and the average patient only visits 1.30 facilities. Our approach

[27]We drop the certain quarters in which a facility enters, exits, moves or rapidly expands or contracts. See appendix B for further details. Patients matched to one of these facilities during this time-period are considered to be matched to the outside option. Thus, the value of the outside option is the inclusive value of going to one of these facilities or of choosing peritoneal or home dialysis.

Table 1: Facility Sample

	All facilities	Ownership		
		Fresenius and Davita	Other chains	Independent
Facility				
N	553	377	114	78
Facility-year	2093	1418	385	290
Number of patients				
Mean	108.6	113.0	100.2	98.2
Std. dev	46.8	46.3	38.9	54.9
Number of stations				
Mean	22.3	22.3	22.0	22.5
Std. dev	7.6	7.2	7.2	9.6

Notes: Sample of all facility-year observations, as described in table B.1. The number of patients for a facility is the daily average of enrolled patients undergoing hemodialysis.

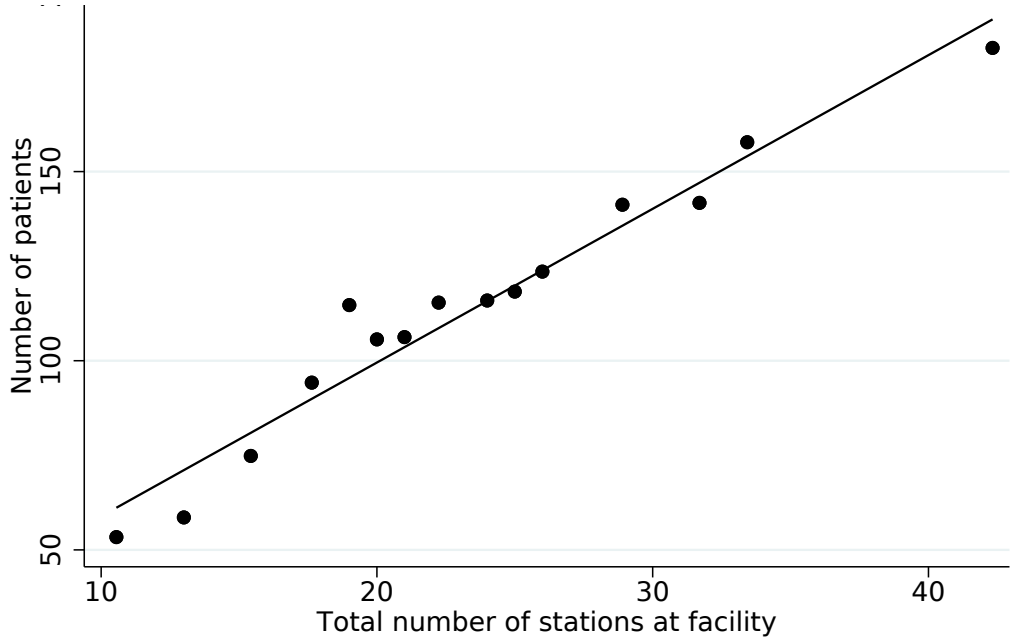
is consistent with facility moves being unexpected, say due to residential moves or other changes that are unexpected at the time when the patient begins dialysis.

4.3 Description of Sample and Choices

Table 1 describes the hemodialysis facilities in our sample. There are 552 facilities, most of them owned by one of the two large chains, Fresenius and DaVita. These and the vast majority of other facilities are for-profit and freestanding in that they are not associated with a hospital. In fact, these establishments usually focus exclusively on dialysis care and are not directly associated with another hospital or healthcare system. The average facility cares for just under 100 patients at a time, with chains and freestanding facilities caring for more patients per facility. The ratio of the number of stations to the number of patients is approximately five. This ratio is consistent with an average of two four-hour treatments per station per day since most patients require three treatments per week. Indeed, figure 3 shows that the number of patients per station is almost constant at five patients per station over the size distribution of facilities.

Table 2 describes the patient sample, which contains 41,913 new patients in our sample. Most of these patients choose hemodialysis at a facility in our facility sample. The patients are predominantly white, and the incidence of hypertension and diabetes is high. The majority of patients are on Medicare, an HMO or in the waiting period. The HMO group primarily

Figure 3: Patients per Dialysis Station



consists of patients over the age of 65 that are covered by a Medicare Advantage plan. The high share already on a Medicare plan at the start of dialysis is a consequence of the fact that age is a strong correlate of kidney disease. Going forward, we pool all patients who are Medicare eligible. The table also shows that the majority of patients begin dialysis in a freestanding facility. These facilities are not associated with a hospital and most of them are owned by chains. The largest chains are owned by either Fresenius or DaVita.

Table 3 describes the facilities near the patients in our sample and the chosen facility. The average patient has 6.5 facilities within 5 miles of their home zip-code and 17 facilities within 10 miles.[28] The typical patient receives dialysis at a facility with an average distance of 6.8 miles, but the median is lower, at 4.4 miles.

4.4 Evidence on Supply-Side Rationing

We now argue that capacity constraints affect the choice sets of patients. Our argument proceeds in two steps. We start by showing that facilities that have an unusually high caseload at a given point in time relative to their baseline are less likely to accept a new patients for a while. After demonstrating this pattern, we turn to analyzing how the facility where a patient starts treatment is affected by these constraints. To do this, we show that the

[28] Distances are measured between the patient's zip-code centroid and the facility's address.

Table 2: Patient Sample

	All patients	Treated at an in-sample facility	Ownership		
			Fresenius and Davita	Other chains	Independent
Panel A: Patient characteristics					
Patient count					
N	50002	43423	28647	7853	6923
Age					
Mean	63.1	63.7	63.6	64.9	63.1
Std. dev	15.0	14.8	14.8	14.9	15.1
Employed (%)					
Mean	0.1	0.1	0.1	0.1	0.1
White (%)					
Mean	71.3	72.1	73.0	67.4	73.4
Black (%)					
Mean	10.7	10.8	11.1	11.2	9.2
BMI					
Mean	28.4	28.4	28.4	28.4	28.4
Std. dev	7.3	7.4	7.4	7.6	7.5
Diabetes (%)					
Mean	39.6	40.4	40.4	40.0	40.8
Hypertension (%)					
Mean	86.5	86.4	85.8	86.9	88.2
Panel B: Insurance type at admission					
Medicare (%)					
Mean	32.9	32.8	32.4	37.0	29.6
Medicare Advantage (%)					
Mean	24.5	24.6	24.7	24.0	24.5
Medicare waiting period (%)					
Mean	12.5	13.1	12.7	12.8	15.2
Other (%)					
Mean	30.1	29.6	30.2	26.3	30.8

Notes: Sample of patients, as described in patient Table B.2. BMI is Body Mass Index (kg/m^2). Medicare Waiting Period is the 90-day period before Medicare covers hemodialysis. Other represents patients not covered by Medicare. These patients are typically covered by employer group health plans, the Department of Veteran Affairs, and private insurers. Most of them will become eligible for Medicare as a primary payer after 30-33 months.

Table 3: Patient Choices

	Facilities			
	Chosen	Within 5 miles	Within 10 miles	Within 25 miles
Number of facilities				
Mean	---	6.6	18.0	59.4
Std. dev	---	5.2	17.5	55.2
Median	---	5.0	11.0	32.0
Distance to facility				
Mean	6.7	3.2	6.0	14.1
Std. dev	7.3	0.7	1.3	3.1
Median	4.3	3.2	6.1	14.3
95th percentile	21.5	4.4	8.3	18.7
Number of patients at facility				
Mean	124.1	120.9	118.0	114.9
Std. dev	47.9	27.4	24.5	18.5
Median	120.0	121.7	120.1	120.9
Total stations				
Mean	24.1	23.4	23.1	22.8
Std. dev	7.8	4.2	3.3	2.3
Median	24.0	23.2	23.3	23.4
Chain (%)				
Overall mean	87.1	89.7	89.2	88.2
Fresenius	20.4	23.3	22.7	22.2
Davita	48.3	48.6	48.1	48.7

Notes: Sample of patient-facility pairs. Distance is measured in miles from the facility to the centroid of a patient's zip code. The number of patients at a facility is the sum of all patients enrolled at a facility that are undergoing hemodialysis.

distance to the chosen facility is higher if nearby facilities are more constrained. Moreover, the effects of constraints at facilities of different qualities are different. This latter finding suggests that patients also have preferences over our measures of quality.

Effects on flow of new patients

We hypothesize that the current caseload at a facility influences the facility's decision to accept a new patient. Let z_{ij} be a measure of the excess occupancy (relative to a target) in facility j when patient i enters the dialysis market. If our measure of excess occupancy is excludable from the patients' utility, conditional on controls that enter the utility function, then the inflow of new patients into facility j should be conditionally independent of the facility's caseload given these controls. To see this, consider a model without capacity constraints in which $\sigma_{ij} = 1$ for all i and j. In this model, assuming that the patient arrival into the dialysis market is exogenous, the probability that a new patient arrives into facility j is given by the unconditional probability that $u_{ij} > u_{ij'}$ for all j', which is independent of z_{ij}. However, if facilities are less likely to accept a patient when z_{ij} is high, then the in flow of new patients will be negatively correlated with caseload. As discussed in Section 2, Gandhi (2021) presents one micro-foundation for this relationship.

We will test this hypothesis using two sets of dependent variables measuring patient inflow on occupancy and excess occupancy. In the first set, the dependent variable is whether a facility j accepts a new patient on day t. We estimate this set using data from all days a facility is operating during our sample period. The dependent variable in the second set is the number of the days until the next patient begins treatment at facility j. This set is estimated using the subset of days in which a new patient began treatment. The regressions control for either facility-year or facility-month level fixed effects, and cluster standard errors at the facility level. In a subset of regressions, we also control for the average occupancy in other facilities within five miles of facility j.

Facility occupancy is measured as the number of patients being treated on date t at facility j and the excess occupancy is the difference between occupancy and a measure of target occupancy. The measure of excess occupancy is motivated by an examination of the time series of the number of patients at a facility, which reveals that several facilities undergo periods of expansion or contraction. These periods may correspond to investment in capital, increases in staffing or restructuring of the facility's operations and could confound the results. To account for these changes, we need to construct a measure of target capacity given the facility's operational setup on a given day. One way forward would be to use high-frequency data on facility inputs and investments in order to estimate facility capacity. Unfortunately,

labor inputs and capital investment are recorded only annually, and their timing is unknown. Instead, we estimate target occupancy using a regime-switching autoregressive model with a linear trend on the occupancy time series for each facility. The model detects breaks in each facility's occupancy trend to identify points at which the facility's occupancy process changes. We construct the target occupancy on a given date as the expected value on a given day.[29] We do not detect any breaks in trends for 500 of 553 facilities. Conditional on finding a break in the trend, the average number of breaks is 2.02. Thus, while not rare, the breaks in trend are not relevant for the vast majority of facilities. Table B.3 in the appendix shows that our estimate of target occupancy is positively correlated with the (low-frequency) measures of facility inputs available in our dataset, even conditional on facility fixed effects. The daily within-facility standard deviation of excess occupancy is 4.22.

There are three notable findings from the regressions of patient inflows on our measures of occupancy (see table 4). First, controlling for facility-year fixed effects, higher occupancy is negatively correlated with the probability of a new patient beginning dialysis at the facility and positively correlated with the expected waiting time until the next patient (columns 3-6). This relationship is robust to the inclusion of occupancy at other nearby facilities. Although not reported, this negative relationship between occupancy and patient inflow is robust to the inclusion of finer controls, such as facility-quarter or facility-month fixed effects.

Second, we observe that including facility-time controls appears to be important. The results in columns (1) and (2) are analogous to those in columns (3) and (5), but use only facility-specific fixed effects instead of facility-year fixed effects. The estimated relationship between the probability of new patient beginning dialysis and the facility's occupancy is now positive. When combined with the result that facility-time controls yield a robust negative coefficient, it suggests that fluctuations in a facility's target occupancy may be important.

Third, our measure of excess occupancy purges some of the confounding variation in the raw measure of occupancy that resulted in a positive coefficient in column (1). This variation was absorbed in specifications that employed fixed effects at the facility-year or finer levels. Since including a richer set of fixed effects will not be feasible in the non-linear model that we will

[29] Specifically, let $n_{j\tau}$ be the number of patients being treated at facility j on day τ. Assume that $n_{j\tau}$ follows the following time series model with $m \geq 1$ regimes $n_{j\tau} = \alpha_{jk(\tau)} + \beta_{jk(\tau)}\tau + \gamma_{jk(\tau)} n_{j\tau-1} + e_{j\tau}$,where $k(\tau)$ is a weakly increasing function that maps days $\tau = 1, ..., T$ to regimes $k = 1, ..., m$. The disturbance $e_{j\tau}$ has mean zero, constant variance, and follows an ergodic process. This model is consistent with a birth-death process in which departure rates are proportional to n_{jt} and arrival rates are a function of $n_{j\tau} - n^*_{j\tau}$. The target occupancy on date τ is defined as $n^*_{j\tau} = \frac{\alpha_{jk(\tau)} + \beta_{jk(\tau)}\tau}{1-\gamma_{jk(\tau)}}$. We estimate the parameters of this model, which include the dates on which the regimes changes. The regime changes for each facility are estimated using a modified Schwartz criterion proposed in Liu et al. (1997) and analyzed in Bai and Perron (2003). We winsorize $n_{j\tau} - n^*_{j\tau}$ by censoring the top and bottom 5% for each facility j in order to limit the influence of outliers.

ultimately estimate, our empirical specifications will use this measure of excess occupancy in the acceptance policy function.

Fourth, these regressions also speak to the effect of capacity constraints at other facilities close to facility j. There are two opposing forces. Constraints at other facilities close to j can increase the demand for facility j. But, this force can also push facility j to be more selective and turn away less profitable patients because it expects a higher flow of patients, allowing the facility to cream-skim the most desirable patients. Our results show that the number of patients being treated at other facilities close to facility j increases the probability that new patients start treatment at facility j (see columns 8 and 10 in table 4). This evidence weighs in favor of increased demand at the facility rather than the hypothesis that constraints at nearby facilities create a strong enough push for a facility to be more selective, although we cannot rule out this latter possibility because of the offsetting effects. Because our results are consistent with facility strategies that are not responsive to short-term constraints faced by competitors, we will ignore strategic interactions of this nature in our model. This assumption is also made in Gandhi (2021) for tractability, which studies selective patient acceptance in nursing homes.

Effects on where patients are treated

Having shown that capacity constraints affect the inflow of patients, we now investigate the effects of capacity constraints on where patients receive treatment. Figure 4 presents a binscatter indicating that the distance to the chosen facility is increasing in the average excess occupancy of facilities within five miles of the patient's zip-code centroid. This exhibit residualizes fixed effects at the zip-code-quarter level in order to control for confounding trends in the facilities' target occupancy. Again, we find that facility capacity constraints influences outcomes in this market.

Discussion

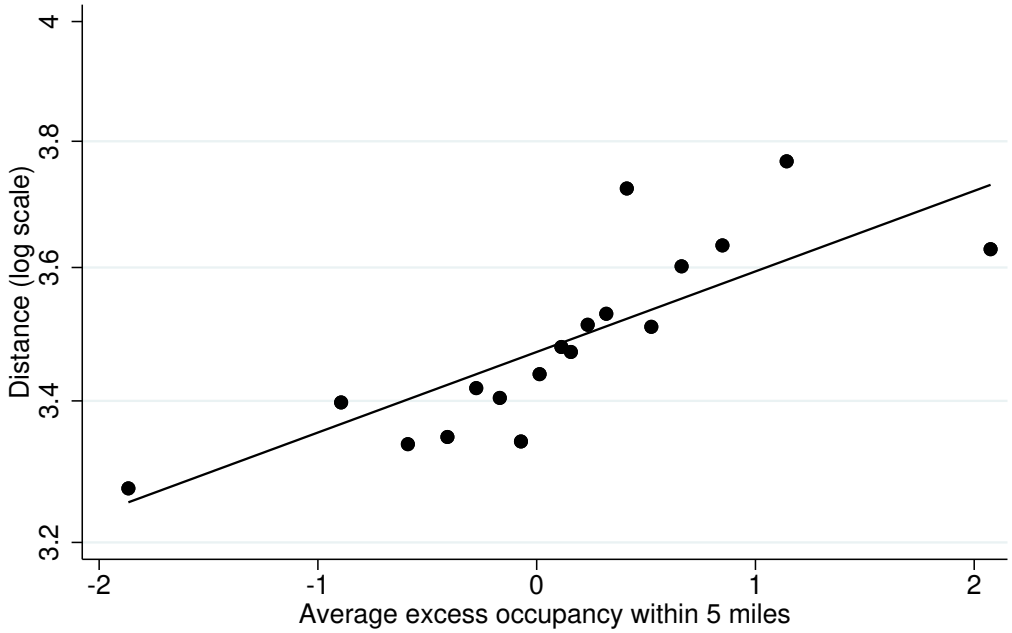
Taken together, the qualitative results indicate that capacity constraints are important drivers of realized matches if fluctuations in occupancy are not correlated with preferences for the facility. The main potential threat is that crowded facilities are undesirable. However, this concern is limited if patients primarily determine their decisions on longer-term crowding than the finer variation that we leverage in these estimates. The annual within-facility autocorrelation in excess occupancy is 0.06, suggesting that utilization on a specific day is not strongly correlated with the long-term occupancy.

Table 4: Evidence of Capacity Constraints

	Any new patient	Log(days to next patient)	Any new patient		Log(days to next patient)		Any new patient		Log(days to next patient)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Occupancy	0.0002**	0.004***	-0.0012***	-0.0012***	0.019***	0.018***				
	(0.0001)	(0.001)	(0.0001)	(0.0001)	(0.002)	(0.002)				
Excess occupancy							-0.0004***	-0.0006***	0.016***	0.016***
							(0.0001)	(0.0001)	(0.002)	(0.002)
Occupancy within 5 miles				-0.0001		0.004*		0.0007***		0.002*
				(0.0001)		(0.002)		(0.0001)		(0.001)
Facility FE	X	X					X	X	X	X
Facily-Year FE			X	X	X	X				
Observations	724,946	35,332	724,946	706,690	35,332	35,332	724,946	706,690	35,332	35,332
R-squared	0.0128	0.112	0.0264	0.0252	0.158	0.158	0.0128	0.0119	0.116	0.116

Notes: Sample of facilities as described in table 1. An observation is a day-facility pair, where the facility is open over the entire sample. Regressions with Log(days to next patient) consider the subset of days on which a facility admitted a new patient. The patients included are described in table 2. $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. Standard errors are clustered at the facility and county level.

Figure 4: Distance to Chosen Facility



Notes: Binscatter with twenty bins, residualized against patient zip-code and quarter-year fixed effects using the estimator in Cattaneo et al. (2021).

Without further institutional context, a potential alternative interpretation of our results is that capacity constraints manifest themselves in increased waiting time rather than a binary accept/reject decision by a facility. Dialysis, however, is a time-sensitive treatment and either delaying or advancing treatment by more than a few days relative to an optimal start can pose substantial health costs (Chan et al., 2014). This feature of the market favors our interpretation over an unobserved waiting time as the instrument rationing demand.

5 Estimates

5.1 Parametric Specification and Estimation

Although the arguments showing Theorem 1 are non-parametric and constructive,[30] there are two important challenges in constructing a non-parametric estimator. In principle, the constructive nature of our argument suggests estimating the market shares in equation (3) and then recover $g(\cdot)$ and the joint distribution of (u, π) in a second step. This problem is challenging because of dimensionality – the share equation has a $J-$dimensional range and at least a $2J-$dimensional domain. This approach has not been shown to work well even is models without choice constraints if there are more than a few products in the market (see also Compiani, 2021, for example). The typical solution to this problem is to directly estimate the distribution of preferences and, in our case, also the distribution of choice sets. Estimating such a model with latent choice sets non-parametrically is exceedingly difficult because the number of potential choice sets is large even for relatively small J.[31] Thus, enumerating all possible latent choice sets in order to compute the likelihood is often computationally infeasible.[32]

Instead, following much of the literature on discrete choice demand models, we parametrize the distribution of (u, π). We chose a specification to aid computation. First, we address

[30] To see why our proof is constructive, note that arguments in Lemma 2 show that $g(\cdot)$ is the solution to differential equation with an initial condition given by the normalizations on $g_j(\cdot)$ and derivatives given by the ratio of the cross partials of $s(\cdot)$, which are data. Lemma 1 yields the joint distribution of (u_i, π_i) given $g(\cdot)$, where the probability mass on any Borel set in the interior of χ is constructed in the proof. In fact, if $s(w, y, z)$ has continuous partial derivaties with respect to y and z, and each $\frac{\partial g_j(\cdot)}{\partial y_j} \neq 0$, then the density of (u_i, π_i) at $(g(y_i), z_i)$ is given by $\frac{\partial^{2J} s_0(w,y,z)}{\partial y_1 ... \partial y_J \partial z_1 ... \partial z_J} / \prod_{j=1}^{J} \frac{\partial g_j(y)}{\partial y_j}$.

[31] Recall that the likelihood of observing agent i in facility j given agent i's observable characteristics (w_i, y_i, z_i) is given by equation (3). The number of terms in this sum is equal to the number choice possible choice sets, which is equal to $2^{|J|}$. With only fourteen facilities, which is approximately the average number of facilities within ten miles for a patient, the number of choice sets is 16,384.

[32] Simulation studies also often limit the number of goods to a small number for these reasons. Abaluck and Compiani (2020), for example, conduct their monte carlo simulations with 10 goods or less.

the curse of dimensionality due to the large number of potential choice sets using a Gibbs sampler (see also Logan et al., 2008; Menzel and Salz, 2013; He et al., 2024). It modifies the sampler from McCulloch and Rossi (1994) with a data augmentation step to accomodate the case with latent choice sets. This will motivate distributional assumptions that admit closed-form solutions of certain conditional distributions. Second, we allow for correlations between preferences and choice sets via unobservables (ω_i in our notation). As mentioned earlier, the prior literature often assumes that choice sets are independent of preferences in order to further simplify computation. Third, we include random coefficients on agent's preferences for facility characteristics, which allows for more flexible substitution patterns.

Based on these considerations, we make the following assumptions on the preferences and acceptance functions:

$$v_{ij} = \delta_j + \beta_d d_i - g\left(d_i, y_{ij}\right) + \beta_i x_j + \varepsilon_{i0} + \varepsilon_{ij} \tag{6}$$

$$\sigma_{ij} = 1\left\{\gamma_j + \alpha d_i - z_{ij} + \nu_{ij} > 0\right\}, \tag{7}$$

where x_j are observed facility characteristics, δ_j and γ_j are facility fixed effects, and $\beta_i, \varepsilon_{i0}$, ε_{ij} and ν_{i0}, ν_{ij} are idiosyncratic shocks. We adopt the normalizations that $g'\left(d_i, y_{ij}\right) = 1$ at $y_{ij} = 1$ and $g\left(d_i, y_{ij}\right) = 0$ at $y_{ij} = 0$ for all d_i, and that the admission index is expressed in units of z_{ij}. As before, d_i is a vector of agent i's characteristics. We parametrize $g\left(\cdot\right)$ as a quadratic function given d_i, with parameters β_g and collect $\beta = (\beta_w, \beta_g)$. The specific observables d, y and z are described in section 5.2.

We allow for unobserved match-specific correlations by allowing for ε_{ij} and ν_{ij} to be jointly normally distributed with mean zero and an estimated covariance matrix Σ. The term ε_{i0} captures individual heterogeneity in preferences for the facilities in the market relative to the outside option. A restriction in our model, relative to the non-parametric identification result, is that we do not allow ν_{ij} and $\nu_{ij'}$ to be correlated with each other nor do we allow for random coefficients on the acceptance functions.[33]

We will use the measure of excess occupancy presented in section 4 as the choice-set shifter, z_{ij}. As noted earlier, our model will be consistent with a micro-foundation of selective admissions practices due to Gandhi (2021). However, the model can acommodate other unspecified reasons why a facility may not be in a patient's choice-set via error terms in the

[33] We found specifications that included such correlations to be difficult to estimate and unstable in our empirical application. This problem did not exist in Monte Carlo simulations. It is possible that the issue may be specific to our empirical setting.

specification of π_{ij}.[34] Decomposing specific reasons for a facility not belonging to patient i's choice set requires additional structure and is therefore beyond the scope of this paper.

The parametric assumptions on the error terms allow us to use a Gibbs sampler for estimation because, under conjugate prior distributions, the conditional distributions of any of the latent error terms and random coefficients given the others can be obtained in closed form. Moreover, the conditional distributions of each of the parameters $(\alpha, \beta, \Sigma, \delta, \gamma)$ given the errors, random coefficients, and the other parameters can be obtained in closed form. The procedure iterates through each of these parameters, obtaining draws from their conditional posteriors to obtain a Markov Chain of draws of $(\alpha, \beta, \Sigma, \delta, \gamma)$. The draws of the chain converge to the posterior distribution, which is asymptotically equivalent to the maximum likelihood estimator (see van der Vaart, 2000, Theorem 10.1 (Bernstein-von-Mises)). Thus, the mean of the chains' draws yields our point estimate and the covariance of the draws consistently estimates the asymptotic covariance. We check for convergence by ensuring that the number of effective draws is large, the potential scale reduction factor is close to 1, and by visually inspecting the chains.

The key modification from McCulloch and Rossi (1994) involves a data augmentation step in order to avoid calculating the likelihood of choices for each possible latent choice set. Given our model, the likelihood of consumer (henceforth patient) i matching with product (henceforth facility) j is equal to the probability of the event that $\pi_{ij} \geq z_{ij}$, $v_{ij} \geq 0$ and that for all $j' \in J_t$, either $\pi_{ij} < z_{ij}$ or $v_{ij} \geq v_{ij'}$. That is, facility j admits patient i, patient i finds facility j acceptable, and every other facility in the market satisfies at least one of two conditions: either it does not admit i or i prefers j to it. To the best of our knowledge, closed-form solutions for this probability are not known. However, the problem is standard and tractable once we condition on either the vector $\pi_i = (\pi_{i1}, \ldots, \pi_{iJ})$ or $u_i = (u_{i1}, \ldots, u_{iJ})$. This is because π_i determines the latent choice set, making the remaining problem a standard discrete choice problem. And, conditional on u_i, i matches with j if and only if $\pi_{ij} > 0$ and $\pi_{ij'} < 0$ for all j' with $u_{ij'} > u_{ij'}$. This set of π_i is a standard orthant. Thus, our sampler will iterate between data augmentation steps for π_i and u_i. Further details on the Gibbs sampler are provided in appendix C.

Our approach differs from most of the literature on estimating models with latent choice sets, which typically simulates latent choice sets and choice probabilites (Honka, 2014; Honka et al., 2017; Gandhi, 2021; Barseghyan et al., 2021a). Even so, simulating the likelihood without

[34] For example, our model can accomodate patient steering by physicians through choice-set formation because π_{ij} may include factors that physicians value but are irrelevant for patient choices. Strictly speaking, however, this model differs from the one in Gaynor et al. (2016) because we require that capacity constraints z_{ij} are marginal for choice-set formation.

introducing simulation bias (Train, 2009) may be computationally demanding in markets with many possible options. For similar reasons related to the dimensionality of the possible choice sets, Barseghyan et al. (2021a) also utilize an estimation procedure that avoids simulating all the latent choice sets, integrating instead over the distribution of preference parameters and evaluating the probabilities of latent choice sets consistent with the observed data.

This estimation procedure yields estimates of the pair of product-specific fixed effects δ_j and γ_j. The questions that we pursue in the empirical example will not require us to estimate the effects of changes in observable product characteristics. Therefore, we do not further develop these fixed effects in terms of product observables x_j and unobservables ξ_j, e.g. $\delta_j = x_j\beta_X + \xi_j$. In such instances, a researcher may be concerned about potential endogeneity of x_j. This endogeneity does not bias the estimates of δ_j. In fact, it is possible to consistently estimate β_X in a second-step if instruments that are mean independent of ξ_j are available. This two-step approach has been used in a number of prior papers estimating demand with micromodels that do not incorporate choice-set constraints (see Goolsbee and Petrin, 2004; Chintagunta and Dubé, 2005; Train and Winston, 2007, for example).

We conducted Monte Carlo exercises to assess the performance of our estimator, and also to study the consequences of estimating a model that mis-specifies prefences or the choice-set formation process. Specifically, we consider variations that omit random coefficients, incorrectly assume that choice sets are unconstrained, or includes z_{ij} in the utility function as a naive correction for constrained choices. As expected, the resulting bias on the remaining parameter estimates is substantial. Perhaps more importantly, the mis-specifications discussed above translate to biases in economic quantities of interest such as the diversion ratios that we consider in further detail in Section 5.2.4 below. The results from these exercises are discussed in Appendix D.

5.2 Estimates

We start by describing and comparing the estimates from various specifications before turning to a discussion of potential biases in section 5.2.3 and implications on diversion ratios in section 5.2.4.

5.2.1 Empirical Specifications

In all the specifications we consider, the unconstrained choice set for each patient is the set of facilities within a 50 mile radius of their home zip-code centroid. The patient's utility for the inside versus the outside option depends on whether the patient has part-time or full-time

employment as it may affect preferences for in-center versus home dialysis, and whether the patient is eligible for Medicare when she begins dialysis. The variable y_{ij} is the distance between the centroid of the patient's zip code and the facility. We specify $g(\cdot)$ as a quadratic function with the coefficient on the linear term normalized to 1. The slope is allowed to depend on employment status and on the population density of the county where the patient lives. The variable z_{ij} is the excess occupancy of facility j when patient i begins dialysis. Fixed effects are included for each facility.

We compare estimates from three specifications. The first specification, which is our preferred specification, models both preferences and acceptance policies (equations 6 and 7). Patient characteristics that affect acceptance policies include whether or not a patient is Medicare eligible when she begins dialysis, bins of body-mass-index, age, diabetic status and hypertension. We also include patient-specific random coefficients for chain and non-chain facilities in the preferences equation. Facility fixed effects are included in both the preferences and acceptance policy equations.

The second specification, which we refer to as the unconstrained demand model, modifies the preferred specification by omitting capacity constraints and setting $\sigma_{ij} = 1$ for all i and j in equation (7). This specification serves as a comparison of the methods in this paper to a standard approach which does not account for latent choice constraints.

Finally, the third specification, which we refer to as the naive model, modifies the second specification by adding a term $\beta_z z_{ij}$ in equation (6), where β_z is to be estimated. There are two interpretations of this specification. The first is that patients do not face choice constraints, but dislike facilities with high values of z_{ij} (if β_z is negative). Since this interpretation does away with capacity constraints, access to desirable facilities is not influenced by supply-side rationing. This implication may not be a good description for ours and several other markets. The second interpretation is that the specification represents a reduced-form approach that corrects for latent choice set constraints. Section 5.2.4 discusses an undesirable feature of this latter interpretation.

5.2.2 Parameter Estimates

Table 5 presents the estimates from the three specifications. As expected, the estimates indicate that the marginal disutility of distance is decreasing with distance. This and several other estimates are robust across specifications. Consistent with the descriptive evidence in section 4.4, the coefficient on excess occupancy in the naive specification is negative.

There are some notable differences between our preferred specification and the rest. First,

Table 5: Parameter Estimates

	Preferred Specification		Unconstrained	Naïve Model
	(1)		(2)	(3)
	Acceptance	Utility	Utility	Utility
hain	6.637	5.345	2.391	2.405
	(5.054)	(0.780)	(0.829)	(0.822)
o Chain	2.207	5.379	2.028	2.048
	(6.076)	(0.840)	(0.879)	(0.872)
iabetes	0.570	0.724	0.809	0.821
	(2.121)	(0.148)	(0.138)	(0.141)
ypertension	-5.873	-0.244	-0.501	-0.502
	(2.708)	(0.201)	(0.191)	(0.194)
MI<20	-4.232	-0.038	-0.232	-0.230
	(1.880)	(0.254)	(0.265)	(0.268)
5<=BMI<30	1.038	-0.367	-0.360	-0.357
	(1.239)	(0.161)	(0.170)	(0.173)
0<=BMI	5.996	0.024	0.266	0.268
	(1.398)	(0.161)	(0.169)	(0.172)
ge	0.470	0.001	0.001	0.001
	(0.198)	(0.000)	(0.000)	(0.000)
ge squared	-0.001	-1.275	-1.399	-1.409
	(0.002)	(0.188)	(0.192)	(0.198)
edicare	-2.411	0.008	0.034	0.035
	(1.656)	(0.026)	(0.027)	(0.027)
edicare Advantage	26.929	-2.652	-1.923	-1.937
	(3.169)	(0.213)	(0.218)	(0.226)
edicare waiting period	8.618	2.183	2.772	2.797
	(1.690)	(0.225)	(0.242)	(0.244)
mployed		-5.304	-5.764	-5.802
		(0.230)	(0.262)	(0.269)
mployed x distance		0.004	0.006	0.006
		(0.007)	(0.006)	(0.006)
opulation density x distance		0.000	0.001	0.001
		(0.000)	(0.001)	(0.001)
istance squared		0.013	0.013	0.013
		(0.000)	(0.000)	(0.000)
xcess Occupancy				-0.042
				(0.003)
tandard deviation of δ_j		2.574	2.414	2.402
		(0.100)	(0.094)	(0.095)
tandard deviation of ε_{i0}		8.715	9.550	9.663
		(0.264)	(0.327)	(0.360)
tandard deviation of ε_{ij}		4.274	4.799	4.796
		(0.042)	(0.028)	(0.028)
tandard deviation of γ_j	38.799			
	(3.950)			
tandard deviation of random coef on Chain		1.827	2.350	2.358
		(0.264)	(0.219)	(0.217)
tandard deviation of random coef on No Chain		0.688	0.704	0.697
		(0.223)	(0.247)	(0.240)
tandard deviation of v_{ij}	37.398			
	(3.314)			
orrelation between ε_{ij} and v_{ij}	-0.118			
	(0.036)			

Notes: All specifications include distance with a coefficient normalized to -1 in the utility equation. Specification (1) includes “excess occupancy" in the acceptance equation. Specific intercepts for Chain and No Chain facilities obviate the need of a constant term. Standard errors in parentheses.

the mean utility of chain and non-chain facilities (at a distance of zero) is higher in our preferred specification than the other specifications. This reflects the idea that some patients in our specification prefer one of the inside option facilities but are forced to an outside option because of capacity constraints at the inside options. Second, the standard deviations of the facility mean utilities, the outside option utility ε_{i0}, and preference shocks ε_{ij} are lower in our preferred specification than the others. This is expected because a model with unconstrained demand would attribute latent choice constraints to unobserved preference heterogeneity, requiring larger shocks in order to rationalize the observed data.

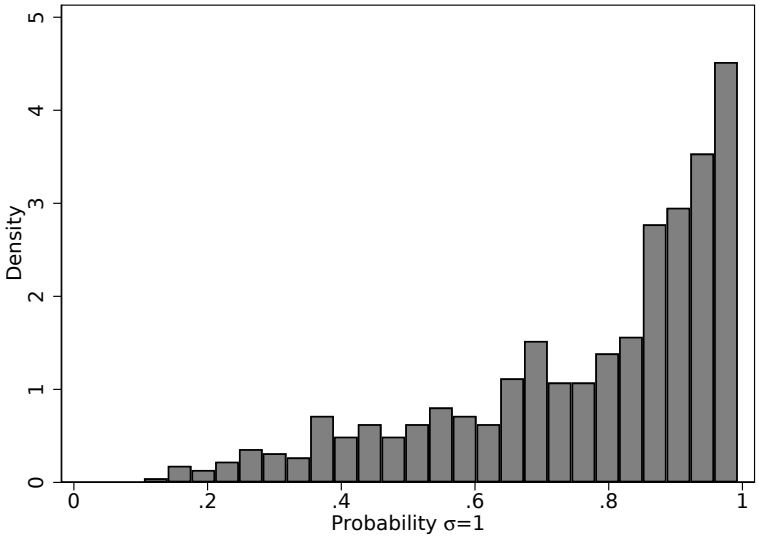
Turning to the acceptance policy function, we find that measures of patient health conditions and insurance status are correlated with acceptance. The propensity of facilities to accept patients increases with the patient's BMI and whether the patient is insured by Medicare Advantage or a private insurer, and therefore, is in the waiting period. Figure 5 shows the estimated distribution of acceptance probabilities for each facility, averaged over all patients for whom the facility is in the patient's choice set. The probability of acceptance is calculated based on the excess occupancy at the facility on the date when the patient begins dialysis. That is, the probability that $\sigma_{ijt} = 1$ is calculated using time-varying characteristics z_{ij} as relevant for patient i. Our results indicate that while the acceptance probabilities are close to 1 for a significant portion of facilities, there are a large number of facilities where the average acceptance probability is much lower than 1. Thus, constraints on choices due to supply-side rationing are non-trivial.

5.2.3 Biases in demand estimates

The capacity constraints estimated above imply a bias in estimated demand using standard approaches. In particular, estimates of demand based on observed market share have the property that, within a market, the product with the highest market share provides the highest indirect utility to the average consumer. Figure 6 shows the estimated relationship between (the log of) market shares and the estimated mean utility (in miles) for our preferred and unconstrained specficiations. The relationship between these two quantities is positive in both specifications, but steeper in the unconstrained specification.[35] This difference occurs both because constraints at desirable facilities can force patients to choose less desirable ones and because the inflow of patients at more desirable facilities can be limited. Therefore, an

[35]Even in the unconstrained specification, we observe dispersion around the central relationship between market shares and mean utility because patient heterogeneity, both in choice sets and in characteristics. For example, not all patients have the same distance to each facility. A strictly monotonic relationship holds in the unconstrained model only conditional on consumer observable characteristics and choice sets (see Berry et al., 2013).

Figure 5: Acceptance Probabilities



analyst who ignores latent choice set constraints may incorrectly deduce a higher desirability for facilities with greater inflows of patients.

The biased relationship between market shares and utility reflects into a bias in the estimated demand for a facility. One way in which demand estimates are biased is that the number of patients for which the facility is the patients' first choice is misestimated. The unconstrained specification equates demand – at fixed values of y and z – to the observed market shares. Figure 7(a) compares the latent demand estimated using the preferred and the unconstrained specifications. It shows that the latent demand for some facilities is higher for some and lower for others. The former bias is clear, as a desirable facility may have to turn away some patients for whom the facility is their first choice. The latter bias occurs because these patients then start treatment at a different facility, increasing the numbers of patients that start treatment there. The results from the naive correction are similar, suggesting that the correction does little to reduce this bias.

Another way to illustrate this bias in demand is to evaluate the estimated willingness to travel for various dialysis facilities. Figure 7(b) compares the average estimated willingness to travel – as compared to taking the outside option – from specifications 1 and 2. Since the proportion of patients for whom a facility is their first choice is a monotonic function of the mean utility (Berry et al., 2013), this figure reflects the same biases as in Figure 7(a).

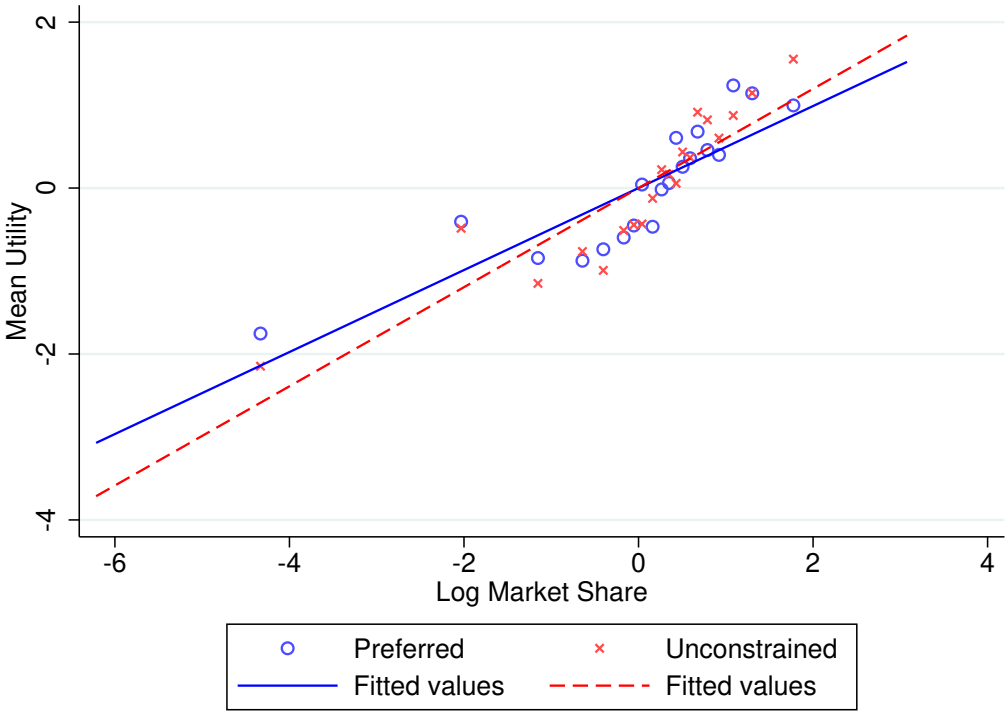
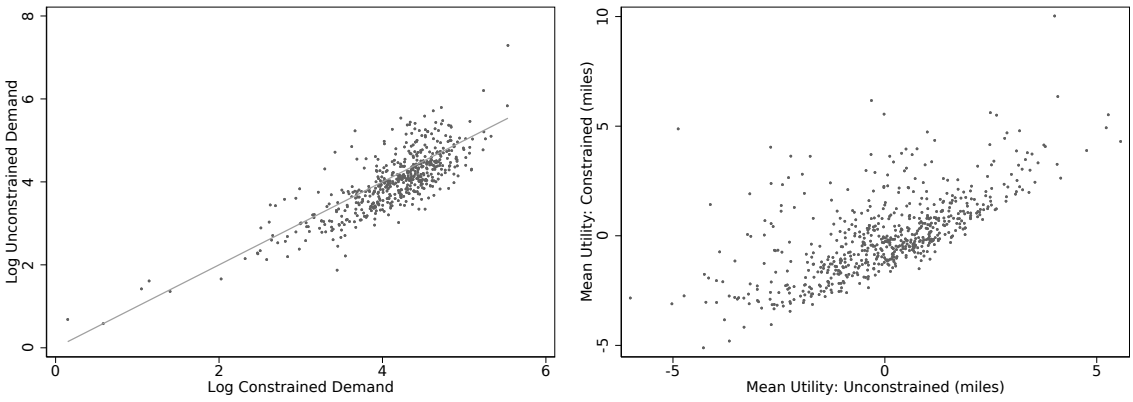


Figure 6: Willingness to Travel and Market Shares



(a) Effects of constraints on demand (b) Comparing estimated mean utilities

Figure 7: Bias in Demand

As before, latent choice constraints feed into biased estimates of demand.[36] Thus, similar sources of bias affect estimates of patient welfare or the desirability of various facilities.

Our model and analysis suggest that the relationship between shares and desirability that is commonly used to estimate demand is suspect in the presence of supply-side rationing, not only in the dialysis industry but also in other settings where market shares may be driven by capacity instead of quality. Because demand often plays a central role in empirical studies, potential biases in demand estimates can propagate into final conclusions.

5.2.4 Implications of choice constraints on diversion ratios

We close this section by noting that there can also be economic grounds on which naive corrections for latent choice constraints are unappealing. We illustrate this point by showing that the naive specification (of the form in specification 3) restricts the comparison between diversion ratios arising from demand-side factors and acceptance decisions.

Specifically, let $s_j(z_i, y_i)$ be the market share of product j, where d_i and t have been dropped from the notation for simplicity, and $z_i = (z_{i1}, \ldots, z_{iJ})$ and $y_i = (y_{i1}, \ldots, y_{iJ})$. The diversion ratio of j with respect to k, in principle depends on whether j loses a customer because of changes in choice constraints, equivalently z, or changes in preferences, equivalently y. The two diversion ratios are

$$\frac{\partial s_k}{\partial z_{ij}} \Big/ \frac{\partial s_j}{\partial z_{ij}} \quad \text{and} \quad \frac{\partial s_k}{\partial y_{ij}} \Big/ \frac{\partial s_j}{\partial y_{ij}}.$$

In our empirical specification, the latter diversion ratio is equivalent to the diversion ratio obtained based on changes in mean utility δ_j.

Notice that there are no a priori reasons why these two diversion ratios need to be the same. To see this, observe that following a marginal change in y_{ij}, product j loses customers that are indifferent between j and another good. The consumers that switch between k and j following a change in y_{ij} are consumers that (i) are indifferent between j and k, (ii) have both j and k in their choice sets, and (iii) do not have any other more preferable options in their choice set. Contrast this with consumers that switch between these two products following a change in z_{ij}. These consumers (i) strictly prefer j to k, (ii) are on the margin of being accepted by j, and (iii) do not have any other more preferable options in their choice set. Notice that the first two requirements select consumers on different dimensions – on the preference margin following changes in y_{ij} and on the acceptance margin following changes in z_{ij}. Thus, the diversion ratios on these two margins may be different.

[36] Figure E.2 in the Appendix homes in on this point by showing the difference in estimated mean utility for facility j and the probability that $\sigma_{ijt} = 1$ for facility j.

Figure 8: Diversion Ratios

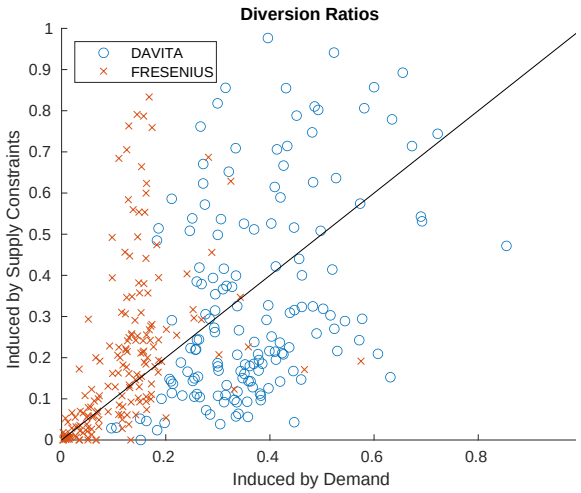


Figure 8 compares these two types of diversion ratios using our preferred specification. Each point in the figure represents an independent facility j in California, where we sum the diversion ratios over all facilities k that are either run by DaVita or Fresenius, the two largest dialysis chains in the US. As can be seen, these two diversion ratios are substantially different across the two margins. The diversion with respect to demand factors is usually higher than diversion with respect to factors affecting supply constraints, with larger diversion with respect to demand for DaVita than for Fresenius. These differences speak to whether competitive incentives to strategically choose capacity or quality are more predominant in the market.

In contrast to the differences measured here, naive corrections can directly restrict these two diversion ratios to be identical, even under flexible functional forms. Consider the generalized version of specification 3 in which we assume that $\sigma_{ij} = 1$ for all i and j, and we set

$$v_{ij} = u_j(\omega_i) - g(z_{ij}, y_{ij}).$$

Assume that $\omega_i \perp z_{ij}$, $u(\omega_i) = (u_1(\omega_i), \ldots, u_J(\omega_i))$ admits a density, and $g_j(\cdot)$ is differentiable with respect to the first two arguments. Thus, the observed market share for product j is $s_j(z_{ij}, y_{ij}) = P(v_{ij} > v_{ij'} | z_{ij}, y_{ij})$. And, notice that $\frac{\partial s_l}{\partial z_{ij}} / g_z(z_{ij}, y_{ij}) = \frac{\partial s_l}{\partial y_{ij}} / g_y(z_{ij}, y_{ij})$ for $l \in \{j, k\}$. Therefore, $\frac{\partial s_k}{\partial z_{ij}} / \frac{\partial s_j}{\partial z_{ij}} = \frac{\partial s_k}{\partial y_{ij}} / \frac{\partial s_j}{\partial y_{ij}}$ and all the points on figure 8 would be restricted to lie on the 45-degree line.[37] Restrictions, such as this one, can have important implications

[37] An alternative approach for obtaining differences in diversion ratios between factors affecting supply

and go beyond the biases in estimated quantities described above.[38]

6 Conclusion

Consumers often face restricted choice sets for reasons other than monetary budget constraints. Examples include information or search frictions, preferences of the other side in two-sided matching markets, and selective admission practices. These constraints are usually unobserved to the analyst. We developed a unified model for analyzing discrete choice demand in the presence of latent constraints on choice sets that encompasses many of the models discussed earlier.

We show how to point identify the joint distribution of preferences and latent choice constraints in the presence of two sets of observable shifters, one that influences preferences and the other that influences choice sets. Each set of shifters must be excluded from the other side of the model. Relative to the prior literature, our approach achieves point identification while placing minimal restrictions on functional forms, on the statistical dependence between choice sets and preferences, and allows for the endogeneity of product characteristics. The cost is that our results require access to the shifters mentioned above. However, we show that our results are sharp in the sense that additional restrictions on the model are necessary for identification if either set of shifters are not available.

As an illustrative example, we estimate the demand for hemodialysis facilities. The data shows clear evidence of supply-side rationing – facilities with a higher than usual occupancy are less likely to admit new patients, and patients that begin dialysis when nearby centers are constraints are observed to travel further away. Next, we use patient enrollment outcomes to estimate a joint model of preferences and supply-side rationing using a Gibbs sampler. Our results show that ignoring supply-side constraints when present can lead to significant bias in estimates and yield misleading answers to important economic quantities.

Our approach stops at specifying a reduced-form for the supply-side acceptance decision. This reduced form immediately yields a structural object in certain models, such as in empirical models of two-sided matching (Agarwal, 2015; He et al., 2024). The reduced-form yields a first-stage estimate in models with more complex supply-side behavior. For example, Gandhi

constraints and demand constraints would be to introduce random coefficients that interact with some of these factors, but not others. While it is plausible that such preference heterogeneity is present, it is less clear whether differing competitive incentives for choosing capacity and quality are solely intermediated through demand instead of also through capacity constraints.

[38] The Monte Carlo exercises discussed in Appendix D also show the bias arising from this functional form on diversion ratios.

(2021) interprets acceptance probabilities as conditional choice probabilities (Hotz and Miller, 1993) when estimating a dynamic model of selective admission practices. Fleshing out this link between the reduced-form model that we identify and a structural model of acceptance policies is left for future research, but it is important for evaluating some counterfactuals that involve changes in equilibrium supply-side behavior.

References

Abaluck, Jason and Abi Adams-Prassl, "What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses," *The Quarterly Journal of Economics*, 2021, pp. 1611–1663.

_ **and Giovanni Compiani**, "A Method to Estimate Discrete Choice Models that is Robust to Consumer Search," *Working Paper*, 2020.

Agarwal, Nikhil, "An Empirical Model of the Medical Match," *American Economic Review*, 2015, *105*, 1939–78.

_ **and Paulo Somaini**, "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism," *Econometrica*, 2018, *86*, 391–444.

Alba, Joseph W, J Wesley Hutchinson, and John G Lynch, *Memory and Decision Making*, Prentice-Hall,

Allen, Roy and John Rehbeck, "Identification With Additively Separable Heterogeneity," *Econometrica*, 2019, *87*, 1021–1054.

Allende, Claudia, "Competition Under Social Interactions and the Design of Education Policies," *Working Paper*, 2019.

Azevedo, Eduardo M. and Jacob Leshno, "A Supply and Demand Framework for Two-Sided Matching Markets," *Journal of Political Economy*, 2016, *124*, 1235–1268.

Bai, Jushan and Pierre Perron, "Computation and analysis of multiple structural change models," *Journal of Applied Econometrics*, 2003, *18*, 1–22.

Barseghyan, Levon, "Identification and Inference for Pure Random Coefficients Models with Limited Consideration," *Working Paper*, 2022.

_ **, Francesca Molinari, and Matthew Thirkettle**, "Discrete Choice under Risk with Limited Consideration," *American Economic Review*, 2021, *111*, 1972–2006.

_ **,** _ **, Maura Coughlin, and Joshua C. Teitelbaum**, "Heterogeneous Choice Sets and Preferences," *Econometrica*, 2021, *89*, 2015–2048.

Berry, Steven and Ariel Pakes, "The Pure Characteristics Demand Model," *International Economic Review*, 2007, *48*, 1193–1225.

_ **and Philip Haile**, "Nonparametric Identification of Differentiated Products Demand Using Micro Data," *Working Paper*, 8 2020.

Berry, Steven T., "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 1994, *25*, 262.

_ **, Amit Gandhi, and Philip A. Haile**, "Connected Substitutes and Invertibility of Demand," *Econometrica*, 2013, *81*, 2087–2111.

Berry, Steven T and Philip A Haile, "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 2014, *82*, 1749–1797.

Berry, Steven T. and Phillip A. Haile, "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," 2010.

_ **, James Levinsohn, and Ariel Pakes**, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, *63*, 841 – 890.

Billingsley, Patrick, *Probability and Measure*, 3 ed., John Wiley and Sons., 1995.

Block, H and J Marshak, *Random Orderings and Stochastic Theories of Responses*, Stanford University Press,

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff, "Analyzing the determinants of the matching of public school teachers to jobs: Disentangling the preferences of teachers and employers," *Journal of Labor Economics*, 2013, *31*, 83–117.

Butters, Gerard R., "Equilibrium Distributions of Sales and Advertising Prices," *Review of Economic Studies*, 1977, *44*, 465–491.

Cattaneo, Matias D., Richard K. Crump, Max Farrell, and Yingjie Feng, "On Binscatter," 2021.

Chade, Hector and Lones Smith, "Simultaneous Search," *Econometrica*, 2006, *74*, 1293–1307.

Chan, Kevin E., Ravi I. Thadhani, and Franklin W. Maddux, "Adherence Barriers to Chronic Dialysis in the United States," *Journal of the American Society of Nephrology*, 11 2014, *25*, 2642–2648.

Chernozhukov, Victor and Christian Hansen, "An IV Model of Quantile Treatment Effects," *Econometrica*, 2005, *73*, 245–261.

Ching, Andrew T., Fumiko Hayashi, and Hui Wang, "Quantifying the Impact of Limited Supply: The Case of Nursing Homes," *International Economic Review*, 2015, *56*, 1291–1322.

Chintagunta, Pradeep K. and Jean-Pierre Dubé, "Estimating a Stockkeeping-Unit-Level Brand Choice Model that Combines Household Panel Data and Store Data," *Journal of Marketing Research*, 2005.

Compiani, Giovanni, "Market Counterfactuals and the Specification of Multi-Product Demand: A Nonparametric Approach," *Working Paper*, 2021.

Conlon, Christopher T. and Julie Holland Mortimer, "Demand Estimation under Incomplete Product Availability," *American Economic Journal: Microeconomics*, 2013, *5*, 1–30.

Dafny, Leemore S., David Cutler, and Christopher Ody, "How Does Competition Impact Quality of Care? A Case Study of the U.S. Dialysis Industry," *Working Paper*, 2018.

Dagsvik, John K., "Aggregation in Matching Markets," *International Economic Review*, 2000, *41*, 27–58.

de Palma, André, Nathalie Picard, and Paul Waddell, "Discrete Choice Models with Capacity Constraints: An Empirical Analysis of the Housing Market of the Greater Paris Region," *Journal of Urban Economics*, 2007, *62*, 204–230.

Department of Health and Human Services, "Medicare and Medicaid Programs; Conditions for Coverage for End-Stage Renal Disease Facilities," 2008.

Diamond, W. and N. Agarwal, "Latent indices in assortative matching models," *Quantitative Economics*, 2017, *8*, 685–728.

Dinerstein, Michael and Troy Smith, "Quantifying the Supply Response of Private Schools to Public Policies," *American Economic Review*, 2021, *111*, 3376–3417.

Dubé, Jean-Pierre, Ali Hortaçsu, and Joonhwi Joo, "Random-Coefficients Logit Demand Estimation with Zero-Valued Market Shares," *Marketing Science*, 2021.

Eliason, Paul, "Market Power and Quality: Congestion and Spatial Competition in the Dialysis Industry," *Working Paper*, 2019.

Eliason, Paul J, Benjamin Heebsh, Ryan C McDevitt, and James W Roberts, "How Acquisitions Affect Firm Behavior and Performance: Evidence from the Dialysis Industry," *The Quarterly Journal of Economics*, 2020, *135*, 221–267.

Eliaz, K. and R. Spiegler, "Consideration Sets and Competitive Marketing," *The Review of Economic Studies*, 2011, *78*, 235–262.

Fack, Gabrielle, Julien Grenet, and Yinghua He, "Beyond Truth-Telling: Preference Estimation with Centralized School Choice and College Admissions," *American Economic Review*, 2019, *109*, 1486–1529.

Gandhi, Ashvin, "Picking Your Patients: Selective Admissions in the Nursing Home Industry," *Working Paper*, 2021.

Gaynor, Martin, Carol Propper, and Stephan Seiler, "Free to Choose? Reform, Choice, and Consideration Sets in the English National Health Service," *American Economic Review*, 2016, *106*, 3521–3557.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, *Bayesian Data Analysis*, 3 ed., CRC Press, 2014.

Goeree, Michelle Sovinsky, "Limited Information and Advertising in the U.S. Personal Computer Industry," *Econometrica*, 2008, *76*, 1017–1074.

Goolsbee, Austan and Amil Petrin, "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 2004.

Grieco, Paul L. E. and Ryan C. McDevitt, "Productivity and Quality in Health Care: Evidence from the Dialysis Industry," *The Review of Economic Studies*, 2017, *84*, 1071–1105.

He, YingHua, Shruti Sinha, and Xiaoting Sun, "Identification and Estimation in Many-to-one Two-sided Matching without Transfers," *Econometrica*, 5 2024, *92*, 749–774.

Heiss, Florian, Daniel McFadden, Joachim Winter, Amelie Wuppermann, and Bo Zhou, "Inattention and Switching Costs as Sources of Inertia in Medicare Part D," *American Economic Review*, 2021, *111*, 2737–2781.

Hickman, William and Julie Holland Mortimer, *Demand Estimation with Availability Variation*, Edward Elgar Publishing Ltd.,

Ho, Kate, Joseph Hogan, and Fiona Scott Morton, "The impact of consumer inattention on insurer pricing in the Medicare Part D program," *The RAND Journal of Economics*, 2017, *48*, 877–905.

Honka, Elisabeth, "Quantifying search and switching costs in the US auto insurance industry," *The RAND Journal of Economics*, 2014, *45*, 847–884.

_ , Ali Hortaçsu, and Maria Ana Vitorino, "Advertising, Consumer Awareness, and Choice: Evidence from the U.S. Banking Industry," *The RAND Journal of Economics*, 2017, *48*, 611–646.

Hortaçsu, Ali, Seyed Ali Madanizadeh, and Steven L. Puller, "Power to Choose? An Analysis of Consumer Inertia in the Residential Electricity Market," *American Economic Journal: Economic Policy*, 2017, *9*, 192–226.

Hotz, V Joseph and Robert A Miller, "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 1993, *60*, 497–529.

Kepler, John, Valeri V. Nikolaev, Nicholas Scott-Hearn, and Christopher R. Stewart, "Quality Transparency and Healthcare Competition," *SSRN Electronic Journal*, 2021.

Lee, Anne, Claire Gudex, Johan V. Povlsen, Birgitte Bonnevie, and Camilla P. Nielsen, "Patients' views regarding choice of dialysis modality," *Nephrology Dialysis Transplantation*, 2008, *23*, 3953–3959.

Lewbel, Arthur, "Endogenous Selection or Treatment Model Estimation," *Journal of Econometrics*, 2007, *141*, 777–806.

Liu, Jian, Shiying Wu, and James V. Zidek, "On Segmented Multivariate Regression," *Statistica Sinica*, 1997, *7*, 497–525.

Logan, John Allen, Peter D Hoff, and Michael A Newton, "Two-Sided Estimation of Mate Preferences for Similarities in Age, Education, and Religion," *Journal of the American Statistical Association*, 2008, *103*, 559–569.

Manski, Charles F., "The structure of random utility models," *Theory and Decision*, 1977, *8*, 229–254.

Matzkin, Rosa L., "Nonparametric identification and estimation of polychotomous choice models," *Journal of Econometrics*, 1993, *58*, 137–168.

_ , *Nonparametric identification*, Vol. 6B, Elsevier,

McCulloch, Robert and Peter E Rossi, "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 1994, *64*, 207–240.

McFadden, Daniel, *Econometric Models of Probabilistic Choice*, The MIT Press, 1981.

Menzel, Konrad, "Large Matching Markets As Two-Sided Demand Systems," *Econometrica*, 2015, *83*, 897–941.

_ **and Tobias Salz**, "Robust Decisions For Incomplete Structural Models Of Social Interactions," *Working Paper*, 2013.

Neilson, Christopher, "Targeted Vouchers, Competition Among Schools, and the Academic Achievement of Poor Students," *Working Paper*, 2020.

Newey, Whitney K. and James L. Powell, "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 2003, *71*, 1565–1578.

Petrin, Amil, "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 2002, *110*, 705–729.

Roberts, John H. and James M. Lattin, "Development and Testing of a Model of Consideration Set Composition," *Journal of Marketing Research*, 1991, *28*, 440.

Roth, Alvin E. and Marilda A. Oliveira Sotomayor, *Two-Sided Matching: : A Study in Game Theoretic Modeling and Analysis*, Cambridge University Press, 1990.

Swait, Joffre and Moshe Ben-Akiva, "Incorporating random constraints in discrete models of choice set generation," *Transportation Research Part B: Methodological*, 1987, *21*, 91–102.

Train, Kenneth E., *Discrete Choice Methods with Simulation*, Cambridge University Press, 2009.

_ **and Clifford Winston**, "Vehicle Choice Behavior and the Declining Market Share of U.S. Automakers," *International Economic Review*, 2007.

U. S. Renal Data System, "2020 USRDS Annual Data Report: Epidemiology of kidney disease in the United States — End Stage Renal Disease," 2020.

van der Vaart, A. W., *Asymptotic Statistics*, Cambridge University Press, 2000.

Weitzman, Martin L., "Optimal Search for the Best Alternative," *Econometrica*, 1979, *47*, 641–654.

Wollmann, Thomas, "How to Get Away with Merger: Stealth Consolidation and Its Real Effects on US Healthcare," *Working Paper*, 2022.

Appendix

A Proofs

A.1 Proof of Lemma 1

Because ties are allowed, it must be that

$$s_{jt}\left(d_i, y_i, z_i\right) \leq \sum_{O \in \mathcal{O}} P\left(O_i = O, j \in \arg\max_{k \in O} v_{ikt} \middle| t, d_i, y_i, z_i\right)$$

The inequality follows because $c_{ij} = 1$ only if $j \in \arg\max_{k \in O_i} v_{ikt}$. Conditioning on d_i and dropping it from the notation, we rewrite preferences as

$$v_{ij} = u_j\left(\omega_i\right) - g_{ij}$$

and we treat g_{ij} as observable. Consumer i remains unmatched if for every facility $j \in O_i$ $u_j\left(\omega_i\right) < g_{ij}$ and only if for every facility $j \in O_i$ $u_j\left(\omega_i\right) \leq g_{ij}$. Similarly, facility $j \in O_i$ if $\pi_j\left(\omega_i\right) < z_j$ and only if $\pi_j\left(\omega_i\right) \leq z_j$. Let $s_0\left(g, z\right)$ be the share of consumers that are unmatched conditional on g and z, define $\bar{s}_0\left(\bar{g}, \bar{z}\right)$ as $\lim_{(g,z)\downarrow(\bar{g},\bar{z})} s_0\left(g, z\right)$, where $(g, z) \downarrow (\bar{g}, \bar{z})$ if there exists a sequence $g_n > \bar{g}$ and $z_n > \bar{z}$ with $g_n \to \bar{g}$ and $z_n \to \bar{z}$. If $s_0\left(g, z\right)$ is continuous at $(\bar{g}, \bar{z})$, $\bar{s}_0\left(g, z\right) = s_0\left(g, z\right)$; otherwise, $\bar{s}_0\left(g, z\right) > s_0\left(g, z\right)$. By assumption (1) and by set inclusion,

$$\begin{aligned}
\lim_{(g,z)\downarrow(\bar{g},\bar{z})} s_0\left(g, z\right) &\geq \lim_{(g,z)\downarrow(\bar{g},\bar{z})} P\left(\cap_j \left\{u_j\left(\omega_i\right) < g_j \vee \pi_j\left(\omega_i\right) < z_j\right\}\right) \\
&\geq P\left(\cap_j \left\{u_j\left(\omega_i\right) \leq \bar{g}_j \vee \pi_j\left(\omega_i\right) \leq \bar{z}_j\right\}\right).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\lim_{(g,z)\downarrow(\bar{g},\bar{z})} s_0\left(g, z\right) &\leq \lim_{(g,z)\downarrow(\bar{g},\bar{z})} P\left(\cap_j \left\{u_j\left(\omega_i\right) \leq g_j \vee \pi_j\left(\omega_i\right) \leq z_j\right\}\right) \\
&= P\left(\cap_j \left\{u_j\left(\omega_i\right) \leq \bar{g}_j \vee \pi_j\left(\omega_i\right) \leq \bar{z}_j\right\}\right),
\end{aligned}$$

where the inequality follows from set inclusion and the equality follows because the probability of a sequence of nested events converges to the probability of the limiting event (Billingsley, 1995, Theorem 2.1). Thus,

$$\bar{s}_0\left(\bar{g}, \bar{z}\right) = \lim_{(g,z)\downarrow(\bar{g},\bar{z})} s_0\left(g, z\right) = P\left(\cap_j \left\{u_j\left(\omega_i\right) \leq \bar{g}_j \vee \pi_j\left(\omega_i\right) \leq \bar{z}_j\right\}\right).$$

Let $\mathcal{B}_\chi$ be the collection of sets that are a Cartesian product of half-open intervals of the form $B = \{(u,\pi) : \underline{u} < u \leq \bar{u}, \underline{\pi} < \pi \leq \bar{\pi}\}$ with $B \subseteq \chi$. Consider some $B \in \mathcal{B}_\chi$ and let $\underline{g} = \underline{u}$, $\overline{g} = \overline{u}$,$\underline{z} = \underline{\pi}$ and $\overline{z} = \overline{\pi}$. Define g^j such that $g_k^j = \bar{g}_k$ for $j = k$ and $g_k^j = \underline{g}_k$ for $j \neq k$. Likewise, define $\bar{z}^j$ such that $z_k^j = 1\{j = k\}\bar{z}_k + 1\{j \neq k\}\underline{z}_k$. Define:

$$\Lambda_1(g,z) \equiv \left[\bar{s}_0\left(g^1,z\right) - \bar{s}_0(g,z)\right] - \left[\bar{s}_0\left(g^1,z^1\right) - \bar{s}_0\left(g,z^1\right)\right],$$

and for $j > 1$,

$$\Lambda_j(g,z) \equiv \left[\Lambda_{j-1}\left(g^j,z\right) - \Lambda_{j-1}(g,z)\right] - \left[\Lambda_{j-1}\left(g^j,z^j\right) - \Lambda_{j-1}\left(g,z^j\right)\right].$$

Observe that each $\Lambda_j\left(\underline{g},\underline{z}\right)$ is identified. We will now calculate $\Lambda_J(\underline{g},\underline{z})$. To do this, observe that $\bar{s}_0(g^1,\underline{z}) - \bar{s}_0\left(\underline{g},\underline{z}\right)$ is equal to

$$P\left(\left\{\underline{g}_1 < u_1(\omega_i) \leq \bar{g}_1 \wedge \pi_1(\omega_i) > \underline{z}_1\right\} \cap_{k>1} \left\{u_k(\omega_i) \leq \underline{g}_k \vee \pi_k(\omega_i) \leq \underline{z}_k\right\}\right).$$

Similarly, $\bar{s}_0(g^1,z^1) - \bar{s}_0\left(\underline{g},z^1\right)$ equals

$$P\left(\left\{\underline{g}_1 < u_1(\omega_i) \leq \bar{g}_1 \wedge \pi_1(\omega_i) > \overline{z}_1\right\} \cap_{k>1} \left\{u_k(\omega_i) \leq \underline{g}_k \vee \pi_k(\omega_i) \leq \underline{z}_k\right\}\right).$$

By set inclusion, the probability

$$P\left(\left\{\underline{g}_1 < u_j(\omega_i) \leq \bar{g}_1 \wedge \underline{z}_1 < \pi(\omega_i) \leq \bar{z}_1\right\} \cap_{k>1} \left\{u_k(\omega_i) \leq \underline{g}_k \vee \pi_k(\omega_i) \leq \underline{z}_k\right\}\right)$$

is equal to $\Lambda_1\left(\underline{g},\underline{z}\right)$. By an identical argument and induction, for any $j > 1$, we have that $\Lambda_j\left(\underline{g},\underline{z}\right)$ equals

$$P\left(\cap_{k\leq j}\left\{\underline{g}_j < u_j(\omega_i) \leq \bar{g}_j \wedge \bar{z}_j < \pi(\omega_i) \leq \bar{z}_j\right\} \cap_{k>j} \left\{u_j(\omega_i) \leq \underline{g}_k \vee \pi(\omega_i) \leq \underline{z}_k\right\}\right).$$

In particular,

$$\begin{aligned}\Lambda_J\left(\underline{g},\underline{z}\right) &= P\left(\cap_j\left\{\underline{g}_j < u_j(\omega_i) \leq \bar{g}_j \wedge \bar{z}_j < \pi(\omega_i) \leq \bar{z}_j\right\}\right) \\ &= P\left((u(\omega_i),\pi(\omega_i)) \in B\right).\end{aligned}$$

Thus, we can identify the probability that $(u(\omega_i),\pi(\omega_i))$ belongs to any set $B \in \mathcal{B}_\chi$, i.e., sets that are a Cartesian product of half-open intervals and are subsets of the interior of the support of (g,z).

We will show that conditional cumulative distribution function of (u_i, π_i) given $(u_i, \pi_i) \in \chi$, $P\left(u_i \leq \bar{u}, \pi_i \leq \bar{\pi} \middle| (u_i, \pi_i) \in \chi\right)$, is identified. There are two cases. The first case is when $P\left((u_i, \pi_i) \in \chi\right) > 0$. Then, we have that

$$P\left(u_i \leq \bar{u}, \pi_i \leq \bar{\pi} \middle| (u_i, \pi_i) \in \chi\right) = P\left((u_i, \pi_i) \in \bar{B} \cap \chi\right) / P\left((u_i, \pi_i) \in \chi\right)$$

where $\bar{B} = \{(u, \pi) : u \leq \bar{u}, \pi \leq \bar{\pi}\}$. It would suffice to show that we can identify $P\left((u_i, \pi_i) \in \bar{B} \cap \chi\right)$ and $P\left((u_i, \pi_i) \in \chi\right)$. In the second case, $P\left((u_i, \pi_i) \in \chi\right) = 0$. In this case, we will still be able to identify $P\left((u_i, \pi_i) \in \chi\right)$, but notice that the statement is vacuous and thus completes the proof.

To identify $P\left((u_i, \pi_i) \in \chi\right)$, we will show that $\chi = \bigcup_{k=1}^{\infty} B'_k$ for a countable collection of $B'_k \in \mathcal{B}_\chi$ and $B'_k \cap B'_{k'} = \emptyset$. This would imply that $P\left((u_i, \pi_i) \in \chi\right) = \sum_{k=1}^{\infty} P\left((u_i, \pi_i) \in B'_k\right)$ is identified since each term in the summand is identified. Towards this, we first show that there exists a countable collection of half-open cartesian products of intervals $B_k = \{(u, \pi) : \underline{u}_k < u \leq \bar{u}_k, \underline{\pi}_k < \pi \leq \bar{\pi}_k\} \in \mathcal{B}_\chi$ such that $\chi = \bigcup_{k=1}^{\infty} B_k$. To do this, let $x \in \chi$ and note that there exist vectors of rational numbers $\underline{u}_k, \bar{u}_k, \underline{\pi}_k$ and $\bar{\pi}_k$ such that

$$x \in B_k = \{(u, \pi) : \underline{u}_k < u \leq \bar{u}_k, \underline{\pi}_k < \pi \leq \bar{\pi}_k\}$$

and $B_k \subseteq \chi$. Since the set of rational numbers is countable, we have that there exists a countable collection of B_k with $\chi = \bigcup_{k=1}^{\infty} B_k$ and $B_k \subseteq \chi$. Now, notice that for any two elements of this collection B_k and $B_{k'}$, $B_k \cap B_{k'} \in \mathcal{B}_\chi$. And, $B_k \backslash B_{k'}$ is a union of at most $2^{2J} - 1$ sets in $\mathcal{B}_\chi$. Therefore, there exists an at most a countable number of disjoint sets $B'_k \in \mathcal{B}_\chi$ such that $\bigcup_k B'_k = \bigcup_k B_k = \chi$. Hence, $P\left((u_i, \pi_i) \in \chi\right)$ is identified.

Next, we show that $P\left((u_i, \pi_i) \in \bar{B} \cap \chi\right)$ is identified. Notice that $\bar{B} \cap \chi = \bigcup_k \left(\bar{B} \cap B'_k\right)$. Since $\bar{B} \cap B'_k \in \mathcal{B}_\chi$, the quantity $P\left((u_i, \pi_i) \in \bar{B} \cap B'_k\right)$ is identified. Since $B'_k \cap B'_{k'} = \emptyset$, we have that $P\left((u_i, \pi_i) \in \bar{B} \cap \chi\right) = \sum_k P\left((u_i, \pi_i) \in \bar{B} \cap B'_k\right)$ is identified. Hence, the conditional cumulative distribution function of (u_i, π_i) conditional on $(u_i, \pi_i) \in \chi$ is identified.

A.2 Primitive Conditions for Assumption 3

Condition on d_i and drop it from the notation for simplicity. Fix $\{j, k\}$. For each y_i, define the set

$$U_{jk}(y_i, O_i) = \left\{ u(\omega_i) : \min_{l \in \{j,k\}} \{u_l(\omega) - g_l(y_{il})\} \geq \max_{l \in O_i \backslash \{j,k\}} \{u_l(\omega) - g_l(y_{il})\} \right\}.$$

Definition 2. The pair of goods $\{j,k\}$ is relevant at characteristics (y_i,z_i) and choice set O if

$$P\left(O,u\left(\omega_i\right)\in U_{jk}\left(y_i,O\right)\middle|\, z_i\right)>0.$$

Proposition 2. *Suppose assumption 1 is satisfied. If (i) the pair of goods $\{j,k\}$ is relevant at characteristics (y_i,z_i) and choice set O_i for some $O_i\in\mathcal{O}$, (ii) the distribution of*

$$u_j\left(\omega\right)-u_k\left(\omega\right)$$

conditional on $u\left(\omega\right)\in U_{jk}\left(y_i,O_i\right)$ and O_i admits a density f_{jk}, (iii) $f_{jk}\left(g_j\left(y_{ij}\right)-g_k\left(y_{ik}\right)\right)>0$, and (iv) for each O and all y in a neighborhood of y_i, $P\left(\left|\arg\max_{j\in O}\left\{u_j\left(\omega\right)-g_j\left(y_{ij}\right)\right\}\right|>1\middle|\, O,y\right)=0$ then (i) $g_j\left(y_{ij}\right)$ is differentiable if and only if $s_k\left(y_i,z_i\right)$ is differentiable with respect to y_{ij}, (ii) the sign of $\frac{\partial s_k\left(y_i,z_i\right)}{\partial y_{ij}}$ coincides with the sign of $\frac{\partial g_j\left(y_{ij}\right)}{\partial y_{ij}}$ provided that these derivatives exist, and (iii) a symmetric relation exists between $g_k\left(y_{ik}\right)$ and $s_j\left(y_i,z_i\right)$. Consequently, j and k are strict substitutes if and only if $g_j\left(y_{ij}\right)$ and $g_k\left(y_{ik}\right)$ are differentiable with non-zero derivatives.

Proof. Fix specific values of y_i and z_i. Observe that

$$\begin{aligned}
s_j\left(y_i,z_i\right)&=\sum_{O\in\mathcal{O}}P\left(c_{ij}=1\middle|\, O,y_i,z_i\right)P\left(O\middle|\, y_i,z_i\right)\\
&=\sum_{O\in\mathcal{O}}P\left(j\in\arg\max_{l\in O}u_l\left(\omega\right)-g_l\left(y_{il}\right)\middle|\, O,z_i\right)P\left(O\middle|\, z_i\right)
\end{aligned}$$

since requirement (iv) implies that $\arg\max_{l\in O}\left\{u_l\left(\omega\right)-g_l\left(y_{il}\right)\right\}$ has at most one element with probability 1 and assumption 1 allow us to drop the conditioning on y_i. Equation 3 implies that

$$\begin{aligned}
&\frac{\partial s_j\left(y_i,z_i\right)}{\partial y_{ik}}\\
=&\sum_{O\in\mathcal{O}}\frac{\partial P\left(j\in\arg\max_{l\in O}u_l\left(\omega\right)-g_l\left(y_{il}\right)\middle|\, O,z_i\right)}{\partial y_{ik}}P\left(O\middle|\, z_i\right)\\
=&\sum_{O\in\mathcal{O}}\left.\frac{\partial P\left(u_j\left(\omega_i\right)-\bar{g}_{ij}\geq u_k\left(\omega_i\right)-\bar{g}_{ik}\middle|\, O,u\left(\omega_i\right)\in U_{jk}\left(y_i,O\right),z_i\right)}{\partial g_{ik}}\right|_{\bar{g}_{ik}=g_k\left(y_{ik}\right)}\\
&\frac{\partial g_k\left(y_{ik}\right)}{\partial y_{ik}}P\left(O,u\left(\omega_i\right)\in U_{jk}\left(y_i,O\right)\middle|\, z_i\right)\\
=&\frac{\partial g_k\left(y_{ik}\right)}{\partial y_{ik}}\sum_{O\in\mathcal{O}}\frac{\partial\int_{g_{ij}-g_{ik}}^{\infty}f_{jk}\left(v\right)dv}{\partial g_{ik}}P\left(O,u\left(\omega_i\right)\in U_{jk}\left(y_i,O\right)\middle|\, z_i\right)\\
=&\frac{\partial g_k\left(y_{ik}\right)}{\partial y_{ik}}\sum_{O\in\mathcal{O}}f_{jk}\left(g_{ij}-g_{ik}\right)P\left(O,u\left(\omega_i\right)\in U_{jk}\left(y_i,O\right)\middle|\, z_i\right)
\end{aligned}$$

where the derivatives in the summands exist since f_{jk} is a density. The hypotheses ensure the existence of $O_i \in \mathcal{O}$ such that its corresponding summand is strictly positive. Thus, if $g_k(y_{ik})$ is differentiable, $\frac{\partial s_j(y_i, z_i)}{\partial y_{ik}}$ exists and it has the same sign as $\frac{\partial g_k(y_{ik})}{\partial y_{ik}}$. Conversely, if $g_k(y_{ik})$ is not differentiable, the limit $\frac{g_k(y_{ik}) - g_k(y_{ik}+\Delta)}{\Delta}$ as $\Delta \to 0$ does not exist; thus, $\frac{\partial s_j(y_i, z_i)}{\partial y_{ik}}$ does not exist. This completes the proof of parts (i) and (ii). Part (iii) follow immediately from a symmetric argument. □

Corollary 2. *Suppose assumption 1 is satisfied. If there exists $z_i^* \in Z$ such that (i) $\cup_{O:\{j,k\}\subseteq O} P(O|z_i^*) > 0$, and (ii) for each each O with $\{j,k\} \subseteq O$ and $P(O|z_i^*) > 0$, the joint distribution of $(u_{ij})_{j\in O}$ conditional O on has full support on an open neighborhood $B \subseteq \mathbb{R}^{|O|}$ of $(g_j(y_{ij}))_{j\in O}$ and is absolutely continuous with respect to Lebesgue measure on B, then the functions $s_j(y_i, z_i^*)$ and $s_k(y_i, z_i^*)$ are differentiable at y_{ik} and y_{ij} respectively with non-zero derivatives if and only if $g_j(y_{ij})$ and $g_k(y_{ik})$ are differentiable at y_{ij} and y_{ik} with non-zero derivatives.*

As another corollary, we state stronger but simpler to interpret conditions.

Corollary 3. *Suppose assumption 1 is satisfied. If the joint distribution of u_i conditional on each O admits a density conditional on each O and there exists O with $\{j,k\} \subseteq O$ and $P(O|z_i^*) > 0$ for some z_i^*, then the functions $s_j(y_i, z_i^*)$ and $s_k(y_i, z_i^*)$ are strictly increasing and differentiable at y_{ik} and y_{ij} respectively if and only if $g_j(y_{ij})$ and $g_k(y_{ik})$ are strictly increasing and differentiable at y_{ij} and y_{ik}.*

A.3 Proof of Lemma 2

The proof of lemma 2 requires the following intermediate result.

Lemma 3. *Suppose that assumption 1 holds and $|J| > 1$. If j and k are strict substitutes in y at some (d_i, y_i, z_i^*) in the support of the data and $g_j'(d_i, y_i) \neq 0$, then (i) $g_k'(d_i, y_i) \neq 0$, (ii) the sign of $g_k'(d_i, y_i)$ coincides with the sign of $\frac{\partial s_k(d_i, y_i, z_i^*)}{\partial y_{ij}}$, and (iii)*

$$\frac{g_k'(d_i, y_i)}{g_j'(d_i, y_i)} = \frac{\partial s_j(d_i, y_i, z_i^*)}{\partial y_{ik}} \Big/ \frac{\partial s_k(d_i, y_i, z_i^*)}{\partial y_{ij}},$$

which implies that $\frac{g_k'(d_i, y_i)}{g_j'(d_i, y_i)}$ is identified and bounded.

Proof. Because j and k are strict substitutes in y at (d_i, y_i, z_i^*) $\frac{\partial s_j(y_i, d_i, z_i^*)}{\partial y_{ik}}$ and $\frac{\partial s_k(y_i, d_i, z_i^*)}{\partial y_{ij}}$ exist and are non-zero. For notational simplicity, we omit z_i^*, d_i, y_l and g_l for $l \notin \{j,k\}$ from the notation as they are fixed throughout the proof.

Since $P\left(c_{ij}=1|\, O_i, t, d_i, y_i=y, z_i\right) = P\left(c_{ij}=1|\, O_i, t, d_i, y_i=y', z_i\right)$ if $g\left(y\right)=g\left(y'\right)$, equation (3) and Assumption 1 implies that there exists a function $\hat{s}\left(\cdot\right):\mathbb{R}^2\to\mathbb{R}^2$ such that

$$s_k\left(y_{ik}, y_{ij}\right) = \hat{s}_k\left(g_k\left(y_{ik}\right), g_j\left(y_{ij}\right)\right).$$

Moreover, the function $\hat{s}_k\left(g_k, g_j\right)$ is weakly increasing in g_j and weakly decreasing in g_k.

The proof consists of four steps. The first step shows that the function $\hat{s}_k\left(g_k, g_j\right)$ is differentiable with respect to g_j at $g_k = g_k\left(y_{ik}\right)$ and $g_j = g_j\left(y_{ij}\right)$. Therefore, we can use the chain rule to calculate the cross partials of $s_k\left(y_{ik}, y_{ij}\right)$ and $s_j\left(y_{ik}, y_{ij}\right)$. The second step proves part (i): the derivative of $g_k\left(\cdot\right)$ at y_{ik} is not zero. The third step shows symmetry of the cross-partial derivatives $\frac{\partial \hat{s}_j\left(g_k, g_j\right)}{\partial g_k} = \frac{\partial \hat{s}_k\left(g_k, g_j\right)}{\partial g_j}$ without requiring continuity of $\frac{\partial \hat{s}_j\left(g_k, g_j\right)}{\partial g_j}$ and $\frac{\partial \hat{s}_k\left(g_k, g_j\right)}{\partial g_k}$, a key requirement for Young's Theorem. The fourth and final step applies the chain rule and employs the symmetry of cross-partial derivatives to derive parts (ii) and (iii).

First step: For any $\Delta \neq 0$,

$$\frac{\hat{s}_k\left(g_k, g_j\left(y_{ij}+\Delta\right)\right) - \hat{s}_k\left(g_k, g_j\left(y_{ij}\right)\right)}{g_j\left(y_{ij}+\Delta\right) - g_j\left(y_{ij}\right)} = \frac{s_k\left(y_{ik}, y_{ij}+\Delta\right) - s_k\left(y_{ik}, y_{ij}\right)}{\Delta} \Big/ \frac{g_j\left(y_{ij}+\Delta\right) - g_j\left(y_{ij}\right)}{\Delta}.$$

The limit of the right-hand side as $\Delta\to 0$ exists because $\frac{\partial s_k\left(y_{ik}, y_{ij}\right)}{\partial y_{ij}}$ and $\frac{\partial g_j\left(y_{ij}\right)}{\partial y_{ij}}$ exist, and the latter is non-zero. Thus, the limit on the left hand side as $\Delta\to 0$ also exists and it is finite. Moreover, $\frac{\partial s_k\left(y_{ik}, y_{ij}\right)}{\partial y_{ij}} \neq 0$, and weak monotonicity of $\hat{s}_k\left(g_k, g_j\right)$ with respect to g_j implies that

$$\frac{\partial \hat{s}_k\left(g_k, g_j\left(y_{ij}\right)\right)}{\partial g_j} = \frac{\partial s_k\left(y_{ik}, y_{ij}\right)}{\partial y_{ij}} \Big/ \frac{\partial g_j\left(y_{ij}\right)}{\partial y_{ij}} > 0, \tag{8}$$

where the strict inequality follows because each term in the RHS is non-zero. By a symmetric argument, $\frac{\partial g_k\left(y_{ik}\right)}{\partial y_{ik}} \neq 0$ implies that $\frac{\partial \hat{s}_j\left(g_k, g_j\right)}{\partial g_k}$ exists at $g_k = g_k\left(y_{ik}\right)$ and $g_j = g_j\left(y_{ij}\right)$. We will show in the second step below that $\frac{\partial g_k\left(y_{ik}\right)}{\partial y_{ik}} \neq 0$ without assuming that $\frac{\partial \hat{s}_j\left(g_k, g_j\right)}{\partial g_k}$ exists.

Second step: Consider $\Delta > 0$. The difference $\hat{s}_j\left(g_k+\Delta, g_j\right) - \hat{s}_j\left(g_k, g_j\right)$ is equal to the mass of consumers whose match switches from k to j due to an increase in g_k. Since the distribution of choice sets O is independent of y, and therefore $g\left(y\right)$, the switchers have both k and j in their choice sets at both $\left(g_k+\Delta, g_j\right)$ and $\left(g_k, g_j\right)$. These consumers would switch to j if, instead of g_k increasing by Δ, g_j decreased by the same amount. Thus, by set inclusion,

$$0 \leq \hat{s}_j\left(g_k+\Delta, g_j\right) - \hat{s}_j\left(g_k, g_j\right) \leq \hat{s}_k\left(g_k, g_j\right) - \hat{s}_k\left(g_k, g_j-\Delta\right). \tag{9}$$

By definition of $\frac{\partial s_j(y_{ij},y_{ik})}{\partial y_{ik}}$,

$$\frac{\partial s_j(y_{ij},y_{ik})}{\partial y_{ik}} = \lim_{\Delta\downarrow 0}\left(\frac{\hat{s}_j(g_k(y_{ik}+\Delta),g_j)-\hat{s}_j(g_k(y_{ik}),g_j)}{g_k(y_{ik}+\Delta)-g_k(y_{ik})}\times\frac{g_k(y_{ik}+\Delta)-g_k(y_{ik})}{\Delta}\right)\neq 0. \tag{10}$$

The limit on the right-hand side exists because $\frac{\partial s_j(y_{ij},y_{ik})}{\partial y_{ik}}$ is well-defined. Taking the absolute value of the terms in parenthesis and using the inequalities in (9) yields

$$\begin{aligned}&\lim_{\Delta\downarrow 0}\left|\frac{\hat{s}_j(g_k(y_{ik}+\Delta),g_j)-\hat{s}_j(g_k(y_{ik}),g_j)}{g_k(y_{ik}+\Delta)-g_k(y_{ik})}\times\frac{g_k(y_{ik}+\Delta)-g_k(y_{ik})}{\Delta}\right|\\ &\leq\lim_{\Delta\downarrow 0}\frac{\hat{s}_k(g_k(y_{ik}),g_j)-\hat{s}_k\left(g_k(y_{ik}),g_j-\tilde{\Delta}\right)}{\tilde{\Delta}}\times\left|\frac{g_k(y_{ik}+\Delta)-g_k(y_{ik})}{\Delta}\right|\end{aligned}$$

where $\tilde{\Delta}=g_k(y_{ik}+\Delta)-g_k(y_{ik})$. Both terms converge as $\Delta\downarrow 0$: the first one converges to $\frac{\partial\hat{s}_k(g_k,g_j)}{\partial g_j}$ and the second one to the absolute value of $\frac{\partial g_k(y_{ik})}{\partial y_{ik}}$. Therefore, $\frac{\partial g_k(y_{ik})}{\partial y_{ik}}\neq 0$ because otherwise, $\frac{\partial s_j(y_{ij},y_{ik})}{\partial y_{ik}}=0$ contradicting equation (10). This proves part (i).

Third step: The arguments above imply that $\frac{\partial\hat{s}_j(g_k,g_j)}{\partial g_k}$ exists at $g_k=g_k(y_{ik})$ and $g_j=g_j(y_{ij})$. By a symmetric argument to the one that yields equation (9), for any $\Delta>0$

$$0\leq\hat{s}_k(g_k,g_j+\Delta)-\hat{s}_k(g_k,g_j)\leq\hat{s}_j(g_k,g_j)-\hat{s}_j(g_k-\Delta,g_j). \tag{11}$$

Dividing (9) and (11) by Δ and taking the limit $\Delta\downarrow 0$, yields:

$$0<\frac{\partial\hat{s}_j(g_k,g_j)}{\partial g_k}=\frac{\partial\hat{s}_k(g_k,g_j)}{\partial g_j}. \tag{12}$$

Fourth step: We have shown that $\hat{s}_k(g_k,g_j)$ is differentiable with respect to g_j and that $\hat{s}_j(g_k,g_j)$ is differentiable with respect to g_k. Applying the chain rule yields:

$$\frac{\partial s_j(y_{ij},y_{ik})}{\partial y_{ik}}=\frac{\partial\hat{s}_j(g_k(y_{ik}),g_j)}{\partial g_k}\times\frac{\partial g_k(y_{ik})}{\partial y_{ik}} \tag{13}$$

and

$$\frac{\partial s_k(y_{ij},y_{ik})}{\partial y_{ij}}=\frac{\partial\hat{s}_k(g_k,g_j(y_{ij}))}{\partial g_j}\times\frac{\partial g_j(y_{ij})}{\partial y_{ij}}. \tag{14}$$

Parts (ii) and (iii) follow immediately from equations (12), (13), and (14). □

We are now ready to prove lemma 2. Fix d_i and omit it from notation. Let j be the reference good and recall the normalization that $\left|g_j'(y_0)\right|=1$ and $g_j(y_0)=0$ for some y_0. Take

any pair (k, y_k) such that there is a path connecting it with (j, y_0) where j is the reference good and y_0 is the value for which we have normalized $\left|\frac{\partial g_j(d_i, y_0)}{\partial y}\right| = 1$. Let this path be $(j, y_0) = (m_0, y_1), (m_2, y_2), ..., (m_n, y_n) = (k, y_k)$ where for all $l = 2, ..., n$, m_l and m_{l-1} are strict substitutes in y at some (d_i, y_i, z_i^*) in the support of the data with $y_{im_l} = y_l$ and $y_{im_{l-1}} = y_{l-1}$. Lemma 3 implies that $\frac{g'_{m_l}(y_l)}{g'_{m_{l-1}}(y_{l-1})}$ is identified for each $l \in \{2, ...n\}$. Moreover, $g'_{m_l}(y_l)$ and $g'_{m_{l-1}}(y_{l-1})$ are bounded and non-zero. Thus, $g'_k(y_k) = \frac{g'_k(y_k)}{g'_j(y_0)} = \prod_{l=2}^{n} \frac{g'_{m_l}(y_l)}{g'_{m_{l-1}}(y_{l-1})}$ is identified. Since $g_k(y_0) = 0$ and $g_k(\cdot)$ is continuously differentiable, $g_k(y_k) = \int_{y_0}^{y_k} g'_k(\tau)\, d\tau$ is identified as the argument above and assumption 3 imply that $g'_k(\tau)$ is identified for almost all τ in the support of y_{ik}. [39]

A.4 Proof of Proposition 1

To simplify notation, we drop the conditioning on d_i. Since the function $g(\cdot)$ is known, in a minor abuse of notation we write $g = g(y)$ and $s(g) = \{s_j(g)\}_{j \in J}$. We also drop z_i from the notation because its support is a singleton. With this simplification, the function $s_j(g)$ can be re-written as follows:

$$\begin{aligned}
s_j(g) &= \sum_{O \in \mathcal{O}} P\left(O, j \in \arg\max_{k \in O} u_k - g_k \middle| g\right) \\
&= \sum_{O \in \mathcal{O}} \int 1\left\{j \in \arg\max_{k \in O} u_k - g_k\right\} P(O|u, g)\, f_U(u)\, du \\
&= \sum_{O \in \mathcal{O}} \int 1\left\{j \in \arg\max_{k \in O} u_k - g_k\right\} P(O|u)\, f_U(u)\, du \\
&= \sum_{O \in \mathcal{O}} \int 1\left\{j \in \arg\max_{k \in O} u_k - g_k\right\} \left(\int_{O^c} P(O|u)\, f_U(u)\, du_{O^c}\right) du_O \\
&= \sum_{O \in \mathcal{O}} \int_{g_j}^{\infty} \left(\int_{-\infty}^{u_j - g_j + g_k} ... \int_{-\infty}^{u_j - g_j + g_{k'}} h_O(u_O)\, du_{O - \{j\}}\right) du_j,
\end{aligned}$$

where $O^c = J \backslash O$, $u_O = (u_j)_{j \in O}$, $u_{O^c} = (u_j)_{j \in J \backslash O}$ and $h_O(u_O) = \int P(O|u)\, f_U(u)\, du_{O^c}$. The third equality follows from assumption 1 whereas the others simply re-write the problem. Since $s(g)$ is the only observable when the support of z is a singleton, under assumption 1, identification of the model is equivalent to identification of $P(O|u)$ and $f_U(u)$.

We use a standard definition of identification (Matzkin, 2007). Define a model as a collection of admissible structures $\{P(\cdot|\cdot), f_U(\cdot)\}$. A pair of structures is observationally equivalent if they yield the same observable market share functions $s(\cdot)$. In particular, since the func-

[39] Footnote 15 of HSS refers to a previous version of our paper that employed a more restrictive version of assumption 3.

tions $\{h_O(\cdot)\}_{O\in\mathcal{O}}$ determine the functions $s_j(g)$, two structures that yield the same functions $h_O(\cdot)$ are also observationally equivalent. Thus, the function $f_U(\cdot)$ is identified if and only if for any pair of observationally equivalent admissible structures $\{P(\cdot|\cdot), f_U(\cdot)\}$ and $\left\{\tilde{P}(O|\cdot), \tilde{f}_U(\cdot)\right\}$, $f_U(\cdot) = \tilde{f}_U(\cdot)$.

To complete the proof of the proposition, define admissible structures as pairs $\{P(\cdot|\cdot), f_U(\cdot)\}$ such that (i) $f_U(u)$ is a density, (ii) $0 < \tilde{P}(O|u) < 1$ for all $O \in \mathcal{O}$ and all $u \in \mathbb{R}^{|J|}$, and (iii) the choice set probabilities add to one for each u: $\sum_{O\in\mathcal{O}} \tilde{P}(O|u) = 1$. The first conditions follow from the assumptions in the proposition. The second and third conditions ensure that $P(O|u)$ is a proper probability for any pair (O, u). The distribution of indirect utilities is not identified if there are two observationally equivalent admissible structures $\{P(\cdot|\cdot), f_U(\cdot)\}$ and $\left\{\tilde{P}(O|\cdot), \tilde{f}_U(\cdot)\right\}$ with $f_U(\cdot) \neq \tilde{f}_U(\cdot)$. The following lemma shows that this is the case under the hypothesis of the proposition.

Lemma 4. *If for the admissible structure $\{P(\cdot|\cdot), f_U(\cdot)\}$ there exists an open set $B \subset \mathbb{R}^{|J|}$ and a choice set $O \subsetneq J$ such that for all $u \in B$, $f_U(u) > 0$ and $P(O|u) > \kappa > 0$, then there exist an alternative admissible structure $\left\{\tilde{P}(\cdot|\cdot), \tilde{f}_U(\cdot)\right\}$ with $f_U(\cdot) \neq \tilde{f}_U(\cdot)$ and for all u_O,*

$$h_O(u_O) = \int P(O|u) f_U(u) \, \mathrm{d}u_{O^c} = \int \tilde{P}(O|u) \tilde{f}_U(u) \, \mathrm{d}u_{O^c}.$$

Proof. Fix an open set $U \subset \mathbb{R}^{|J|}$, a choice set $O \subsetneq J$ such that for all $u \in U$, $f_U(\boldsymbol{u}) > 0$ and $P(O|\boldsymbol{u}) > \kappa > 0$. These quantities exist by assumption. Let $R = \prod_{j\in\mathcal{J}} \left[\underline{u}_j, \bar{u}_j\right] \subset U$ be a closed cartesian product of $|J|$ intervals, one for each good. Define an arbitrary absolutely continuous function $c(u_{O^c})$ such that (i) $c(u_{O^c}) \neq 0$, (ii) $\|c(u_{O^c})\|_\infty < \frac{\kappa}{2}$, (iii) $c(u_{O^c}) = 0$ for $u_{O^c} \notin R_{O^c}$, where $R_{O^c} = \prod_{j\in O^c} \left[\underline{u}_j, \bar{u}_j\right]$ denotes the product of the intervals in R corresponding to the products in O^c.

Define a family of functions $\{a_{O'}(u)\}_{O'\in\mathcal{O}}$ as follows. Let $a_{O'}(u) = 0$ for $O' \neq O$ and

$$a_O(u) = 1\{u \in R\} \left[c(u_{O^c}) - \frac{\int_{R_{O^c}} c(u_{O^c}) f_U(u) \, \mathrm{d}u_{O^c}}{\int_{R_{O^c}} f_U(u) \, \mathrm{d}u_{O^c}}\right].$$

Note that each $\|a_O(u)\| < \kappa$, and that

$$\int a_O(u) f(u) \, du_{O^c} = \int_{R_{O^c}} \left[c(u_{O^c}) - \frac{\int_{R_{O^c}} c(u_{O^c}) f_U(u) \, du_{O^c}}{\int_{R_{O^c}} f_U(u) \, du_{O^c}}\right] f(u) \, du_{O^c} = 0.$$

Moreover, for every $O' \subset O$

$$\int a_O(u) f(u) du_{O'^c} = \int \int a_O(u) f(u) du_{O^c} du_{O \setminus O'} = 0.$$

Define the alternative structure as

$$\tilde{f}(u) = (1 - a_O(u)) f(u)$$
$$\tilde{P}(O'|u) = \frac{P(O'|u) - a_{O'}(u)}{1 - a_O(u)}$$

for every $O' \in \mathcal{O}$. Now we verify that $\left\{\tilde{P}(\cdot|\cdot), \tilde{f}(\cdot)\right\}$ is an admissible structure. First, $\tilde{f}(u)$ is a density because $(1 - a_O(u)) f(u) \geq 0$ and

$$\int (1 - a_O(u)) f(u) du = 1 - \int_O \int_{O^c} a_O(u) f(u) du_{O^c} du_O = 1.$$

Second, the choice set probabilities satisfy $0 < \tilde{P}(O'|u) < 1$ for all $O' \in \mathcal{O}$.Third, the choice set probabilities add to one for each u:

$$\sum_{O' \in \mathcal{O}} \tilde{P}(O'|u) = \frac{\sum_{O' \in \mathcal{O}} P(O'|u) - a_O(u)}{1 - a_O(u)} = 1.$$

Now we verify that the alternative structure is observationally equivalent to the original one. Note that $\int_{O'^c} \tilde{P}(O'|u) \tilde{f}(u) du_{O'^c} = \int_{O'^c} P(O'|u) f(u) du_{O'^c} = h_{O'}(u_{O'})$ for all $O' \neq O$. And, finally

$$\begin{aligned}
\int_{O^c} \tilde{P}(O|u) \tilde{f}(u) du_{O^c} &= \int_{O^c} (P(O|u) - a_O(u)) f(u) du_{O^c} \\
&= \int_{O^c} P(O|u) f(u) du_{O^c} - \int_{O^c} a_O(u) f(u) du_{O^c} \\
&= h_O(u_O).
\end{aligned}$$

□

A.5 Identification across Markets

We show results analogous to those in Proposition 2 for non-separable models. These results follow Theorem 2 in Berry and Haile (2010). Let

$$\delta_{jt} = \tilde{u}_j(x_{jt}, \xi_{jt}) \equiv \text{med}\ (u_{ijt}|\, x_{jt}, \xi_{jt}),$$

and let $f_{\delta_j}(\cdot|\, x_{jt}, r_{jt})$ be the conditional density of δ_j, where r_{jt} are a set of instruments. Fix $\varepsilon_\tau > 0$ and $\varepsilon_f > 0$, small. For $\tau \in (0,1)$, let $\mathcal{L}_j(\tau)$ be the convex hull of functions $m_j(\cdot,\tau)$ such that for all r_{jt}, $P(\delta_{jt} \leq m_j(x_{jt},\tau)|\, r_{jt}) \in [\tau - \varepsilon_\tau, \tau + \varepsilon_\tau]$, and for all x_{jt}, $m_j(x_{jt},\tau) \in s_j(x_{jt}) \equiv \left\{\delta : f_{\delta_j}(\delta|\, x_{jt}, r) \geq \varepsilon_f,\ \forall r \text{ with } f_X(x_{jt}|\, r) > 0\right\}$.

Assumption 8. $\xi_{jt} \perp r_{jt}$

Assumption 9. *For all j and $\tau \in (0,1)$, (i) for any bounded function $B_j(x,\tau) = m_j(x,\tau) - \tilde{u}_j(x,\tau)$ with $m_j(\cdot,\tau) \in \mathcal{L}_j(\tau)$ and $\varepsilon_{jt} \equiv \delta_{jt} - \tilde{u}_j(x_{jt},\tau)$, $E[B_j(x_{jt},\tau)\,\psi_j(x_{jt},r_{jt},\tau)|\, r_{jt}] = 0$ a.s. only if $B_j(x_{jt},\tau) = 0$ a.s. for $\psi_j(x,r,\tau) = \int_0^1 f_{\varepsilon_j}(\sigma B_j(x,\tau)|\, x, r)\,\mathrm{d}\sigma > 0$. (ii) the density $f_{\varepsilon_j}(e|\, x, w)$ of ε_{jt} is continuous and bounded for all $e \in \mathbb{R}$, and (iii) $\tilde{u}_j(x_{jt},\tau) \subset s_j(x_{jt})$ for all x_{jt}.*

Proposition 3. *(Berry and Haile, 2014; Chernozhukov and Hansen, 2005). If δ_{jt} is identified and assumptions 8 and 9 are satisfied, then the functions $\tilde{u}(\cdot)$ and ξ_{jt} are identified for each j and t.*

Proof. Follows from theorem 4 in Chernozhukov and Hansen (2005) since δ_{jt} is identified. □

An analogous results holds for identification of $\tilde{g}_j$ since

$$g_{jt} = \tilde{g}_j(x_{jt}, \zeta_{jt})$$

is known. Here, we switch g_{jt} for δ_{jt} and $\tilde{g}_j(\cdot)$ for $\tilde{u}_j(\cdot)$.

B Data Appendix

The data reported here have been supplied by the United States Renal Data System (USRDS) and the Centers for Medicare & Medicaid Services (CMS). These sources provide us with data on all dialysis facilities and the near universe of kidney patients in the US. Patient characteristics include the residence zip-code, co-morbidities and the facility that they attend. For each facility, we observe their address, ownership status and the number of stations. Patients and facilities are uniquely identified by a USRDS generated identifier that can be used to link records across separate datasets. We geocode patient zip-codes and facility addresses to calculate the straight line distance between a given facility and a patient's zip-code centroid.

We will retain copies of the data until permitted by our Data Use Agreement with the United States Renal Data System (USRDS). Researchers interested in using our dataset should directly contact USRDS to obtain permission.

B.1 Data Description

Our data on patient profiles and treatment history come from the USRDS Researcher Standard Analysis File (SAF) which combines information from ESRD claims filed to CMS and data from the Consolidated Renal Operations in a Web-Enabled Network System (CROWN), a mandatory data system used by dialysis facilities to collect information on all patients, regardless of payer type. The main SAF datasets used in this analysis are Medical Evidence (medevid), which includes patient health information like co-morbidities and the whether a nephrologist was already caring for a patient when dialysis commenced, Treatment History (rxhist), where we obtain the sequence of facilities in which a patient was treated, Payer History (payhist) for insurance information, Residence History for the residence zip code and the Facility dataset from the USRDS.

Though the patient information is sourced from claims, facility data come from the CMS Annual Facility Survey and the CMS Facility Compare dataset maintained separately by CMS. These includes identifiers for the facility, years of operation, profit status, chain status, and setting status. The facility and patient identifiers allow us to link the patient information from claims and the facility information from Facility Compare, providing a complete overview of the patient-facility interaction.

We also geocoded facility addresses and obtained the geocodes for the centroid of each patient's zip code. These coordinates are used to estimate the distance from the facility to the patient, calculated as the distance from the patients' reported zip code centriod to the facility. Geo-coordinates are obtained via queries sent to the Google Maps API; these queries have as an input the facility addresses included in the Facility Compare dataset provided by CMS and return as an output the associated longitude and latitude for each facility. Zip-code centroids are also obtained using Google Maps.

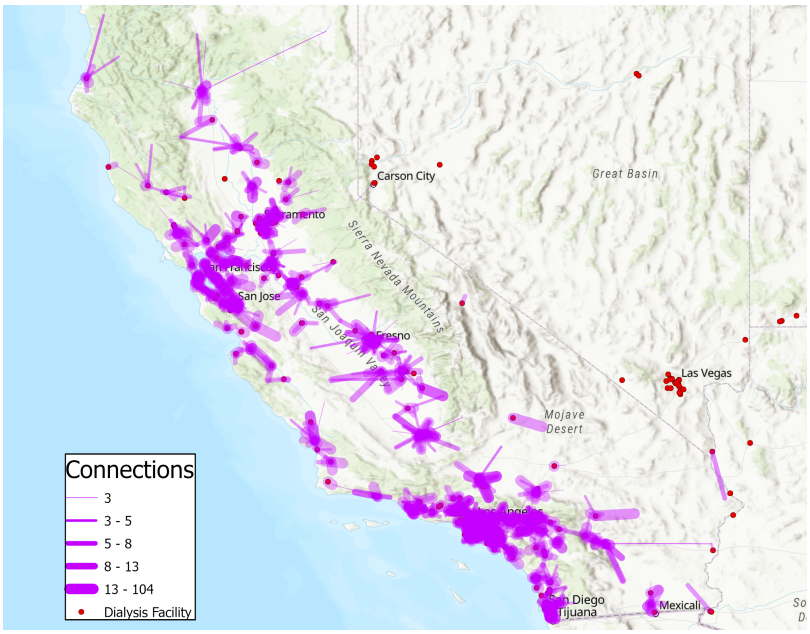
We use the Treatment History files to construct the number of patients receiving care at each facility at a given point in time. This file contains the start date and the end date of each patient's treatment at each facility where they receive care. We use this information to compute the number of patients undergoing in-patient hemodialysis at each facility on each day during our sample period. These calculations will include all patients, irrespective of whether they are in the sample of patients that we use to estimate our model (see section B.2.2 below).

B.2 Sample Selection

We consider first-time admissions in California facilities between Jan 1, 2015 and December 31, 2018. As mentioned in the main text, moving costs and other considerations can be important in subsequent stays, which complicates the analysis. Nonetheless, the first facility a patient chooses is consequential as the median and average patient is treated at 1 and 1.30 facilities respectively.

California is essentially an isolated market, with few outgoing or incoming patient-facility connections across its state borders. Figure B.1 shows the linkages between all facilities in the US and zip-code centroids in California. The thickness of each edge connecting a facility with a zip-code centroid indicates the number of patients residing in a zip-code that started dialysis at a given facility. We omit edges with fewer than three patients. Only in rare instances does a patient living in California attend a facility outside the state. When they do, our approach will treat the patient as choosing the outside option.

Figure B.1: California Connections



B.2.1 Facility Sample Selection

Table B.1 describes the facility sample. All facilities in California during our sample period were successfully geocoded. From this universe of facilities, we restrict attention to facilities that focus on in-center care and are non-pediatrics. Both variables are calculated using the

admissions data for facilities during our sample period; a facility is said to focus on in-facility care if more than 50% of its admitted patients enroll in facility-based hemodialysis. We classify a facility as pediatrics if the average age of the patients they admit is less than or equal to 18. Patients living in California who receive dialysis but do not attend one of these facilities are considered as being treated at a composite outside option.

We restricted to facilities that focus on non-pediatric and in-center care for two reasons. First, we want to focus on the interactions for individuals that are going to facilities to receive treatment, as opposed to receiving home dialysis in which case the distance to the facility is not as salient in the patient's choice of facility. Only a small minority of patients receive home dialysis and are likely selected on health condition and income. Second, we restrict to non-pediatric facilities because the baseline differences in co-morbidities and clinical indications for pediatric and adult dialysis can be substantial, creating significantly different needs and operational setups for pediatric facilities.

We only include the quarters for which the facility operation was relatively stable, excluding periods around entry, exit, capacity changes, or moves as these events could substantially affect a facility's demand and acceptance policies. In particular, we include in the inside option facility-quarters in years with no changes in the number of stations or address. We remove the quarter of and the quarter after a facility entered. Similarly, we remove the quarter before and the quarter of a facility exit.

Table B.1: Facility Sample

Restrictions	Facilities
Restricted to 2015 - 2018 and California	721
Restricted to facilities with geocoordinates	721
Restricted to facilities specializing in facility-based hemodialysis and are non-pediatric	640
Facilities with at least one stable quarter	553

B.2.2 Patient Sample Selection

Table B.2 describes the patient sample. We make three major restrictions on the patient sample, starting from the universe of patients with a residential zip-code in California that started dialysis in the years 2015 - 2018. First, and analogously to the focus on non-pediatric facilities, we keep only adults in our sample, defined as at least 18 years of age when they first started dialysis. Second, we drop patients for whom we weren't able to compute a distance to the facility attended; practically, this means that we drop a handful of patients for whom we did not observe a valid zip-code. These two restrictions together result in a couple hundred

patients being dropped from our sample. The biggest cut in the sample comes from dropping patients that chose facilities greater than 50 miles from their reported zip-code centroid. Based on an inspection of these observations, we suspect that the residential zip-code is incorrectly recorded for these patients. One indication is that the 95th percentile of distance, conditional on the chosen facility being is less than 50 miles away, is less than 20 miles.

Table B.2: Patient Sample

Restriction	Patients
Restricted to 2015 - 2018 and California	53,074
Restricted to adults (>=18 years old)	52,768
Restricted to admissions with distance between patient and facility	52,751
Restricted to those that chose a facility within 50 miles	50,002

B.2.3 Target Capacity

Table B.3 presents estimates of a regression of the estimated target capacity on facility inputs measured annually, controlling for facility fixed effects. The result shows that univariate regressions of facility inputs are positively correlated with target capacity. This includes both capital and labor inputs. The relationship holds even though (i) target capacity varies at a higher frequency level than the recorded inputs and (ii) the inputs are measured only annually.

Table B.3: Correlation Between Target Capacity and Facility Inputs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Total Number of Dialysis Stations	-0.006***								-0.019***
	(0.002)								(0.004)
Late Shift		-0.016							-0.063
		(0.029)							(0.044)
Registered Nurses on staff full-time			0.014						0.001
			(0.009)						(0.021)
Licensed Practical/Visiting Nurses FTime				0.060					0.058
				(0.048)					(0.051)
Patient Care Technicians on staff FTime					0.006				0.005
					(0.006)				(0.017)
Advanced Practice Nurses on staff FTme						0.107			0.096
						(0.131)			(0.135)
Dieticians on staff full-time							0.087		-0.144
							(0.074)		(0.158)
Social Workers on staff full-time								0.215***	0.381***
								(0.062)	(0.140)
Constant	0.093***	-0.049***	-0.135**	-0.081***	-0.110	-0.056***	-0.136*	-0.265***	0.043
	(0.030)	(0.013)	(0.056)	(0.029)	(0.070)	(0.014)	(0.075)	(0.066)	(0.042)
Observations	2,061	2,038	2,061	2,061	2,061	2,061	2,061	2,061	2,038
R-squared	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.002	0.005

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

C Estimation Appendix: Gibbs Sampler

Our sampler starts with an initial guess for the parameters $(\alpha, \beta, \Sigma, \delta, \gamma)$ and for the latent variables $(\beta_i, \varepsilon_{i0}, v_i, \pi_i)$ for every i. We denote this guess by $\theta^{(0)}$. For each draw k, we perform the following steps:

1. Data augmentation:

   (a) Draw the latent acceptance index $\pi_{ij}|\theta^{(k-1)}$ for every i and j in the sample. The posterior distribution of π_{ij} conditional on all the parameters $\theta^{(k-1)}$ is normal. If i was allocated to facility j, then we draw π_{ij} from the conditional posterior truncated by $\pi_{ij} \geq z_{ij}$. If i was allocated to facility $j^* \neq j$ and $v_{ij}^{(k-1)} > v_{ij^*}^{(k-1)}$, then we draw π_{ij} from the conditional posterior truncated by $\pi_{ij} < z_{ij}$. Otherwise,

we draw it from the conditional posterior without any truncation. Let $\pi^{(k)}$ denote the vector of draws and let $O_i^{(k)}$ be $\{j \in J : \pi_{ij} \geq z_{ij}\}$.

(b) Draw the latent utility $v_{ij}|\,\theta^{(k-1)}, \pi^{(k)}$ for every i and j. The posterior distribution of v_{ij} conditional on all the parameters $\theta^{(k-1)}$ and on $\pi^{(k)}$ is normal. Let j^* be the facility chosen by i. Draw v_{ij^*} from the conditional posterior truncated at $v_{ij^*} \geq \max_{j \in O_i^{(k)}/\{j^*\}} v_{ij}$. Denote it by $v_{ij^*}^{(k)}$. Then, draw v_{ijt} for $j \in O_i^{(k)} \setminus \{j^*\}$ from the conditional posterior truncated at $v_{ij} \leq v_{ij^*}^{(k)}$. Lastly, draw v_{ij} for $j \notin O_i^{(k)}$from its unconditional posterior without any truncation. Let $v^{(k)}$ denote the vector of draws.

2. Seemingly unrelated Bayesian regression: with the draws of $v^{(k)}$and $\pi^{(k)}$ and for fixed value of $\delta_j^{(k-1)}, \gamma_j^{(k-1)}, \beta_i^{(k-1)}$ and $\varepsilon_{i0}^{(k-1)}$; the equations above form a system of seemingly unrelated regressions. The posterior distributions of the parameters α, β are normal and the posterior distribution of Σ is inverse Wishart. We draw these parameters and obtain the resulting residuals $\hat{\varepsilon}_{ij}^{(k)}$ and $\hat{\nu}_{ij}^{(k)}$.

3. Update random effects:

   (a) Draw $\beta_i|\hat{\varepsilon}_{ij}^{(k)}, \hat{\nu}_{ij}^{(k)}, \Sigma^{(k)}$. The posterior distribution of γ_i conditional on the residuals $\hat{\varepsilon}_{ij}^{(k)}$ and $\hat{\nu}_{ij}^{(k)}$ and the previous variance draw $\Sigma^{(k)}$ is normal. We draw β_i from this conditional posterior. Let $\beta_i^{(k)}$ denote these draws and obtain the updated residuals$\bar{\varepsilon}_{ij}^{(k)} = \hat{\varepsilon}_{ij}^{(k)} + \beta_i^{(k)} x_j - \beta_i^{(k-1)} q_j$.

   (b) Draw $\varepsilon_{i0}|\hat{\varepsilon}_{ij}^{(k)}, \hat{\nu}_{ij}^{(k)}, \Sigma^{(k)}$. The posterior distribution of ε_{i0} conditional on the residuals $\bar{\varepsilon}_{ij}^{(k)}$ and $\hat{\nu}_{ij}^{(k)}$ and the previous variance draw $\Sigma^{(k)}$ is normal. We draw ε_{i0} from this conditional posterior. Let $\varepsilon_{i0}^{(k)}$ denote these draws and obtain the updated residuals$\tilde{\varepsilon}_{ij}^{(k)} = \bar{\varepsilon}_{ij}^{(k)} + \hat{\varepsilon}_{i0}^{(k-1)} - \hat{\varepsilon}_{i0}^{(k)}$.

   (c) Draw $\gamma_j|\tilde{\varepsilon}_{ij}^{(k)}, \hat{\nu}_{ij}^{(k)}, \Sigma^{(k)}$. The posterior distribution of γ_j conditional on the residuals $\tilde{\varepsilon}_{ij}^{(k)}$ and $\hat{\nu}_{ij}^{(k)}$ and the previous variance draw $\Sigma^{(k)}$ is normal. We draw γ_j from this conditional posterior. Let $\gamma_j^{(k)}$ denote these draws and obtain the updated residuals$\tilde{\nu}_{ij}^{(k)} = \hat{\nu}_{ij}^{(k)} + \gamma_j^{(k-1)} - \gamma_j^{(k)}$.

   (d) Draw $\delta_j|\tilde{\varepsilon}_{ij}^{(k)}, \tilde{\nu}_{ij}^{(k)}, \Sigma^{(k)}$. The posterior distribution of δ_j conditional on the residuals $\tilde{\varepsilon}_{ij}^{(k)}$ and $\tilde{\nu}_{ij}^{(k)}$ and the previous variance draw $\Sigma^{(k)}$ is normal. We draw δ_j from this conditional posterior. Let $\delta_j^{(k)}$ denote these draws.

4. Update the variance of the random effects:

(a) Draw $\sigma^2_{\varepsilon 0}|\varepsilon^{(k)}_{i0}$. The posterior distribution of $\sigma^2_{\varepsilon 0}$ conditional on $\sigma^2_{\varepsilon 0}$ is inverse-gamma. Similarly, draw $\sigma^2_\beta|\beta^{(k)}_i$, $\sigma^2_\gamma|\gamma^{(k)}_i$ and $\sigma^2_\delta|\delta^{(k)}_i$.

5. Finally, collect all parameter draws in step k and denote them by $\theta^{(k)}$.

We specify a set of diffuse conjugate priors to each set of parameters, following recommendations in McCulloch and Rossi (1994). The priors for $\alpha, \beta, \delta, \gamma$ are normal with zero mean and covariance equal to the identity matrix times a large constant: 1000. The prior of Σ is an inverse Wishart with a 2×2 identity matrix as its scale matrix and 3 degrees of freedom. Similarly, the priors of $\sigma^2_{\varepsilon 0}$, σ^2_β, σ^2_γ and σ^2_δ are three independent inverse-gamma distributions with scale and shape parameters equal to 1/2. These priors are uninformative relative to the size of our dataset and thus, the estimation results are unlikely to change substantially should we make them even less precise.

We start a chain from a random starting points and run the Gibbs sampler for 4 million draws, discarding the first million draws. We summarize the draws for each parameter and verify that the Potential Scale Reduction Factor for each parameters is close to one, which indicates that letting the chain run for longer is not likely to change the results (Gelman et al., 2014).

D Monte Carlo Exercises

This section presents Monte Carlo evidence to assess the properties of the Gibbs sampler described in the main text, and to assess bias arising from model mis-specification. Our experiments focus on a single market with $J = 5$ products and vary the number of consumers in the market, $N \in \{5000, 20000\}$.

To simulate a dataset, we begin by simulating observed characteristics. Consumer and product locations are drawn uniformly at random from a unit square to generate distances x_{ij}; an observable preference shifter y_{ij} is drawn from a standard normal; a choice-set shifter z_{ij} is drawn from the Poisson distribution with parameter 10; a consumer-specific binary observable d_i is drawn from the Bernoulli distribution with parameter 0.5.

Next, we then simulate indirect utilities and choice sets by drawing

$$\begin{aligned} v_{ij} &= \delta_j + \beta_i x_{ij} - y_{ij} + \varepsilon_{i0} + \varepsilon_{ij}, \\ \sigma_{ij} &= 1\left\{\gamma_j + \alpha_i w_i + \nu_{i0} + \nu_{ij} > z_{ij}\right\} \end{aligned}$$

where ε_{i0}, ν_{i0} and $(\epsilon_{ij}, \nu_{ij})$ are mutually independent (multivariate) normal distributions with

mean zero and variance $\sigma^2_{\varepsilon 0}$, $\sigma^2_{\nu 0}$, and Σ respectively; the random coefficients β_i and α_i are normally distributed, mutually independent of each other and other random variables in the model with means and variances $\left(\bar{\beta}, \sigma^2_\beta\right)$ and $(\bar{\alpha}, \sigma^2_\alpha)$ respectively; and the facility fixed-effects γ_j and δ_j are generated from independent mean-zero normal distributions with variances σ^2_γ and σ^2_δ respectively. These latent variables provides an a product that each consumer is matched with.

We repeat this simulation procedure to produce 100 datasets that are then used to estimate the model using a Gibbs' sampler. Our sampler uses 1 million iterations, a burn-in of 25% of the chain, and one-in-ten thinning. For each dataset, we estimate four different models:

1. The correct specification
2. The "No Random Coefficients" model, which sets $\beta_i = \bar{\beta}$ and $\alpha_i = \bar{\alpha}$ for all i
3. The model with "Choice Set Shifter in Utility," which sets $\sigma_{ij} = 1$ and $v_{ij} = \delta_j + \beta_i x_{ij} + \beta_z z_{ij} - y_{ij} + \varepsilon_{i0} + \varepsilon_{ij}$,
4. The "Unconstrained Demand" model, which sets $\sigma_{ij} = 1$.

The second model assess the importance of random coefficients whereas the third and fourth assess whether mis-specification by omitting choice-set constraints are important, whether with or without the "naive" correction in the third model.

The estimated parameters and the coverage of the 95% confidence sets are presented in Tables D.4 and D.5 respectively for the case with 5000 and 20000 patients. As expected, the correct specification exhibits appropriate coverage of the true parameters. The omission of random coefficients not only creates a substantial bias in the coverage of $\bar{\beta}$ and $\bar{\alpha}$, but also in other parameters such as $\sigma_{\nu 0}$. Models that omit choice-set constraints are particularly problematic with extremely low coverage ratios.

Table D.4: Monte Carlo Summary with 5000 patients

	True Value	Correct Specification (1)			No Random Coefficient (2)			Choice Set Shifter in Utility (3)			Unconstrained Demand (4)		
		Bias	RMSE	95% cov	Bias	RMSE	95% cov	Bias	RMSE	95% cov	Bias	RMSE	95% cov
Mean γ	10	-0.138	0.724	82	-0.133	0.735	83	---	---	---	---	---	---
Mean α	-2	0.014	0.095	94	0.019	0.102	91	---	---	---	---	---	---
Mean δ	10	-0.001	0.446	91	0.069	0.457	91	6.923	7.043	0	-6.409	6.739	0
Mean β	-2	0.005	0.077	97	0.187	0.212	28	0.303	0.323	11	0.312	0.333	15
Coef on z_{ij} in Utility	0	---	---	---	---	---	---	-1.440	1.465	0	---	---	---
Sd ν	1	-0.014	0.094	97	0.049	0.130	93	---	---	---	---	---	---
Sd ε	1	-0.032	0.109	97	0.992	1.002	0	2.720	2.761	0	4.089	4.161	0
Corr (ν,ε)	0	0.075	0.256	99	0.018	0.155	95	---	---	---	---	---	---
Sd ε_0	1.22	0.041	0.127	94	0.188	0.274	80	1.620	1.686	2	3.553	5.218	4
Sd ν_0	1.22	0.033	0.087	92	0.167	0.187	55	---	---	---	---	---	---
Sd δ	1	-0.090	0.290	91	-0.089	0.289	93	0.721	0.955	52	0.607	0.839	61
Sd γ	1.41	-0.213	0.425	88	-0.213	0.428	86	---	---	---	---	---	---
Sd β	2	-0.025	0.104	96	---	---	---	-0.238	0.326	68	-0.336	0.490	65
Sd α	1	-0.098	0.173	94	---	---	---	---	---	---	---	---	---

Notes: Bias is the difference between true parameter and mean estimates. RMSE is the root mean squared error of the estimates. The 95% coverage probability is the number of simulations (out of 100) for which the true parameter lies in the 95% credible interval derived from the Gibbs sampler.

Table D.5: Monte Carlo Summary with 20000 patients

	True Value	Correct Specification (1)			No Random Coefficient (2)			Choice Set Shifter in Utility (3)			Unconstrained Demand (4)		
		Bias	RMSE	95% cov	Bias	RMSE	95% cov	Bias	RMSE	95% cov	Bias	RMSE	95% cov
Mean γ	10	0.023	0.610	94	0.043	0.628	96	---	---	---	---	---	---
Mean α	-2	0.000	0.038	96	-0.003	0.044	97	---	---	---	---	---	---
Mean δ	10	-0.027	0.393	94	0.031	0.386	94	6.762	6.844	0	-6.050	6.201	0
Mean β	-2	-0.002	0.046	92	0.173	0.185	6	0.301	0.312	2	0.300	0.310	2
Coef on z_{ij} in Utility	0	---	---	---	---	---	---	-1.389	1.404	0	---	---	---
Sd ν	1	-0.005	0.053	93	0.071	0.097	76	---	---	---	---	---	---
Sd ε	1	-0.007	0.066	95	1.037	1.041	0	2.654	2.680	0	3.878	3.927	0
Corr (ν,ε)	0	0.027	0.146	95	-0.029	0.076	94	---	---	---	---	---	---
Sd ε_0	1.22	0.011	0.074	93	0.143	0.194	64	1.560	1.611	1	3.586	4.192	0
Sd ν_0	1.22	0.010	0.051	94	0.157	0.164	11	---	---	---	---	---	---
Sd δ	1	-0.043	0.250	94	-0.048	0.256	95	0.776	0.963	47	0.641	0.820	57
Sd γ	1.41	-0.131	0.384	93	-0.135	0.388	92	---	---	---	---	---	---
Sd β	2	-0.006	0.064	94	---	---	---	-0.184	0.220	39	-0.206	0.266	42
Sd α	1	-0.032	0.086	92	---	---	---	---	---	---	---	---	---

Table D.6: Monte Carlo Diversion Ratio

	Average True Value	Correct Specification (1)	No Random Coefficient (2)	Choice Set Shifter in Utility (3)	Unconstrained Demand (4)
Demand Side	0.221				
Mean Bias		0.000	0.027	0.016	0.011
RMSE		0.003	0.035	0.095	0.093
Supply Side	0.625				
Mean Bias		0.024	-0.001	-0.388	---
RMSE		0.473	0.551	1.480	---

Notes: Demand side diversion ratio is defined as $\frac{\partial s_k}{\partial y_{ij}} / \frac{\partial s_j}{\partial y_{ij}}$. Supply side diversion ratio is defined as $\frac{\partial s_k}{\partial z_{ij}} / \frac{\partial s_j}{\partial z_{ij}}$.

Perhaps an economically more important estimand on which to compare the specifications are the estimated diversion ratios. The "demand-side" diversion ratios are computed using marginal changes in y_{ij} and the "supply-side" diversion ratios are computed using marginal changes in z_{ij}. The mean bias and the root mean squared errors are reported in Table D.6. As expected, the mean bias and the RMSE are the lowest for the correct specification. The omission of random coefficients does increase the size of the biases and the RMSE, but less so than misspecified models that omit choice-set constraints altogether.

E Appendix of Exhibits

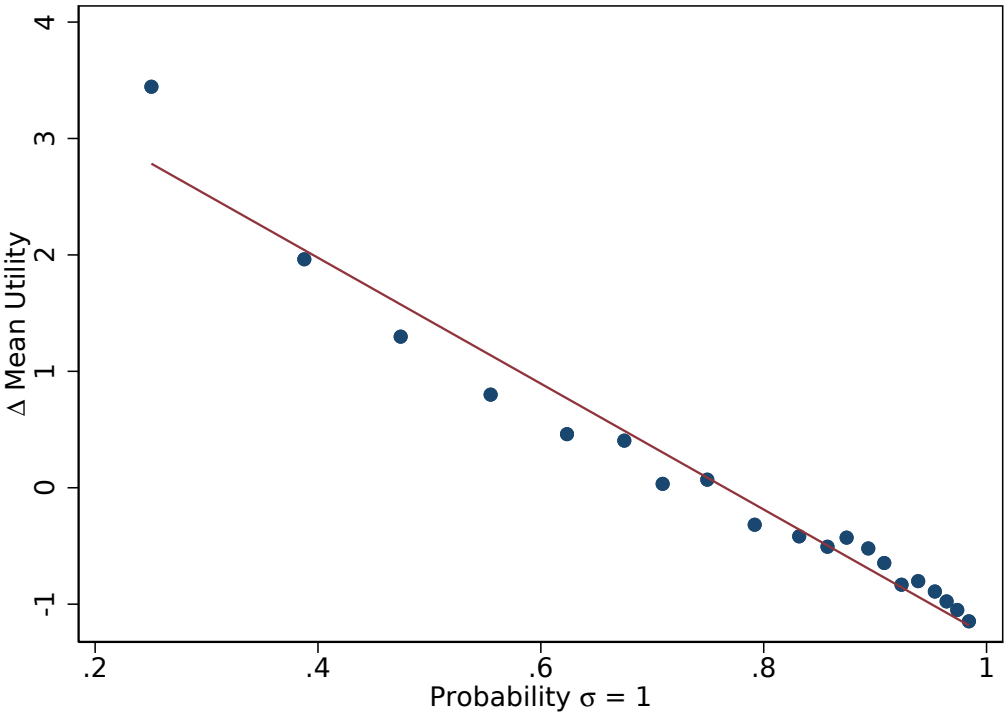


Figure E.2: Mean Utility vs Acceptance Probability