Productivity and Quality in Health Care: Evidence from the Dialysis Industry*

Paul L. E. Grieco† Ryan C. McDevitt‡

July 2014

Abstract

We show that healthcare providers face a tradeoff between increasing the number of patients they treat and improving their quality of care. To measure the magnitude of this quality-quantity tradeoff, we estimate a model of dialysis provision that explicitly incorporates a center's endogenous choice of treatment quality and allows for unobserved differences in productivity across centers. We find that, holding inputs and productivity fixed, a center can increase patient loads by 1.6 percent by reducing their quality standards such that the expected rate of septic infections increases by 1 percentage point, an 8 percent increase for the average center. Our approach provides unbiased estimates of productivity, whereas traditional methods misattribute lower-quality care to greater productivity.

JEL: D24, I1, L2
Keywords: productivity; quality variation; health care

*We thank Russell Cooper, Matt Grennan, Darius Lakdawalla, David Rivers, Mark Roberts, and Frederic Warzynski for their helpful comments. We also thank participants at the 2012 International Industrial Organization Conference (Arlington, VA), the 2012 FTC Microeconomics Conference (Washington, DC), the 2013 Meetings of the Econometric Society (San Diego, CA), the 2013 Annual Conference of the European Association for Research in Industrial Economics (Evora, Portugal), and the 2014 CIBC Conference on Firm-Level Productivity (London, ON) as well as seminar participants at Columbia University, Drexel University, the University of North Carolina, and the University of Toronto.

†The Pennsylvania State University, Department of Economics, paul.grieco@psu.edu
‡Duke University, The Fuqua School of Business, ryan.mcdevitt@duke.edu

1 Introduction

Rising healthcare expenditures have motivated spending reforms such as Medicare's prospective payment system, which ties reimbursements to a fixed amount per service irrespective of a provider's actual costs. Although such initiatives aim to limit wasteful healthcare expenses, they may inadvertently result in lower-quality care if providers cut costs by reducing the quality of their treatments. As such, measuring the tradeoff between the number and quality of treatments is crucial for understanding the impact of any potential policy change. Our paper examines this tradeoff explicitly and provides an empirical framework for measuring its magnitude within health care.

A prominent setting where such a tradeoff may be particularly acute is outpatient dialysis treatments, a process that cleans the blood of patients with end-stage renal disease (ESRD), more commonly referred to as kidney failure. Several features of this industry make it an appealing empirical setting to evaluate the relationship between productivity and quality in health care. First, dialysis treatments follow a straightforward process related to stations and staff, which allows us to closely approximate a facility's production function. Second, we observe centers' input levels (i.e., staffing and machines) and production (i.e., patient loads), which allows us to cleanly identify the relationship between inputs and outputs. Third, facilities have observable differences in outcomes that relate directly to the quality of care they provide (e.g., infection and death rates), which allows us to connect a center's inputs and outputs to its treatment quality. Fourth, payments for treatment are largely uniform due to Medicare's prospective payment system and do not depend on treatment quality, making it possible for us to isolate the effects of quality provision from price discrimination.[1] Finally, payments to dialysis facilities comprise a substantial portion of Medicare's expenditures each year — over $20 billion in 2011, or 6% of total Medicare spending — making it an important area for policy analysis.

Identifying a quality-quantity tradeoff among dialysis providers requires us to first understand the incentives centers face to provide high-quality care. Most directly, dialysis centers have an incentive to minimize the costs of treating patients under Medicare's prospective payment

[1] In 2012, Medicare instituted a Quality Incentive Program (QIP) for dialysis centers that reduces reimbursements by 2 percent if centers do not adhere to a quality standard for average hemoglobin levels and urea reductions rates, two measures of the effectiveness of dialysis treatment. However, although it is considered a novel attempt to incorporate quality standards into the Prospective Payment System, the QIP does not account for infection rates — clearly an important measure of treatment quality — in its measurement system. Furthermore, the QIP was not in effect for the timespan covering the data used in our analysis.

system, which may include providing low-quality — and hence less-costly — care. Counteracting this incentive, however, are plausible motivations for providing high-quality treatments. For instance, centers must report quality statistics to Medicare and face intermittent inspections by state regulators; Ramanarayanan & Snyder (2011) argue that these publicly available reports have a causal impact on the quality of care provided by dialysis centers.[2] In addition, patients have some choice over their dialysis providers and nephrologists make referrals based, in part, on a center's effectiveness, potentially leading centers to compete for patients by providing higher-quality care (Dai 2012). Finally, non-profit centers may have objectives for providing high-quality care unrelated to maximizing profits (Sloan 2000). In keeping with these studies, we find that states' inspection rates of facilities, centers' time since last inspection, and nephrologists' referral rates all correlate with centers' rates of infection, suggesting that centers respond to incentives to provide high-quality care. Accordingly, we include these factors in our model of quality provision, as they provide an important source of variation that allows us to identify centers' quality-quantity tradeoff.

After establishing centers' motivations for providing high-quality care, determining whether dialysis centers do, in fact, face a costly tradeoff between quality and quantity — and quantifying its magnitude — then requires overcoming a key empirical challenge: providers' endogenous choices with respect to inputs and quality may bias estimates of the quality-quantity tradeoff. That is, because centers' input choices and targeted levels of quality are not exogenously assigned, estimating the relationship between quality and quantity becomes confounded by unobserved differences in productivity, such as managerial ability or patient characteristics, which are observable to the center but not to the researcher.[3] As higher levels of productivity effectively shift out a center's production possibilities frontier, the center becomes able to simultaneously treat more patients *and* provide better care; at the extreme, a positive correlation between quality and quantity may result. Even at modest levels of dispersion, this correlation will bias reduced-form estimates of the quality-quantity tradeoff and lead researchers to underestimate facilities' true costs of improving treatment quality.

To recover the cost of providing higher-quality care in a consistent manner, we build on the structural methods for estimating center-level production functions first proposed by Olley

[2]Using a regression discontinuity approach, they show that centers who are rated just below the threshold between "worse than expected" and "as expected" on annual CMS reports have improved performance on quality metrics in the following year relative to those who narrowly surpass the threshold.

[3]While we control for observable differences in patient characteristics, unobservable differences may still affect centers' input and quality choices.

& Pakes (1996), and later extended by Levinsohn & Petrin (2003), Ackerberg et al. (2006), Gandhi et al. (2013), and others. Conceptually, we adapt these methods to incorporate a "quality-choice" stage that comes after a center's choices of labor and capital inputs. That is, after acquiring capital and training workers, a manager observes her center's expected level of productivity and chooses the quality of care to provide by, for example, stipulating guidelines for the length of treatment or cleanliness of equipment. Accommodating these endogenous quality choices in our estimation is a necessary adjustment for healthcare settings because providers under a prospective payment system may inadvertently appear more productive by treating many patients ineffectively, whereas policy makers have concerns over both productivity and efficacy.

Because we do not directly observe centers' choices regarding quality, we instead use observable measures of patient outcomes as a proxy for what those choices must have been — if high-quality care is more likely to result in better health outcomes, those outcomes are valid proxies for quality choices. Proceeding in this manner, however, presents several empirical challenges. First, health outcomes will depend not just on the choices made by centers, but also on underlying patient characteristics. To account for this, we use center-level patient characteristics to control for key sources of variation in the patient population that could affect realized health outcomes. Second, centers themselves have more information about their patients than researchers do. To the extent that unobservable differences in patient characteristics make treatment easier or more difficult, they result in an additional explanation for the differences in productivity across centers. As discussed above, we use a structural model of dialysis provision to address this issue, allowing for unobserved and heterogeneous productivity across centers. Finally, health outcomes depend on many factors beyond just the quality of care provided by centers, including a large random component, which introduces attenuation bias into standard estimation techniques. In light of this, we employ multiple measures of health outcomes (in our case, derived from centers' septic infection and mortality rates) and use an instrumental variable approach to recover the impact of quality choices on output.

From our analysis, we find a substantial quality-quantity tradeoff for dialysis treatments: a center can increase its patient load by 1.6 percent by reducing the quality of its treatments to allow a 1 percentage point increase in its expected septic infection rate, holding input levels and productivity constant. Equivalently, holding the number of treated patients constant but allowing a one standard deviation increase in a center's targeted infection rate decreases its costs

by the equivalent of five full-time employees.

In an extension of our model, we allow for a heterogeneous quality-quantity tradeoff across providers by making the production frontier's slope a function of each center's scale or input mix. Our results suggest that, although a center's scale appears to have little effect on the magnitude of the quality-quantity tradeoff, the tradeoff does vary based on the capital-labor ratio. We find that a high capital-to-labor ratio leads to a steeper-than-average tradeoff (that is, centers can increase output more for a given decrease in quality), while centers with relatively more labor have a flatter tradeoff. Although intuitive, the finding that the quality-quantity tradeoff depends on a center's input mix and scale is novel (to the best of our knowledge) in the healthcare literature. Finally, we consider the differences in productivity among non-profit centers, the industry's two dominant for-profit chains — DaVita and Fresenius — and other for-profits. Allowing for the productivity process to vary non-parametrically by ownership-type reveals little difference across centers, in contrast to other studies that have found that for-profit healthcare providers tend to have higher productivity (Kessler & McClellan 2002).

In addition to providing relevant policy analysis, this paper also contributes to the growing literature in empirical industrial organization on the estimation of production functions. These methods have a long history in economics, with much prior work focused on econometric issues related to selection and simultaneity bias.[4] In light of this, more recent work has developed structural techniques that use centers' observed input decisions to control for unobserved productivity shocks and overcome endogeneity problems.[5] We extend these methods to incorporate observable measures of output quality into the production function, which is necessary for healthcare applications. To our knowledge, we are the first to apply these methods to a healthcare setting with the goal of measuring a quality-quantity tradeoff.[6] Our work also connects to the literature on firms' quality choices within regulated industries (Joskow & Rose 1989, Crawford & Shum 2007), as measuring the tradeoff is central to understanding the full impact of regulations.

The remainder of our paper continues in the following section with a description of the outpatient dialysis industry and our data sources. Section 3 develops our structural model for

[4]See Syverson (2011) for a recent review.

[5]See, for example, Olley & Pakes (1996), Ackerberg et al. (2006), and Levinsohn & Petrin (2003).

[6]Romley & Goldman (2011) consider quality choices among hospitals using a revealed-preference approach rather than outcome-based quality measures. Gertler & Waldman (1992) estimate a quality-adjusted cost function for nursing homes. Lee et al. (2012) use a structural approach to measure the impact of healthcare IT on hospital productivity, but do not consider output quality.

estimating a production function in the presence of an endogenous quality choices, while Section 4 outlines our methods for bringing the model to the data. Section 5 presents our estimation results. Finally, Section 6 concludes with a discussion of our findings' implications for policy analysis.

2 Empirical Setting and Data Description

The demand for dialysis treatments comes from patients afflicted with end-stage renal disease (ESRD), a chronic condition characterized by functional kidney failure that results in death if not treated properly. Patients with ESRD effectively have only two treatment options, a kidney transplant or dialysis. Due to the long wait-list for transplants, however, nearly all ESRD patients must at some point undergo dialysis, a medical process that cleans the blood of waste and excess fluids. Patients can receive different dialysis modalities, with hemodialysis, a method that circulates a patient's blood through a filtering device before returning it to the body, constituting 90.4 percent of treatments (Center for Medicare and Medicaid Services). The typical dialysis regimen calls for three treatments per week lasting 2 to 5 hours each, with the duration dictated by a nephrologist to meet clinical thresholds. Although individual patient characteristics, such as the severity of ESRD, influence treatment lengths, treatment frequency rarely deviates from the standard protocol of three sessions per week.[7]

Patients receiving dialysis in the United States primarily do so at free-standing dialysis facilities, which collectively comprise over 90 percent of the market (USRDS 2010).[8] Medicare's ESRD program, instituted by an act of Congress in 1973, covers the majority of these patients; notably, all patients with ESRD become eligible for Medicare coverage, regardless of age, and the program now includes over 400,000 individuals. Today, Medicare spends more than $20 billion a year on dialysis care — approximately $77,000 per patient annually — which constitutes more than six percent of all Medicare spending despite affecting fewer than one percent of Medicare patients (ProPublica 2011).

Beginning in 1983, Medicare has paid dialysis providers a fixed, prospective payment — the "composite rate" — for each outpatient treatment delivered, up to a maximum of three sessions

[7] Generally, Medicare reimbursements are limited to three sessions per week. Hirth (2007) has argued that this limit may lead to inadequate dialyzing of some patients and failure to experiment with alternative dialysis schedules (such as shorter but more frequent sessions).

[8] Other options for receiving dialysis include hospital emergency rooms and in-home treatments.

per week per patient. Initially, the payment rate did not adjust for quality, length of treatment, dialysis dose, or patient characteristics, though Medicare began to adjust payments based on patient characteristics in 2005. Many have speculated that this payment structure affects the quality of dialysis treatments, such as Hirth (2007) who states, "Research on the relationship between payment for dialysis and the quality and nature of the process is not definitive, but there is evidence that practices such as dialyzer reuse, staffing reductions, and scheduling inflexibilities (fewer dialysis stations per patient) were encouraged by financial pressures."

Dialysis treatments require constant supervision by trained medical professionals, as patients must remain connected to a station for several hours to filter impurities and remove excess fluid from their blood. Prior to treatment, staff connect the machine to a patient by inserting two lines into a vascular access and assess his condition. During treatment, staff must continually monitor patients to evaluate their condition (e.g., blood pressure) and to treat symptoms that arise (e.g., hypotension). Following treatment, staff disconnect the patient from the station and assess his condition a final time before discharge; they then clean and sanitize machines in preparation for the next patient. As a result of this hands-on care, the cost per patient treated necessarily increases with the average amount of time devoted to treatments and cleaning. Labor costs, which consist largely of nurses and technicians' wages, reflect this, accounting for approximately 70 to 75 percent of a facility's total variable costs (Ford & Kaserman 2000).

Centers employ different types of labor, with registered nurses (RNs) comprising the majority of staff. Technicians, who have less-extensive training than RNs, also treat patients but can do so with only a high-school diploma and in-house training (though they must eventually pass a state or national certification test). Notably, centers cannot quickly react to changes in productivity by hiring more workers due to persistent nurse shortages and the additional training and certification required to become a dialysis nurse. As an example of this, for-profit dialysis chain Fresenius claims in an internal report that, "In practical terms, nurse staffing turnover is a costly proposition because of the training required to bring new hires up to speed."[9] Centers also must have board-certified physicians as medical directors, though often have no physician on site. Medicare does not mandate a specific staffing ratio for dialysis centers, although some states do.

In addition to staffing levels, another significant decision for dialysis facilities is the number of stations to have in operation. Centers vary widely in terms of size, ranging from 1 to 80

[9]See "FMS Pathways: Nursing Shortage."

stations. Based on industry reports, a typical dialysis station costs $16,000 and has a useful life of approximately seven years (Imerman & Otto 2004).

Along with labor and capital decisions, centers must also choose how much effort to put towards providing high-quality care, the central focus of our study. Quality in this setting can mean many things, from the effectiveness of dialysis in reducing urea from blood to the comfort of patients during treatment. We focus on a single dimension of quality, a patient's risk of contracting a septic infection, as such infections are particularly costly and life-threatening for patients. Nationally, the rate of hospitalization for septic infections while undergoing dialysis is over 12 percent per year.

Importantly, centers can allocate their resources in ways that affect patients' risk of infection. For example, infections related to dialysis stem in large part from the exposure of a patient's blood during treatment, making the cleanliness of the dialysis center and its stations a key factor. Because dialysis sessions require up to one hour of preparation and cleaning, the center has considerable control over its targeted infection rate, as health professionals who follow straightforward procedures can virtually eliminate their patients' risk of contracting infections (Patel et al. 2013, Pronovost et al. 2006).[10]

Reducing the risk of infection, however, comes with the opportunity cost of treating fewer patients due to the resource constraints of the facility, which may ultimately reduce the center's profits. That is, because a facility's reimbursement per treatment does not vary with its duration under Medicare's prospective payment system, a facility's profit per treatment decreases as treatment and cleaning times — and, hence, labor costs — increase. In essence, the tradeoff faced by centers stems from their choice to either improve treatment quality or decrease costs.[11]

An extensive medical literature has examined these tradeoffs in health care more generally, mostly from an accounting perspective (Weinstein & Stason 1977). Morey et al. (1992), for example, found that a 1% increase in the quality of care increased hospital costs by 1.3%; Jha et al. (2009) found that low-cost hospitals had slightly worse risk-adjusted outcomes for common medical conditions; and Laine et al. (2005) found that efficient wards had issues with quality

[10]There may also be differences in the quality of dialysis stations in regard to how efficiently they can be cleaned, although this is not highlighted in industry reports or CDC guidelines, which emphasize thorough cleaning and sterilization of machines, the use of appropriate disinfectants, and monitoring and appropriate cleanup of blood and other fluid spills (CDC 2001).

[11]Critics allege that facilities may sacrifice their quality of care in pursuit of efficiency, turning over three to four shifts of patients a day. And while policy makers contend that technicians should not monitor more than four patients at once, patient-to-staff ratios exceed this guideline in many facilities. At the extreme, inspection reports allege that some clinics have allowed patients to soil themselves rather than interrupt dialysis (ProPublica 2011).

Table 1: Summary Statistics.

Variable	Mean	St. Dev.
Patient Years	50.856	31.913
FTE Staff	13.484	7.769
Net Hiring	0.182	3.868
Zero Net Hiring	0.127	0.333
Stations	18.612	7.877
Zero Net Investment	0.923	0.266
Septic Infection Rate	12.504	6.399
Death Rate Ratio	1.041	0.405
Number of Centers	4,270	
Number of Center-Years	18,295	

for conditions that require time-consuming nursing procedures. In this paper, we consider the tradeoff using a structural model of production intended to control for possible confounding factors, such endogenous quality decisions and the measurement error that arises from using observable outcomes as proxies for centers' unobservable choices.

2.1 Data Sources

Our primary dataset comes from the Centers for Medicare and Medicaid Services (CMS) which contracts with the University of Michigan's Kidney Epidemiology and Cost Center to compile customized reports for each dialysis facility across the country. In December 2010, ProPublica, a non-profit organization dedicated to investigative journalism, obtained these reports under the Freedom of Information Act and posted them online. We systematically downloaded all individual reports covering 2004-2008 and constructed a usable dataset. The data include detailed center-level information on aggregated patient (e.g., age, gender, co-morbid conditions, etc.) and facility (e.g., number of stations and nurses, years in operation, etc.) characteristics.

Table 1 presents selected summary statistics from the data, and several variables deserve note. First, Medicare analyzes individual patient records and calculates the number of patient-years attributable to each center (e.g., a patient treated at a center for six months is accounted for as one half of a patient-year). We use this variable as our measure of output, as it provides an accurate record of dialysis provision that accounts for partial years of service due to death, transfers, transplants, newly diagnosed patients, and so forth.[12] We also use the number of

[12]Since treatment is mostly standardized at three treatments per week and the goal of dialysis is to clean the blood,

full-time equivalent (a weighted mix of full-time and part-time) employees at each center and the number of dialysis stations as our measures of labor and capital inputs, respectively. In terms of capital stock, the average number of dialysis stations used by a center is 18, making the purchase of a new machine a significant investment; reflecting this, centers have zero net investment for 92 percent of the center-year observations in the data. In terms of hiring, centers, on average, increase their staff by the equivalent of one full-time employee each year, while 12.7 percent of centers have no net change in employment.

We use a center's hospitalization rate from septic (blood) infections as our primary measure of quality, which averages 12.5 percent per year and has a standard deviation of over 6 percent. The high infection rate reflects the severe vulnerability of the ESRD patient population relative to the general population, for whom septic infections are rare outside of hospital settings.[13] Although some of this variation is due to deliberate choices made by centers regarding their quality of care, other factors outside of a center's control also influence infections. For this reason, we control for the characteristics of each center's patient mix, which we discuss at length in Section 4. We also allow for centers' quality choices to depend on their unobserved productivity, which may be related to the susceptibility of their patients to infections that is observed by the center but not the researcher. Moreover, the number of infections may have a large random component beyond patient characteristics and quality choices, so we must account for this measurement error in the model. To do so, we use the ratio of deaths to expected deaths as an alternative measure of quality.[14] In our regressions, we use the septic infection rate as our preferred measure of quality because it is the quality outcome most closely tied to centers' actions during dialysis treatments, whereas deaths may have many other causes besides infection from a dialysis machine.

2.2 Incentives for Quality

Before moving to the structural model, we first consider whether variation in quality across centers stems from deliberate managerial choices. As discussed above, the question of whether external forces influence dialysis facilities' incentives for providing high-quality care has been investigated in previous studies (Ramanarayanan & Snyder 2011, Sloan 2000). To illustrate this

we do not consider differences in treatment times as output variation.

[13]For some perspective on this vulnerability, in 2009 remaining life expectancy for a 50-54 year old in the US general population was 27.1 years, while for the population of ESRD patients it was 7.1 only years (USRDS 2013).

[14]The center-level expected death rate is calculated by Medicare using individual patient characteristics.

in our own data, we consider a series of fixed-effects regressions of the septic infection rate on plausibly exogenous drivers of quality and patient characteristics. These regressions use within-center variation to show that centers facing stronger incentives to provide higher-quality care have better outcomes, suggesting that centers do, in fact, have an ability to adjust their level of quality when it is in their best interest to do so.

First, we look at the time since a facility was last inspected by a state authority.[15] As centers with many reported violations face the possibility of losing certification, a recently inspected facility has an incentive to correct any reported deficiencies. Reflecting this, Column (1) of Table 3 shows that the infection rate increases with the time since inspection by a statistically significant 0.15 percentage points per year, or about 11.8% at the mean. Note that the most likely source of endogeneity bias for this regression — that regulators target centers with poor outcomes for more-frequent inspections — works against this result, making it a conservative estimate.

Our second test uses variation in inspection rates across states, which differ based on funding and other local regulations — and in that sense not confounded by an individual center's unobserved characteristics. As shown in Column (2), we find that an increase in the state inspection rate is associated with a lower infection rate: centers in states with an inspection rate one standard deviation above the mean have a 0.1 percentage point lower infection rate, on average, which is about 1% lower at the mean.

Finally, we consider the possibility that the referral rates of nephrologists might affect treatment quality.[16] Presumably, nephrologists will not refer patients to a center with a poor record of quality, and they may also serve as a check on facilities after referral (e.g., a nephrologist will act as an advocate for his patients should they receive poor-quality treatments). Column (3) shows that a higher rate of referrals is associated with a lower infection rate, with the effect similar in magnitude to that found for the state inspection rate.

Although each of these regressions suffers from its own particular shortcomings (e.g., both referrals from nephrologists and the likelihood of inspection are potentially endogenous with respect to centers' quality choices), taken together they provide consistent evidence that plausibly exogenous drivers of quality do, in fact, influence a facility's provision of quality. Note also

[15] We know only the year of the last inspection, so if a center is inspected this year there are "zero" years since its last inspection.

[16] Unfortunately, the referral rate is available for only 3 years of our 5 year panel, severely reducing the number of observations in the data; for this reason, we do not include it in our baseline model.

Table 2: Quality Drivers.

Variable	Mean	St. Dev.	N
Time Since Inspection	1.634	1.813	18,221
State Inspection Rate	29.348	11.648	18,295
% Patients Referred by Nephrologist	69.445	20.346	11,372

Table 3: Infection Rate Fixed Effects Regressions.

	(1)	(2)	(3)	(4)	(5)
Time Since Inspection	0.148			0.141	0.142
	(0.031)			(0.031)	(0.047)
State Inspection Rate		-0.010		-0.007	-0.002
		(0.005)		(0.007)	(0.694)
% Patients Referred by Nephrologist			-0.009		-0.009
			(0.004)		(0.004)
Patient Characteristics	Yes	Yes	Yes	Yes	Yes
Center Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	18221	18295	11372	18221	11342
R^2	0.572	0.573	0.662	0.573	0.663

that each specification controls for facility-level fixed effects, so any time-invariant institutional factors (e.g., the facility is in a region with sicker patients) are accounted for in the regressions.

That infection rates depend on incentives for quality provides us with an important source of variation we will use to identify the quality-quantity tradeoff. In short, we will compare centers with similar productive capacities but different incentives to provide high-quality treatments to estimate the relationship between quantity and quality, with our structural model explicitly accounting for the different incentives centers face to provide high-quality treatments, as we discuss in Section 4.2.

3 A Model of the Quality-Quantity Tradeoff in Dialysis

To measure the relationship between a center's productivity and its treatment quality, we propose and estimate a structural model of dialysis provision. In doing so, we account for both the standard endogeneity problems associated with using observed input choices to estimate production functions and the additional problem introduced by a center's endogenous choice of treatment quality. The complication related to endogenous quality decisions stems from the

unobserved (to the econometrician) choice made by centers that receive positive shocks to their productivity: they may choose either to treat more patients, or to treat current patients more intensively. If highly productive centers choose to provide higher-quality care for their patients, naïve estimates of the quality-quantity tradeoff will be biased, leading us to underestimate the true cost of delivering high-quality care.

To control for this potential source of bias, we extend the work of Olley & Pakes (1996) and Ackerberg et al. (2006) by incorporating centers' endogenous quality targets. Because we only observe noisy measures of quality in our data, we also control for measurement error in quality choices, proxied for by center-level hospitalization rates for septic infection. Specifically, the attenuation bias introduced by measurement error in quality choices would cause us to underestimate the magnitude of the quality-quantity tradeoff, which we correct for using an instrumental variable approach.

3.1 The Production Technology

We model the provision of dialysis treatments as a stochastic two-output production process, where the outputs are the number of patients treated and the quality of treatments. As centers face a tradeoff between quality and output, they operate under a production possibilities frontier relating the number of patients they treat and the level of quality they provide. Formally, we define centers' production possibilities frontier as

$$T(\tilde{y}_{it}, \tilde{q}_{it}) \leq F(k_{it}, \ell_{it}, \omega_{it}). \tag{1}$$

The production function $F(\cdot)$ is the most familiar part of this constraint; it governs how the center's inputs of (log) capital, k_{it}, and (log) labor, ℓ_{it}, as well as the center's unobserved assessment of its own productivity, ω_{it}, determine its overall capacity for production. In estimation, the number of stations is our measure of capital and the full-time equivalent number of nurses and technicians is our measure of labor. The unobserved productivity term, ω_{it}, is intended to account for all factors observable to the center but not to the econometrician that impact its production possibilities, such as the center's square footage, managerial ability, labor or capital quality, or patient characteristics; this last source of unobserved productivity is particularly important in a healthcare setting such as dialysis where patient sorting may induce large differences in each center's ability to treat patients. For example, highly educated

patients may follow treatment protocols more closely and therefore require less attention from technicians while being treated. Although our data will allow us to control for a number of key patient characteristics, some will remain unobserved and must be captured by ω_{it}. We follow the productivity literature in modeling ω_{it} as a scalar representing Hicks-neutral total factor productivity.[17]

The transformation function, $T(\cdot)$, determines how the center divides its productive capacity between two output goals: the number of patients treated and the infection risk of those patients. The first output, $\tilde{y}_{it}$, is the center's targeted (log) number of patient-years for the period. The second output, $\tilde{q}_{it}$, represents the quality of its treatments, which we model as a scalar index representing the center's targeted infection rate.[18]

For a variety of reasons, the center is unable to perfectly anticipate its output or infection rate. Instead, it chooses targets $(\tilde{y}, \tilde{q})$ rather than what we ultimately observe as output and infection outcomes (y, q). For output, this is a standard approach in the productivity literature, where an "unanticipated" productivity shock or measurement error term is typically included in the production process. For quality, this unanticipated shock between the targeted and realized infection rate is closely tied to the nature of infections, which are highly stochastic in nature. While centers can implement procedures to reduce infections (i.e., choose a lower $\tilde{q}$), many factors outside the center's control also influence the observed infection rate q. For example, the ability of a patient's immune system to fight off a particular bacteria may depend on his or her previous exposure, which is less likely to be known by the center's manager when she sets her quality standards. Put differently, although the center is able to take actions to reduce the risk of infections, $\tilde{q}$, ultimately the tradeoff between such actions and the ability to treat patients is governed by $T(\cdot)$. Recovering this relationship between quantity and quality is the primary goal of our econometric analysis.

[17] Ideally, we would include multiple dimensions of unobserved productivity — for example, separate terms for labor or capital productivity, or separate terms for producing output or quality. Recent work by Doraszelski & Jaumandreu (2012) and Zhang (2014), building in part on Gandhi et al. (2013), propose using a center's first-order conditions of profit maximization to allow for multiple dimensions of heterogeneity. Unfortunately, since we do not explicitly model center objectives, this approach cannot be applied in our setting.

[18] Note that "quality" here reflects how carefully the center acts to reduce the risk of infection, not the infection rate itself. As we discuss in Section 2, "quality" can have many dimensions for patients, such as the likelihood of becoming sick, the amount of time spent waiting for treatments, the convenience of the center's operating hours, or even having televisions available during treatments. Despite this, we focus on one specific dimension of quality, low septic infection risk, that is arguably the most prominent dimension of quality due to its severe impact on patients' well-being.

3.2 The Timing of Dialysis Center Decision Making

In their seminal paper, Olley & Pakes (1996) use capital investment as a proxy for unobserved productivity, arguing that centers with greater productivity, all else equal, will make larger investments. Given this intuition, differences in investments will reflect differences in productivity. Although a natural assumption for an industry such as telecommunications equipment, this approach is not appropriate for dialysis centers because investment in new stations is too infrequent, as over 90 percent of the center-year observations in the data have zero investment. In light of this, we instead use centers' hiring decisions, which provide a more natural proxy in our setting. Because industry reports suggest that nurses and technicians employed by dialysis centers require extensive training and credentialing, costs and time lags affect centers' hiring and layoff decisions. Therefore, we regard labor as a dynamic variable, which allows us to use a center's (net) hiring decision to recover ω_{it}.[19]

In contrast to labor choices, a center can quickly adjust the quality of care it provides. For example, to improve quality, a manager could advise her center's staff to take extra precautions when treating patients, or to reduce quality by placing less emphasis on cleanliness and more on speed (Pronovost et al. 2006). At the same time, even though a center can dictate these policy changes more quickly than it can make hiring or investment changes, a lag still exists between a manager's decision about quality and its actual implementation.

A center's manager makes investment, hiring, and quality choices based on her center's capital stock, labor productivity, and a vector of other observable characteristics, x_{it}. Importantly, the components of x_{it} may affect the center's policy function even though they do not impact production directly, and may include the extent of competition in the market, the center's taste for quality via it's non-profit and ownership status, and other differences in incentives discussed in Section 2.2. This leads to the timing assumptions of our model:

1. *Quality Choices Made.* Centers begin the period knowing their current levels of capital, k_{it}, and labor, ℓ_{it}, as well as a vector of observable state variables, x_{it}, which affect the center's preferences but not its productive capacity. (For example, it observes whether it is a for-profit or non-profit, and the characteristics of the surrounding market.) In the model, centers also observe ω_{it}^q, their productivity this period given their information set

[19]Note that this assumption conflicts with OP's conception of labor representing an immediately flexible input, though the distinction fits our setting. Although centers have some flexibility to increase nurses' hours rather than hire additional staff, their ability to do so is constrained by states' overtime regulations (Bae et al. 2012).

at the beginning of the period. Based on this information, the center choose its targeted level of output and quality for the period, $(\tilde{y}, \tilde{q})$.

2. *Production Occurs.* Based on its chosen target, the center treats patients and observes realized outcomes for patient loads and infections, (y, q). The center also updates its beliefs about its productivity, ω^h.

3. *Hiring and Investment Choices Made.* After observing production, the center's state is updated to reflect what has been learned about its productivity, becoming $(k_{it}, \ell_{it}, x_{it}, \omega^h_{it})$. With this information, the center decides on hiring, h, and investment, i; newly hired workers and invested capital become available at the start of period $t+1$.

4. *New State Realized.* In line with the literature, we assume centers' expectations of productivity follow an exogenous Markov process between periods t and $t+1$,

$$E[\omega^q_{i,t+1}|I_{i,t}] = E[\omega^q_{i,t+1}|\omega^h_{i,t}],$$

where I_{it} represents center i's information set at the end of period t. Also following the literature, we assume this process is stochastically increasing in $\omega^h_{i,t}$ (Pakes 1994) and that the state variable x_{it} moves according to an exogenous Markov process (similar to De Loecker 2011).

In this setting, unobserved productivity encompasses any factor that allows a center to treat more patients given its observable characteristics and quality target. For instance, a center's patients may follow treatment protocols more closely than other centers' patients do, which then frees the center either (i) to treat more patients because it devotes less time to dealing with complications that arise, or (ii) to spend additional time treating existing patients more intensively, which ultimately improves outcomes but does not appear in raw productivity measures, such as output-to-labor ratios.

3.3 The Center's Quality Decision

The center enters each period's quality-choice stage with a state variable (k, ℓ, x, ω^q). Based on its expected productive capacity, it chooses its targeted level of output and quality for the period, $(\tilde{y}, \tilde{q})$. We assume that demand for dialysis is inelastic, which fits with the tight capacity in the industry (which we model though the production frontier) and wait-lists for treatment in

many markets. After making its decision, production occurs and the center observes the number of patients served and its infection rate for the period. In addition, the center updates its beliefs about its productivity. Specifically, it will observe three new variables,

$$y = \tilde{y} + \varepsilon^y,$$
$$q = \tilde{q} + \varepsilon^q,$$
$$\omega^h_{it} = \omega^q_{it} + \varepsilon^\omega_{it},$$

the first two of which will determine period payoffs, while the last one is the center's updated productivity. We assume that $(\varepsilon^y_{it}, \varepsilon^q_{it}, \varepsilon^\omega_{it})$ are mean zero and uncorrelated with the information available to the center as it makes its quality choice, though the components of this vector may be correlated with each other. That is, conditional on both ε^y_{it} and ε^q_{it} being positive, we would expect the center to raise its assessment of its own productivity.[20]

Because the center will not learn about $(\varepsilon^y, \varepsilon^q, \varepsilon^\omega)$ until it treats patients, it must optimize its quality choice under uncertainty. Because we assume quality choices are fully flexible and that the quality and output outcomes do not affect future states, the center's quality-choice problem does not have dynamic links. As such, the center chooses its expected quality and output to solve the static problem,

$$\begin{aligned} \pi(k, \ell, x, \omega^q) = \max_{\tilde{y}, \tilde{q}} \; & E[\rho(y, q, k, \ell, x)] \\ \text{subject to: } & T(\tilde{y}, \tilde{q}) \leq F(k, \ell, \omega^q) \\ & y = \tilde{y} + \varepsilon^y \\ & q = \tilde{q} + \varepsilon^q. \end{aligned} \tag{2}$$

Here, $\rho(\cdot)$ represents the center's return from its output and infection rate in the current period given its state variables.[21] As dialysis centers' objectives are difficult to model directly, we remain agnostic as to the precise form of this function. For instance, even for-profit centers may see value in treating patients as effectively as possible rather than strictly maximizing per-period profits. Moreover, centers face potential tort litigation or additional oversight if patients have

[20] Without loss of generality, we could allow productivity within the period to evolve according to an unknown stochastically increasing Markov process. Letting it evolve according to a random walk is notationally convenient because $E[\omega^h|\omega^q] = \omega^q$.

[21] Note that the center payoff is determined by observed outcomes, not its underlying quality, which is difficult to directly observe.

particularly poor outcomes.

We assume that the per-period payoffs incorporating these considerations can be summarized by $\rho(\cdot)$, which is increasing in the center's two outcomes, y (output) and q (quality). The state vector x may play a critical role in determining how centers view the relative importance of each outcome. For example, given the structure of the prospective payment system, for-profit centers may place a higher priority on y relative to q than non-profits. Alternatively, if a center has recently been inspected — or faces a looming inspection — it may place a greater emphasis on quality this period. Allowing $\rho(\cdot)$ to depend on x enables centers with the same productive capacity to have different quality policies within the model, and this variation in centers' policies provides identifying power for estimating the quality-quantity tradeoff.[22]

Our assumption that the number of patients treated in the current period does not affect the state of the center in subsequent periods is common in the literature. Our assumption that current levels of quality have no dynamic implications is stronger, owing to the possibility of long-lasting reputation effects; however, one could imagine accounting for the effects of reputation through per-period profits (e.g., the center immediately pays for the discounted future costs of low-quality performance). Extending the model to allow for a long-run reputation would require an additional state variable and a precise model of how quality affects reputation.

Before moving to the firm's hiring choice, the following lemma establishes that the return to labor is increasing in productivity, which will be important for establishing the invertibility of the hiring policy (in Proposition 1 below) and motivates using hiring as a proxy for productivity as part of our estimation strategy.

Lemma 1. *The center's expected per-period return to labor is increasing in ω^q; that is, $\frac{\partial \pi}{\partial \ell}$ is increasing in ω^q.*

We provide the proof in the appendix. Intuitively, increases in both ℓ and ω^q relax the production constraint, which, due to non-satiation, must always bind if the center is acting optimally. This binding constraint implies that the return to increasing ℓ is increasing in any variable whose only effect is to relax the constraint further, such as ω^q.

3.4 The Center's Hiring and Investment Problem

After production, the center makes its hiring and investment decisions for the following period.

[22] Without variation in x, identification would rely exclusively on the timing assumption that centers choose quality when observing ω^q but decide to hire when observing ω^h.

The Bellman equation for this choice is

$$V^h(k,\ell,x,\omega^h) = \max_{i,h} -c(i,h) + \beta E[V^q(k+i,\ell+h,x',\omega^{q'})|k,\ell,\omega^h,i,h], \tag{3}$$

where i is net investment and h is net hiring. The function $c(\cdot)$ captures adjustment costs for investment and hiring.[23]

Our decision to model hiring with a lag reflects the institutional detail that training and other adjustment costs are significant in the dialysis industry relative to the difficulty of altering current workers' on-the-job incentives to strive for either higher output or higher quality. We follow the literature in assuming that hiring costs are differentiable and convex, except possibly with a fixed adjustment cost at $h = 0$ where a "zone of inactivity" in which the center does not adjust its staffing level for a range of productivity levels may obtain.[24]

The function $V^q(\cdot)$ represents the value of the center at the start of the period,

$$V^q(k,\ell,x,\omega^q) = \pi(k,\ell,x,\omega^q) + E[V^h(k,\ell,x,\omega^h)|k,\ell,x,\omega^q].$$

We adopt this slightly cumbersome notation because the center's perception of its own productivity evolves over the course of the period from ω^q to ω^h as a result of the center observing its own production process.

Based on the lumpiness of investment in this industry, we assume that the choice of next period's capital is discrete. By contrast, we view the hiring choice as effectively continuous. This seems reasonable given the number of nurses in the industry and the ability to adjust nurses' hours from period to period. Under these assumptions, the following proposition establishes that, for a given level of investment, a one-to-one relationship exists between ω^h and the center's hiring choice, $h(k,\ell,x,\omega^h)$.

Proposition 1. *For any fixed investment level ι, the center hiring function $h(k,\ell,x,\omega^h)$ is*

[23] We can also allow $c(i,h)$ to be zero, in which case time-to-build is the only hiring and investment friction.

[24] Clearly, a fixed adjustment cost at zero means that we cannot invert the hiring function at $h = 0$, and these observations must be dropped. However, under the model, this truncation only affects efficiency. On the other hand, unanticipated zones of inactivity (say, a maximum allowable level of hiring) have the potential to bias our estimates. The discussion on possible failure of the investment proxy in Levinsohn & Petrin (2003, 321) applies to our hiring proxy. See also (Pakes 1994, Remark 2). Recall that in our setting hiring is zero in 12.7 percent of firm-year observations (Table 1).

invertible with respect to ω^h *on the domain* $\{(k, \ell, x, \omega^h) : i(k, \ell, x, \omega^h) = \iota\}$,

$$\omega^h = h_\iota^{-1}(k, \ell, x, h).$$

The proof of this theorem makes use of results in Theorem 1 from Pakes (1994) and Appendix C from De Loecker (2011). We show that, given Lemma 1, our problem can be written in such a way that we can apply Theorem 1 from Pakes (1994) directly, where hiring is the inverting variable instead of investment. We have the added complication, however, of controlling for centers' discrete investment choices: if a center invests in a new station, the cost of this new investment may lead the center to hire fewer nurses than it might in a situation where it had lower productivity but did not choose to invest. To account for this possibility directly with our data, we can isolate cases when centers make the same investment choice (e.g., add one new station) and conclude that those within a given investment tier that hire more workers must have higher productivity. Furthermore, because $i = 0$ in over 92% of the observed periods in our data, any complication related to this point will be comparatively mild.

Our strategy for controlling for unobserved productivity relies on using hiring choices as proxy, which is grounded in the relationship between net hiring and productivity in the data. Here, labor productivity (i.e., output-per-worker) and hiring have a positive correlation of 0.36, which is robust to considering only positive or negative hiring observations. While ω^h represents total factor productivity, it nevertheless provides some reassurance that the posited relationship between hiring and productivity exists for readily observable productivity measures.

4 Estimation

We use the model above to estimate the underlying parameters of the production function and recover each center's unobserved productivity in every period, adopting the following parsimonious functional forms to describe the transformation and production functions,

$$T(\tilde{y}_{it}, \tilde{q}_{it}) = \tilde{y}_{it} + \alpha_q \tilde{q}_{it} \tag{4}$$

$$F(k_{it}, \ell_{it}, \omega_{it}^q) = \beta_k k_{it} + \beta_\ell \ell_{it} + \omega_{it}^q. \tag{5}$$

In short, we follow the common practice in the literature of assuming a Cobb-Douglas production function, where ω_{it} is a Hicks-neutral technology shifter. For the transformation function, we

also assume a Cobb-Douglas-like specification that parameterizes the production possibilities frontier by assuming that reducing the infection rate 1 percentage point (i.e., increasing $\tilde{q}_{it}$ by 1) will reduce expected output by a factor of α_q, which is constant across centers.

This specification allows us to connect a center's quality target to its observable outcomes in a direct manner. By increasing the effort it puts towards providing high-quality treatments, the center incurs additional costs but increases the likelihood of delivering better treatment outcomes — that is, the center may treat fewer patients with the same level of inputs. On the other hand, a change in inputs or productivity shifts the production possibilities frontier but does not alter the relative transformation between outputs. For instance, a center with healthier patients recognizes that its production frontier has shifted outwards, but still faces a tradeoff between treating more patients at a given level of quality or providing higher-quality care for a given number of patients.

In the data, we do not observe centers' expected output and quality. Instead, we observe realized patient loads and infection rates, which are subject to both measurement error and unanticipated shocks. To account for this, we assume that observed output is $y_{it} = \tilde{y}_{it} + \varepsilon^y_{it}$ and the observed infection rate is $q_{it} = \tilde{q}_{it} + \varepsilon^q$. Substituting these into (1), we arrive at the linear equation

$$y_{it} = -\alpha_q q_{it} + \beta_k k_{it} + \beta_\ell \ell_{it} + \omega^q_{it} - \alpha\varepsilon^q_{it} - \varepsilon^y_{it}. \tag{6}$$

Estimating (6) by ordinary least squares with data on (y, q, k, ℓ) would imply the composite error term is $\omega^q_{it} - \alpha\varepsilon^q_{it} - \varepsilon^y_{it}$, making two sources of bias immediately apparent — one due to ω_{it}, and the other due to ε^q_{it}.

First, we have the well-known endogeneity problem associated with estimating production functions: because ω_{it} is observed by the center but not the econometrician, it may be correlated with the center's capital and labor choices. Second, our approach adds an additional endogeneity problem, as ω^q_{it} may also affect the center's quality target. As a result, OLS estimates of (6) are inconsistent. Some classical methods of correcting for this endogeneity include finding instruments for capital, labor, and quality, or assuming productivity is fixed over time (i.e., $\omega^q_{it} = \omega_i$) and using a fixed-effects estimator (Mundalk 1961). In application, these approaches have had limited success. Although input prices would seem to be appropriate instruments for capital and labor choices, they often have weak predictive power and the data can be difficult to obtain. A valid instrument for quality targets that is uncorrelated with unobserved productivity

would be even more challenging to find. Furthermore, while the fixed-effects assumption is relatively easy to implement, it is quite strong and would not resolve the endogeneity problems if changes in productivity are responsible for changes in input (or, in our case, quality) choices.

To address these issues in a manufacturing context, Olley & Pakes (1996) propose an explicit structural approach to estimate the production process that uses observed center decisions as proxies for unobserved productivity shocks, with the basic ideas behind their method extended further by Levinsohn & Petrin (2003) and Ackerberg et al. (2006).[25] We adapt this approach to a healthcare context. In our setting, productivity differences may be due to unobserved differences in inputs and management practices, but also due to unobservable differences in patient characteristics that could make achieving a given level of quality more difficult.

A second source of bias results from the error term, ε_{it}^{q}. Although this error is unanticipated by the center, it is, by definition, correlated with our proxy for treatment quality, the observed infection rate q_{it}. This form of classical measurement error will lead to attenuation bias, moving our estimate of α_q towards zero. We will address this issue by instrumenting for q_{it} with a second proxy for treatment quality, the center's "unexpected" death rate, which we construct as the actual death rate over the expected rate. If the unobservable (to the researcher) factors that cause infections are uncorrelated with those that cause death, then the instrument is valid and we can consistently estimate α_q.[26] In the event that they are correlated and our instrument is invalid, our estimate of α_q remains biased towards zero and is best understood as a lower bound, making our results conservative.

Estimation proceeds in three steps. First, because we do not observe quality directly, we derive an appropriate proxy for quality based on center-level outcomes. Second, we specify the observed policy shifters, x_{it}, which we include in the center's hiring function. Finally, we adapt the standard two-stage estimation strategy to incorporate an endogenous quality choice with a noisy proxy.

[25] A second approach to production function estimation comes from the dynamic panel literature (e.g., Blundell & Bond 2000); Ackerberg et al. (2006) provides a comparison of these approaches.

[26] It is possible that the unobservable factors related to contracting an infection are correlated with the center's death rate. Note, however, that the unobservable factors from the researcher's perspective are observable to the center (e.g., a patient with AIDS is both more likely to contract an infection and to die) are accounted for in our model through ω_{it}, and not the unanticipated quality shock, ε_{it}^{q}, and so would not induce such correlation. Because our results ultimately show a strong quality-quantity tradeoff, our results are robust to this potential confound.

4.1 Proxy for the Quality Target

Although we do not observe treatment quality directly, the data contain information on patient outcomes that are correlated with a center's choices on this dimension. In particular, we focus on the center's infection rate as an indicator of quality. This is only an imperfect measure, however, because variation in the infection rate may be due to differences in patient characteristics across centers rather than centers' deliberate quality choices. To account for this, we control for center-level averages of several patient characteristics that influence infection rates. Specifically, we use the (negative) residual from a regression of infection rates on patient characteristics as our proxy for patient quality; this residual represents the variation in infection rates that remains unexplained after controlling for observable differences in the patient pool, and therefore serves as a proxy for the center's targeted quality level.

We control for several observable patient characteristics that influence a center's infection rate beyond its quality decision, with summary statistics displayed in Table 4. Most notably, we include controls for patients' vascular access type, which can be either an arteriovenous (AV) fistula, AV graft, or venous catheter. A patient's vascular access method influences his likelihood of developing a blood infection, as those with an AV fistula are significantly less likely to experience clots or infections. In addition to a patient's vascular access, other characteristics have been shown clinically to affect treatment outcomes. Because centers' patient loads vary in terms of these characteristics, we also include controls for patients' (i) average number of comorbid conditions, (ii) average duration of ESRD, (iii) average age, (iv) gender distribution, and (v) average hemoglobin levels.[27] Putting these center-level average patient characteristics together into the vector z_{it}, we estimate

$$f_{it} = z_{it}\gamma - q_{it},$$

where f_{it} is the realized infection rate at center i in period t. The residuals from this regression reflect the center's relative infection rate after controlling for observable patient characteristics, which we then use as our measure of center quality.

Even after controlling for observable characteristics, some unobservable differences in patient health may remain, part of which may be observable to centers as they make their quality choices. Within our model, we interpret these unobservable differences as differences in ω_{it}

[27] Low hemoglobin levels are associated with anemia and pose health risks for dialysis patients.

Table 4: Patient Characteristics Summary Statistics.

Variable	Mean	St. Dev.
Avg. Patient Age	61.518	4.381
Pct. Female	45.798	8.333
Pct. AV Fistula	43.016	13.477
Avg. Comorbid Conditions	3.026	0.826
Avg. Duration of ESRD	4.089	0.953
Avg. Hemoglobin Level	11.882	0.332
Number of Center-Years	18,221	

across centers, which we can address, along with other unobservable differences in productivity (e.g., management ability or unobserved quality of inputs), using a control function approach to correct for these differences.

As discussed above, we account for expectational or measurement error in our specification of the production function by including ε_{it}^{q} and instrumenting with a second outcome variable. Specifically, we use the ratio of actual death rates to Medicare's estimates for each center's expected death rate that it constructs using individual patient characteristics (individual-level characteristics are not released to protect patient privacy). Medicare uses this ratio as an indicator of center quality in its own reports, and we include this measure as a second noisy proxy for a center's quality. For this instrument to be valid, the variation in it which is unrelated to center quality should be uncorrelated with the variation in the infection rate that is unrelated to quality, which fits our setting because, although infections do raise the risk of death, they are the primary cause of fewer than 10 percent of patient deaths overall (USRDS 2013). By contrast, over one-third of patient deaths stem from cardiovascular issues, which may be related to center quality through the effectiveness of monitoring for hypotension and other complications that arise during treatment. Conversely, roughly 80 percent of patients hospitalized with a septic infection survive. We include the death rate ratio as an instrument rather than as the primary proxy because it is less directly tied to the quality choices made by dialysis centers (e.g., cleaning protocols) than the septic infection rate.[28]

[28]Results are qualitatively robust when the roles of the primary proxy and the instrument are reversed.

4.2 Controlling for Policy Shifters

To invert the hiring function and recover each center's productivity, we must explicitly control for the factors other than productivity that affect hiring. As such, we include the following sources of variation in x in our specification:

Ownership Status Centers differ in their ownership type, with roughly 87.7 percent operating as for-profit entities and the remainder as non-profit. Among the for-profit centers, two major chains dominate, with DaVita owning roughly 28 percent of centers nationwide and Fresenius 31 percent.[29] A center's ownership structure may affect its polices related to hiring and treatment quality, as non-profit centers could, on average, target a different weighting of quality over quantity. We therefore control for this distinction by including a dummy variable for the center's ownership status in x_{it}. This also enables us to allow for, and subsequently analyze, differences in productivity across ownership types.

Competition Because demand for dialysis treatments is local, the extent of competition a center faces may affect its hiring and quality choices. For instance, centers in highly competitive markets may choose to improve quality or increase staff levels to attract patients. We include the level of competition each center faces in x_{it} in the form of dummy variables for having 0, 1, 2, or 3 or more competitors in an hospital service area (HSA).[30] We assume that entry is realized at the beginning of the period, so the center observes its competitors when making its quality and hiring choices.

Quality Incentive Shifters Based on the reduced-form regressions in Section 2.2, we include in x the number of years since a center's last inspection and its state's inspection rate as proxies for the incentives centers face for providing higher-quality treatments. We do not include the referral rate of nephrologists because it is available for only 3 of the 5 years in our data, which would severely limit our sample size.

[29] These averages are taken across all years in the data.

[30] Following the healthcare literature, we use hospital service areas (HSA) as our market definition for dialysis centers. The Dartmouth Atlas determines HSA boundaries based on Medicare data for patients' actual hospital choices, and therefore serve as a well-suited market definition because they explicitly incorporate patients' travel patterns in a way that geographic boundaries such as counties or MSAs would not.

4.3 Two-Step Estimation

To recover the parameters of the production frontier, we first note that $\omega^h_{it} = \omega^q_{it} + \varepsilon^\omega_{it}$, so we can rewrite (6) as

$$y_{it} = -\alpha_q q_{it} + \beta_k k_{it} + \beta_\ell \ell_{it} + \omega^h_{it} - \varepsilon^\omega_{it} - \alpha\varepsilon^q_{it} - \varepsilon^y_{it}.$$

Because $(\varepsilon^\omega_{it}, \varepsilon^q_{it}, \varepsilon^y_{it})$ are revealed to the center after it makes it's quality choice and are uncorrelated with the center's information set at the time quality and output choices are made, they do not impose an endogeneity problem. Because centers' expectations about ω^h_{it} are a function of ω^q_{it}, however, we must still control for ω^h_{it}. From Proposition 1, we know that a center's expectation about its productivity at the time of hiring can be recovered by inverting the center's hiring policy at a fixed investment level such that

$$\omega^h_{it} = h^{-1}_{i_{it}}(h_{it}, k_{it}, \ell_{it}, x_{it}). \tag{7}$$

Substituting (7) into (6), we arrive at our first-stage estimation equation,

$$\begin{aligned} y_{it} &= -\alpha_q q_{it} + \beta_k k_{it} + \beta_\ell \ell_{it} + h^{-1}_{i_{it}}(h_{it}, k_{it}, \ell_{it}, x_{it}) - \varepsilon^\omega_{it} - \alpha\varepsilon^q_{it} - \varepsilon^y_{it}. \\ &= -\alpha_q q_{it} + \Phi_{i_{it}}(h_{it}, k_{it}, \ell_{it}, x_{it}) + \varepsilon_{it}, \end{aligned} \tag{8}$$

where $\varepsilon_{it} = -\varepsilon^\omega_{it} - \alpha\varepsilon^q_{it} - \varepsilon^y_{it}$ and $\Phi(h_{it}, k_{it}, \ell_{it}, x_{it}) = \beta_k k_{it} + \beta_\ell \ell_{it} + h^{-1}(h_{it}, k_{it}, \ell_{it}, x_{it})$. Due to invertibility requirements, we only have usable observations of (8) whenever hiring is non-zero.[31] Moreover, because the function $h^{-1}_i(\cdot)$ depends on the level of investment, we must estimate a separate $\Phi_i(\cdot)$ for each investment level. In practice, investment is zero 92 percent of the time, and we drop other investment levels and estimate (8) using observations where the center did not invest. Dropping observations where hiring is zero and investment is non-zero collectively reduce the size of the dataset by 19 percent. If the model is correctly specified, this truncation does not bias our results. Comparing the dropped observations to those used in the first stage, centers with dropped center-years are slightly smaller on average, but have similar health outcomes (infection rates and ratios of deaths to expected deaths). Running the

[31] Because there are likely adjustment costs to hiring, $h^{-1}_i(\cdot)$ is not well defined when hiring is zero (multiple productivity levels may lead to zero net hiring). We follow the productivity literature and drop observations of zero hiring when estimating the first stage.

descriptive analysis presented in Table 3 on the truncated sample also produces qualitatively similar results.

Finally, notice that the optimal policy for quality is $q_{it} = q(k_{it}, \ell_{it}, x_{it}, \omega^q_{it})$, whereas the optimal hiring policy is $h_{it} = h(k_{it}, \ell_{it}, x_{it}, \omega^h_{it})$. Therefore, the difference between ω^q_{it} and ω^h_{it} provides the variation needed to separately identify α_q.

Although the approach above handles the endogeneity of ω^q_{it}, we still have attenuation bias because ϵ_{it} and q_{it} are correlated through ε^q_{it}. To address this, we use a second noisy measure of quality as an instrument in the second stage of a three-stage estimation procedure following Robinson (1988).[32] First, we estimate $\widehat{E}[y|h_{it}, k_{it}, \ell_{it}, x_{it}, i_{it}]$ and $\widehat{E}[q|h_{it}, k_{it}, \ell_{it}, x_{it}, i_{it}]$ non-parametrically using local linear regression.[33] In doing so, we are careful to account for the possible discontinuity of these functions at $h_{it} = 0$ by considering positive and negative hiring observations separately.[34] We then estimate $\hat{\alpha}_q$ with the linear instrumental variables regression,

$$y_{it} - \widehat{E}[y|h_{it}, k_{it}, \ell_{it}, x_{it}, i_{it}] = -\alpha_q(q_{it} - \widehat{E}[q|h_{it}, k_{it}, \ell_{it}, x_{it}, i_{it}]) + \varepsilon_{it},$$

where we instrument for q_{it} with a second noisy measure of quality. In practice, we use the ratio of expected to actual deaths as this instrument, as discussed in Section 4.1. Finally, we recover $\hat{\Phi}_i(\cdot)$ from the nonparametric estimation

$$y_{it} + \hat{\alpha}_q q_{it} = \Phi_{i_{it}}(h_{it}, k_{it}, \ell_{it}, x_{it}) + \varepsilon_{it}.$$

We recover the remaining parameters in a subsequent stage. Note that, given any $\beta = (\beta_k, \beta_\ell)$, we can compute an estimate of unobserved productivity for each center-year that has non-zero hiring from

$$\hat{\omega}_{it}(\beta) = \hat{\Phi}_{i_{it}}(h_{it}, k_{it}, \ell_{it}, x_{it}) - \beta_k k_{it} - \beta_\ell \ell_{it}.$$

[32] An alternative approach, following Ackerberg et al. (2006), would have estimated y_{it} as a non-parametric function of $(q_{it}, h_{it}, k_{it}, \ell_{it}, x_{it}, i_{it})$ and then estimated α_q together with (β_k, β_ℓ) in the second stage. This would have the advantage of removing the requirement that q_{it} be flexibly chosen during the quality stage. However, the first stage estimation would be a nonparametric instrumental variables regression, introducing significant complications due to the high dimensionality of the problem.

[33] Bandwidths are chosen using the rule of thumb proposed by Scott (1992), which is a generalization of Silverman (1986) to the multivariate case. Results are robust to using alternative bandwidths. We have also experimented with the method of sieves which yields qualitatively similar results.

[34] That is, only negative hiring observations are used in the local linear regression when $h_{it} < 0$, and only positive hiring observations are used when $h_{it} > 0$; Not doing this would raise the possibility of inconsistent estimates of these expectations near zero hiring. Recall that at $h_{it} = 0$, the hiring function is not invertible, and these observations are dropped. Hiring is negative in roughly 40 percent of our observations.

Because ω_{it} follows a Markov process, we have

$$\omega_{it} = g(\omega_{it-1}) + \xi_{it}, \tag{9}$$

where g is a non-parametric function of ω_{it-1} and ξ_{it} is a shock to productivity between time $t-1$ and t that is independent of the center's time-t information set.[35] Thus, for any given $\beta = (\beta_k, \beta_\ell)$, we can estimate $g(\cdot)$ from the equation

$$y_{it} + \hat{\alpha}_q q_{it} - \beta_k k_{it} - \beta_\ell \ell_{it} = g(\hat{\omega}_{it-1}(\beta)) + \eta_{it}(\beta),$$

which follows from substituting the production function from (6) into the innovation of productivity from (9), where $\hat{\alpha}_q$ is the consistent estimator of α_q recovered in the first stage.[36]

At the true value of β, $\eta_{it}(\beta) = \varepsilon_{it} + \xi_{it}$, and so $\eta_{it}(\beta)$ is uncorrelated with the time-t labor and capital variables by construction and β can be consistently estimated using the moment conditions

$$E\begin{bmatrix} \eta_{it}(\beta) k_{it} \\ \eta_{it}(\beta) \ell_{it} \end{bmatrix} = 0. \tag{10}$$

We use (10) to estimate $\hat{\beta}$ via GMM, which can then be used to recover estimates of center-level productivity. Finally, standard errors are calculated using the block bootstrap, which accounts for statistical uncertainty in recovering the quality proxy, as well as both stages of the estimation process.

5 Results

We present results from our baseline model and two extensions, one which allows for the slope of the production frontier to vary based on a center's capital and labor, and another that allows for heterogeneous transitions for productivity.

[35] We use a fifth order polynomial sieve to approximate $g(\cdot)$, results are robust to using other orders.

[36] We can estimate this equation using each observation that follows an observation used in the first stage. While it might seem more straightforward to recover $g(\cdot)$ by regressing $\hat{\omega}_{it}(\beta)$ on $\hat{\omega}_{it-1}(\beta)$, this would require using only observations where consecutive periods of hiring are non-zero (and investment is zero), reducing the available data even further and introducing a potential selection problem since we would be censoring on a left-hand side variable (although our results are robust to this approach). We thank David Rivers for pointing this out to us.

5.1 Baseline Model

We present our estimates of the production and transformation functions in the first column of Table 5. As a point of comparison for our structural estimates, we include ordinary least squares (OLS) and fixed effects (FE) estimates in the next two columns. Finally, we include results from a specification that excludes the effect of quality choices, which are effectively estimates of a standard one-output production function.

We first consider the baseline production function parameters β_k and β_ℓ. Estimates of these parameters are strikingly different across methods, though similar with regards to whether or not quality choices are included in the model. The comparison with the OLS and FE estimates is instructive for several reasons. First, OLS does not control for endogenous input choices. Because OLS relies on cross-sectional variation in stations to identify the labor and capital coefficients, it must ignore the possibility of productivity differences across centers, resulting in a substantially higher labor coefficient and, consequently, the suggestion of increasing returns to scale. We believe that the finding of increasing returns to scale is due to endogeneity bias from unobserved productivity, as more productive centers are likely both to use more stations and employ more staff.

The FE procedure, by contrast, allows for productivity differences across centers but assumes that these differences remain constant over time; that is, the FE estimates identify the capital and labor coefficients on the basis of year-to-year changes in centers' inputs. Using this approach, both the capital and labor coefficients fall substantially relative to the OLS results. We believe this is primarily due to two factors. First, relying on only year-to-year variation makes measurement error in both capital and labor inputs a more prominent concern. Because stations and employees remain fairly stable over time, measurement error for hiring and investment decisions biases these coefficients towards zero;[37] this is especially an issue for capital because of infrequent investment. A second potential reason for the discrepancy between the OLS and FE approaches is that capital and labor differences in the cross-section may proxy for unobserved time-invariant characteristics (e.g., floorspace) that the FE specification captures through the productivity term.

In contrast to OLS and FE, estimates of our model yield a coefficient on labor of 0.24 and a

[37] For example, if a new station was installed in June of 2002, it will first be reported in 2003, but the difference in the number of patients served in 2002 versus 2003 will underreport the impact of the new station that actually came online for the second half of 2002.

Table 5: Transformation and Production Estimates.

	With Quality			Without Quality		
	Model	OLS	FE	Model	OLS	FE
Quality, $-\alpha_q$	-0.0155	-0.0026	-0.0017			
	(0.0037)	(0.0007)	(0.0004)			
Capital, β_k	0.5204	0.4548	0.2608	0.5331	0.4572	0.2634
	(0.0437)	(0.0212)	(0.1040)	(0.0401)	(0.0212)	(0.1037)
Labor, β_ℓ	0.2436	0.6848	0.1985	0.2556	0.6834	0.1978
	(0.0316)	(0.0163)	(0.0135)	(0.0312)	(0.0164)	(0.0135)

coefficient on capital of 0.52, which suggest decreasing returns to scale overall. To review, our structural specification employs a Markov process for productivity and uses both cross-section and time-series variation to identify the parameters, while at the same time using centers' hiring choices to identify unobserved productivity. The relatively larger weight of capital in this specification fits well with our understanding of the production process. Although hiring more employees may allow a center to treat more patients by speeding up the transition of stations from one patient to the next, the number of patients being treated by the center at any given time is necessarily bounded by the number of available stations. While the labor coefficient is small relative to many previous studies, it is in line with some micro studies (e.g., De Loecker 2011), albeit for different industries. We know of no other study on the dialysis industry which could serve as a benchmark for comparison. The finding of decreasing returns to scale may reflect omitted inputs, such as floor-space and other forms of physical capital not related to the number of dialysis stations, which are presumably captured by the productivity term.

We next turn to the primary focus of the paper, the estimates of the quality-quantity tradeoff in the transformation function, α_q.[38] All three specifications provide evidence of a statistically significant quantity-quality tradeoff, though the magnitude of the effect is much larger in the structural model than with either the OLS or FE methods. The smaller impact of quality on output in the OLS and FE specifications likely stems from endogeneity and attenuation bias. Because the OLS specification does not control for differences in productivity, an estimate of α_q in this setup will be biased towards zero. While the FE approach controls for time-invariant productivity, if centers' changes in quality choices are positively correlated with changes in their productivity, the FE estimate of α_q will also be biased downwards. This effect, coupled with the

[38]Note that, we report "$-\alpha_q$" in the tables, incorporating the negative sign in (8).

Table 6: Robustness Checks.

	Baseline	No Patient Controls	No Center Controls	No Quality Instrument
Quality, $-\alpha_q$	-0.0155	-0.0126	-0.0148	-0.0108
	(0.0037)	(0.0032)	(0.0037)	(0.0008)
Capital, β_k	0.5204	0.5174	0.5065	0.5233
	(0.0437)	(0.0435)	(0.0423)	(0.0418)
Labor, β_ℓ	0.2436	0.2453	0.2276	0.2467
	(0.0316)	(0.0312)	(0.0202)	(0.0307)

attenuation bias already discussed above, drives the estimates of the quality-quantity tradeoff towards zero.

The coefficient of 0.0155 from the structural model indicates that, holding inputs fixed, a center that improves its quality enough so that its targeted infection rate falls by 1 percentage point would need to reduce overall patient hours by 1.55 percent. Equivalently, a center could increase its output 1 percent by reducing quality such that its targeted infection rate increases 0.65 percentage points, holding inputs and productivity fixed. Alternatively, we can measure the cost of providing high-quality treatments in units of labor: a center can reduce its infection rate by 1 percent while maintaining its current level of output by increasing labor 6.4 percent. Given that the average center employs approximately 13 full-time-equivalent nurses, this roughly equates to expanding employment by an additional 0.83 full-time workers. Moreover, reducing the targeted infection rate by a full standard deviation (6.3 percentage points) would cost the equivalent of roughly five additional full-time workers for the average center.

In Table 6, we consider several robustness checks of the baseline results, which are repeated in the first column. The second column drops controls for patient characteristics, simply useing the infection rate itself as a proxy for quality targets instead. This has a minimal effect on the production function parameters but decreases the quality-quantity tradeoff, a result that may stem from the fact that our measure of quality is now contaminated by unobserved factors previously controlled for through patient characteristics. The third column drops from the hiring function center characteristics related to for-profit status and competition, meaning that centers have the same hiring policy across these characteristics.[39] This causes all of the production coefficients to decline slightly, though they remain within the confidence interval. Finally, the

[39]We maintain controls in x for time since health inspection and the state inspection rate, although results are also qualitatively robust to eliminating these measures.

fourth column does not instrument for the quality proxy, but instead simply uses OLS to estimate the first stage. We see a substantial decrease in the estimates of the quality-quantity tradeoff, which suggests that instrumenting for quality is effectively controlling for attenuation bias. In all cases, the effect of quality declines slightly, though our estimate of a significant quality-quantity tradeoff remains robust to various model specifications.

5.2 Heterogeneity in the Quality-Quantity Tradeoff

Our baseline model assumes that the slope of the production frontier is homogeneous across centers. While a reasonable starting point, the slope of the frontier may vary across centers. In particular, we would like to know whether the slope of the frontier changes depending on a center's scale or capital-labor ratio. One might expect that adding nurses and technicians, holding the number of stations fixed, could make it easier to reduce infections compared to adding more machines. To investigate this possibility, we consider a generalized form of (6) which allows for the slope of the frontier to depend on the center's labor and capital inputs,[40]

$$y = -(\alpha_q q_{it} + \alpha_{qk} q_{it} k_{it} + \alpha_{q\ell} q_{it} \ell_{it}) + \beta_k k_{it} + \beta_\ell \ell_{it} + \omega^q_{it} + \epsilon_{it}. \tag{11}$$

With this specification of the production frontier, the distinction between the transformation function and the production function is no longer straightforward, though the production frontier itself is still well defined. The quality-quantity tradeoff for a center is now $\alpha_q + \alpha_{qk} k_{it} + \alpha_{q\ell} \ell_{it}$, and the return to capital and labor is now likewise dependent on the center's quality choice.

Table 7 presents the results for this specification of our model, as well as the OLS and FE approaches. Although the OLS and FE approaches are statistically insignificant for the most part, our model indicates that the slope of the production frontier is strongly related to the capital-labor ratio. In particular, adding stations makes the quantity-quality tradeoff steeper, while adding labor flattens it. In other words, the differential impact of adding stations expands the production frontier relatively more in the quantity direction as compared to hiring more employees. This result corresponds well with our description of the industry: an additional station can be used to expand output with the same number of nurses and technicians, though the risk of infection increases as fewer nurses are available to monitor and clean machines.

Interestingly, the coefficients on α_{qk} and $\alpha_{q\ell}$ sum to almost zero, which suggests that the

[40] Here in a slight abuse of notation we let ϵ_{it} collect all the unanticipated error terms.

Table 7: Flexible Production Frontier.

	Model	OLS	FE
Quality, $-\alpha_q$	-0.0082	0.0057	-0.0069
	(0.0252)	(0.0047)	(0.0025)
Quality x Capital, α_{qk}	-0.0364	0.0008	0.0023
	(0.0108)	(0.0025)	(0.0012)
Quality x Labor, , $\alpha_{q\ell}$	0.0407	-0.0045	-0.0004
	(0.0103)	(0.0021)	(0.0011)
Capital, β_k	0.4451	0.4569	0.2589
	(0.0603)	(0.0212)	(0.1056)
Labor, β_ℓ	0.2969	0.6830	0.1975
	(0.0496)	(0.0163)	(0.0134)

quality-quantity tradeoff is insensitive to scale.[41] Instead, it appears that the capital-labor ratio is the key factor in determining the tradeoff.

Overall, the average slope of the production frontier is -0.0115, which is similar to that found in the baseline model of Table 5, though slightly smaller in magnitude. The result is not statistically significant, however, due to the much larger standard errors in this model. The lack of precision is at least partially due to the high correlation between capital and labor, which is further exacerbated when both are interacted with our quality measures. Moreover, about 20 percent of centers — those with the lowest capital-to-labor ratios — are estimated to have a production frontier with a positive slope, which violates the model. This could be due to specification error or attenuation bias from the use of proxies to control for quality choices. Therefore, while the results of this specification are instructive, we take our baseline estimates as our primary estimate of the quality-quantity tradeoff across the industry.

5.3 Heterogeneity in the Productivity Process

Finally, we extend our baseline model so that the productivity process depends on both ω_{t-1} and center characteristics, allowing them to differ based on for-profit status and whether the center belongs to one of the two major chains in the industry, Fresenius and DaVita. Several other analyses of the healthcare industry have found that for-profits tend to be more productive than non-profits (Kessler & McClellan 2002). Our analysis considers whether this stylized fact holds in the dialysis industry after controlling for centers' endogenous quality choices.

[41] Formally, the model does not reject the hypothesis that $\alpha_{yq} + \alpha_{y\ell} = 0$; the p-value for the test is 0.67.

Table 8: Heterogeneity in Productivity Process

	Baseline	Separable	Nonpara.
Quality, $-\alpha_q$	-0.0155	-0.0155	-0.0155
	(0.0037)	(0.0037)	(0.0037)
Capital, β_k	0.5204	0.5130	0.4980
	(0.0437)	(0.0452)	(0.0500)
Labor, β_ℓ	0.2436	0.2375	0.2353
	(0.0316)	(0.0306)	(0.0282)

Table 9: Average Productivity Transition Difference by Center Type.

	Separable	Nonpara.
For-Profit	0.0302	-0.0270
	(0.0084)	(0.0260)
Frenesius	0.0090	-0.0055
	(0.0075)	(0.0254)
DaVita	0.0306	0.0078
	(0.0078)	(0.0267)

Specifically, we consider two alternatives to (9). First, we allow center-type, p, which can be either non-profit, independent for-profit, DaVita-owned, or Fresenius-owned, to shift the level of the production process,

$$\omega_{it} = \delta_p + g(\omega_{t-1}).$$

Second, we consider a non-parametric approach and estimate a separate productivity process for each center-type,

$$\omega_{it} = g_p(\omega_{t-1}).$$

We do so because institutionalized management practices at the chain level may influence the manner in which productivity evolves within a center.

We present the results from estimates of these two alternative specifications in Table 8, with the baseline results repeated for comparison purposes. Of course, $\hat{\alpha}_q$ remains the same across all specifications because it is estimated in the first stage. More importantly, we see that the production function estimates are robust to alternative specifications of the productivity process.

We examine type-heterogeneity in center productivity in Table 9, which shows the average productivity differences of the three for-profit types relative to the non-profit base case. For the separable case, this is simply the estimate of δ_p, while for the non-parametric case it is the

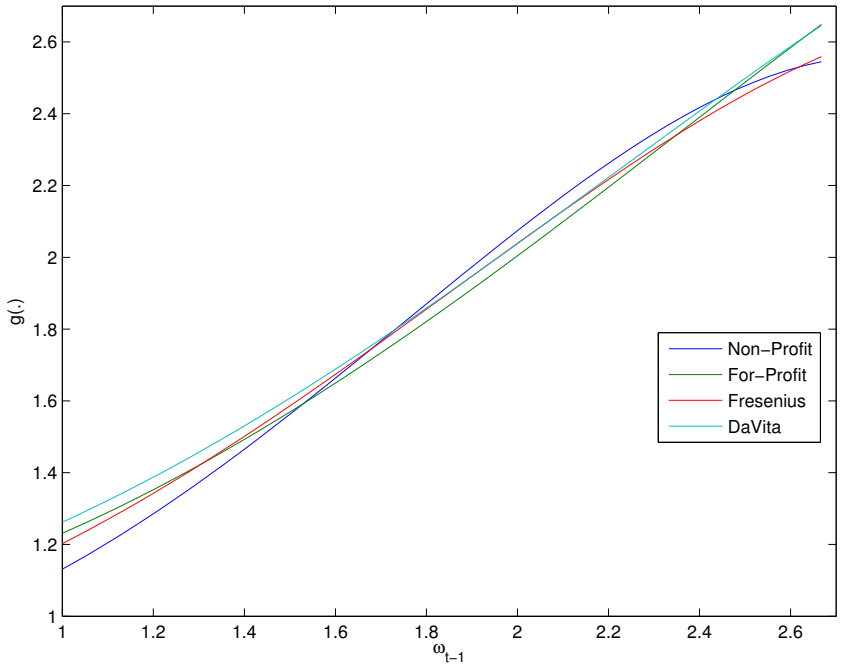

Figure 1: Productivity Transition by Center Type.

average difference in the $g_p(\cdot)$ functions across all centers. Interestingly, while the separable case indicates that for-profit centers are more productive than non-profits, this result is not robust to allowing for a fully nonparametric specification of the transition process. Some intuition for this result can be found in Figure 1, which plots the estimates of the $g_p(\cdot)$ functions for all four types. While the three for-profit types are all very similar, the non-profit type appears slightly steeper. This indicates that low-productivity non-profits are more likely to also have lower productivity in the following period relative to for-profit centers, and vice versa for high-productivity non-profits. Although these differences are not statistically significant, they do suggest that the differences in productivity between for-profit and non-profit centers are more complex than a simple level shifter would indicate.[42]

[42]Pointwise standard errors are not included in the figure for clarity, but are available from the authors by request.

6 Conclusion

Because dialysis treatments comprise a large — and growing — expense for Medicare, controlling their costs will likely concern policy makers for the foreseeable future. By estimating center-level production functions that incorporate endogenous quality choices, we quantify the tradeoff that dialysis centers face between treating more patients and providing higher-quality care. Understanding this relationship is crucial for designing effective policies that promote the proper balance of efficiency and efficacy.

A back-of-the-envelope calculation demonstrates how our analysis can inform policy decisions. In the first of two approaches, we benchmark the cost of reducing infections by calculating the number of patients a center would have to forgo in order to prevent one infection (in expectation). Under this scenario, the median number patient-years a center must forgo to eliminate one infection each year is 1.5. As industry studies suggest that the cost of hemodialysis treatment is between $45,000-55,000 per patient annually (Lee et al. 2002), this suggests the opportunity cost of preventing one infection per year is roughly $75,000. Alternatively, we could consider the possibility that centers treat the same number of patients but reduce infections by hiring more staff. From this perspective, the median increase in staff required to eliminate one infection would be 1.8 full-time-equivalent employees. Although compensation varies based on staff qualifications and location, if we assume compensation ranges from $35,000-50,000, this would suggest preventing one infection costs $63,000-90,000.[43] Under either approach, the opportunity cost of one infection to a center is approximately $75,000.

With this estimate in hand, we can compare the opportunity cost of preventing infections for a dialysis center to the cost of treating septic infections in a hospital. Although hospitalization costs vary widely depending on the severity of the infection, a recent study estimates that the hospitalization of a hemodialysis patient for an infection costs, on average, $25,000 (Ramanathan et al. 2007). Therefore, tighter quality regulation will improve social welfare if society's non-hospitalization cost of infection (i.e., the increased risk of death and disutility of the infected patient) is greater than $50,000, but will reduce welfare if it is less than this amount. While admittedly speculative, this analysis serves as a guidepost for policy makers seeking to regulate dialysis providers along this dimension.

[43] BLS (2014) reports that the median salary of all licensed practical and vocational nurses is $41,000. Our labor measure includes nurses and technicians so this salary estimate is likely an upper bound. On the other hand, this salary estimate does not account for non-salary compensation.

More broadly, our work informs policy discussions by showing that, while productivity dispersion is extensive within the industry, cost-cutting initiatives may result in centers reducing the quality of care they provide. Because dialysis resembles other healthcare settings, these findings illustrate the challenges of introducing policies intended to minimize costs while maintaining high standards of care.

References

Ackerberg, D. A., Caves, K. & Frazer, G. (2006), Structural identification of production function. UCLA, Deloitte and Touche, Rotman School of Management.

Bae, S.-H., Brewer, C. S. & Covner, C. T. (2012), 'State mandatory overtime regulations and newly licensed nurses' mandatory and volunatry overtime and total work hours', *Nursing Outlook* **60**(2), 60–71.

BLS (2014), Occupational outlook handbook, Technical report, Bureau of Labor Statistics.

Blundell, R. & Bond, S. (2000), 'Gmm estimation with persistent panel data: an application to production functions', *Econometric Reviews* **19**, 321–340.

CDC (2001), 'Recomendations for preventing transimission of infections among chronic hemodialysis patients', *Morbidity and Mortality Weekly Report* **50**(RR05), 1–43.

Crawford, G. S. & Shum, M. (2007), 'Monopoly quality degradation and regulation in cable television', *Journal of Law and Economics* **50**(1), 181–219.

Dai, M. (2012), Specialization and differentiation in outpatient dialysis. Drexel Lebow.

De Loecker, J. (2011), 'Product differentiation, multi-product firms and estimating the impact of trade liberalization on productivity', *Econometrica* **79**(5), 1407–1451.

Doraszelski, U. & Jaumandreu, J. (2012), Measuring the bias of technological change. CEPR.

Ford, J. M. & Kaserman, D. L. (2000), 'Ownership structure and the quality of medical care: evidence from the dialysis industry', *Journal of Economic Behavior & Organization* **43**, 279–293.

Gandhi, A., Navarro, S. & Rivers, D. (2013), On the identification of production functions: How heterogeneous is productivity? University of Wisconsin-Madison and University of Western Ontario.

Gertler, P. J. & Waldman, D. M. (1992), 'Quality-adjusted cost functions and policy evaluation in the nursing home industry', *Journal of Political Economy* pp. 1232–1256.

Hirth, R. A. (2007), 'The organization and financing of kidney dialysis and transplant care in the united states of america', *International Journal of Health Care Finance and Economics* **7**(4), 301–318.

Imerman, M. & Otto, D. (2004), Preliminary Market and Cost Analysis of a Five-Station Hemodialysis Facility in Marengo, Iowa. Iowa State University.

Jha, A. K., Orav, E. J., Dobson, A., Book, R. A. & Epstein, A. M. (2009), 'Measuring efficiency: the association of hospital costs and quality of care', *Health Affairs* **28**(3), 897–906.

Joskow, P. L. & Rose, N. L. (1989), 'The effects of economic regulation', *Handbook of industrial organization* **2**, 1449–1506.

Kessler, D. P. & McClellan, M. B. (2002), 'The effects of hospital ownership on medical productivity', *RAND Journal of Economics* pp. 488–506.

Laine, J., Finne-Soveri, U. H., Björkgren, M., Linna, M., Noro, A. & Häkkinen, U. (2005), 'The association between quality of care and technical efficiency in long-term care', *International Journal for Quality in Health Care* **17**(3), 259–267.

Lee, H., Manns, B., Taub, K., Ghali, W. A., Dean, S., Johnson, D. & Donaldson, C. (2002), 'Cost analysis of ongoing care of patients with end-stage renal disease: The impact of dialysis modality and dialysis access', *American Journal of Kidney Diseases* **40**(3), 611–622.

Lee, J., McCullough, J. S. & Town, R. J. (2012), The impact of health information technology on hospital productivity. National Bureau of Economic Research.

Levinsohn, J. & Petrin, A. (2003), 'Estimating production functions using inputs to control for unobservables', *The Review of Economic Studies* **70**(2), pp. 317–341.
URL: *http://www.jstor.org/stable/3648636*

Morey, R. C., Fine, D. J., Loree, S. W., Retzlaff-Roberts, D. L. & Tsubakitani, S. (1992), 'The trade-off between hospital cost and quality of care: An exploratory empirical analysis', *Medical Care* pp. 677–698.

Mundalk, Y. (1961), 'Empirical production function free of management bias', *Journal of Farm Economics* **43**, 44–56.

Olley, G. S. & Pakes, A. (1996), 'The dynamics of productivity in the telecommunications equipment industry', *Econometrica* **64**(6), pp. 1263–1297.
URL: *http://www.jstor.org/stable/2171831*

Pakes, A. (1994), The estimation of dynamic structural models: Problems and prospects, *in* J. J. Laffont & C. Sims, eds, 'Advances in Econometrics: Proceedings of the 6th World Congress of the Econometric Society', Vol. II, pp. 171–259.

Patel, P. R., Yi, S. H., Booth, S., Bren, V., Downham, G., Hess, S., Kelly, K., Lincoln, M., Morrissette, K., Lindberg, C., Jernigan, J. A. & Kallen, A. (2013), 'Bloodstream infection rates in outpatient hemodialysis facilities participating in a collaborative prevention effort: A quality improvement report', *American Journal of Kidney Disease* **62**(2), 322–330.

Pronovost, P., Needham, D., Berenholtz, S., Sinopoli, D., Chu, H., Cosgrove, S., Sexton, B., Hyzy, R., Welsh, R., Roth, G. et al. (2006), 'An intervention to decrease catheter-related bloodstream infections in the icu', *New England Journal of Medicine* **355**(26), 2725–2732.

ProPublica (2011), 'Dialysis: High costs and hidden perils of a treatment guaranteed to all'.
URL: *http://www.propublica.org/series/dialysis*

Ramanarayanan, S. & Snyder, J. (2011), Reputations and firm performance: Evidence from the dialysis industry. UCLA Anderson.

Ramanathan, V., Chiu, E. J., Thomas, J. T., Khan, A., Dolson, G. M. & Darouiche, R. O. (2007), 'Healthcare costs associated with hemodialysis catheter-related infections: A single-center experience', *Infection Control and Hospital Epitemiology* **28**(5), 606–609.

Robinson, P. (1988), 'Root-n-consistent semiparametric regression', *Econometrica* **56**(4), 931–954.

Romley, J. A. & Goldman, D. P. (2011), 'How costly is hospital quality? a revealed-preference approach', *The Journal of Industrial Economics* **59**(4), 578–608.
URL: *http://dx.doi.org/10.1111/j.1467-6451.2011.00468.x*

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley series in probability and mathematical statistics, John Wiley and Sons, Ltd., New York.

Silverman, B. W. (1986), *Density Estimaton for Statistics and Data Analysis*, Monographs on statistics and applied probability, Chapman and Hall, London.

Sloan, F. (2000), Not-for-profit ownership and hospital behavior, *in* A. Culyer & J. Newhouse, eds, 'Handbook of Health Economics', Vol. 1B, Elsevier Science B.V., Amsterdam.

Syverson, C. (2011), 'What determines productivity?', *Journal of Economic Literature* **49**(2), 326–65.

USRDS (2010), 2010 Annual Data Report, Technical report, United States Renal Data System, Minneapolis, MN.

USRDS (2013), 2013 annual data report: Atlas of end stage renal disease in the united states, Technical report, US Renal Data System, National Institutes of Health, Bethesda, MD.

Weinstein, M. C. & Stason, W. B. (1977), 'Foundations of cost-effectiveness analysis for health and medical practices.', *The New England journal of medicine* **296**(13), 716–721.

Zhang, H. (2014), Biased technology and contribution of technological change to economic growth: Firm-level evidence from china. University of Hong Kong.

A Proofs

Proof of Lemma 1 *The center's expected period-return to labor is increasing in ω^q; that is, $\frac{\partial \pi}{\partial \ell}$ is increasing in ω^q.*

Proof. Because center payoffs are increasing both y and q (i.e., the center has non-satiable payoffs), we know that the center will choose $(\tilde{y}, \tilde{q})$ to solve the following problem where the production constraint binds:

$$\begin{aligned} \pi(k, \ell, x, \omega^q) = \max_{\tilde{y}, \tilde{q}} \; & E[\rho(y, q, k, \ell, x)] \\ \text{subject to: } & T(\tilde{y}, \tilde{q}) = F(k, \ell, \omega^q) \\ & y = \tilde{y} + \varepsilon^y \\ & q = \tilde{q} + \varepsilon^q. \end{aligned}$$

Totally differentiating π with respect to ℓ, the return to an increase in labor is,

$$\frac{d\pi}{d\ell} = E\left[\rho_y \frac{d\tilde{y}}{d\ell} + \rho_q \frac{d\tilde{q}}{d\ell} + \rho_\ell\right],$$

where ρ_x represents the partial derivative of ρ with respect to x and the total derivatives with respect to $\tilde{y}$ and $\tilde{q}$ are the center's optimal policy change for a change in ℓ. We know both are weakly positive — with at least one strictly positive — because an increase in ℓ relaxes the production constraint through an increase in $F(\cdot)$, and $\rho(\cdot)$ is increasing in both y and q. To see that this is increasing in ω^q, note that an increase in ω^q also relaxes the production constraint. Differentiating again with respect to ω^q yields

$$\frac{d^2\pi}{d\ell d\omega^q} = E\left[\rho_y \frac{d\tilde{y}}{d\ell}\frac{d\tilde{y}}{d\omega^q} + \rho_q \frac{d\tilde{q}}{d\ell}\frac{d\tilde{q}}{d\omega^q}\right].$$

Non-satiation again ensures that both terms are weakly positive and at least one is strictly positive. □

Proof of Proposition 1 *For any fixed investment level κ, the center hiring function $h(k, \ell, x, \omega^h)$ is invertible with respect to ω^h on the domain $\{(k, \ell, x, \omega^h) : i(k, \ell, x, \omega^h) = \iota\}$,*

$$\omega^h = h_\iota^{-1}(k, \ell, x, h).$$

Proof. We will apply Theorem 1 from Pakes (1994) while accounting for three differences which complicate our model. First, following Lemma 1 from Pakes (1994), we note the the inclusion of a discrete choice of capital investment does not alter our ability to use the center's first-order condition with respect to hiring; we must simply substitute the (observed) optimal investment choice ι into the first-order condition such that

$$c_h(\iota, h) + \beta E V_h(k + \iota, \ell + h, x', \omega^{q'}) | k, \ell, \omega^h, \iota, h] = 0.$$

Second, because x evolves according to an exogenous stochastic process, we can use the insight found in Appendix C of De Loecker (2011) that additional exogenous variables do not alter the invertibility property. The only remaining difference between this problem and the traditional investment problem described by Olley & Pakes (1996) is that our productivity process evolves intra-period between the quality and investment stages. However, because $Pr(\omega^{q'}|\omega^h)$ and $Pr(\omega^h|\omega^q)$ are both stochastically increasing in ω^h and ω^q (the former by assumption, and the

latter because it is a random walk), we know that $Pr(\omega^{h'}|\omega^h)$ is also stochastically increasing. We can thus write a single Bellman equation for a center at the time of the hiring decision as

$$V(k,\ell,x,\omega^q,\xi,\omega^h) = \max_{i,h} -c(i,h,k,\ell)+\pi(k,\ell,x,\omega^q)+\xi+\beta E[V(k',\ell',x',\omega^{q'},\xi,\omega^{h'})|k,\ell,x,\omega^h,i,h].$$

Note here that today's realized profits from the quality stage are $\pi(k,\ell,x,\omega^q)+\xi$, where ξ is uncorrelated with the agent's information set at the time of the quality choice (or any time before the quality choice), but is known at the time of the hiring decision since production outcomes are already revealed; that is, they are sunk with respect to today's hiring decision. Note also that ω^q and ξ represent two additional state variables, but they both evolve exogenously. Moreover, conditional on ω^h, they are uncorrelated with future draws of ω^q and ξ, which is why they do not appear in the final expectation term. Finally, using Lemma 1 and the fact that $Pr(\omega^{q'}|\omega^h)$ is stochastically increasing, we know $E[\frac{\partial\pi(k',\ell',x'\omega^{q'})}{\partial\ell}|k,\ell,x,\omega^h]$ is increasing in ω^h.

Following De Loecker (2011), group $k^* = (k,\ell,x,\omega^q,\xi)$, meaning that the policy function can be written as $h(k^*,\omega^h)$. We can now directly apply Pakes (1994, Lemma 3) where $c(h,\iota,k^*)$ stands for $c(x,k)$ (recall ι is the optimal capital investment decision); $\pi(\omega^q,k^*) = \pi(k,\ell,x,\omega^q)+\xi$ for $\pi(\omega,k)$; and the choice variable is h (hiring), rather than x, which was continuous capital investment in Pakes (1994).

□