Demand Analysis under Latent Choice Constraints*†

Nikhil Agarwal and Paulo Somaini‡

April 19, 2022

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 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, Francesca Molinari, Aviv Nevo, Peter Reiss and participants at several seminars and conferences for helpful feedback and discussions. We also thank Felipe Barbieri, 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 a 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 function that determines choice sets. 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 accomodates 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 incentives induce the 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 are inapplicable as the largest market share product need not be the one preferred by the largest number 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.

As an illustrative application, we apply our methods to the kidney dialysis market in Cali-

[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 be endogenously different.

fornia. 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 accomodate 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 that is excluded from choice-set constraints is the distance between the facility and the patient's residence. This variable is excluded from patient profitability 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., 2020). 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 profitability 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 73.0%, 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 also consider alternative models which 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., 2020); 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., 2020; Aguiar et al., 2022) or on the dependence between preferences and choice-sets (c.f. Menzel,

[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.

2015; Abaluck and Compiani, 2020; Abaluck and Adams-Prassl, 2021). We discuss these differences 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 (2010) 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., 2022). 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 been based on the discrete choice models discussed above, which abstract away from supply-side rationing.

The empirical model for our application is closest to models of selective admission practices in nursing homes (Ching et al., 2015; Gandhi, 2021), although these papers do not formally consider the identification of the empirical model. Other empirical models that our identification analysis covers includes 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); 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 accomodates 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 futher 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. For consistency of terminology, 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 and can only choose or match with at most one product $j \in J_t$ on the other side. 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, and consumers are free to choose an outside option, denoted with 0. Each consumer i participates in only one market. As will be clear in section 2.2, our model is relevant for several settings.

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

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

where w_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_{i0t} to zero for each i and t; for some known value y_0 and a fixed j in each t, we set $\left|\frac{\partial g_{jt}}{\partial y}\left(w_i, y_0\right)\right| = 1$ for all w_i; and we set $g_{jt}\left(w_i, y_0\right) = 0$ for every j, t and w_i. The restrictions on v_{i0t} and the partial derivative of $g_{jt}\left(\cdot\right)$ are familiar location and scale normalizations. The restriction that $g_{jt}\left(w_i, y_0\right) = 0$ is without loss because a constant shift in $g_{jt}\left(\cdot\right)$ can be subsumed in $u_{jt}\left(\cdot\right)$.

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 $u_{jt}\left(\cdot\right)$ and $g_{jt}\left(\cdot\right)$ 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 w_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 w_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

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

denote this latent acceptance decision, where $\sigma_{jt}\left(w_i, \omega_i, z_{ij}\right) = 1$ denotes that consumer i was accepted by product j in market t. We refer to the function $\sigma_{jt}\left(\cdot\right)$ 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.

The term z_{ij} is a consumer-product specific observable scalar characterstic that affects the decision of the product to accept the consumer that is excluded from the consumer's utility. As opposed to w_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 w_i.

We assume that each consumer is matched with one of her most preferred products that accepts her. Formally, each consumer's (latent) choice set is given by the set of products that accept the consumer:

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

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 $c_{ij} = 1$ only if $j \in \arg\max_{k \in O_i} v_{ikt}$ and $\sum_{j \in O_i} c_{ij} = 1$. 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 accomodates 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 w_i.*

This assumptionplaces 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 w_i as well as on product-market specific characteristics through the function $g_{jt}(\cdot)$. Second, it implies that unobserved determinants of preferences are independent of z_i given w_i. Thus, z_i is an instrument that shifts choice sets without affecting the distribution of preferences. 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.

We assume that the random vector ω_i is independent and identically distributed 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 (w_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 assumption on the data generating process and assumption 1 imply that the share of consumers with observables (w_i, y_i, z_i) that are matched with product j in market t is given by

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

The information in the data consists only of these market shares for each value of (w_i, y_i, z_i) in its support.

The shares $s_{jt}(\cdot)$ can be re-written as

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

The first term in the summand is the probability that a consumer with attributes (w_i, y_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 (w_i, z_i). Because we depart from the often-used assumption that preferences and choice sets are independent, the first term depends on z_i, which affects the distribution of ω conditional on O. Assumption 1 allows us to omit the conditioning on

y_i when writing the second.

As is clear from equation (3), 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 yield invalid mean utility measures whenever relevant latent choice set constraints are ignored (c.f. Berry, 1994; Berry et al., 1995, 2013). Below, we describe some specific cases of our general model where constrained choice sets are central to analysis.

2.2 Examples

We start by showing that our preference model is general enough to accomodate 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 unobservables ξ_{jt} (e.g. Berry et al., 1995; Petrin, 2002):

$$v_{ijt} = w_i' \Gamma x_{jt} + x_{jt}\beta_i + y_{ij} + \xi_{jt} + \varepsilon_{ij}.$$

We can nest this specification by setting $u_{jt}(w_i, \omega_i) = w_i' \Gamma x_{jt} + x_{jt}\beta_i + \xi_{jt} + \varepsilon_{ij}$, $\omega_i = (\beta_i, \varepsilon_{i1}, \ldots, \varepsilon_{iJ})$ and $g_{jt}(w_i, y_{ij}) = -y_{ij}$. 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 characterics w_i and observable product characteristics x_{jt}. Our identification results will accomodate most commonly used distributional assumptions on ε_{ij}, including those that yield the familiar logit or nested-logit models (Train, 2009). We can also accomodate 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 accomodates 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:

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

where $NPV_{jt}\left(\omega_i, w_i\right)$ denotes the present value of variable profits from patient i at facility j, and $V\left(z_{ij}+1\right) - V\left(z_{ij}\right)$ is the change in the continuation value given an caseload of z_{ij} at the time of arrival of patient i. The terms w_i and ω_i denote observable and unobservable characteristics of patient i. The term $V_j\left(z_{ij}\right) - V_j\left(z_{ij}+1\right)$ 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, w_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 encompases 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 payscales (e.g. Dagsvik, 2000; Agarwal, 2015). Let $e_{jt}(w_i, \omega_i, z_{ij})$ 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, w_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 $O_i = \{j \in J_t : e_{jt}(w_i, \omega_i, z_{ij}) \geq p_{jt}\} \cup \{0\}$. Thus, in our notation:

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

The identification of a similar model of two-sided matching was recently studied in He et al. (2020). Our results will place fewer functional form 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 $\sigma_{jt}\left(w_i, \omega_i, z_{ij}\right) = 1$. Since w_i and ω_i are arguments in $u_{jt}\left(\cdot\right)$, 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 advertizing expenditure and a consumer's exposure to advertizing to construct a variable analogous to z_{ij}. Another example is Gaynor et al. (2016), who model 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.[4] Our framework allows for a weaker version of defaults 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}, to influence 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}.[5] 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 advertizing 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

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

[5] 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 $\sigma_{jt}(\cdot)$.

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). However, when the dataset does not record 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 shifterz_{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.

3 Identification and Estimation

We start by studying identification of the joint distribution of σ_{it} and v_{it} and the necessity of choice-set shifters in sections 3.1 and 3.2 respectively. These results use only within-market variation and we therefore drop explicity conditioning on t. In section 3.3 we derive results exploiting both within and across market variation and reintroduce explicit conditioning on t..

3.1 Identification within a market

Our main result – Theorem 1 – shows identification of the joint distribution of indirect utilities v_{ij} and acceptance decisions σ_{ij} given observable agent characteristics (w_i, y_i, z_i) for each market as well as the function $g_j(\cdot)$. Identifying these quantities is sufficient for conducting a number of counterfactuals and calculations of interest. For example, identifying the joint distribution of v_i and is sufficient for identifying changes in demand in response to changes in y_i and to perform welfare analysis if $g_j(w_i, y_i)$ is an appropriate numeriare. Identification of σ_{ij} yields product-specific acceptance policy functions, which can be used to infer acceptance payoffs for a product in several of our examples.[6]

We start by introducing some notation and assumptions that we use throughout this subsection:

Assumption 2. *The function $\sigma_j(w_i, \omega_i, z_{ij})$ is non-increasing in z_{ij}. Moreover, for all j, w_i and ω_i, $\lim_{z\to-\infty}\sigma_j(w_i, \omega_i, z) = 1$ and $\lim_{z\to\infty}\sigma_j(w_i, \omega_i, z) = 0$.*

[6]For example, the acceptance policy function directly yields preferences in the case of static two-sided matching models. In the case of a dynamic acceptance policy, we may use Hotz and Miller (1993) inversion to recover payoffs.

Monotonicity requires that product j is more likely to accept consumer i if the value of z_{ij} is lower. This assumption is natural in the examples discussed in section 2.2 above. In addition, we assume that the acceptance decision changes from 0 to 1 for some value of z for each value of the other variables. This latter restriction rules out certain models – e.g. inattention with strong defaults discussed in section 2.2 – where the choice-set shifter for a set of the products is irrelevant.

Define the cut-off quantity, $\pi_j(w_i, \omega_i) = \sup\{z : \sigma_j(w_i, \omega_i, z) = 1\}$. Under assumption 2, the function $\pi_j(\cdot)$ determines product j's acceptance policy for almost every z since $z < \pi_j(w_i, \omega_i)$ implies $\sigma_j(w_i, \omega_i, z) = 1$, and $z > \pi_j(w_i, \omega_i)$ implies $\sigma_j(w_i, \omega_i, z) = 0$. However, the acceptance policy function can take any value when $z = \pi_j(w_i, \omega_i)$.

We will build our main result (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).

3.1.1 Identification with known $g(\cdot)$

Identification in the case when $g(\cdot)$ is known can be achieved without any further assumptions:

Lemma 1. *Fix w_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 w_i. The joint distribution of (u_i, π_i) conditional on $(u_i, \pi_i) \in \chi$ and w 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 (π, u) 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 an agent is not matched to any of the products in the market. This probability, which is observed, is equal to the probability that either $u_{ij} < g_{ij}$ or $\pi_{ij} < z_{ij}$, 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. The point $(\bar{z}_1, \bar{g}_1)$ collects the first components of $\bar{z}$ and $\bar{g}$. 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 (π_1, u_1)

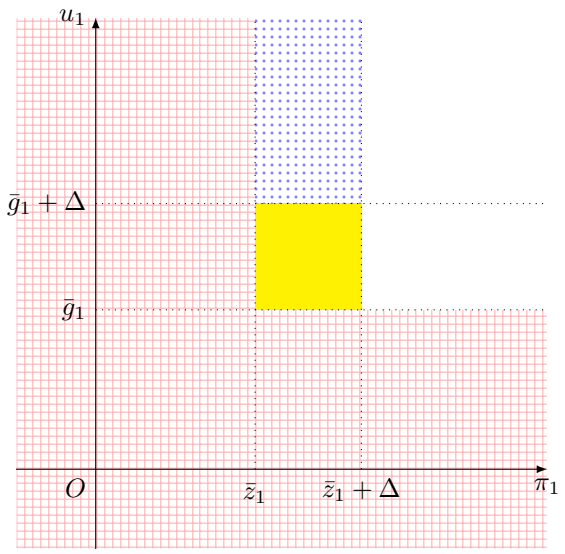


Figure 1: Two Dimensions

falls within this region is equal to the increase in the probability from an agent remaining unmatched at $(\bar{g}, \bar{z})$ to remaining unmatched when $\bar{z}_1$ is increased by Δ. This is because product 1 is not in the choice set at $\bar{z}_1 + \Delta$. Since this increase in probability is observed, we can determine the probability that (π_1, u_1) belongs to the set $[\bar{z}_1, \bar{z}_1 + \Delta] \times [\bar{g}_1, \infty)$. Using a similar argument and subtracting observed probabilities, we can determine the probability that (π_1, u_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 agents on the margin between choosing the sole inside option and the outside option, and variation in $\bar{z}_1$ affects the match of agents that are on the margin of being acceptable for product 1. Thus, the two pertubations 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 (π_1, u_1). This is because the shaded yellow box from figure 1 is the projection on the $u_1 - \pi_1-$hyperplane. In higher dimensions, this region represents a tube with a square cross-section. The yellow region in figure 2 illustrates this set projected on the $u_1 - u_2 - \pi_2$ hyperplane for a particular value of $(\bar{z}, \bar{g})$. Observe that this region conditions on the event that $u_2 < \bar{g}_2$ or $\pi_2 < \bar{z}_2$ in order to focus on the set of agents that would not be matched with product 2 if $\bar{z}_1$ or $\bar{g}_1$ were perturbed.

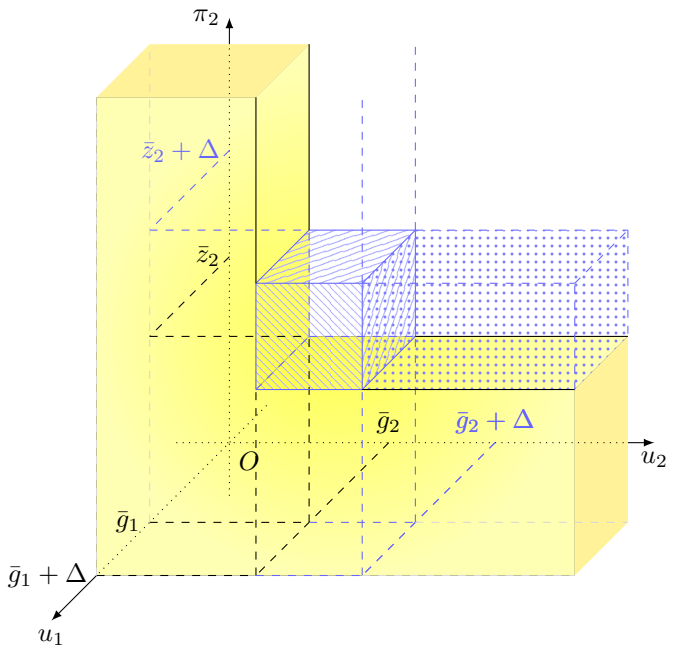


Figure 2: Three Dimensions

Our approach uses mathematical induction to extend this argument to higher dimensions, ultimately recovering the joint distribution of (π, u). 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 agents such that $(u_1, \pi_1) \in [\bar{z}_1, \bar{z}_1 + \Delta] \times [\bar{g}_1, \bar{g}_1 + \Delta]$. 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 $(\pi_2, u_2) \in [\bar{z}_2, \bar{z}_2 + \Delta] \times [\bar{g}_2, \bar{g}_2 + \Delta]$ and $(u_1, \pi_1) \in [\bar{z}_1, \bar{z}_1 + \Delta] \times [\bar{g}_1, \bar{g}_1 + \Delta]$. 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 (π, u) belongs to the set $\prod_{j=1}^{J} [\bar{z}_j, \bar{z}_j + \Delta] \times [\bar{g}_j, \bar{g}_j + \Delta]$. This probability, for arbitrarily small Δ, yields the density of (π, u) if it exists. The proof formalizes this intuition without requiring that (π, u) 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 z and g for acceptance decisions and preferences respectively. Assumption 1 implies that each shift leaves the joint distribution of (π, u)

unchanged. And, since π simply re-writes the vector of acceptance decisions σ, the result implies the identification of acceptance decisions jointly with the distribution of indirect utilities, u.

This argument is inspired by those pertaining to two-sided matching in He et al. (2020), which relaxes restrictions of preference heterogeneity (e.g. Diamond and Agarwal, 2017) or on tail behavior on unobservables (e.g. Menzel, 2015). Relative to lemma 1, the results in He et al. (2020) rely on a non-primitive rank condition and apply to a more restrictive specification of preferences. Our model incorporates endogenous product characteristics, and allows for non-separable, multi-dimensional unobserved heterogeneity.[7] It allows for general random utility models, including models in which there is perfect correlation between (components of) u and π, and models that yield non-differentiable market share functions. In addition, we only need local variation in the shifters to identify the distribution of (u, π).

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, $g_j(w_i, y_{ij}) = y_{ij}$ (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 (w_i, y_i, z_i) if $\frac{\partial}{\partial y_{ik}} s_j(w_i, y_i, z_i)$ and $\frac{\partial}{\partial y_{ij}} s_k(w_i, y_i, z_i)$ exist and are strictly positive.

Thus, our notion of substitution between two goods requires the existence of cross-partials that are strictly positive. Moreover, it assumes that if j is a strict substitute for k, then the reverse is also true.[8]

Let $\Sigma(w_i, y_i, z_i)$ be a matrix with one in the (j, k) entry if the pair of goods are strict substitutes at (w_i, y_i, z_i), but zero otherwise. Note that the definition above implies that $\Sigma(w_i, y_i, z_i)$ is symmetric. Observe that $\Sigma(w_i, y_i, z_i)$ defines an undirected graph where the

[7] Specifically, He et al. (2020) assume $v_{ijt} = u_{jt}(w_i) - g_{jt}(y_{ij}) + \omega_{ijt}$, $\sigma_{ijt} = 1\{\pi_{jt}(w_i) - h_{jt}(z_{ij}) + \eta_{ijt} > 0\}$. Their results assume a rank condition on the matrix of derivatives of market shares with respect to each of the observable characteristics (Condition 3.4, He et al., 2020). One interpretation of lemma 1 is that it provides a primitive condition for their results in a more general model. In appendix B, He et al. (2020) 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.

[8] 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.

nodes are the set of products and an edge exists between j and k if and only if the corresponding element in $\Sigma(w_i, y_i, z_i)$ is 1. Finally, define $\Sigma(w_i, y_i) = \vee_{z \in Z} \Sigma(w_i, y_i, z)$. The graph of $\Sigma(w_i, y_i)$ has an edge between j and k if and only if there exists z in its support such that the $j - k$ entry of $\Sigma(w_i, y_i, z)$ is equal to 1.

Assumption 3. *For every w_i and all but a finite set of y_i in its support, the graph of $\Sigma(w_i, y_i)$ has a path connecting any pair of products.*

This assumption on the substitution patterns is related to but stronger than the one in Berry et al. (2013). Yet, it is both testable and significantly weaker than a requirement in which all pairs of goods are substitutes. 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 subtitute 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 from j to k in $\Sigma(w_i, y_i)$ if there is a third good, l, such that the pairs (j, l) and (l, k) are strict substitutes at some z_i.

Assumption 3 is closely related to monotonicity of $g_j(w_i, \cdot)$ in the second argument. Equation (1) and assumption 1 together imply that the market share of each good k is weakly increasing in y_{ij} if $g_j(w_i, y_{ij})$ is weakly increasing in y_{ij}.[9] Moreover, in several models, goods j and k will be strict substitutes if and only if $g_j(w_i, y_{ij})$ and $g_k(w_i, y_{ik})$ are differentiable and strictly increasing in y_{ij} and y_{ik} respectively. One such model occurs when every pair of products $\{j, k\}$ belongs to some choice set O_i with non-zero probability and the conditional distribution of u_i given O_i admits a density with full-support on $\mathbb{R}^J$ (see corollary 4 in the appendix). If this requirement holds for every pair (j, k) then there is an edge between all pairs of goods in $\Sigma(w_i, y_i)$.

However, the example above rules out important cases such as a model with vertical preferences or, more generally, the pure characteristics model. Proposition 3 in the appendix shows weaker conditions under which goods j and k are strict substitutes if and only if each

[9] We can also consider the case where $g_j(w_i, y_{ij})$ is decreasing in the second argument by requiring that the cross-partials in assumption 3 are strictly negative. This case is covered in our analysis by redefining the observed covariate y_{ij} to be $-y_{ij}$.

$g_j(w_i, y_{ij})$ is strictly increasing and differentiable in y_{ij}. The conditions require that whenever the pair of goods $\{j, k\}$ belong to the choice set O_i with non-zero probability and that the joint distribution of indirect utilities is such that consumers substitute smoothly between the goods. Thus, it provides weaker conditions under which the is an edge between goods j and k in the graph of $\Sigma(w_i, y_i)$. 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 Proposition 3 in the appendix.

While assumptions 1 and 2 have allowed for atoms in the joint distribution of (u_i, π_i), assumption 3 restricts the number of atoms to be finite. Specifically, if the distribution of u_i (conditional on w_i) has an atom at $g(w_i, y_i)$, then $s(w_i, y_i, z_i)$ may not be differentiable with respect to y_i at that value even if the function $g(w_i, y_i)$ is differentiable. In this case, goods j and k may not be strict substitutes. We view this restriction as mild.

Finally, our proof requires the following weak support and regularity conditions:

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

(ii) For each w_i and j, the function $g_j(w_i, y_j)$ is continuously differentiable in y_j with $g'_j(w_i, y_j) \neq 0$ for all y_j.

Part (i) places a weak requirement on the support of Y that is used mostly for tractability. Part (ii) implies that the functions $g_j(w_i, y_j)$ are smooth and strictly monotone with respect to the second argument. Together with assumption 3, this restriction implies that $g_j(w_i, y_{ij})$ is strictly increasing in y_{ij}. We are now ready to show that each function $g_j(\cdot)$ 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 $g_j(w_i, \cdot)$ is identified for all $y_j \in Y_j$.*

Proof. See appendix A.3. □

The argument first identifies the ratio of $g'_k(w_i, y_{ik})$ and $g'_j(w_i, y_{ij})$ for goods j and k with an edge in $\Sigma(w_i, y_i, z_i)$ for some z_i. Consider the inclusive value of a consumer conditional on (w_i, z_i, y_i). Dropping the conditioning on w_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. We first use the envelope theorem to show 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.[10] Taking the partial derivative of this equation with respect to y_{ik} yields that

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

This derivative exists and is non-zero under assumption 3. 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

$$\frac{g_k'(y_{ik})}{g_j'(y_{ij})} = \frac{\partial s_j(y_i)}{\partial y_{ik}} \Big/ \frac{\partial s_k(y_i)}{\partial y_{ij}} \tag{4}$$

is identified.

If all pairs of goods are strict substitutes at all values of (y_i, z_i) (for each w_i), we could directly use the normalizations that $g_j(y_0) = 0$, $\left|g_j'(y)\right| = 1$ 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 $(k = j_0, j_1, \ldots, j_n = j)$ in $\Sigma(y_i)$ between any pair of goods j and k. 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 $g_{jt}(w_i, \cdot)$ can be identified when $J > 1$ under the assumptions outlined earlier. In the case when $|J| = 1$, we can assume without loss that $g_j(w_i, \cdot)$ is known as long as it is monotonic since the outside option is normalized to zero.

This result, which shows the identification of $g(\cdot)$, allows us to achieve identification without relying on quasi-linear special regressors. This differentiates our approach from that

[10] 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 3 – that imply the required sufficient conditions on the cross-partials of $V^*(g(y_i))$.

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.[11] 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.

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 $g_j(w, \cdot)$ 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 $g(w, Y) \times Z = \prod_{j=1}^{J} g_j(w, Y_j) \times Z$, where $g_j(w, Y_j)$ is the image of the set Y_j under $g_j(w, \cdot)$ and Z is the support of the random vector z_i.*

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

It is worth noting that the techniques used in this section rely only on local variation in the shifters y_i and z_i. Thus, we do not lean on "identification at infinity" arguments. 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. Our proofs show that 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 $int\,(g(w_i, Y) \times Z)$, the joint distribution of u_i and π_i conditional on w_i is identified.*

[11] 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 4). Abaluck and Adams-Prassl (2021) uses departures from Slutsky symmetry combined with specific models of consideration to identify incomplete consideration sets.

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. agents who like one tend to also like the other one. Correlation between v_{ji} and $v_{j'i}$ suggests that products j and j' tend to prefer the same set of agents. Moreover, correlation between u_{ij} and v_{ji} suggests that agents tend to prefer products that are likley to admit them.

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.[12] 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 be informative about 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 $g(w_i, y_i)$ 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.*

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

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

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, 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.

This failure occurs even though we allow for the shifter of preferences to have full support on its domain. The main requirement is that choice sets cannot be complete for all u. As discussed above, the distribution of preferences is over-identified under the remaining assumptions if latent choice sets are complete. Of course, the under-identification issue would be more severe if the support of $g(w_i, y_i)$ is more limited or if $g(\cdot)$ were unknown.

This result implies that complete choice sets are essentially the only case when the other restrictions of our model are sufficient for identification. Since simply allowing for incomplete latent choice sets results in under-identification, the results indicate that the conditions in theorem 1 are sharp.

The 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 Aguiar et al. (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 correlated with unobservable demand shocks (Berry, 1994; Berry et al., 1995). Such correlation may occur because products may choose some characteristics based on demand shocks

that are unobservable to the econometrician. For example, products may set prices 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).

Our identification arguments in the presence of such endogeneity closely follow Berry and Haile (2010). We now assume that equation (1) can be written as follows

$$v_{ijt} = u\left(x_{jt}, \xi_{jt}\left(w_i\right), w_i, \omega_i\right) + g\left(x_{jt}, \zeta_{jt}\left(w_i, y_{ij}\right), w_i, y_{ij}\right),$$

where $\xi_{jt}\left(w_i\right)$ and $\zeta_{jt}\left(w_i, y_{ij}\right)$ are scalar unobservables, x_{jt} denotes a vector of observable product characterstics that are potentially correlated with $\xi_{jt}\left(w_i\right)$ and $\zeta_{jt}\left(w_i, y_{ij}\right)$, and $u\left(\cdot\right)$ and $g\left(\cdot\right)$ are unknown functions. We will reintroduce the market index t to remind the reader that we will allow for both within and cross-market variation in the results below.

The combination of the assumptions that (i) the unobservables are scalars and (ii) the functions $u\left(\cdot\right)$ and $g\left(\cdot\right)$ are not indexed by j and t, makes this specification more restrictive than the one in equation (1). Yet, this model is more general than commonly estimated demand models because it allows $g\left(x_{jt}, w_i, \zeta_{jt}\left(w_i, y_{ij}\right), y_{ij}\right)$ to be both non-linear in y_{ij} and interact with product-level observables and unobservables $\left(x_{jt}, \zeta_{jt}\left(w_i, y_{ij}\right)\right)$.

Our goal is to identify the joint distribution of

$$v_{it} | \, w_i, y_i, \left\{x_{jt}, \xi_{jt}\left(w_i\right), \zeta_{jt}\left(w_i, y_{ij}\right)\right\}_j,$$

by exploiting variation both across products and across markets. The joint distribution that we previously identified conditioned on the market's identity t and implicitly 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}, $\xi_{jt}\left(w_i\right)$ and $\zeta_{jt}\left(w_i, y_{ij}\right)$. Knowledge of these distributions is sufficient for identification of several quantities of interest. These include identification of 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 $\xi_{jt}\left(w_i\right)$ and $\zeta_{jt}\left(w_i, y_{ij}\right)$. Once we recover these unobservables, we can get the joint distrubiton of v_{it} conditional on the full vector of observed and unobserved characteristics. Since the unobservables $\xi_{jt}\left(w_i\right)$ and $\zeta_{jt}\left(w_i, y_{ij}\right)$ only affect the utilities of product j in market t, it will be sufficient to work with the marginal distribution of $u_{ijt} = u\left(x_{jt}, \xi_{jt}\left(w_i\right), w_i, \omega_i\right)$ conditional on $w_i, x_{jt}, \xi_{jt}\left(w_i\right)$ and the function $g\left(x_{jt}, \zeta_{jt}\left(w_i, y_{ij}\right), w_i, y_{ij}\right)$. Recall that corollary 1 and lemma 2 imply identification of the marginal distribution of u_{ijt} and the function $g_{jt}\left(w_i, y_{ij}\right)$ for a fixed value of w_i and

market t, which conditions on w_i and $x_{jt}, \xi_{jt}(w_i), \zeta_{jt}(w_i, y_{ij})$. The remaining challenge is that the unobservables $\xi_{jt}(w_i)$ and $\zeta_{jt}(w_i, y_{ij})$ may be correlated with x_{jt}.

We now adapt arguments from Berry and Haile (2010) to show identification in the simplest case, relegating extensions under less restrictive functional form assumptions to appendix A.5. To maintain the focus on product-level observables and unobservables we condition on agent's characteristics w_i and y_i, and omit them from notation, referring to ξ_{jt} and ζ_{jt} instead of functions with these arguments. We make the following assumption:

Assumption 5. *(i) Additive separability of product unobservables in mean utility:* $E[u_{ijt}|x_{jt}, \xi_{jt}] = \tilde{u}(x_{jt}) + \xi_{jt}$.

(ii) Additive separability of product unobservables in $g(\cdot)$*:* $g(x_{jt}, \zeta_{jt}) = \tilde{g}(x_{jt}) + \zeta_{jt}$.

(iii) Availability of instruments: $E[(\xi_{jt}, \zeta_{jt})|r_{jt}] = 0$ *for all* j *and* r_{jt}.

(iv) Completeness condition: For any function $B(x_{jt})$ *with* $E[B(x_{jt})|r_{jt}] = 0$ *a.e. in* r_{jt} *implies that* $B(x_{jt}) = 0$ *a.e. in* x_{jt}.

Assumption 5 makes three types of restrictions. First, parts (i) and (ii) require that the mean of u_{itj} conditional on (x_{jt}, ξ_{jt}) and the function $g(\cdot)$ are additively separable in ξ_{jt} and ζ_{jt} respectively. We can relax these requirements to instead place analogous restrictions on known transformations of u_{ijt} and $g(\cdot)$. Second, part (iii) imposes an exclusion restriction by assuming mean-independence of the instrument r_{jt}. Third, part (iv) imposes a completeness condition, which an infinite dimensional analog of the familiar rank condition (see Newey and Powell, 2003).

As was shown in Berry and Haile (2010), this assumption and knowledge of the distribution of u_{ijt} is sufficient to identify $\tilde{u}(\cdot)$ and $\tilde{g}$:

Proposition 2. *(Newey and Powell, 2003; Berry and Haile, 2010). If the marginal distribution of* u_{ijt} *and* $g_{jt}(\cdot)$ *are identified for every* j *and* t*, and Assumption 5 is satisfied, then the functions* $\tilde{u}(\cdot)$ *and* $\tilde{g}(\cdot)$*, and the unobservables* ξ_{jt} *and* ζ_{jt} *are identified for each* j *and* t.

Proof. Follows from assumption 5 and Proposition 2.1 in Newey and Powell (2003). □

Now we present the a result that shows identification of the distribution of v_{it} in the presence of latent choice constraints and endogenous product characteristics. The result follows from Corollary 1 and Proposition 2 We reintroduce w_i in the notation for completeness.

Corollary 2. *For every* w_i *in its support and every* t*, assume that the hypotheses of corollary 1 and assumption 5 are satisfied. Then, the joint distribution of* v_{it} *conditional on* x_t, ξ_t, w_i *and* y_i *is identified for every tuple* (x_t, ξ_t, w_i, y_i) *in its support.*

Berry and Haile (2010) also present a second result that allows for unobservables ξ_{jt} and ζ_{jt} to enter non-linearly at the expense of the stronger assumption of independence of the instrument. We replicate their arguments in the appendix A.5.

Finally, observe that while we addressed endogeneous characteristics only on the demand side, an analogous approach can be used to introduce endogenous characteristics into the supply side, by setting $\sigma_{ijt} = \sigma\left(x_{jt}, \chi_{jt}\left(w_i\right), w_i, \omega_i, z_{ij}\right)$, where $\chi_{jt}\left(w_i\right)$ is unobserved. This would require restrictions analogous to those in assumption 5 to be made on the distribution of π_{ijt} instead of on u_{ijt}. We do not flesh out these details as they are repetitive with the arguments already presented.

These results, which 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. This assumption may not be reasonable in markets with selective school admissions. Our framework, to our knowledge, is the first to accomodate both these features.

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: Centers for Medicare and Medicaid 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 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, 2021). 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, idiosyncracies 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.[13] In our sample, 77.9% of patients are treated at only one facility and the average patient only visits 1.22 facilities. Our approach 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.

[13] 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.

Table 1: Facility Sample

	All facilities	Ownership		
		Fresenius and Davita	Other chains	Independent
Facility				
N	552	377	114	77
Facility-year	2101	1418	385	298
Number of patients				
Mean	108.5	113.0	100.2	97.8
Std. dev	46.7	46.3	38.9	54.6
Number of stations				
Mean	22.3	22.3	22.0	22.9
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.

Figure 3: Patients per Dialysis Station

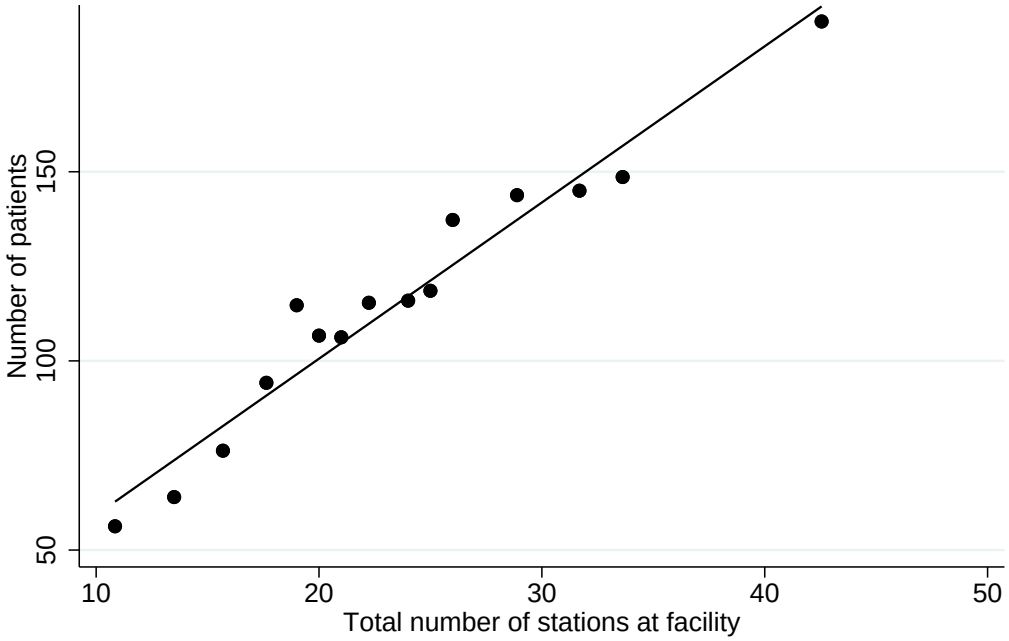


Table 2 describes the patient sample, which contains 41,913 new patients in our sample. Most of these patients choose hemodialysis at a facility in our facility sample. The patients are predominantly white, and the incidence of hypertension and diabetes is high. The majority of patients are on Medicare, an HMO or in the waiting period. The HMO group primarily 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.[14] 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 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 occupancy in facility j when patient i enters the dialysis market. If this measure of 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

[14] 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	33563	28629	18246	5586	4797
Age					
Mean	63.1	63.8	63.7	65.0	63.1
Std. dev	15.0	14.8	14.8	14.7	15.0
Employed (%)					
Mean	0.1	0.1	0.1	0.1	0.1
White (%)					
Mean	71.0	71.6	73.1	66.1	72.3
Black (%)					
Mean	10.4	10.6	10.6	11.5	9.7
BMI					
Mean	28.5	28.4	28.5	28.3	28.5
Std. dev	7.3	7.5	7.5	7.5	7.5
Diabetes (%)					
Mean	39.6	40.5	40.4	40.2	41.4
Hypertension (%)					
Mean	86.6	86.4	85.7	86.9	88.4
Panel B: Insurance type at admission					
Medicare (%)					
Mean	33.3	33.1	32.7	37.1	30.0
Medicare Advantage (%)					
Mean	24.5	24.6	25.0	24.0	24.0
Medicare waiting period (%)					
Mean	12.3	13.0	12.3	13.5	15.1
Other (%)					
Mean	29.9	29.3	30.1	25.4	30.9

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	---	6.5	17.8	59.1
Std. dev	---	5.2	17.2	54.7
Median	---	5.0	12.0	32.0
Distance to facility				
Mean	6.8	3.2	6.0	14.1
Std. dev	7.4	0.7	1.3	3.1
Median	4.4	3.2	6.1	14.3
95th percentile	21.8	4.4	8.3	18.7
Number of patients at facility				
Mean	119.1	120.3	117.1	114.3
Std. dev	46.7	27.3	24.4	18.4
Median	115.0	121.0	119.3	120.7
Total stations				
Mean	23.7	23.5	23.2	22.8
Std. dev	6.9	4.3	3.3	2.3
Median	24.0	23.3	23.2	23.4
Chain (%)				
Overall mean	87.5	89.9	89.1	88.2
Fresenius	17.9	22.3	22.0	21.4
Davita	49.4	48.9	48.4	49.0

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.

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. An examination of the time series of the number of patients at a facility 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.[15] We do not detect any breaks in trends for 498 of 552 facili-

[15]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 outlier estimates of excess occupancy.

ties. Conditional on finding a break in the trend, the average number of breaks is 2.17. 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 4.26.

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 begining 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 begining dialysis and the facility's occupancy is now positive. When combined with the result that facility-time controls yield a robust negative co-efficient, 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 co-efficient 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 ultimately estimate, our empirical specifications will use this measure of excess occupancy in the acceptance policy function.

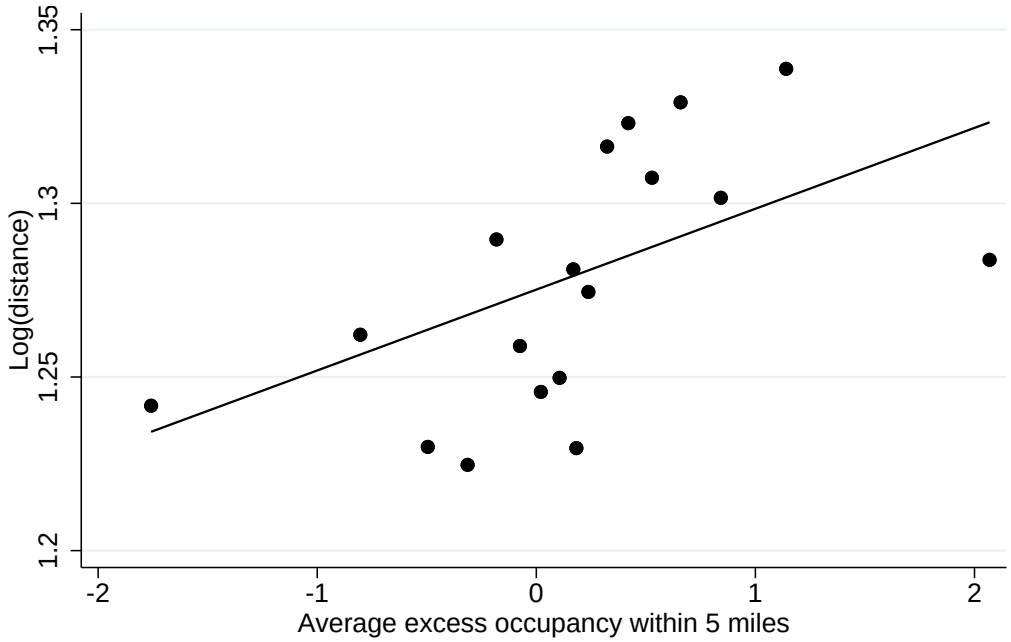
Fourth, these regressions also speak to the effect of capacity constraints at other facilities close to facility j. There are two opposing forces. Contraints at other facilities close to j can increase the demand for facility j. But, this force can also push facility j to be more selective and turn away 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

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	0.0001	0.004***	-0.0008***	-0.0008***	0.017***	0.016***				
	(0.0001)	(0.001)	(0.0001)	(0.0001)	(0.002)	(0.002)				
Excess occupancy							-0.0003**	-0.0003**	0.017***	0.016***
							(0.0001)	(0.0001)	(0.002)	(0.002)
Occupancy within 5 miles				-0.0001		0.004		0.0003***		0.003**
				(0.0001)		(0.003)		(0.0001)		(0.001)
Facility FE	X	X					X	X	X	X
Facily-Year FE			X	X	X	X				
Observations	708,969	23,666	708,969	708,969	23,666	23,666	708,969	708,969	23,666	23,666
R-squared	0.0237	0.122	0.0418	0.0418	0.159	0.159	0.0237	0.0237	0.125	0.126

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 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).

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 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 market outcomes if fluctuations in occupancy are not correlated with preferences for the facility. The main potential threat is that crowded facilities may be 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 correlation in excess occupancy is 0.07, suggesting that utilization on a specific day is not strongly correlated with the long-term occupancy.

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. 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 our identification results are non-parametric, estimating models with latent choice sets non-parametrically can be exceedingly difficult. Estimating the distribution of preferences non-parametrically is challenging to begin with, even without constraints on choice sets, if there are more than a few products in the market (see Compiani, 2021, for example). Latent choice sets only add to this difficulty because the number of potential choice sets is large even for relatively small J.[16] Thus, enumerating all possible latent choice sets in order to compute the likelihood is often computationally infeasible.[17]

Our parametric specification is chosen in light of this issue. First, we address 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., 2020). It modifies the sampler from McCulloch and Rossi (1994) with a data augmentation step to accomodate the case

[16] 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.

[17] 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.

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.

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

$$v_{ij} = \delta_j + \beta_w w_i - g\left(w_i, y_{ij}\right) + \varepsilon_{i0} + \varepsilon_{ij} \tag{5}$$

$$\sigma_{ij} = 1\left\{\eta_j + \alpha w_i - z_{ij} + \nu_{ij} > 0\right\}, \tag{6}$$

where δ_j and η_j are facility fixed effects, and ε_{i0}, ε_{ijt} and ν_{i0}, ν_{ij} are idiosyncratic shocks. We adopt the normalizations that $g'\left(w_i, y_{ij}\right) = 1$ at $y_{ij} = 1$ and $g\left(w_i, y_{ij}\right) = 0$ at $y_{ij} = 0$ for all w_i, and that the admission index is expressed in units of z_{ij}. As before, w_i is a vector of agent i's characteristics. We parametrize $g\left(\cdot\right)$ as a quadratic function given w_i, with parameters β_g and collect $\beta = (\beta_w, \beta_g)$. The specific observables w, 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.[18]

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 given the others can be obtained in closed form. Moreover, the conditional distributions of each of the parameters $(\alpha, \beta, \Sigma, \delta, \eta)$ given the errors 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 $(\alpha, \beta, \Sigma, \delta, \eta)$. The mean of this chain is asymptotically equivalent to the maximum likelihood estimator (see van der Vaart, 2000, Theorem 10.1 (Bernstein-von-Mises)). 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.

[18] 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 that we used to test our code. It is possible that the issue may be specific to our empirical setting.

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 agent i matching with 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 $\pi_{ijt} < z_{ij}$ 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 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 introducing simulation bias (Train, 2009) may be computationally demanding in markets with many possible options.

5.2 Estimates

5.2.1 Parameter estimates

In all the specifications we consider, 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 withthe 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. The unconstrained choice set for each patient is the set of facilities within a 50 mile radius of their home zip-code centroid.

We compare estimates from three specifications. The first specification, which is our preferred specification, models both preferences and acceptance policies (equations 5 and 6). 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. Facility fixed effects are included in the acceptance policy equation.

The second specification, which we refer to as the unconstrained demand model, omits capacity constraints by setting $\sigma_{ij} = 1$ for all i and j in equation (6). 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 by adding a term γz_{ijt} in equation (5), where γ 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_{ijt} (if γ 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 several markets. The second interpretation is that the specification represents a reduced-form approach that corrects for latent choice set constraints. Section 5.2.3 discusses an undesirable feature of this interpretation.

We start by describing and comparing the estimates that result from these specifications, before turning to a discussion of potential biases in section 5.2.2. 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 co-efficient on excess occupancy in the naive specification is negative.

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

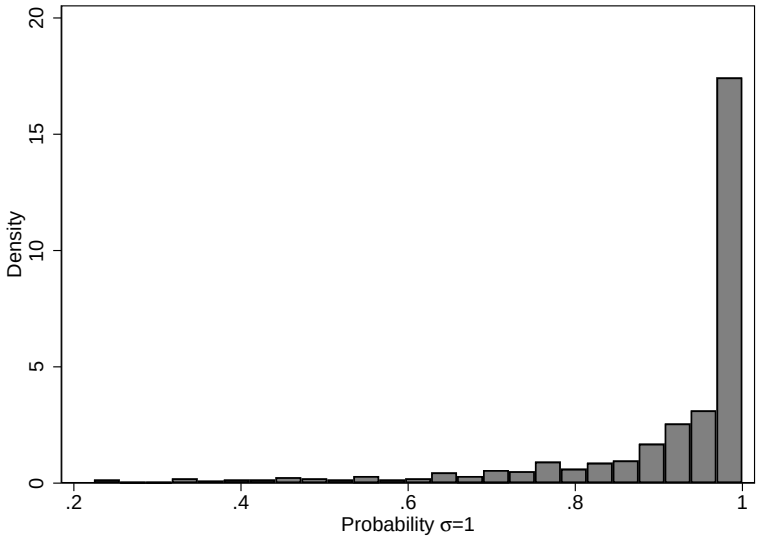
Turning to the acceptance policy function, we find that measures of patient health conditions and insurance status are correlated with acceptance. Figure 5 shows the estimated distribution of aceptance probabilities for each facility, averaged over all patients for whom the facility is in the patient's 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

Table 5: Parameter Estimates

	Preferred Specification (1)		Unconstrained (2)	Naïve (3)
	Acceptance	Utility	Utility	Utility
Diabetes	7.868	0.654	1.066	1.104
	(1.142)	(0.206)	(0.204)	(0.206)
Hypertension	10.051	-1.488	-0.980	-0.998
	(1.524)	(0.275)	(0.281)	(0.292)
BMI<20	2.605	-0.068	0.057	0.058
	(1.369)	(0.392)	(0.398)	(0.405)
25<=BMI<30	-0.397	-0.109	-0.144	-0.143
	(0.845)	(0.245)	(0.245)	(0.253)
30<=BMI	-1.076	0.477	0.422	0.439
	(0.947)	(0.248)	(0.246)	(0.258)
Age	-0.388	0.000	0.001	0.001
	(0.141)	(0.000)	(0.000)	(0.000)
Age squared	0.004	-2.392	-2.161	-2.224
	(0.001)	(0.284)	(0.279)	(0.297)
Medicare	5.521	0.080	0.059	0.061
	(1.193)	(0.039)	(0.040)	(0.040)
Medicare Advantage	-7.750	-2.145	-2.631	-2.708
	(1.525)	(0.334)	(0.318)	(0.336)
Medicare waiting period	-1.475	3.600	3.610	3.715
	(1.078)	(0.352)	(0.350)	(0.372)
Employed		-6.878	-7.035	-7.212
		(0.386)	(0.363)	(0.428)
Employed x distance		0.002	-0.002	-0.002
		(0.008)	(0.008)	(0.008)
Population density x distance		0.003	0.002	0.002
		(0.001)	(0.001)	(0.001)
Distance squared		0.013	0.013	0.013
		(0.000)	(0.000)	(0.000)
Excess Occupancy				-0.053
				(0.004)
Mean of δ_j		4.062	1.853	1.935
		(1.159)	(1.175)	(1.217)
Standard deviation of δ_j		3.036	3.204	3.186
		(0.127)	(0.118)	(0.119)
Standard deviation of ε_{i0}		11.445	11.772	12.136
		(0.470)	(0.401)	(0.570)
Standard deviation of ε_{ij}		4.274	4.775	4.769
		(0.044)	(0.033)	(0.033)
Mean of η_j	20.437			
	(4.791)			
Standard deviation of η_j	32.142			
	(3.601)			
Standard deviation of ν_{ij}	21.850			
	(2.105)			
Correlation between ε_{ij} and ν_{ij}	-0.138			
	(0.042)			

Notes: All specifications include distance with a coefficient normalized to -1 in the utility equation. Specification (1) includes “Excess occupancy" in the acceptance equation. Standard errors in parentheses.

Figure 5: Acceptance Probabilities



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 is non-trivial.

5.2.2 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.[19] 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

[19]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).

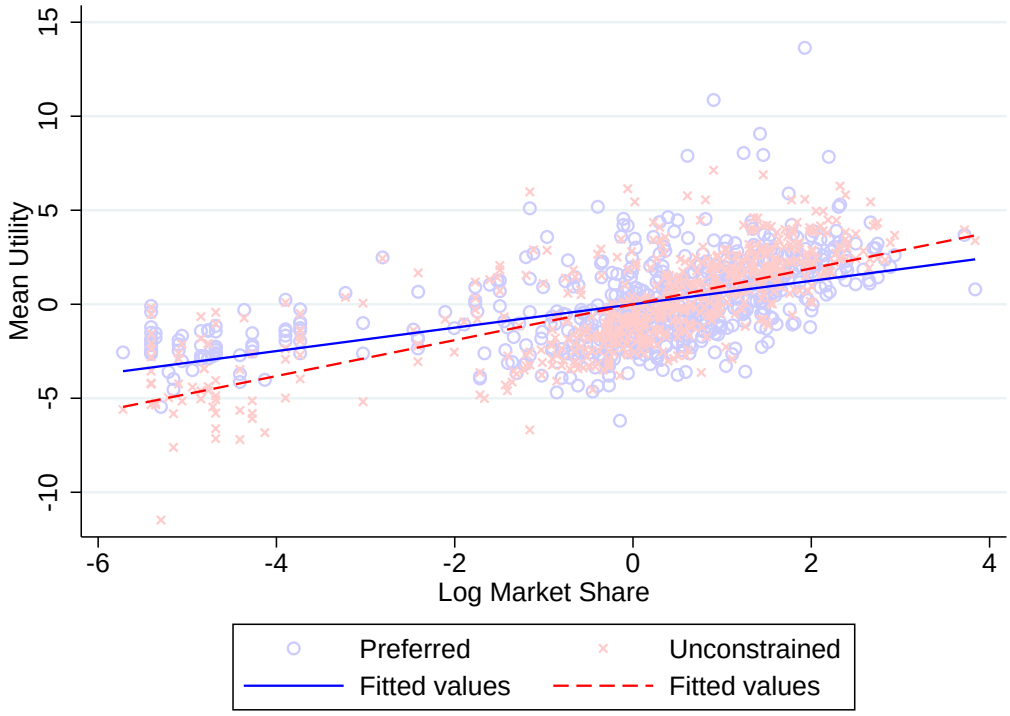
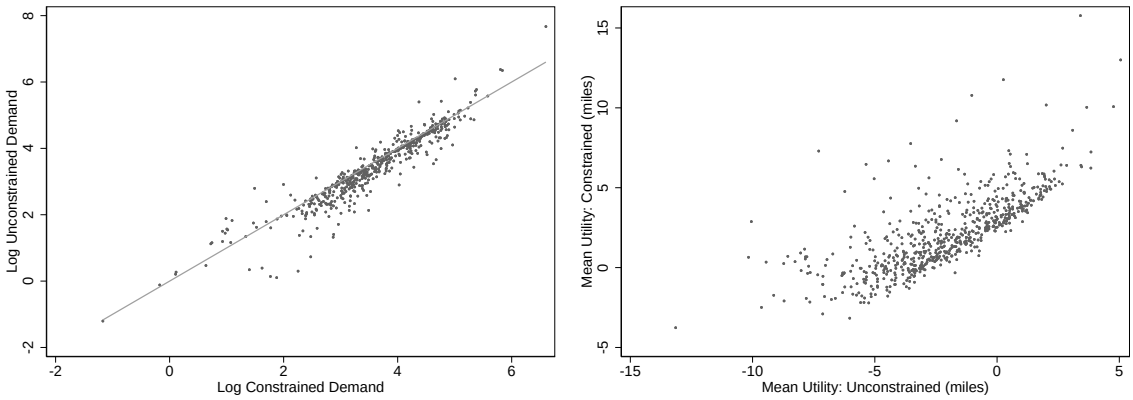


Figure 6: Willingness to Travel and Market Shares

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 they do 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



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

Figure 7: Bias in Demand

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). As before, latent choice constraints feed into biased estimates of demand.[20] 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.

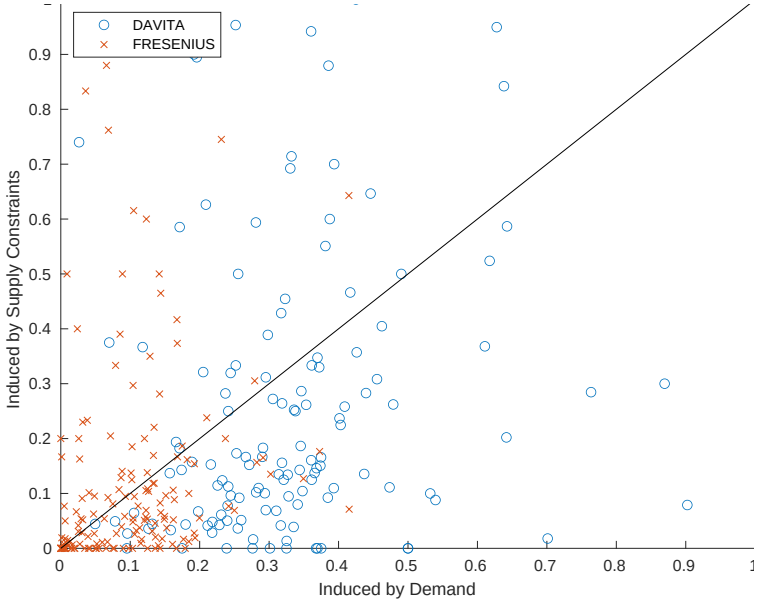
5.2.3 Implications of choice constraints on diversion ratios

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 w_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

[20] Figure D.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



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.

Figure 8 compares these two types of diversion ratios using our preferred specification. Each point in the figure represents a 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.[21] Restrictions, such as this one, can have important implications and go beyond the biases in estimated quantities described above.

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

[21] An alternative approach for obtaining differences in diversion ratios between factors affecting supply constraints and demand constraints would be to introduce random co-efficients 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.

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., 2020). The reduced-form yields a first-stage estimate in models with more complex supply-side behavior. For example, Gandhi (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* (7), 1939–78.

_ **and Paulo Somaini**, "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism," *Econometrica*, 2018, *86* (2), 391–444.

Aguiar, Victor, Levon Barseghyan, and Francesca Molinari, "Discrete Choice Models with Heterogeneous Preferences and Consideration," *Working Paper*, 2022.

Alba, Joseph W, J Wesley Hutchinson, and John G Lynch, "Memory and Decision Making," in Thomas S. Robertson and Harold H. Kassarjian, eds., *Handbook of Consumer Behavior*, Englewood Cliffs, NJ: Prentice-Hall, 1991, chapter 1, pp. 1–49.

Allen, Roy and John Rehbeck, "Identification With Additively Separable Heterogeneity," *Econometrica*, 2019, *87* (3), 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* (5), 1235–1268.

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

Barseghyan, Levon, Francesca Molinari, and Matthew Thirkettle, "Discrete Choice under Risk with Limited Consideration," *American Economic Review*, 2021, *111* (6), 1972–2006.

_ , _ , Maura Coughlin, and Joshua C. Teitelbaum, "Heterogeneous Choice Sets and Preferences," *Econometrica*, 2021, *89* (5), 2015–2048.

Berry, Steven and Ariel Pakes, "The Pure Characteristics Demand Model," *International Economic Review*, 2007, *48* (4), 1193–1225.

Berry, Steven T., "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 1994, *25* (2), 262.

_ , Amit Gandhi, and Philip A. Haile, "Connected Substitutes and Invertibility of Demand," *Econometrica*, 2013, *81* (5), 2087–2111.

_ and Philip A. Haile, "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," *Working Paper*, 2010.

Berry, Steven T and Philip A Haile, "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 2014, *82* (5), 1749–1797.

Berry, Steven T., James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, *63* (4), 841 – 890.

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

Block, H and J Marshak, "Random Orderings and Stochastic Theories of Responses," in I Olkin, S Ghurye, W Hoeffding, W.G. Mado, and H.B. Mann, eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford, CA: Stanford University Press, 1960, pp. 97–132.

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* (1), 83–117.

Butters, Gerard R., "Equilibrium Distributions of Sales and Advertising Prices," *Review of Economic Studies*, 1977, *44* (3), 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* (5), 1293–1307.

Chernozhukov, Victor and Christian Hansen, "An IV Model of Quantile Treatment Effects," *Econometrica*, 2005, *73* (1), 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* (4), 1291–1322.

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* (4), 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* (1), 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* (2), 204–230.

Department of Health and Human Services: Centers for Medicare and Medicaid Services, "Medicare and Medicaid Programs; Conditions for Coverage for End-Stage Renal Disease Facilities, 42 Federal Register," 2008.

Diamond, W. and N. Agarwal, "Latent indices in assortative matching models," *Quantitative Economics*, 2017, *8* (3), 685–728.

Dinerstein, Michael and Troy Smith, "Quantifying the Supply Response of Private Schools to Public Policies," *American Economic Review*, 2021, *111* (10), 3376–3417.

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* (1), 221–267.

Eliaz, K. and R. Spiegler, "Consideration Sets and Competitive Marketing," *The Review of Economic Studies*, 2011, *78* (1), 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* (4), 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* (11), 3521–3557.

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

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

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* (3), 1071–1105.

He, Yinghua, Shruti Sinha, and Xiaoting Sun, "Identification and Estimation in Many-to-One Two-sided Matching without Transfers," *Working Paper*, 2020.

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* (9), 2737–2781.

Hickman, William and Julie Holland Mortimer, "Demand Estimation with Availability Variation," in Emek Basker, ed., *Handbook on the Economics of Retailing and Distribution*, Cheltenham, UK: Edward Elgar Publishing Ltd., 2016, chapter 13, pp. 306–339.

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* (4), 877–905.

Honka, Elisabeth, "Quantifying search and switching costs in the US auto insurance industry," *The RAND Journal of Economics*, 2014, *45* (4), 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* (3), 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* (4), 192–226.

Hotz, V Joseph and Robert A Miller, "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 1993, *60* (3), 497–529.

Kepler, John D., Valeri V. Nikolaev, Nicholas Scott-Hearn, and Christopher R. Stewart, "Quality Transparency and Healthcare Competition," *Working Paper*, 2022.

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* (12), 3953–3959.

Lewbel, Arthur, "Endogenous Selection or Treatment Model Estimation," *Journal of Econometrics*, 2007, *141* (2), 777–806.

Liu, Jian, Shiying Wu, and James V. Zidek, "On Segmented Multivariate Regression," *Statistica Sinica*, 1997, *7* (2), 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* (482), 559–569.

Manski, Charles F., "The structure of random utility models," *Theory and Decision*, 1977, *8* (3), 229–254.

Matzkin, Rosa L., "Nonparametric identification and estimation of polychotomous choice models," *Journal of Econometrics*, 1993, *58* (1-2), 137–168.

_ , "Nonparametric identification," in James J. Heckman and Edward E. Leamer, eds., *Handbook of Econometrics*, Vol. 6B, Amsterdam: Elsevier, 2007, chapter 73, pp. 5307–5368.

McCulloch, Robert and Peter E Rossi, "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 1994, *64* (1-2), 207–240.

McFadden, Daniel, "Econometric Models of Probabilistic Choice," in Charles F. Manski and Daniel L. McFadden, eds., *Structural Analysis of Discrete Data and Econometric Applications*, Cambridge: The MIT Press, 1981, chapter 5.

Menzel, Konrad, "Large Matching Markets As Two-Sided Demand Systems," *Econometrica*, 2015, *83* (3), 897–941.

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

Milgrom, Paul and Ilya Segal, "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 2002, *70* (2), 583 – 601.

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* (5), 1565–1578.

Petrin, Amil, "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 2002, *110* (4), 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* (4), 440.

Roth, Alvin E. and Marilda A. Oliveira Sotomayor, *Two-Sided Matching: : A Study in Game Theoretic Modeling and Analysis*, Cambridge: 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* (2), 91–102.

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

U. S. Renal Data System, "2020 USRDS Annual Data Report: Epidemiology of kidney disease in the United States — End Stage Renal Disease," National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 2020.

U.S. Renal Data System, "2021 USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States," National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 2021.

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

Weitzman, Martin L., "Optimal Search for the Best Alternative," *Econometrica*, 1979, *47* (3), 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

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

The inequality follows because $c_{ij} = 1$ only if $j \in \arg\max_{k \in O_i} v_{ikt}$. Conditioning on w_i and droping 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 proability 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 $(u_i, \pi_i) \in \chi$, $P\left(u_i \leq \bar{u}, \pi_i \leq \bar{\pi} | \left(u_i, \pi_i\right) \in \chi\right)$, is identified. There are two cases. The first case is when $P\left(\left(u_i, \pi_i\right) \in \chi\right) > 0$. Then, we have that

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

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

To identify $P\left(\left(u_i, \pi_i\right) \in \chi\right)$, 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 $P\left(\left(u_i, \pi_i\right) \in \chi\right) = \sum_{k=1}^{\infty} P\left(\left(u_i, \pi_i\right) \in B_k'\right)$ 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, $P\left(\left(u_i, \pi_i\right) \in \chi\right)$ is identified.

Next, we show that $P\left(\left(u_i, \pi_i\right) \in \bar{B} \cap \chi\right)$ 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 $P\left(\left(u_i, \pi_i\right) \in \bar{B} \cap B_k'\right)$ is identified. Since $B_k' \cap B_{k'}' = \emptyset$, we have that $P\left(\left(u_i, \pi_i\right) \in \bar{B} \cap \chi\right) = \sum_k P\left(\left(u_i, \pi_i\right) \in \bar{B} \cap B_k'\right)$ 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 w_i and drop it from the notation for simplicity. Fix $\{j, k\}$. For each y_i, define the set

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

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 3. *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 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 and strictly increasing.

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,

$\frac{\partial s_j(y_i,z_i)}{\partial y_{ik}}$ exists and is strictly positive. By a symmetric argument $\frac{\partial s_k(y_i,z_i)}{\partial y_{ij}}$ exists and is strictly positive. Conversely, if $\frac{\partial s_j(y_i,z_i)}{\partial y_{ik}}$ exists and is strictly positive, then $\frac{\partial g_k(y_{ik})}{\partial y_{ik}}$ exists and is strictly positive. □

Corollary 3. *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 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}.*

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

Corollary 4. *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 assumptions 1, 3 and 4 hold and $|J| > 1$. If the (j,k) element of $\Sigma(w_i, y_i)$ is equal to 1, then*

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

where z is chosen such that the (j,k) element of $\Sigma(w_i, y_i, z_i^*)$ is equal to 1 and (w_i, y_i, z_i^*) is in the support of the data. Thus, $\frac{g_k'(w_i,y_i)}{g_j'(w_i,y_i)}$ is identified, bounded and strictly positive.

Proof. Because z_i^* and w_i are held constant, we drop them from the notation for simplicity. Define

$$V^*(g) = E\left[\sum_j v_{ij}c_{ij} \middle| g(y_i) = g\right].$$

Let $C_O = \left\{c \in \{0,1\}^J : \sum_j c_j \leq 1, c_j = 0 \text{ if } j \notin O\right\}$. Observe that

$$\begin{aligned} V^*(g) &= \sum_{O \in \mathcal{O}} E\left(\max_{j \in O} u_j(\omega_i) - g_j \middle| O, g\right) P(O|g) \\ &= \sum_{O \in \mathcal{O}} E\left(\max_{c \in C_O} \sum_j c_j \cdot (u_j(\omega_i) - g_j) \middle| O\right) P(O), \end{aligned}$$

where assumption 1 allowed us to drop the conditioning on g. Observe that $\sum_j c_j \cdot (u_j(\omega_i) - g_j)$ is an affine function of g and thus equidifferentiable with a continuous derivative. By Theorem 3 of Milgrom and Segal (2002), $\max_{c \in C_O} \sum_j c_j \cdot (u_j(\omega_i) - g_j)$ are left- and right-differentiable in each g_j.Boundedness of c_j and differentiation under the integral sign yields that

$$\begin{aligned} \frac{\partial V^*}{\partial g_j}(g(y_i)) &= -\sum_{O \in \mathcal{O}} P(O) E(c_{ij}|O) \\ &= -s_j(y_i), \end{aligned}$$

where the left- and right- partial derivatives of $V^*(\bar{g})$ with respect to g_j are equal because $s_j(y_i)$ is continuous at y_i with respect to y_{ij} and $g_j'(y_{ij}) \neq 0$ (see Theorem 3 of Milgrom and Segal (2002)). The dependence of $s_j(\cdot)$ on z_i^* and w has been subsumed since we have conditioned on it. Thus, we have that for any y, $s_j(y) = -\frac{\partial V^*}{\partial g_j}(g(y))$.
Differentiating the expression $s_j(y) = -\frac{\partial V^*}{\partial g_j}(g(y))$ with respect to y_k for $k \neq j$ yields

$$\frac{\partial s_j(y)}{\partial y_k} = -\frac{\partial^2 V^*(g(y))}{\partial g_j \partial g_k} g_k'(y_k). \tag{7}$$

Note that assumption 3 states that the partial derivative $\frac{\partial s_j(y)}{\partial y_k}$ exists and it is strictly positive and assumption 4 states that $g_k'(y_k)$ exists, it is not zero and it is finite. Thus, it follows that the cross partial also exists. Assumption 3 implies that neither the cross-partial derivative of V^* nor the derivative $g_k'(\cdot)$ are zero. By Young's theorem, $\frac{\partial^2 V^*(g(y))}{\partial g_j \partial g_k} = \frac{\partial^2 V^*(g(y))}{\partial g_k \partial g_j}$. Therefore, for any two values of y_j and y_k, the ratio

$$\frac{g_k'(y_k)}{g_j'(y_j)} = \frac{\partial s_j(y)}{\partial y_k} \Big/ \frac{\partial s_k(y)}{\partial y_j}$$

since $\frac{\partial^2 V^*(g(y))}{\partial g_j \partial g_k}$ is non-zero everywhere. Assumption 3 implies that this ratio is strictly positive and bounded. □

We are now ready to prove lemma 2. Fix w_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 $\tilde{y} = (y_1, ..., y_{j-1}, y_0, y_{j+1}, ...y_J) \in Y^*$where $Y^* = \{y \in Y : \Sigma(y_i) = 1\}$. Let $(k = j_0, j_1, j_2, \cdots, j_{n-1}, j_n = j)$ be a shortest path in the graph of $\Sigma(\tilde{y})$ that connects k to the reference good j. This path exists by assumption 3. Lemma 3 implies that $\frac{g'_{j_i}(\tilde{y}_{j_i})}{g'_{j_{i-1}}(\tilde{y}_{j_{i-1}})}$ is identified for each $i \in \{1, 2, ...n\}$. Moreover, $\frac{g'_{j_i}(\tilde{y}_{j_i})}{g'_{j_{i-1}}(\tilde{y}_{j_{i-1}})} \in (0, \infty)$. Thus, $g'_k(y_k) = \frac{g'_k(y_k)}{g'_j(y_0)} = \prod_{i=1}^n \frac{g'_{j_i}(\tilde{y}_{j_i})}{g'_{j_{i-1}}(\tilde{y}_{j_{i-1}})}$ is identified for every $k \neq j$ and $y_k \in Y_k$, where Y_k denotes the projection of the set Y on the k-th dimension. 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 can be used to identify $g'_k(\tau)$ for almost all $\tau \in Y_k$ because Y_k is an interval and the set $Y - Y^*$ has a finite number of elements. Moreover, each $g_k(\cdot)$ is strictly increasing since $g'_k(y_k) > 0$. We now turn to identification of $g_j(\cdot)$. Take any $y \in Y^*$ with $y_j \neq y_0$ and let k be a good such that that (j, k) element of $\Sigma(w_i, y)$ is 1. Lemma 3 and the fact that $g'_k(y_k)$ is identified imply that $g'_j(y_j)$ is identified for almost all $y_j \in Y_j$. Again, because Y is rectangular and $g'_j(\cdot)$ is continuous, we have that $g_j(y_j)$ is identified by the fundamental theorem of calculus.

A.4 Proof of Proposition 1

To simplify notation, we drop the conditioning on w_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 functions $\{h_O(\cdot)\}_{O\in\mathcal{O}}$ 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

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

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 6. $\xi_{jt} \perp r_{jt}$

Assumption 7. *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 $\psi_j(x, r, \tau) = \int_0^1 f_{\varepsilon_j}(\sigma B_j(x, \tau) | x, r)\, d\sigma > 0$. (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 4. *(Berry and Haile, 2010; Chernozhukov and Hansen, 2005). If δ_{jt} is identified and assumptions 6 and 7 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

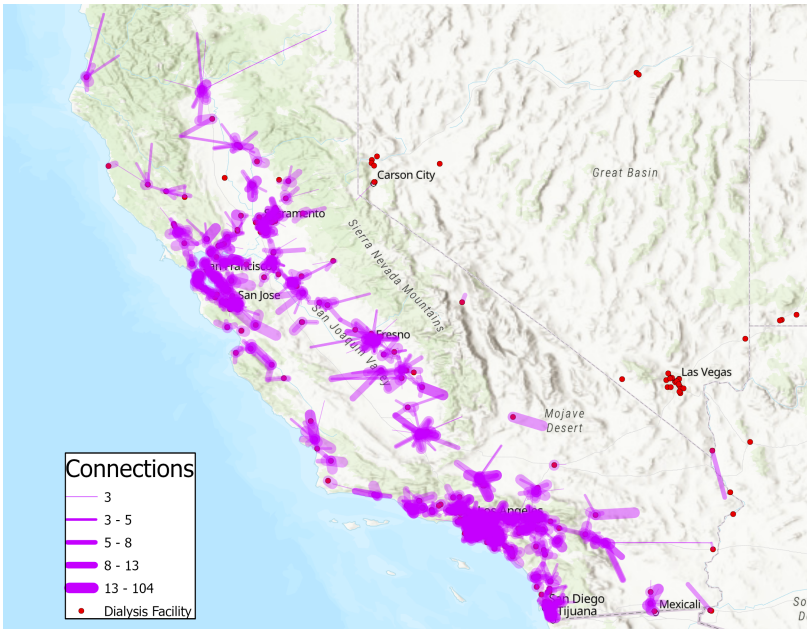
each facility at a given point in time. This file contains the start date and the end date of each patient's treatment at each facility where they receive care. We use this information to compute the number of patients undergoing in-patient hemodialysis at each facility on each day during our sample 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 1.22 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	641
Facilities with at least one stable quarter	552

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	35,559
Restricted to adults (>=18 years old)	35,408
Restricted to admissions with distance between patient and facility	35,397
Restricted to those that chose a facility within 50 miles	33,563

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	-0.005***								-0.017***
	(0.002)								(0.005)
Late Shift		0.019							-0.039
		(0.028)							(0.042)
Registered Nurses on staff full-time			0.022**						0.013
			(0.011)						(0.023)
Licensed Practical/Visiting Nurses FTime				0.054					0.043
				(0.048)					(0.051)
Patient Care Technicians on staff FTime					0.007				-0.002
					(0.007)				(0.017)
Advanced Practice Nurses on staff FTme						0.098			0.095
						(0.134)			(0.136)
Dieticians on staff full-time							0.121		-0.140
							(0.079)		(0.159)
Social Workers on staff full-time								0.260***	0.424***
								(0.068)	(0.142)
Constant	0.064*	-0.051***	-0.176***	-0.075**	-0.125*	-0.052***	-0.167**	-0.307***	-0.010
	(0.037)	(0.013)	(0.065)	(0.029)	(0.072)	(0.015)	(0.080)	(0.071)	(0.048)
Observations	2,041	2,038	2,041	2,041	2,041	2,041	2,041	2,041	2,038
R-squared	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.003	0.006

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 $(\alpha, \beta, \Sigma, \delta, \eta)$ and for the latent variables $(v_i, \varepsilon_{i0}, \pi_i)$ 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 $\varepsilon_{i0}^{(k-1)}$; the equations above form a system of seemingly unrelated regressions. The posterior distributions of the parameters $\alpha, \beta, \delta, \eta$ 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 $\varepsilon_{i0}|\hat{\varepsilon}_{ij}^{(k)}, \hat{\nu}_{ij}^{(k)}, \Sigma^{(k)}$. The posterior distribution of ε_{i0} conditional on the residuals $\hat{\varepsilon}_{ij}^{(k)}$ 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$\tilde{\varepsilon}_{ij}^{(k)} = \hat{\varepsilon}_{ij}^{(k)} + \hat{\varepsilon}_{i0}^{(k-1)} - \hat{\varepsilon}_{i0}^{(k)}$.

   (b) Draw $\eta_j|\tilde{\varepsilon}_{ij}^{(k)}, \hat{\nu}_{ij}^{(k)}, \Sigma^{(k)}$. The posterior distribution of η_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 η_j from this conditional posterior. Let $\eta_j^{(k)}$ denote these draws and obtain the updated residuals$\tilde{\nu}_{ij}^{(k)} = \hat{\nu}_{ij}^{(k)} + \eta_j^{(k-1)} - \eta_j^{(k)}$.

   (c) 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 $\sigma^2_{\eta}|\eta^{(k)}_{i}$ 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 $\alpha, \beta, \delta, \eta$ 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_{η} 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 Appendix of Exhibits

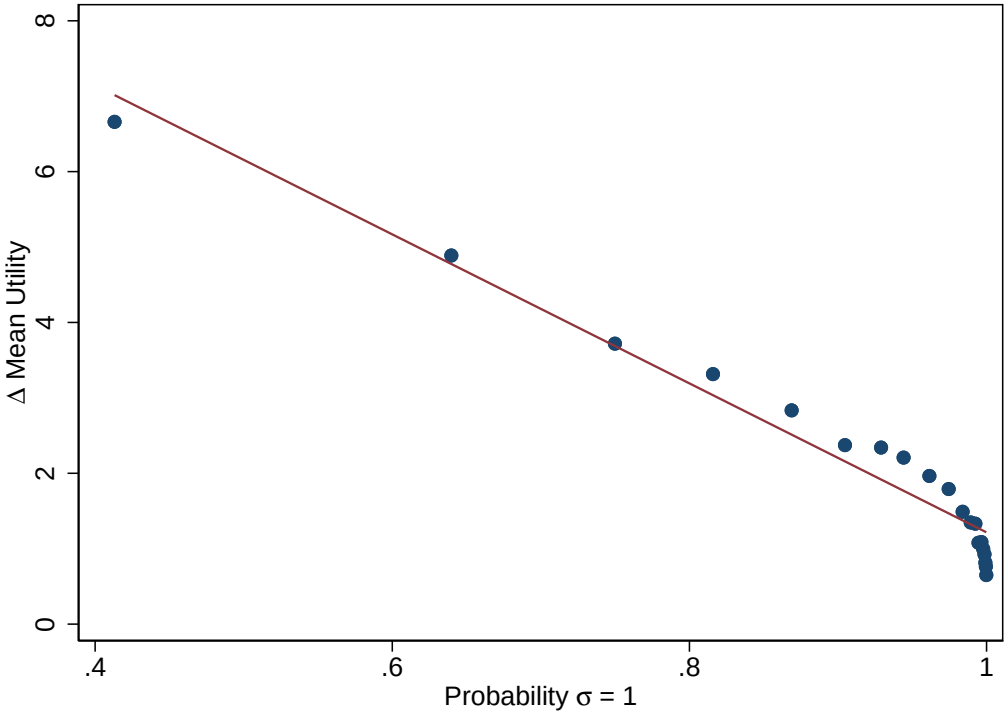


Figure D.2: Mean Utility vs Acceptance Probability