Review of Economic Studies (2017) **84**, 1071–1105 doi:10.1093/restud/rdw042

Advance access publication 12 September 2016

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

PAUL L. E. GRIECO
The Pennsylvania State University
and
RYAN C. MCDEVITT
Duke University

First version received April 2013; *final version accepted June* 2016 (*Eds.*)

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 centre's unobservable and endogenous choice of treatment quality while allowing for unobserved differences in productivity across centres. We find that a centre that reduces its quality standards such that its expected rate of septic infections increases by 1 percentage point can increase its patient load by 1.6%, holding productivity, capital, and labour fixed; this corresponds to an elasticity of quantity with respect to quality of −0.2. Notably, our approach provides estimates of productivity that control for differences in quality, whereas traditional methods would misattribute lower-quality care to greater productivity.

Key words: Productivity, Quality variation, Health care

JEL Codes: D24, I1, L2

1. INTRODUCTION

Healthcare providers face a tradeoff between increasing the number of patients they treat and maintaining high standards of care. This tension between the quality and quantity of treatments lies at the centre of many recent policy initiatives, such as Medicare's prospective payment system that ties reimbursements to a fixed amount per service irrespective of a provider's actual costs. Although these initiatives aim to limit wasteful healthcare expenses, they may inadvertently result in less-effective 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, though doing so in a way that properly accounts for differences in productivity across providers poses several econometric challenges. This article examines the quality-quantity tradeoff explicitly and provides an empirical framework for measuring its magnitude. Applied to the dialysis industry, we find that the cost of improving

treatment quality is substantial, whereas previous measurement techniques would show only a small tradeoff.

Centres that provide outpatient dialysis treatment—a process that cleans the blood of patients with kidney failure—face a particularly acute quality-quantity tradeoff, and several institutional details make it a natural setting for studying the relationship between productivity and quality within 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. Secondly, we observe centres' input levels (*i.e.* staffing and machines) and production (*i.e.* patient loads), which allows us to cleanly identify the transformation of inputs into outputs. Thirdly, facilities have observable differences in outcomes that relate directly to the quality of care they provide (*e.g.* infections and mortality), which allows us to connect a centre's inputs and outputs to its treatment quality. Fourthly, 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 distinguish quality choices from price discrimination.[1] Finally, payments to dialysis facilities are substantial, at over $20 billion in 2011 and 6% of total Medicare spending, which allows us to implement our model on an important area for policy analysis.

Identifying a quality-quantity tradeoff among dialysis providers requires us first to understand the incentives centres face for providing high-quality care. Most directly, dialysis centres have an incentive to minimize treatment costs under Medicare's prospective payment system, which may lead them to provide low-quality—and hence less-costly—care. As an example, a centre can treat more patients if it spends less time cleaning machines after each use, although doing so increases the risk of patients acquiring infections. Counteracting the incentive to increase patient loads by reducing treatment quality are several motivations for maintaining high standards. Perhaps most notably, centres must report quality statistics to Medicare and face intermittent inspections by state regulators, with Ramanarayanan and Snyder (2011) finding that these reports have a causal impact on the quality of care provided by dialysis centres.[2] In addition, patients have some choice over their dialysis providers and nephrologists make referrals based, in part, on a centre's effectiveness, potentially leading centres to compete for patients by providing higher-quality care (Dai, 2014). Finally, non-profit centres may have objectives for providing high-quality care unrelated to maximizing profits (Sloan, 2000). In keeping with these studies, we find that the frequency of inspections, the number of years since a centre was last inspected, and nephrologists' referral rates are all correlated with centres' infection rates. In recognition of this institutional detail, our model of centres' quality choices includes such incentive shifters, as they provide an important source of variation that allows us to identify a centre's quality-quantity tradeoff.

After establishing centres' motivations for providing high-quality care, determining whether dialysis centres do, in fact, face a costly tradeoff between quality and quantity requires us to overcome two key econometric challenges: the presence of unobserved differences in productivity across centres and the difficulty of directly measuring centres' quality choices.

As it relates to the first challenge, unobserved differences in productivity may bias estimates of the quality-quantity tradeoff: because centres choose their targeted levels of quality, estimating the relationship between quality and quantity becomes confounded by unobserved differences

1. In 2012, Medicare instituted a Quality Incentive Program (QIP) for dialysis centres that reduces reimbursements by 2% if centres do not adhere to a quality standard for average haemoglobin levels and urea reductions rates, two measures of the effectiveness of dialysis. 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. The QIP was not in effect for the timespan covering the data used in our analysis.

2. Using a regression discontinuity approach, they show that centres rated just below the threshold between "worse than expected" and "as expected" on annual CMS reports have improve their performance in the following year relative to those that narrowly surpass the threshold.

in productivity, such as staff skill levels, managerial ability, or patient characteristics, which are observable to the centre but not to the researcher.[3] As greater productivity effectively shifts out a centre's production possibilities frontier, the centre becomes able both to treat more patients *and* to provide better care. At the extreme, high levels of unobserved productivity may even generate a positive correlation between quality and quantity; this correlation would bias reduced-form estimates of the quality-quantity tradeoff and lead researchers to underestimate facilities' true costs of improving their quality standards.

To obtain consistent cost estimates for providing higher-quality care, we build on the structural methods for estimating firm-level production functions first proposed by Olley and Pakes (1996), and later extended by Levinsohn and Petrin (2003), Ackerberg *et al.* (2015), Gandhi *et al.* (2016), and others. Conceptually, we adapt these methods to incorporate a "quality-choice" stage that comes after a centre's choices of labour and capital inputs. That is, after acquiring capital and training workers, a manager observes her centre's expected level of productivity and dictates an optimal level of quality by, *e.g.* stipulating guidelines for the cleanliness of equipment or the amount of time between shifts. Incorporating these endogenous quality choices into our estimation technique is a necessary adjustment for healthcare settings such as dialysis because providers would otherwise appear more productive when they are instead treating many patients ineffectively.

To address the second main econometric challenge—that we do not directly observe centres' quality choices—we use observable measures of patient outcomes as proxies for what those choices must have been. Our approach is based on the assumption that, if high-quality care is more likely to result in better health outcomes, those outcomes are valid proxies for unobserved quality choices. Proceeding in this manner, however, presents two additional complications. First, health outcomes depend not just on the choices made by centres, but also on patients' underlying characteristics. To account for this, we use centre-level patient characteristics to control for key sources of variation in the patient population that could influence health outcomes. Secondly, health outcomes depend on many factors beyond just the quality of care provided by centres, 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 centres' septic (blood) 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 centre can increase its patient load by 1.6% by reducing the quality of its treatments such that the expected septic infection rate increases by 1 percentage point, holding input levels and productivity constant. This corresponds to an elasticity of quantity with respect to quality of -0.185 at the median, with an inter-quartile range from -0.127 to -0.251. Equivalently, holding the number of patients constant but allowing a one standard deviation increase in the targeted infection rate would reduce a centre's costs by the equivalent of five full-time employees, which almost 40% of the staff at an average centre.

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 centre's scale or input mix, we find that the tradeoff depends on a centre's capital–labour ratio. Specifically, a high capital-to-labour ratio corresponds to a steeper-than-average tradeoff (*i.e.* centres can increase output more for a given decrease in quality), while centres with relatively more labour have a flatter tradeoff. Although intuitive, our finding that the quality-quantity tradeoff depends on a centre's input mix and scale is, to our knowledge, novel in the healthcare literature.

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

In addition to informing policy discussions surrounding productivity in health care, our article also contributes to the growing literature in empirical industrial organization related to the estimation of production functions. These methods have a long history in economics, with much prior work focused on selection and simultaneity bias.[4] In light of this, more recent work has developed structural techniques that use centres' 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 and Rose, 1989; Crawford and Shum, 2007), as measuring the tradeoff is central to understanding the full impact of regulations.

The remainder of our article continues in the following section with a description of the outpatient dialysis industry and our data sources. Section 3 proposes a structural model for estimating a production frontier in the presence of unobserved productivity differences and an endogenous quality choice. Section 4 shows how variation in our data can be used to identify the model. Section 5 outlines the implementation of our estimator for the parameters of the model, while Section 6 presents our estimation results. Finally, Section 7 concludes with a discussion of our findings' implications for policy analysis.

2. THE DIALYSIS INDUSTRY AND DATA

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 managed properly. Patients with ESRD 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 procedure that cleans the blood of waste and excess fluids. Patients can receive different dialysis modalities, with haemodialysis, a method that circulates a patient's blood through a filtering device before returning it to the body, constituting 90.4% of treatments (Centre for Medicare and Medicaid Services). The typical dialysis regimen calls for three treatments per week lasting 2–5 hours each, with the duration dictated by a nephrologist to meet clinical thresholds. Although treatment lengths depend on individual patient characteristics, such as the severity of ESRD, treatment frequency rarely deviates from the standard protocol of three sessions per week.[7]

Patients receiving dialysis in the U.S. primarily do so at free-standing dialysis facilities, which collectively comprise over 90% of the market (USRDS, 2010).[8] Medicare's ESRD programme, instituted by an act of Congress in 1973, covers 90% of these patients; notably, all patients with ESRD become eligible for Medicare coverage, regardless of age, and the programme now includes

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

5. See, *e.g.* Olley and Pakes (1996), Ackerberg *et al.* (2015), and Levinsohn and Petrin (2003).

6. Romley and Goldman (2011) consider quality choices among hospitals using a revealed-preference approach rather than outcome-based quality measures. Gertler and Waldman (1992) estimate a quality-adjusted cost function for nursing homes. Lee *et al.* (2013) use a structural approach to measure the impact of healthcare IT on hospital productivity, but do not consider output quality.

7. Generally, Medicare reimbursements are limited to three sessions per week. Hirth (2007) argues that this limit may lead to inadequate dialyzing for some patients.

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

over 400,000 individuals. Today, Medicare spends more than $20 billion a year on dialysis care—approximately $77,000 per patient annually—which represents more than 6% of all Medicare spending despite constituting fewer than 1% of Medicare patients (ProPublica, 2011).

Beginning in 1983, Medicare has paid dialysis providers a fixed, prospective payment per treatment—the "composite rate"—up to a maximum of three sessions per week per patient. Initially, payments 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, which were approximately $135 per session throughout the course of our study. 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 requires constant supervision by trained medical professionals, as patients must remain connected to a station for several hours while being treated. 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 and treat any complications 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 sterilize machines in advance of the next patient. As a result of this labour-intensive and hands-on care, the cost per patient treated necessarily increases with the average amount of time devoted to treatments and cleaning. Labour costs, which consist largely of nurses and technicians' wages, reflect this, accounting for approximately 70%–75% of a facility's total variable costs (Ford and Kaserman, 2000).

Centres' employees have varying skill levels, with registered nurses (RNs) comprising the majority of staff. Technicians, who have less-extensive training than RNs, also treat patients and can do so with only a high-school diploma and in-house training (although they must eventually pass a state or national certification test). Notably, centres 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 (usually a nephrologist), but commonly have no physician on site. Medicare does not mandate a specific staffing ratio for dialysis centres, although some states do.

In addition to staffing levels, another significant decision for dialysis facilities is the number of stations to operate. Each station can serve only one patient at a time, and stations must be thoroughly disinfected after each use. Centers vary widely in terms of size, ranging from 1 to 80 stations. Based on industry reports, a typical dialysis station costs $16,000 and has a useful life of approximately seven years (Imerman and Otto, 2004).

Along with labour and capital decisions, centres 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 removing urea from blood to the comfort of patients during treatment. We focus on a dimension of quality directly related to improving patients' health, a patient's risk of contracting a septic infection, as such infections are particularly

9. See "FMS Pathways: Nursing Shortage", http://bridgesights.com/hillwriter/Nursing%20Shortage.pdf

costly and life-threatening for patients. As an alternative measure of quality, we also consider centres' excess mortality rates, described below.

Centers can allocate their labour and capital resources in ways that improve patients' health outcomes, but doing so comes with the opportunity cost of treating fewer patients. For example, infections stem in large part from the exposure of a patient's blood during dialysis, making the cleanliness of the centre and its stations a key determinant of health outcomes. Because dialysis sessions require up to one hour of preparation and cleaning, the centre has considerable control over its targeted infection rate, as health professionals who follow straightforward procedures can virtually eliminate their patients' risk of contracting infections (Pronovost *et al.*, 2006; Patel *et al.*, 2013).[10]

Centers may also have some limited scope for shortening patients' treatment times, which also would represent a reduction in treatment quality at the margin.[11] In an international survey of dialysis centres, Tentori *et al.* (2012) found that the average treatment time in the U.S. was 214 minutes, with a 95% confidence interval of 197 to 231 minutes across centres.[12] This study also finds that longer treatment times are associated with lower infection rates and better rates of survival.

Reducing the risk of infection and mortality with either longer cleaning times or longer treatment times comes with the opportunity cost of treating fewer patients due to the resource constraints of the facility, which may ultimately reduce the centre's profits. That is, because a facility's reimbursement per treatment does not vary with the treatment's duration or thoroughness of cleaning under Medicare's prospective payment system, a facility's profit per treatment decreases as treatment and cleaning times—and, hence, costs per patient—increase. In essence, the tradeoff faced by centres stems from their choice to either improve treatment quality or decrease costs.[13]

An extensive medical literature has examined these tradeoffs in health care more generally, mostly from an accounting perspective (Weinstein and Stason, 1977). Morey *et al.* (1992), *e.g.* found that a 1% increase in a hospital's quality of care increased 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 struggled to maintain high-quality standards for conditions that require time-consuming nursing procedures. Building on this literature, we consider the tradeoff using a structural model of healthcare provision that controls for possible confounding factors, such as centres' endogenous quality decisions and the measurement error that arises from using observable outcomes as proxies for centres' unobservable choices.

To study the relationship between the quality and quantity of treatments, we use a dataset derived from annual reports compiled by the Centers for Medicare and Medicaid Services (CMS) 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

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

11. Although patient treatment times are dictated by a nephrologist, slight deviations between the prescribed treatment time and the actual treatment time are at the discretion of the centre.

12. Average treatment times in the U.S. were the shortest of the 12 countries surveyed in the study.

13. 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 at 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
Excess mortality	1.041	0.405
Number of centres	4,270	
Number of centre-years	18,295	

Notes: Summary statistics from the dialysis facility reports. The unit of observation is a centre-year. *Patient Years* is the annualized number of patients treated at a centre. *FTE Staff* is the number of full-time equivalent nurses and technicians at the centre. *Net Hiring* is the annual change in FTE staff. *Zero Net Hiring* is a dummy variable equal to 1 if a centre had no net change in FTE staff that year. *Stations* is the number of dialysis stations at the centre. *Zero Net Investment* is a dummy variable equal to 1 if a centre had no net change in stations that year. *Septic Infection Rate* is the percentage of a centre's patients that contracted a septic infection that year. *Excess Mortality* is the ratio of actual deaths to the number of deaths projected by CMS based on the centre's patient characteristics.

reports covering 2004–8 and constructed a usable dataset. The data include detailed centre-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, with several variables deserving note. First, CMS analyses individual patient records and calculates the number of patient-years attributable to each centre (*e.g.* a patient treated at a centre for six months is counted 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.[14] We also use the number of full-time equivalent (a weighted mix of full-time and part-time) employees at each centre as our measure of labour and the number of dialysis stations as our measure of capital. The average centre has 13.5 full-time-equivalent, patient-facing employees and increases its staff by the equivalent of one full-time employee each year (although 12.7% of centres have no net change in employment in a given year). In terms of capital stock, the average number of dialysis stations used by a centre is 18, making the purchase of a new machine a significant investment; reflecting this, centres have zero net investment for 92% of the centre-year observations in the data.

We use a centre's hospitalization rate from septic infections as our primary measure of quality, which averages 12.5% per year and has a standard deviation of over 6%. Finally, in addition to the mortality rate at the centre, CMS includes an expected mortality rate calculated from data on individual patient characteristics (only centre-average characteristics are publicly available). We construct our excess mortality variable as the ratio of realized mortality to expected mortality; this ratio is very close to one on average, but has substantial variation across centres.

In Table 2, we report several factors that may influence a centre's decision to favour quality over quantity, while in Section 5.2 we examine whether any of these proposed shifters are indeed correlated with centres' quality measures. First, states intermittently inspect dialysis centres to

14. Since treatment is mostly standardized at three treatments per week and the goal of dialysis is to clean the blood, we do not consider differences in treatment times as output variation.

TABLE 2
Potential quality drivers

Variable	Mean	St. Dev.	N
State Inspection Rate	29.335	11.640	18,221
Time Since Inspection	1.634	1.813	18,221
% Patients Referred by Nephrologist	69.441	20.346	11,342
Competitors	7.826	13.095	18,221
For Profit	0.883	0.321	18,221

Notes: Summary statistics from the dialysis facility reports. The unit of observation is a centre-year. *State Inspection Rate* is the percentage of centres that were inspected in the centre's state that year. *Time Since Inspection* is the number of years since a facility was last inspected by regulators. *% Patients Referred by Nephrologist* is the percentage of a centre's patients who were referred by a nephrologist. *Competitors* is the number of competing facilities in an HSA. *For Profit* is a dummy variable equal to 1 if a centre is for-profit.

ensure that they meet regulated standards of care. States vary in terms of inspection rates, as the average state inspects 29.3% of its centres each year, with a standard deviation of 11.6%. Because a centre that faces a greater likelihood of inspection may take measures to improve its quality, those in states with more frequent inspections may have a stronger incentive to maintain high quality standards. Similarly, a centre that has not been inspected in many years may have more unaddressed quality issues. The average number of years since a centre was last inspected is 1.6 years, with a standard deviation of 1.8.

In addition, centres that rely more on referrals from nephrologists may have a greater incentive to maintain high-quality standards, as a nephrologist may advocate on his patient's behalf in the event he receives poor treatment or refer patients to better-performing facilities. The average centre receives 69.4% of its patients through referral, with a standard deviation of 20.3. Furthermore, competition may discipline centres' quality choices, as those that face the prospect of losing patients to competitors may take steps to improve their quality. The average centre has nearly eight competing centres in its market—defined as a hospital service area[15]—with a standard deviation of 13.1. Finally, for-profit centres may favour quantity over quality due to their explicit mandate to maximize profits. Nearly 90% of dialysis centres are for-profit.

3. A MODEL OF THE QUALITY-QUANTITY TRADEOFF IN DIALYSIS

To estimate the quality-quantity tradeoff in dialysis, we develop a model that accounts for both the standard endogeneity problems associated with using observed input choices to estimate production functions and the additional problem introduced by a centre's endogenous choice of treatment quality. The complication related to endogenous quality choices stems from the unobserved (to the econometrician) choice made by centres that receive positive shocks to their productivity: they may choose either to treat more patients, or to treat their current patients more intensively.

15. Following the healthcare literature, we use hospital service areas (HSA) as our market definition for dialysis centres. The Dartmouth Atlas determines HSA boundaries based on CMS data for patients' actual hospital choices, and therefore serves 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.

3.1. *The production technology*

In each period, 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 the treatments. Given this setup, a centre faces a production frontier governing the number of patients it treats and the quality of care it provides, with the frontier dictated by its staff, stations, and productivity. Formally, we define a centre's production possibilities frontier as

$$T(\tilde{y}, \tilde{q}) \leq F(k, \ell, \omega). \tag{1}$$

The production function $F(\cdot)$ is the most familiar part of this constraint; it governs how the centre's (log) capital, k, (log) labour, ℓ, and unobserved productivity, ω, determine its overall capacity for providing treatments.

The unobserved productivity term, ω, accounts for all of the factors that impact a centre's production possibilities that are observable to the centre but not to the econometrician, such as a centre's square footage.[16] Because ω affects firms' quality and investment decisions, it also reflects firms' "anticipated" productivity from their assessment of unobservable staff quality (*e.g.* how efficient their staff are at cleaning machines and taking care of patients), unobservable patient characteristics (*e.g.* how well patients follow treatment protocols), and managerial ability. Differences in centres' patient populations are particularly important in a healthcare setting such as dialysis where variation in patient characteristics may lead to large differences in each centre'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.[17] Importantly, we will allow ω to change over time, so that centres' perceptions of their own productivity can evolve as they hire new staff, acquire new capital, or simply gain operational experience.

Our approach, which is common in the production function literature, assumes these differences in unobserved productivity can be summarized by a single index, this implies that all of the differences impact the production frontier in the same way. This assumption holds in our setting because all centres face the same choice following an increase in productivity: they can either treat more patients (*e.g.* by shortening cleaning or treatment times), or they can treat current patients more thoroughly. For instance, if a centre's patients follow treatment protocols more closely than other centres' patients do, then this would free the centre either (i) to treat more patients because it devotes less time to dealing with complications that arise, or (ii) to spend additional time intensively cleaning machines and advising patients, which ultimately improves outcomes but would not appear in a raw productivity measure like output-to-labour ratios. Similarly, better trained staff may be more efficient at cleaning machines or evaluating patients, making both labour and capital more productive. Finally, better designed stations may be easier to clean, allowing the centre either to treat more patients or clean stations more thoroughly in the same amount of time. Given that the fundamental tradeoff we study is based on how much time a centre allocates to treating each patient (including station cleaning times), we believe that allowing all heterogeneity to be summarized by a scalar index is a sensible simplification in an

16. Below, we use a superscript to denote productivity at a particular point in time during a period of the model. Here, we use ω without superscript to denote an argument of the production function. The timing of the model and the evolution of ω is presented in the following section.

17. Although our data will allow us to control for a number of key patient characteristics, some will remain unobserved and therefore must be captured by ω.

industry where "better stations", "better patients", and "better staff", all have the potential to make other factors more productive.[18]

The transformation function, $T(\cdot)$, determines how the centre can divide its productive capacity between the two targeted outputs, the number of patients treated and the quality of these treatments. The first output, $\tilde{y}_{jt}$, is the centre's targeted (log) number of patient-years for the period. The second output, $\tilde{q}_{jt}$, represents the quality of its treatments and is a scalar index representing the centre's targeted infection rate. We assume that $T(\cdot)$ is differentiable and increasing in both outputs, and in estimation we will adopt a parametric form for $T(\cdot)$.[19]

The centre is unable to choose either its patient load or its infection rate directly. For instance, some patients will either die for reasons outside of a centre's control, receive transplants and no longer need dialysis, or transfer to other centres; conversely, a centre's patient load may increase unexpectedly when new patients arrive at random intervals. Patients' infections are also difficult to predict because they will depend not just on the centre's cleaning protocol, but also on whether infectious bacteria exist in the centre. Therefore, instead of modelling the centres as perfectly choosing their quality and quantity, we instead assume they choose targets $(\tilde{y}, \tilde{q})$ that are stochastically related to the observed number of patients and infections. For quantity, 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. Although centres can implement procedures to reduce infections, such as by devoting more time to cleaning machines, many factors outside the centre's control also influence the realized infection rate. 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 centre's manager when she sets quality standards.

3.2. *The timing of dialysis centre decision making*

We now turn to how the centre makes its choices regarding quality, hiring, and investment. Although we are primarily interested in the centre's choice of quality, the hiring decision is an important part of the model because we will adapt the insights of Olley and Pakes (1996) to control for a centre's unobserved productivity.

At the start of the period t, each centre j observes its state, $(k_{jt}, \ell_{jt}, x_{jt}, \omega^q_{jt})$. The vector x_{jt} contains observed variables that affect the centre's incentives for providing high-quality care but that do not directly affect its productive capacity (such as a state's inspection rate). The assumption that the variables in x_{jt} do not affect the production process directly makes them useful excluded variables for tracing out the production frontier, as we discuss in Section 4. In practice, we consider a variety of incentive shifters in x_{jt}, including a centre's for-profit status, its

18. Ideally, we would include multiple dimensions of unobserved productivity—*e.g.* separate terms for labour or capital productivity, or separate terms for producing output or quality. Recent work by Doraszelski and Jaumandreu (2016) and Zhang (2014), building in part on Gandhi *et al.* (2016), propose using a centre's first-order conditions of profit maximization to allow for multiple dimensions of heterogeneity. Unfortunately, since we do not explicitly model centre's objectives, this approach cannot be applied in our setting.

19. Note that "quality" here reflects how carefully the centre acts to reduce the risk of infection and improve health outcomes, 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 centre's operating hours, or even having televisions available during treatments. Despite this, we focus on one specific outcome, low septic infection risk, which is arguably the most prominent dimension due to its severe impact on patients' well-being and its direct connection to quality choices.

state's inspection rate, and the time since it was last inspected.[20] Finally, the centre's assessment of its own productivity, ω_{jt}^q, is not observed by the econometrician. We denote this assessment with a superscript q since it is the productivity centres consider when making their quality decision. As we explain below, a centre's assessment of its productivity will change during the period as it observes its true production and infection rates.

The period consists of the following sequence of events:

(1) *Quality and Quantity Choices Made.* Based on its initial state, the centre chooses its targeted level of quantity and quality for the period, $(\tilde{y},\tilde{q})$, such that it maximizes the objective function described below subject to the production constraint (1).
(2) *Production Occurs.* Based on its chosen target, the centre treats patients and observes realized outcomes for patient loads and infections, (y,q). The centre also updates its beliefs about its productivity to ω^h.
(3) *Hiring and Investment Choices Made.* After observing production, the centre's state is updated to reflect what has been learned about its productivity, becoming $(k_{jt},\ell_{jt},x_{jt},\omega_{jt}^h)$. With this information, the centre 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 that centres' expectations of productivity follow an exogenous Markov process between periods t and $t+1$,

$$E[\omega_{j,t+1}^q|I_{j,t}]=E[\omega_{j,t+1}^q|\omega_{j,t}^h],$$

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

Based on these timing assumptions, centres make decisions at two distinct points during a period. First, they decide how to allocate their effort between quality and quantity. Second, they decide how to adjust capital and labour for the future after observing their current period's production. The key assumption is that a centre can quickly adjust its allocation to quality versus quantity after observing ω^q, whereas it must make its input decisions for the next period before that period's ω^q is revealed. This restriction reflects the relative ease with which dialysis centres can allocate time across tasks relative to adjusting labour or capital, as centres must train and certify staff before they can begin treating patients. For example, to adjust targets to favour quality over quantity, a manager could advise her centre's staff to take extra precautions when treating patients. Alternatively, a centre may reduce quality in favour of quantity by placing less emphasis on cleanliness and more on speed (Pronovost *et al.*, 2006). At the same time, even though a centre can dictate these policy changes more quickly than it can make hiring or investment changes, a lag still exists between a centre's decision about quality and its actual implementation. Within the model, this is reflected in the fact that the centres observe their productivity change during production (from ω_{jt}^q to ω_{jt}^h) but cannot react to it by altering their quality choices. This seems reasonable in our context given that health outcomes cannot be observed until after treatment is provided.

20. We discuss the appropriateness of excluding these variables from the production process in Section 5.2. We also show that the model is robust to changes in the set of incentive shifters.

3.3. *The center's quality decision*

Prior to production, the centre chooses its targeted level of output and quality for the period, respectively $(\tilde{y},\tilde{q})$. The centre's payoff is determined by actual outcomes, which are unknown to the centre when it makes its quality-quantity choice. We assume quality choices are fully flexible from period to period and that quality and output do not affect future states, implying that the centre's quality-choice problem does not have dynamic links.[21] The centre chooses $(\tilde{y},\tilde{q})$ to optimize the static problem,[22]

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

The payoff function $\rho(\cdot)$ represents the returns a centre receives from its realized output and infection rate in the current period given its state variables. As dialysis centres' objectives are difficult to model directly, we remain agnostic as to the precise form of this function other than to assume that it is increasing in both output and quality. For example, one might think of the firm's payoff as

$$\rho(y,q,k,\ell,x) = \pi y - Px(1-q) - G(k,\ell), \tag{3}$$

where π represents the marginal profit from treating a patient; P represents a penalty from being found negligent due to an infection that occurs during an inspection year; the incentive shifter x (or a function of x) in this example may represent the chance of being inspected given state inspection activity; $(1-q)$ is the realized rate of infection;[23] and $G(k,\ell)$ are the fixed costs of operating a centre of a given size. It is clear in this simple setup that, because x changes the relative payoffs of y and q, it should affect the optimal allocation of resources—and hence provide variation in $\tilde{q}(k,\ell,x,\omega^q)$ that is independent of the production function variables (k,ℓ,ω^q).

Although (3) is a plausible start for specifying a payoff function, an advantage of our method is that it does not require an explicit form for $\rho(\cdot)$, which will be challenging to derive in healthcare settings. In addition to being an industry with a sizable number of not-for-profit firms, the health-economics literature in general has struggled with how to model the objectives of doctors given that the Hippocratic oath, though only informally enforceable, implies some non-profit motive driven by patients' welfare. Moreover, factors such as torts, malpractice insurance, social reputation, and professional sanction through medical boards all play a role in shaping doctors' incentives. Untangling these incentives would be interesting, but is well beyond the scope of our article. However, we can still estimate the key *technological* parameters of $T(\cdot)$

21. Our assumption that the number of patients treated in the current period does not affect the state of the centre 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 centre 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.

22. For notational clarity, we suppress firm and time indices where they are clear from the context throughout the remainder of the article.

23. We say we use infection rates as a proxy for quality. As they are negatively correlated, it might be more intuitive to say we use the negative infection rate, but this quickly becomes tedious and we avoid doing so. The two approaches are clearly isomorphic.

without directly modeling the *incentives* encapsulated in $\rho(\cdot)$. That is, we only use the fact that we observe some variable x that shifts these incentives on the margin.

Returning to the centre's quality-quantity problem, the expectation in (2) is taken over $(\varepsilon^y, \varepsilon^q)$, which represent the random deviations from the centre's targeted quantity and quality levels and its observed outcomes. By construction, these variables are mean zero and uncorrelated with the centre's information set when it makes its quality decision.

We will assume that the demand for dialysis is inelastic in the sense that a reduction in $\tilde{q}$ will lead to an increase in $\tilde{y}$ through the production frontier, which will result in a corresponding increase in the number of patients treated. This fits with the tight capacity in the industry, as reflected by wait-lists for treatment and the opening of new centres in many markets.[24]

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

Lemma 1. *The centre's expected per-period return to labour is increasing in ω^q; i.e. $\frac{\partial \pi}{\partial \ell}$ 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 centre 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 setting its quality-quantity target, treatments occur and the centre observes the number of patients it treats and its infection rate for the period. In addition, the centre updates its beliefs about its productivity according to $\omega^h = \omega^q + \varepsilon^\omega$ and learns the realization of the shocks $(\varepsilon^y_{jt}, \varepsilon^q_{jt}, \varepsilon^\omega_{jt})$.[25] Importantly, the components of this vector may be correlated with each other. That is, conditional on both ε^y_{jt} and ε^q_{jt} being positive, we would expect the centre to raise its assessment of its own productivity.

After observing patient loads and infection rates, the centre makes its hiring and investment decisions for the following period. This choice has dynamic implications owing to the time it takes to install new machines and train new workers. The Bellman equation for this choice is

$$V^h(k,\ell,x,\omega^h) = \max_{i,h} -c(i,h) + \beta E[V^q(k+i,\ell+h,x',\omega^{q'})|k,\ell,\omega^h,i,h], \tag{4}$$

where i is net investment and h is net hiring, while the function $c(\cdot)$ captures adjustment costs for investment and hiring.[26]

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

24. In Section 6.2, we perform a robustness check by dropping centres in markets where the ratio of stations to the general population is particularly high, as these are the centres where a reduction in quality is most likely to lead to slack capacity rather than increased output.

25. 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 $E[\omega^h|\omega^q] = \omega^q$.

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

workers' on-the-job incentives to strive for either more 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 $h=0$.[27]

The function $V^q(\cdot)$ represents the value of the centre 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 notation because the centre's perception of its own productivity evolves over the course of the period from ω^q to ω^h as the centre observes 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. In contrast, we view the hiring choice as effectively continuous. This seems reasonable given a centre's option to adjust nurses' hours from period to period and that we observe part-time staff in the data. 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 centre's hiring choice, $h(k,\ell,x,\omega^h)$.

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

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

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 centres' discrete investment choices: if a centre invests in a new station, the cost of this new investment may lead the centre 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 within our data, we can isolate cases where centres make the same investment choice (*e.g.* keep the number of stations constant) and conclude that centres 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.

4. IDENTIFYING THE QUALITY-QUANTITY TRADEOFF

For our empirical application, we will adopt the following parsimonious functional forms to describe the transformation and production functions:

$$T(\tilde{y}_{jt},\tilde{q}_{jt})=\tilde{y}_{jt}+\alpha_q\tilde{q}_{jt}, \tag{5}$$

$$F(k_{jt},\ell_{jt},\omega^q_{jt})=\beta_k k_{jt}+\beta_\ell \ell_{jt}+\omega^q_{jt}. \tag{6}$$

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

27. Such a fixed adjustment cost would lead to a "zone of inactivity" in which the centre does not adjust its staffing level for a range of productivity levels. 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 and 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% of firm-year observations (Table 1).

assume a Cobb–Douglas-like specification that parameterizes the production possibilities frontier by assuming that a reduction in the infection rate of 1 percentage point (*i.e.* increasing $\tilde{q}_{jt}$ by 1) will reduce expected output by a factor of α_q, which is constant across centres. These functional forms can be rationalized by assuming that centres optimally allocate their productive capacity between the two tasks of treating more patients and reducing infection rates. If we assume, without loss of generality, that all centres choose to operate on the production frontier, we can then re-write the centre's static optimization problem as choosing the proportion of resources, λ, to devote purely to output:

$$\begin{aligned} \pi(k,\ell,x,\omega^q) = \max_{\lambda\in[0,1]} & E[\rho(y,q,k,\ell,x)] \\ \text{subject to: } & \tilde{y} = \lambda(\beta_k k + \beta_\ell \ell + \omega^q) \\ & \tilde{q} = \frac{1-\lambda}{\alpha_q}\left(\beta_k k + \beta_\ell \ell + \omega^q\right) \\ & y = \tilde{y} + \varepsilon^y \\ & q = \tilde{q} + \varepsilon^q. \end{aligned} \tag{7}$$

This problem yields an optimal policy for allocating resources, $\lambda(k,\ell,x,\omega^q)$, which through the two production constraints generates the centre's targeted levels of output and quality, $\tilde{y}(k,\ell,x,\omega^q)$ and $\tilde{q}(k,\ell,x,\omega^q)$. In a model with two independent outputs, such as cars and boats, λ could be viewed as literally assigning workers and capital to assembly tasks. In our setting, it is better to think of λ as a choice of production speed. A high λ would mean placing an emphasis on running many patients though the centre quickly—and hence less emphasis on cleaning machines and evaluating patients' health before, during, and after treatment. This allows the centre to increase $\tilde{y}$ by sacrificing $\tilde{q}$, with α_q representing the key technological parameter that governs this tradeoff.[28]

We do not have data on how the centre allocates staff and machines, or on how it sets its cleaning and evaluation policies, so we cannot claim to directly observe the centre's choice of λ. Instead, we observe the realized outcomes, y and q, and therefore rewrite the problem focusing on the choice of $\tilde{y}$ and $\tilde{q}$ through λ, adding the two production constraints to yield our production frontier:

$$\begin{aligned} \pi(k,\ell,x,\omega^q) = \max_{(\tilde{y},\tilde{q})} & E[\rho(y,q,k,\ell,x)] \\ \text{subject to: } & \tilde{y} + \alpha_q \tilde{q} = \beta_k k + \beta_\ell \ell + \omega^q \\ & y = \tilde{y} + \varepsilon^y \\ & q = \tilde{q} + \varepsilon^q. \end{aligned} \tag{8}$$

This specification of the centre's quality choice maps directly into the more general model discussed in the previous section and connects a centre's quality target to observable outcomes. By increasing the effort it puts towards providing high-quality treatments, the centre incurs additional costs but increases the likelihood of delivering better outcomes (*e.g.* fewer infections and lower mortality). On the other hand, a change in inputs or productivity shifts the production possibilities

28. Note that while α_q only appears in the restriction for $\tilde{q}$, this is simply a normalization. To be precise, this normalization is that the production function $F(k,\ell,\omega^q) = \beta_k k + \beta_\ell \ell + \omega^q$ is expressed in units of (log) output. If we included a separate parameter for the transfer of "productivity units" to output, only the ratio of this new parameter and α_q would be identified. Of course this ratio would still express exactly the tradeoff in which we are interested.

frontier but does not alter the relative transformation between outputs. For instance, a centre 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.

Our goal is to estimate this production frontier, but this is complicated in that we do not observe centres' 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 $y=\tilde{y}+\varepsilon^y$ and that the observed infection rate is $q_{jt}=\tilde{q}+\varepsilon^q$. Substituting these into (1), we arrive at the linear equation

$$y=-\alpha_q q+\beta_k k+\beta_\ell \ell+\omega^q-\alpha\varepsilon^q-\varepsilon^y. \tag{9}$$

This equation makes the potential sources of bias apparent. Estimating (9) by ordinary least squares with data on (y,q,k,ℓ) would be inconsistent, as the composite error term $\omega^q-\alpha\varepsilon^q-\varepsilon^y$ is correlated with observed quality for two separate reasons, simultaneity due to ω^q and measurement error due to ε^q.

To identify the parameter that governs the quality-quantity tradeoff, α_q, as well as the production function parameters β, we must address three different issues: (i) independent variation of $\tilde{q}$ from the other variables that enter the production frontier; (ii) simultaneity due to the fact that the centre chooses $\tilde{q}$—or equivalently in (7), λ—with knowledge of ω^q; and (iii) attenuation bias due to the fact that we do not directly observe $\tilde{q}$ (centre quality), but instead only observe a noisy proxy, q, the (negative) realized infection rate. We take different approaches to address these issues: (i) our timing assumptions that allow productivity to evolve during production but before hiring (separating ω^q and ω^h) provide independent variation of $\tilde{q}$, while variation in x also provide variation in firms' quality choices (as we describe below); (ii) our dynamic model implies that hiring this period to change the quantity of labour between this period and the next period can be used to construct a control function to deal with simultaneity; and (iii) a second noisy outcome related to the quality of care (our empirical approach uses the ratio of actual mortality to expected mortality) is used as an instrument to address measurement error in the initial proxy outcome, q. For expositional ease, we deal with these problems sequentially, starting with a model where neither simultaneity nor measurement error is an issue, before building up to the full model we estimate in the following section. We focus on recovering the quality-quantity tradeoff parameter α_q, as identification of the β parameters is essentially identical to that presented in Olley and Pakes (1996) and Ackerberg *et al.* (2015). We provide the full estimation details, including estimation of β, in Section 5.

4.1. *The simplest model*

We begin with a highly simplified case of the full model in which we assume (i) the infection rate is a perfect indicator of the quality choice—so $q=\tilde{q}$ and there is no ε^q—and (ii) there is not a heterogeneous productivity process (*i.e.* we drop ω^q and ω^h). With these assumptions, the production frontier constraint (9) becomes

$$y=-\alpha_q q+\beta_k k+\beta_\ell \ell-\varepsilon^y, \tag{10}$$

where we have substituted ε as in eq. (10) $y=\tilde{y}+\varepsilon^y$. Since the error term ε^y is unanticipated by the centre when it makes its quality decision (at the start of this period) and when it determines its capital and labour levels through investment and hiring (at the end of the previous period), there is no simultaneity problem in this version of the model. Similarly, with no ε^q, there is no

measurement error problem. Therefore, as long as we have independent variation among (q,k,ℓ), we can consistently estimate the parameters $(\alpha_q,\beta_k,\beta_\ell)$ using ordinary least squares (OLS). However, the optimization problem that governs the quality choice (or the resource allocation choice) now implies that $q=\tilde{q}(k,\ell,x)$.[29] To see why this is an issue, suppose x did not vary across centres, so we effectively have $q=\tilde{q}(k,\ell)$. In that case, all centres with the same (k,ℓ) *must* make the same quality choice. Furthermore, if the policy function is linear, then (q,k,ℓ) are co-linear and the parameters are not identified. If the centre's policy function $q=\tilde{q}(k,\ell)$ is non-linear, then the parameters are identified with the help of the parametric assumption of the production technology, though such identification off of the functional form of the production frontier is clearly not ideal.

If we reintroduce x, however, the problem is relaxed *even if we do not observe x ourselves.* Now, because the centre's policy function is $q=\tilde{q}(k,\ell,x)$, independent variation is restored and OLS successfully recovers the parameters. It is not necessary to use x as an instrument because there is no endogeneity issue, but x provides a source of variation in q that allows us to identify α_q, which illustrates the importance of including x as an incentive shifter in the structural model.

4.2. *Adding heterogenous productivity*

We next incorporate heterogeneity in unobserved total factor productivity across centres. To illustrate the need for our timing assumptions, we will begin with productivity being constant over the period, as in Olley and Pakes (1996), so $\omega^q=\omega^h=\omega$. We continue to assume that quality is perfectly observed through the infection rate, so we have

$$y=-\alpha_q q+\beta_k k+\beta_\ell \ell+\omega-\varepsilon^y, \tag{11}$$

which introduces a simultaneity issue between q and ω. Because productivity is a persistent Markov process, there is also simultaneity between ω and the centre's inputs (k,ℓ), which are a function of lagged productivity. Olley and Pakes (1996) show how to address this simultaneity issue using the centre's dynamic decisions. Under the assumptions of our dynamic model, Proposition 1 shows that the centre's optimal hiring policy, $h_{jt}(k_{jt},\ell_{jt},x_{jt},\omega_{jt})=\ell_{it+1}-\ell_{jt}$, is invertible in the productivity term for a given level of discrete capital investment (*i.e.* a change in the number of stations).[30] This means we can replace ω_{jt} in (11) with the inverted hiring policy $h^{-1}(h,i,k,\ell,x)$, where we control for the centre's discrete level of investment i as part of this inversion. Of course, β_k and β_ℓ are not separately identified from the non-parametric inverted hiring function, so they are subsumed into a general non-parametric function, resulting in the potential semi-parametric estimating equation

$$y=-\alpha_q q+\Phi(h,i,k,\ell,x)-\varepsilon^y. \tag{12}$$

This equation addresses the simultaneity problem and is essentially the estimating equation from the first stage of Olley and Pakes (1996) adapted to our model.

Ackerberg *et al.* (2015) point out a subtle identification problem with this equation, however, which is very similar to what we discussed in the previous subsection. Recall that the policy function for quality is $\tilde{q}(k,\ell,x,\omega)$, and employing the same invertibility argument for productivity we have

$$q=\tilde{q}(k,\ell,x,h^{-1}(h,i,k,\ell,x)).$$

29. Recall that we have temporarily dropped ω^q from the model.
30. Recall that because q_{jt} is flexible and has no dynamic implications, it does not enter the hiring policy.

While this involves some abuse of notation, the point is that, under these assumptions, there is no independent variation between q and the variables of the non-parametric function $\Phi(\cdot)$ because they are both functions of (h,i,k,ℓ,x). Therefore, α_q is not separately identified in (12). In this case, as opposed to the previous subsection, we cannot appeal to x alone to provide the necessary variation because it enters the hiring policy function, and hence $\Phi(\cdot)$.

Ackerberg *et al.* (2015) consider several modifications of Olley and Pakes (1996) to address this issue. Our approach adopts one which will be compatible with our measurement error concerns below and which we believe fits the dialysis industry well: we make use of the timing assumption that allows for an intra-period evolution of productivity in the form of ω^q and ω^h. Notice that the simultaneity issue is really due to the centre's *perception* of its productivity. It is natural to assume that, having observed output and the infection rates in the current period, the centre's perception of its productivity should evolve before it makes investment and hiring decisions for the next period. This reintroduces independent variation between q and the non-parametric function because the quality policy is $q=\tilde{q}(k,\ell,x,\omega^q)$, where the inverted hiring function yields $\omega^h=h^{-1}(h,i,k,\ell,x)$. The two are related through the assumption that $\omega^h=\omega^q+\varepsilon^\omega$, where ε^ω is uncorrelated with the centre's information set when it makes its quality choice.[31] In this context, ε^ω can be thought of as a "discovering productivity by doing" shock. Although we do not explicitly model a learning process, the shock is meant to reflect the fact that the centre's assessment of its own productivity is likely to change when the task is actually completed. We can now write the first stage as

$$y=-\alpha_q q+\Phi(h,i,k,\ell,x)-\varepsilon^\omega-\varepsilon^y, \tag{13}$$

where the new error term reflects the fact that we replace ω^q using inverted hiring and the difference between productivity at the beginning and end of the period. This composite error term is uncorrelated with q and there is independent variation between q and the non-parametric function $\Phi(\cdot)$.[32]

Ackerberg *et al.* (2015) do not include a variable akin to our x, though doing so offers important advantages. Namely, if x is dropped from the model, then all firms with the same production frontier choose the same quality policy, $q(k,\ell,\omega^q)$. Although perturbations between ω^q and ω^h introduce variation between q and the non-parametric function $\Phi(\cdot)$, it is still the case that if x does not vary, centres facing the same production frontier—*i.e.* the same k,ℓ,ω^q—all choose the same point on the frontier. As a result, the variation between ω^q and ω^h only allows us to identify α_q at a point on the frontier (the optimal point, given k,ℓ,ω^q) and the slope of the entire frontier is identified off of functional form. Again, including incentive shifters in x relaxes this issue, because now centres facing the same production frontier choose different levels of output and quality due to differences in x. Note that, as opposed to the simple model discussed above, in this model it is important that we observe x because it potentially plays a role in the hiring policy.

Of course, x only supplies useful variation if it actually leads firms to shift their quality targets. We rely on a set of incentive shifters that includes the centre's for-profit status, its state's inspection rates, and the time since its last inspection. In Section 5.2, we provide evidence that these variables do, in fact, affect centres' incentives to provide high-quality care.

31. The random walk assumption is for notational convenience. We could allow $\omega^h=E[\omega^h|k,\ell,x,\omega^q]+\varepsilon^\omega$ as long as this expectation were increasing in ω^q. Since we are unable to recover ω^q, the random walk assumption amounts to a normalization.

32. Note that at this point we are still assuming that quality is perfectly observed (*i.e.* no ε^q).

4.3. *Adding measurement error in the quality policy*

Finally, we come to the full model by re-introducing measurement error in q by allowing the infection rate to be a noisy signal of the centre's quality choice. Here, we do not mean to say that the infection rate itself is mis-measured, but instead that the infection rate is determined by both the centre's quality choice and some noise: sometimes bacteria do not infect a patient even if the machine is not properly cleaned, and sometimes the centre follows its protocol but a patient acquires an infection for reasons unrelated to dialysis. This means we observe $q=\tilde{q}(k,\ell,x,\omega^q)+\varepsilon^q$, and the first stage becomes

$$y=-\alpha_q q+\Phi(h,i,k,\ell,x)-\varepsilon^{\omega}-\alpha_q\varepsilon^q-\varepsilon^y. \tag{14}$$

Although this introduces attenuation bias because q and ε^q are correlated, it does *not* reintroduce the simultaneity problem. The reason is that ε^q is not known to the centre when it makes its quality (or resource allocation) choice. Therefore, we follow the literature on classical measurement error to address this issue by instrumenting for the infection rate with a second noisy measure of quality, the (excess) mortality rate.[33] Specifically, we assume $d=d(\tilde{q}(k,\ell,x,\omega^q))+\varepsilon^d$, where $d(\cdot)$ is a monotonic function of quality and ε^d is uncorrelated with ε^q. After controlling for patient characteristics and the centre's quality, we would expect infections and deaths to be uncorrelated at the centre level, and we can then use d as an instrument to identify α_q.

Although the validity of this assumption may appear questionable in the sense that infected patients clearly have a higher death rate, the assumption holds if, after controlling for the centre's quality, individual patient outcomes are uncorrelated. To see this, suppose each patient r of centre j has an individual risk of infection and death related to centre quality, $\tilde{q}_j$, so that the risk of infection is $q_j^r=\tilde{q}_j+\eta^r$ and the risk of death is $d_j^r=d(\tilde{q}_j)+\nu^r$. We allow these individual infection and death shocks to be correlated for a given patient but we assume that (η^r,ν^r) are drawn independently across patients. As a result, centre-level infection and death rates are Poisson binomial distributions of the underlying individual risks. For centres with enough patients, the covariance between centre-level infection and death rates is well approximated by

$$\operatorname{cov}(q_j,d_j)\approx\frac{1}{N^2}\sum_{r,s=1}^{N}\operatorname{cov}(q_j^r,d_j^s)=\operatorname{cov}(\tilde{q}_j,d(\tilde{q}_j))+\frac{1}{N^2}\sum_{r=1}^{N}\operatorname{cov}(\eta_j^r,\nu_j^r),$$

where the final term exploits the fact that $\operatorname{cov}(\eta_j^r,\nu_j^s)=0$ if $r\neq s$. Clearly, the second term vanishes as the number of patents at the centre N grows large. With a mean number of patient years in our sample over 50, it seems reasonable to believe that the correlation in centre-level averages is being primarily driven by centre quality.

We acknowledge that if ε^q and ε^d are correlated, then our estimate of α_q will be biased downward, making our results conservative. This could happen either because N is not large enough, or because patient outcomes are not independent after controlling for $\tilde{q}$ and the observable characteristics of the patient population. For example, if infections and deaths had a stochastic component that was common to all patients in a centre (*e.g.* a virulent outbreak of staph infection at the local hospital), our result would not hold. The most obvious correlated shock to patients' infection and death rates, however, is infections spread through an improperly cleaned dialysis machine, which is directly related to the centre's quality choice when determining how closely

33. This is the mortality rate controlling for the expected number of deaths in the centre based on patient characteristics, as discussed in Section 2.

TABLE 3
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. Haemoglobin Level	11.882	0.332
Number of Center-Years	18,221	

Notes: Summary statistics from the dialysis facility reports. The unit of observation is a centre-year. *Avg. Patient Age* is the average age of patients at a centre. *Pct. Female* is the percentage of a centre's patients who are women. *Pct. AV Fistula* is the percentage of a centre's patients who have an AV fistula. *Avg. Comorbid Conditions* is the average number of comorbid conditions a centre's patients have. *Avg. Duration of ESRD* is the average number of years a centre's patients have had kidney failure. *Avg. Haemoglobin Level* is the average haemoglobin levelof a centre's patients.

to follow cleaning protocols. In Section 6.2, we present a robustness check of the model where we do not use an instrument for quality. The results of this experiment are consistent with the presence of attenuation bias.

5. IMPLEMENTATION

We now turn to the details of the empirical analysis, beginning with the construction of our quality proxy. We then discuss our choice of incentive shifters for x and provide some *prima facie* evidence that centres with similar technologies actually do react to our incentive shifters in the predicted way, thus justifying the inclusion of these variables in the estimation. Finally, we describe the two-step estimator used to estimate the transformation and production functions.

5.1. *Infection rate proxy for quality*

Although we do not observe treatment quality directly, the data contain information on patient outcomes that are correlated with a centre's choices on this dimension. In particular, we focus on the centre'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 centres rather than centres' deliberate quality choices. To account for this, we control for centre-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 centre's targeted quality level.

We control for several observable patient characteristics that influence a centre's infection rate beyond its quality decision, with summary statistics displayed in Table 3. 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 centres' 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 haemoglobin levels.[34] Putting these centre-level average patient characteristics together into the vector z_{jt}, we estimate

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

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

Even after controlling for observable characteristics, some unobservable differences in patient health may remain, part of which may be observable to centres as they make their quality choices. Within our model, we interpret these unobservable differences as differences in ω_{jt} across centres along with other unobservable differences in productivity (*e.g.* management ability or unobserved quality of inputs), and address them using a control function approach.

As discussed above, we account for expectational or measurement error in our specification of the production function by including ε_{jt}^{q} and using a second outcome variable as an instrument, in our case the excess mortality ratio that compares actual death rates at the centre to CMS's expected death rate based on individual patient characteristics.[35] CMS uses this ratio as an indicator of centre quality in its own reports, and we include this measure as a second noisy proxy for a centre's quality.[36]

5.2. *Centres' quality incentive shifters*

An important source of identification for our estimation procedure comes from variation in x, the variables that affect the centre's incentives for providing high-quality treatments. These variables should consist of observable factors that affect centres' incentives for providing high-quality care *without* directly affecting centre productivity. That is, these variables should be excluded from the production frontier (1) but shift outcomes through their effect on centre payoffs, $\rho(\cdot)$. Note that, while the variables in x cannot directly enter into the production process, they may be correlated with centres' unobserved productivity. For example, our assumption rules out the possibility that a non-profit centre could alter its production frontier simply by becoming for-profit while keeping the same staff and management practices in place. At the same time, it does allow for the possibility that low-productivity non-profit centres are more likely to enter into the industry than low-productivity for-profit centres, which would generate a correlation between non-profit status and productivity.[37]

34. Low haemoglobin levels are associated with anaemia and pose health risks for dialysis patients.

35. Individual-level characteristics are not released in our data to protect patient privacy.

36. As discussed earlier, for this instrument to be valid, the variation in excess mortality which is unrelated to centre quality should be uncorrelated with the variation in the infection rate that is unrelated to quality. As discussed above, our assumption will hold approximately even if mortality and infection outcomes are correlated at the individual level as long as outcomes are uncorrelated across patients (after controlling for centre quality). Moreover, although infections do raise the risk of death, they are the primary cause of fewer than 10% of patient deaths overall (USRDS, 2013). In contrast, over one-third of patient deaths stem from cardiovascular issues, which may be related to centre quality through the effectiveness of monitoring for hypotension and other complications that arise during treatment. Conversely, the majority 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 centres (*e.g.* cleaning protocols) than the septic infection rate. Results are qualitatively robust when the roles of the primary proxy and the instrument are reversed.

37. In Section 6.2, we explicitly allow centres' productivity evolution to depend on ownership type, one of our variables in x.

We consider several variables that are likely to change a centres' incentives for providing more treatments versus providing higher-quality care, which we describe below. Although the assumption that these variables do not directly affect the production frontier is not testable, we can nevertheless verify they are correlated with health outcomes.[38] To examine the relationship between quality and the observable variables in our data, we consider a series of fixed-effects regressions of the septic infection rate on plausibly exogenous drivers of quality and patient characteristics in Table 4. The results suggest that centres facing stronger incentives to provide high-quality care have better outcomes, indicating that centres adjust their quality level when it is in their best interest to do so.[39]

First, we look at the time since a facility was last inspected by a state authority, as centres with many reported violations face the possibility of losing certification.[40] Column (1) of Table 4 shows that the infection rate increases a statistically significant 0.15 percentage points for each year since a centre was last inspected, or about 11.8% at the mean. Note that the most likely source of endogeneity bias for this regression—that regulators are more likely to target centres with poor health outcomes for inspections—works against this result, making it a conservative estimate. In addition, it seems unlikely that the number of years since a centre was last inspected would directly affect its production frontier, so we include this variable in x.

We next turn to the state-level inspection rates. This incentive shifter relies on year-to-year variation in funding and other local regulations to exogenously change the probability that a centre is inspected—and hence affects the return to providing high-quality care (which presumably reduces the risk of receiving health citations during an inspection). As shown in Column (2), we find evidence that an increase in the state inspection rate is associated with a lower infection rate: an increase in the inspection rate one standard deviation above the mean is associated with a 0.1 percentage point decline in the infection rate, on average, which is about 1% lower at the mean. As with the number of years since a centre was last inspected, it seems unlikely that the state inspection rate directly affects the production frontier of individual firms, so we also include this as a potential incentive shifter.

Thirdly, we consider the possibility that the referral rates of nephrologists might affect treatment quality. If nephrologists are able to ascertain a centre's quality, then centres that are more dependent on referrals will have stronger incentives to provide high-quality care. Column (3) shows that a higher rate of referrals is associated with a lower infection rate, although this correlation may be due to reverse causality—presumably nephrologists will not refer patients to centres with poor quality records. Moreover, the referral rate is available for only three years of our five year panel, severely reducing the number of observations in the data. For these reasons, we do not include referral rates in x as a part of the structural model.

Fourthly, competition potentially shifts centres' quality incentives because centres in highly competitive markets may choose to improve quality or increase staff levels to attract patients.[41] We define the level of competition each centre faces as a categorical variable for having 0, 1, 2, or 3 or more competitors in the same Hospital Service Area (HSA), and the regressions show little

38. As discussed above, the question of whether external forces influence dialysis facilities' incentives for providing high-quality care has been investigated in previous studies (Sloan, 2000; Ramanarayanan and Snyder, 2011).

39. Of course, these regressions are subject to the critique that they are not robust to across-time changes in productivity; our structural model does control for productivity changes over time.

40. We know only the year of the last inspection, so if a centre is inspected this year there are "zero" years since its last inspection.

41. Although in principle the degree of competition is the outcome of an endogenous dynamic entry game between players, we believe it is reasonable to model it as exogenous to the centre's quality choice as the latter is a flexible decision made on a much shorter time frame by centre managers, whereas the entry and exit decision of centres involves a substantial sunk cost and is the result of a lengthy regulatory process.

TABLE 4
Infection rate and quality incentive shifters

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Years Since Inspection	0.1477				0.1412	0.1413			0.1332
	(0.0309)				(0.0313)	(0.0313)			(0.0292)
State Inspection Rate		−0.0103			−0.0068	−0.0069			−0.0096
		(0.0049)			(0.0050)	(0.0050)			(0.0053)
% Patients Referred by Nephrologist			−0.0092						
			(0.0036)						
1 Competitor				−0.6208		−0.6160			−1.1704
				(0.3759)		(0.3756)			(0.3825)
2 Competitors				−0.1056		−0.0804			−1.1254
				(0.5012)		(0.5008)			(0.5082)
3+ Competitors				−0.2373		−0.1961			−1.7079
				(0.5932)		(0.5928)			(0.6029)
Non-Profit							−1.2878	−1.5946	−1.6137
							(0.2061)	(0.2240)	(0.2239)
DaVita								0.0250	0.0072
								(0.1548)	(0.1546)
Fresenius								−0.8834	−0.9038
								(0.1562)	(0.1561)
Patient characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Center fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
HSA fixed effects	No	No	No	No	No	No	Yes	Yes	Yes
Observations	18,221	18,221	11,342	18,221	18,221	18,221	18,221	18,221	18,221
R^2	0.573	0.573	0.663	0.573	0.573	0.574	0.406	0.408	0.409

Notes: OLS regressions with standard errors in parenthesis. The unit of observation is a centre-year. The dependent variable is a centre's septic infection rate. *Years Since Inspection* is the number of years since a facility was last inspected by regulators. *State Inspection Rate* is the percentage of centres that were inspected in the centre's state that year. *% Patients Referred by Nephrologist* is the percentage of a centre's patients that were referred by a nephrologist. *1 Competitor* is a dummy variable equal to 1 if a centre has 1 competitor in its HSA. *2 Competitors* is a dummy variable equal to 1 if a centre has 2 competitors in its HSA. *3+ Competitors* is a dummy variable equal to 1 if a centre has 3 or more competitors in its HSA. *Non-Profit* is a dummy variable equal to 1 if a centre is non-profit. *DaVita* is a dummy variable equal to 1 if a centre belongs to the DaVita chain. *Fresenius* is a dummy variable equal to 1 if a centre belongs to the Fresenius chain. Patient characteristics include AV fistula, average number of comorbid conditions, average duration of ESRD, average age, gender distribution, and average haemoglobin levels.

evidence that competition influences the quality of care provided by centres whether we include a centre-level or market-level fixed effect. Although monopolists do appear to have higher infection rates in the market fixed-effects specification in column (9), the centre fixed-effects specification in column (4) is weaker and shows a non-monotonicity between 1 and more than 1 competitor. Still, even though competition does not seem to have a strong impact on quality, we include it in x as it is clearly appropriate to exclude competition from the production frontier, and it may affect the centre's hiring policies. We provide a robustness check where the model is estimated without including the competition variable in Section 6.2, and all results remain unchanged.

Finally, we consider the effect of a centre's ownership structure, as roughly 87.7% of centres operate as for-profit entities. Among the for-profit centres, two major chains dominate, with DaVita owning roughly 28% of centres nationwide and Fresenius 31%.[42] A centre's ownership structure may affect its polices related to hiring and treatment quality: non-profit centres may, on average, target a different weighting of quality over quantity, and even different corporate owners may have different views on the risks and benefits of reducing quality to treat more patients. Because there is almost no variation in a given centre's for-profit status over time, we consider the impact of for-profit status with market-level fixed effects in Column (7). We find strong evidence that for-profits have worse health outcomes, with infection rates roughly 1.3 percentage points higher then non-profit centres in identical markets. Column (8) adds controls for membership in the two major chains. It appears that, although DaVita's average infection rate is almost the same as unaffiliated for-profits', Fresenius is somewhat lower. Although it is certainly possible that ownership status is correlated with productivity, we see little reason why the legal status of a centre should directly affect the production frontier. Therefore, we include non-profit status, along with the affiliation with one of the two most prominent chains, in our preferred specification of x. In separate robustness checks, we allow for centres' productivity evolution to depend on its ownership status, and also estimate the model dropping ownership status from x. In both cases, our results on the quality-quantity tradeoff are qualitatively similar.

Although each of the regressions in Table 4 suffers from its own particular shortcomings (*e.g.* both referrals from nephrologists and the likelihood of inspection are potentially endogenous with respect to centres' quality choices), taken together they provide consistent evidence that plausibly exogenous drivers of quality do, in fact, influence a facility's quality choices. Note also that Specifications (1)–(6) control 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. To summarize, we include the following variables in our preferred specification of incentive shifters, x: the number of years since a centre was last inspected, the state inspection rate, the extent of competition, and the centre's ownership structure.

5.3. *Two-step estimation*

To recover the parameters of the production frontier, we first note that $\omega_{jt}^h = \omega_{jt}^q + \varepsilon_{jt}^\omega$, so we can rewrite (9) as

$$y_{jt} = -\alpha_q q_{jt} + \beta_k k_{jt} + \beta_\ell \ell_{jt} + \omega_{jt}^h - \varepsilon_{jt}^\omega - \alpha\varepsilon_{jt}^q - \varepsilon_{jt}^y.$$

Because $(\varepsilon_{jt}^\omega, \varepsilon_{jt}^q, \varepsilon_{jt}^y)$ are revealed to the centre after it makes its quality choice and are uncorrelated with the centre's information set at the time quality and output choices are made, they do not impose an endogeneity problem. Because centres' expectations about ω_{jt}^h are a function of ω_{jt}^q, however, we must still control for ω_{jt}^h. From Proposition 1, we know that a centre's expectation

42. These averages are taken across all years in the data.

about its productivity at the time of hiring can be recovered by inverting the centre's hiring policy at a fixed investment level such that

$$\omega_{jt}^{h} = h^{-1}(h_{jt}, i_{jt}, k_{jt}, \ell_{jt}, x_{jt}). \tag{15}$$

Substituting (15) into (9), we arrive at our first-stage estimating equation,

$$\begin{aligned} y_{jt} &= -\alpha_q q_{jt} + \beta_k k_{jt} + \beta_\ell \ell_{jt} + h^{-1}(h_{jt}, i_{jt}, k_{jt}, \ell_{jt}, x_{jt}) - \varepsilon_{jt}^{\omega} - \alpha\varepsilon_{jt}^{q} - \varepsilon_{jt}^{y} \\ &= -\alpha_q q_{jt} + \Phi(h_{jt}, i_{jt}, k_{jt}, \ell_{jt}, x_{jt}) + \varepsilon_{jt}, \end{aligned} \tag{16}$$

where $\varepsilon_{jt} = -\varepsilon_{jt}^{\omega} - \alpha\varepsilon_{jt}^{q} - \varepsilon_{jt}^{y}$ and $\Phi(h_{jt}, i_{jt}, k_{jt}, \ell_{jt}, x_{jt}) = \beta_k k_{jt} + \beta_\ell \ell_{jt} + h^{-1}(h_{jt}, i_{jt}, k_{jt}, \ell_{jt}, x_{jt})$.

Due to invertibility requirements, we only have usable observations of (16) whenever hiring is non-zero.[43] Moreover, because the function $h^{-1}(\cdot)$ depends on the level of investment, we must estimate a separate $\Phi(\cdot)$ for each investment level. In the data, investment is zero in 92% of observations, therefore, in practice, we estimate the model using only observations where the center did not invest while dropping observations with non-zero investment. Dropping observations where hiring is zero and investment is non-zero collectively reduce the size of the dataset by 19%, but this truncation does not bias our results if the model is correctly specified. Comparing the dropped observations to those used in the first stage, centres with dropped centre-years are slightly smaller on average but have similar quality outcomes (infection rates and ratios of deaths to expected deaths). Running the descriptive analysis presented in Table 4 on the truncated sample also produces qualitatively similar results.

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

Although the approach above handles the endogeneity of ω_{jt}^{q}, we still have attenuation bias because ε_{jt} and q_{jt} are correlated through ε_{jt}^{q}. To address this, we use a second noisy measure of quality as an instrument in the second stage of a three-stage procedure.[44] First following Robinson (1988), we estimate $\widehat{E}[y|h_{jt}, i_{jt}, k_{jt}, \ell_{jt}, x_{jt}]$ and $\widehat{E}[q|h_{jt}, i_{jt}, k_{jt}, \ell_{jt}, x_{jt}]$ non-parametrically using local linear regression.[45] In doing so, we are careful to account for the possible discontinuity of these functions at $h_{jt} = 0$ by considering positive and negative hiring observations separately.[46] We then estimate $\hat{\alpha}_q$ with the linear instrumental variables regression

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

43. Because there are likely adjustment costs to hiring, $h^{-1}(\cdot)$ 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.

44. An alternative approach, following Ackerberg *et al.* (2015), would have estimated y_{jt} as a non-parametric function of $(q_{jt}, h_{jt}, k_{jt}, \ell_{jt}, x_{jt}, i_{jt})$ 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_{jt} be flexibly chosen during the quality stage. However, the first stage estimation would be a non-parametric instrumental variables regression, introducing significant complications due to the high dimensionality of the problem.

45. Bandwidths are chosen using an over-smoothed version of 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.

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

where we instrument for $(q_{jt}-\widehat{E}[q|h_{jt},i_{jt},k_{jt},\ell_{jt},x_{jt}])$ with a second noisy measure of quality. Here, we use the ratio of expected to actual deaths as this instrument, as discussed in Section 5.1.[47]

Finally, we recover $\hat{\Phi}_i(\cdot)$ from the non-parametric estimation

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

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 centre-year that has non-zero hiring from

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

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

$$\omega_{jt}=g(\omega_{jt-1})+\xi_{jt}, \tag{17}$$

where g is a non-parametric function of ω_{jt-1} and ξ_{jt} is a shock to productivity between time $t-1$ and t that is independent of the centre's time-t information set.[48] Thus, for any given $\beta=(\beta_k,\beta_\ell)$, we can estimate $g(\cdot)$ from the equation

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

which follows from substituting the production function from (9) into the innovation of productivity from (17), where $\hat{\alpha}_q$ is the consistent estimator of α_q recovered in the first stage.[49]

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

$$E\begin{bmatrix}\eta_{jt}(\beta)k_{jt}\\ \eta_{jt}(\beta)\ell_{jt}\end{bmatrix}=0; \tag{18}$$

we use (18) to estimate $\hat{\beta}$ via GMM. 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.

6. 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 centre's capital and labour and another that allows for heterogeneous transitions of productivity.

47. The instrument in the second stage of the Robinson procedure is $d_{jt}-\widehat{E}[d|h_{jt},i_{jt},k_{jt},\ell_{jt},x_{jt}]$, where d_{jt} is the ratio of deaths to expected deaths.

48. We use a fifth-order polynomial sieve to approximate $g(\cdot)$; results are robust to using other orders.

49. 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 $\hat{\omega}_{jt}(\beta)$ 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.

TABLE 5
Transformation and production estimates

	With Quality			Without Quality		
	(1) Model	(2) OLS	(3) FE	(4) Model	(5) OLS	(6) 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
	(0.0432)	(0.0210)	(0.1044)	(0.0398)	(0.0210)	(0.1041)
Labour, β_ℓ	0.2436	0.6848	0.1985	0.2556	0.6834	0.1978
	(0.0317)	(0.0162)	(0.0134)	(0.0313)	(0.0162)	(0.0134)

Notes: Main model results with standard errors in parenthesis. Columns (1) and (4) are results from the full structural model. Columns (2) and (5) are estimates using ordinary least squares. Columns (3) and (6) include a centre-level fixed effect to control for productivity.

6.1. *Baseline model*

We now 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 labour and capital coefficients, it must ignore the possibility of productivity differences across centres, resulting in a substantially higher labour 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 centres are likely both to use more stations and employ more staff.

The FE procedure, in contrast, allows for productivity differences across centres but assumes that these differences remain constant over time; *i.e.* the FE estimates identify the capital and labour coefficients on the basis of year-to-year changes in centres' inputs. Using this approach, both the capital and labour 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 labour 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;[50] 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 labour 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 labour of 0.24 and a coefficient on capital of 0.52, which suggest decreasing returns to scale overall. To review,

50. 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.

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 centres' hiring choices to identify unobserved productivity. The relatively larger weight of capital in this specification fits well with our understanding of dialysis: although hiring more employees may allow a centre 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 centre at any given time is necessarily bounded by the number of available stations. While the labour coefficient is small relative to many previous studies, it is in line with some micro studies in different industries (*e.g.* De Loecker, 2011); 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 article, the estimates of the quality-quantity tradeoff in the transformation function, α_q.[51] All three specifications provide evidence of a statistically significant quantity-quality tradeoff, although the magnitude of the effect is much larger in the structural model than with either OLS or FE. 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. And while the FE approach controls for time-invariant productivity, if centres' 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 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 centre that improves its quality enough so that its targeted infection rate falls by 1 percentage point would need to reduce its patient load by 1.55%. Across all centres, the elasticity of quantity with respect to quality—*i.e.* the percent change in patient years corresponding to a percent change in expected infections—is -0.185 at the median, with an inter-quartile range of -0.127 to -0.254. Equivalently, a centre could increase its output 1% 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 labour: a centre can reduce its infection rate by 1% while maintaining its current level of output by increasing labour 6.4%. Given that the average centre employs approximately thirteen 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 of 6.3 percentage points would cost the equivalent of roughly five additional full-time workers for the average centre.

6.2. *Robustness checks*

Table 6 presents a series of robustness checks on our baseline model. In the first column, we do not instrument for the quality proxy with the excess death rate, but instead estimate the first stage assuming that there is no error in our quality proxy. The estimate of the quality-quantity tradeoff falls by about 30%, as would be expected if our quality proxy contained measurement error. This suggests that instrumenting for quality is effectively controlling for attenuation bias.

The second column drops controls for patient characteristics, using the infection rate itself as a proxy for quality targets instead. This has a minimal effect on the production function parameters

51. Note that, we report "$-\alpha_q$" in the tables, incorporating the negative sign in (16).

TABLE 6
Robustness checks

	(1) No quality instrument	(2) No patient char. controls	(3) No competition	(4) No low demand markets
Quality, $-\alpha_q$	−0.0108	−0.0126	−0.0145	−0.0185
	(0.0008)	(0.0032)	(0.0039)	(0.0039)
Capital, β_k	0.5233	0.5174	0.5069	0.5420
	(0.0414)	(0.0430)	(0.0442)	(0.0437)
Labour, β_ℓ	0.2467	0.2453	0.2215	0.2257
	(0.0308)	(0.0313)	(0.0202)	(0.0360)

Notes: Alternative specifications to the baseline results presented in Table 5, Column 1. Standard errors in parenthesis. Column (1) does not instrument for the quality proxy with the excess death rate. Column (2) drops controls for patient characteristics in the quality measure. Column (3) drops competition from x. Column (4) drops centres in markets where the ratio of dialysis stations to the overall population is above the ninetieth percentile.

but reduces the estimated 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 adjusts the baseline specification by dropping the extent of competition from the set of incentive shifters in x. In our baseline estimation, we assume that competition is exogenous in regards to the centre's quality decision. We believe this is reasonable given the timing of the model, as quality and hiring decisions can be adjusted much more rapidly than entry and exit decisions (as discussed in Section 5.2 and footnote 41). One might be concerned, however, that the entry and exit decisions of other firms could depend on a centre's quality decisions, in which case the model would be mis-specified. To investigate this concern, we re-estimate the model without including competition in x. We find only very small changes in the estimates, suggesting that our results are not heavily dependent on inclusion of competition as an exogenous incentive shifter.[52]

The fourth column addresses the possibility that not all centres face inelastic demand for dialysis. Our model of the quantity-quality tradeoff assumes that a centre is able to treat more patients by reducing its quality. In markets where all ESRD patients are already being served, however, there may be no additional patients to treat when a centre reduces its quality. (Note that the unobserved slack demand would bias our estimates of the quality-quantity tradeoff downward, making our estimates conservative.) To investigate this possibility, we re-run our baseline specification while dropping centres in markets where the ratio of dialysis stations to the overall population is above the ninetieth percentile.[53] That is, we drop those markets that are most likely to have slack demand. We find that the estimate of the quality quantity tradeoff rises to 0.0185, which is consistent with a bias from slack demand but well within our baseline confidence interval. Because we do not directly observe measures of excess capacity in our data, we find this robustness check reassuring in that the bias is small—and also conservative in that our estimated quality-quantity tradeoff is understated.

Finally, we extend our baseline model so that the productivity process depends on both $\omega_{j,t-1}$ and centre characteristics, allowing them to differ based on for-profit status and whether the centre 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

52. We have also run a version of the model that removes ownership status from the specification and the results were similarly robust.

53. These results are robust to adjusting this threshold to drop up to one quarter of the data.

TABLE 7
Heterogeneity in productivity process

	(1) Baseline	(2) Separable	(3) 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
	(0.0432)	(0.0446)	(0.0492)
Labour, β_ℓ	0.2436	0.2375	0.2353
	(0.0317)	(0.0307)	(0.0283)

Notes: Alternative specifications to the baseline results presented in Table 5, Column 1, that allow the productivity process to depend on both $\omega_{j,t-1}$ and centre characteristics, allowing them to differ based on for-profit status and whether the centre belongs to one of the two major chains in the industry, Fresenius and DaVita. Standard errors in parenthesis.

than non-profits (Kessler and McClellan, 2002). Our analysis considers whether this stylized fact holds in the dialysis industry after controlling for centres' endogenous quality choices.

Specifically, we consider two alternatives to (17). First, we allow centre-type, p, which can be either non-profit, independent for-profit, DaVita-owned, or Fresenius-owned, to shift the level of the production process,

$$\omega_{jt} = \delta_p + g(\omega_{jt-1}) + \xi_{jt}.$$

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

$$\omega_{jt} = g_p(\omega_{jt-1}) + \xi_{jt}.$$

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

We present the results from estimates of these two alternative specifications in Table 7, with the baseline results repeated for comparison purposes. With $\hat{\alpha}_q$ remaining the same across all specifications because it is estimated in the first stage, we see that the production function estimates are robust to alternative specifications of the productivity process.

6.3. *Heterogeneity in the quality-quantity tradeoff*

Our baseline model assumes that the slope of the production frontier is homogeneous across centres. Although a reasonable starting point, the slope of the frontier may vary across centres based on their size or capital-labour ratio. For instance, one might expect that, for a given number of stations, adding nurses and technicians could make it easier to reduce infections compared to adding more machines. To investigate this possibility, we consider a generalized form of (9) that allows for the slope of the frontier to depend on the centre's labour and capital inputs,[54]

$$y = -(\alpha_q q_{jt} + \alpha_{qk} q_{jt} k_{jt} + \alpha_{q\ell} q_{jt} \ell_{jt}) + \beta_k k_{jt} + \beta_\ell \ell_{jt} + \omega_{jt}^q + \varepsilon_{jt}. \tag{19}$$

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 centre is now $\alpha_q + \alpha_{qk} k_{jt} + \alpha_{q\ell} \ell_{jt}$, and the returns to capital and labour are now likewise dependent on the centre's quality choice.

54. Here in a slight abuse of notation we let ε_{jt} collect all the unanticipated error terms.

TABLE 8
Flexible production frontier

	(1) Model	(2) OLS	(3) FE
Quality, $-\alpha_q$	−0.0082	0.0057	−0.0069
	(0.0254)	(0.0047)	(0.0024)
Quality × Capital, α_{qk}	−0.0364	0.0008	0.0023
	(0.0109)	(0.0025)	(0.0012)
Quality × Labour, $\alpha_{q\ell}$	0.0407	−0.0045	−0.0004
	(0.0104)	(0.0021)	(0.0011)
Capital, β_k	0.4451	0.4569	0.2589
	(0.0605)	(0.0214)	(0.1054)
Labour, β_ℓ	0.2969	0.6830	0.1975
	(0.0515)	(0.0165)	(0.0136)

Notes: Alternative specification allowing the slope of the production frontier to depend on labour and capital inputs. Standard errors in parenthesis.

Table 8 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–labour ratio. In particular, adding stations makes the quantity-quality tradeoff steeper, while adding labour flattens it; *i.e.* the differential impact of adding stations expands the production frontier towards quantity 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 quality-quantity tradeoff is not sensitive to scale.[55] Instead, it appears that the capital–labour ratio is the key factor 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 estimates imply an average elasticity of quantity with respect to quality of −0.148. The coefficient is not statistically significant, however, due to the much larger standard errors. The lack of precision is at least partially due to the high correlation between capital and labour, which is further exacerbated when both are interacted with our quality measures. Furthermore, about 20% of centres—those with the lowest capital-to-labour 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 using proxies to control for quality choices. Therefore, although the results of this specification are instructive, we take our baseline estimates as our primary estimate of the quality-quantity tradeoff within the industry.

7. 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 centre-level production functions that incorporate endogenous quality choices, we quantify the tradeoff that dialysis centres face between treating more patients and providing higher quality care.

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

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 centre would have to forgo in order to prevent one (expected) infection. Under this scenario, the median number of patient-years a centre must give up to eliminate one expected infection per year is 1.5. As industry studies suggest that the cost of haemodialysis treatment is between $45,000 and 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 centres treat the same number of patients but reduce infections by hiring more staff. From this perspective, our results indicate that the median increase in staff required to eliminate one expected 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 to 50,000, this would suggest preventing one infection costs $63,000–90,000.[56] Under either approach, the opportunity cost of one infection to a centre is approximately $75,000.

With this estimate in hand, we can compare the opportunity cost of preventing infections for a dialysis centre 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 haemodialysis 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. Although admittedly speculative, this analysis serves as a guidepost for policy makers seeking to regulate dialysis providers along this dimension.

To provide context for this $50,000 figure, we consider the cost of an infection in terms of life-years lost.[57] According to the facility reports which are the source for our data, the death rate at dialysis facilities was 21.0%, while 17.5% of these deaths were due to infection. This implies the rate of deaths due to infection was roughly 3.7% against an overall septic infection rate of 12.3%.[58] If we assume that all infection-related deaths involve hospitalization, this implies that an infection corresponds to a 29.6% chance of death. Given the overall life expectancy of ESRD patients is 6.2 years (USRDS, 2013), this implies that an infection costs roughly 1.8 expected life years. Willingness to pay for a "quality adjusted life-year" (QALY) is commonly cited as $50,000, although studies attempting to estimate the value of a QALY produce a wide range of results.[59] This benchmark would imply that the cost of infection in life-year terms is $92,000, which suggests that welfare could be increased by increasing quality incentives in dialysis centres.[60]

56. BLS (2014) reports that the median salary of all licensed practical and vocational nurses is $41,000. Our labour 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.

57. We consider life-years lost instead of the value of a statistical life because ESRD patients have much shorter life expectancy than the general population. For an ESRD patient age 60–65 years (the average age in our data set is 61.5), life expectancy was 5.1 years, while life expectancy for the U.S. population is 23.1 years (USRDS, 2013).

58. Our data compiles reports for years 2004–8 and we calculate an infection rate of 12.5% based on centre averages (Table 1).

59. Hirth *et al.* (2000) surveys several studies which can be used to provide QALY estimates. These estimates range from $24,000 to $428,000. The authors conclude that the $50,000 rule of thumb likely understates the societal value of a QALY. We use the benchmark for consistency with the literature while recognizing it may be conservative.

60. Clearly, there is substantial uncertainty about this calculation in both directions. For example, one may believe that quality of life for dialysis patients is lower than the QALY benchmark, reducing the cost of an infection. Alternatively,

The introduction of the Quality Incentive Program introduced by Medicare in 2012, which reduces payments to dialysis centres that fail to meet outcome metrics, is an example of such a policy.

More broadly, our work informs policy discussions by showing that, while productivity dispersion is extensive within the industry, cost-cutting initiatives may result in centres 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.

APPENDIX

A. PROOFS

Proof of Lemma 1 *The centre's expected period-return to labour is increasing in ω^q; i.e. $\frac{\partial \pi}{\partial \ell}$ is increasing in ω^q.*

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

$$\pi(k,\ell,x,\omega^q)=\max_{\tilde{y},\tilde{q}} E[\rho(y,q,k,\ell,x)]$$

$$\text{subject to: } T(\tilde{y},\tilde{q})=F(k,\ell,\omega^q)$$

$$y=\tilde{y}+\varepsilon^y$$

$$q=\tilde{q}+\varepsilon^q.$$

Differentiating π with respect to ℓ, the return to an increase in labour 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 centre'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 centre 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 centre'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

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

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 and 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

low-quality care results in more adverse outcomes than simply deaths from infections, implying the full benefits of increasing quality exceed this figure.

walk), we know that $Pr(\omega^{h'}|\omega^h)$ is also stochastically increasing. We can thus write a single Bellman equation for a centre 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; *i.e.* 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 $c(x,k)$; $\pi(\omega^q,k^*)=\pi(k,\ell,x,\omega^q)+\xi$ for $\pi(\omega,k)$; and the choice variable is h (hiring), rather than x, which was continuous capital investment in Pakes (1994). ||

Acknowledgments. We thank Russell Cooper, Matt Grennan, Darius Lakdawalla, Charles Murry, Joris Pinkse, 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 the Bureau of Economic Analysis, Columbia University, Drexel University, Princeton University, the University of North Carolina, and the University of Toronto.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

ACKERBERG, D. A., CAVES, K. and FRAZER, G. (2015), "Identification Properties of Recent Production Function Estimators", *Econometrica*, **83**, 2411–2451.
BLS (2014), Occupational Outlook Handbook, Technical Report, Bureau of Labour Statistics.
CDC (2001), "Recomendations for Preventing Transimission of Infections Among Chronic Hemodialysis Patients", *Morbidity and Mortality Weekly Report*, **50**, 1–43.
CRAWFORD, G. S. and SHUM, M. (2007), "Monopoly Quality Degradation and Regulation in Cable Television", *Journal of Law and Economics*, **50**, 181–219.
CUTLER, D., DAFNY, L. S. and ODY, C. (2012), "Does Competition Impact the Quality of Health Care: A Case Study of the US Dialysis Industry" (Harvard University and Kellogg School of Management).
DAI, M. (2014), "Product Choice Under Price Regulation: Evidence from the Out-patient Dialysis Market", *International Journal of Industrial Organization*, **32**, 24–32.
DE LOECKER, J. (2011), "Product Differentiation, Multi-Product Firms and Estimating the Imact of Trade Liberalization on Productivity", *Econometrica*, **79**, 1407–1451.
DORASZELSKI, U. and JAUMANDREU, J. (2016), *Measuring the Bias of Technological Change* (University of Pennsylvania and Boston University).
FORD, J. M. and 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. and RIVERS, D. (2016), "On the Identification of Production Functions: How Heterogeneous is Productivity?" (University of Wisconsin-Madison and University of Western Ontario).
GERTLER, P. J. and WALDMAN, D. M. (1992), "Quality-adjusted Cost Functions and Policy Evaluation in the Nursing Home Industry", *Journal of Political Economy*, **100**, 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**, 301–318.
HIRTH, R. A., CHERNEW, M. E., MILLER, E., FENDRICK, A. M. and WEISSERT, W. G. (2000), "Willingness to Pay for a Quality-adjusted Life Year: In Search of a Standard", *Medical Decision Making*, **20**, 332–342.
IMERMAN, M. and 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. and EPSTEIN, A. M. (2009), "Measuring Efficiency: the Association of Hospital Costs and Quality of Care", *Health Affairs*, **28**, 897–906.
JOSKOW, P. L. and ROSE, N. L. (1989), "The Effects of Economic Regulation", *Handbook of Industrial Organization*, **2**, 1449–1506.

KESSLER, D. P. and McCLELLAN, M. B. (2002), "The Effects of Hospital Ownership on Medical Productivity", *RAND Journal of Economics*, **33**, 488–506.
LAINE, J., FINNE-SOVERI, U. H., BJÖRKGREN, M., LINNA, M., NORO, A. and HÄKKINEN, U. (2005), "The Association Between Quality of Care and Technical Efficiency in Long-Term Care", *International Journal for Quality in Health Care*, **17**, 259–267.
LEE, H., MANNS, B., TAUB, K., GHALI, W. A., DEAN, S., JOHNSON, D. and 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**, 611–622.
LEE, J., McCULLOUGH, J. S. and TOWN, R. J. (2013), "The Impact of Health Information Technology on Hospital Productivity", *RAND Journal of Economics*, **44**, 545–568.
LEVINSOHN, J. and PETRIN, A. (2003), "Estimating Production Functions using Inputs to Control for Unobservables", *The Review of Economic Studies*, **70**, 317–341.
MOREY, R. C., FINE, D. J., LOREE, S. W., RETZLAFF-ROBERTS, D. L. and TSUBAKITANI, S. (1992), "The Trade-off Between Hospital Cost and Quality of Care: An Exploratory Empirical Analysis", *Medical Care*, 677–698.
OLLEY, G. S. and PAKES, A. (1996), "The Dynamics of Productivity in the Telecommunications Equipment Industry", *Econometrica*, **64**, 1263–1297.
PAKES, A. (1994), "The Estimation of Dynamic Structural Models: Problems and Prospects", *in* J. J. Laffont and C. Sims, eds, *Advances in Econometrics: Proceedings of the 6th World Congress of the Econometric Society*", Vol. II, 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. and KALLEN, A. (2013), "Bloodstream Infection Rates in Outpatient Hemodialysis Facilities Participating in a Collabourative Prevention Effort: A Quality Improvement Report", *American Journal of Kidney Disease*, **62**, 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**, 2725–2732.
PROPUBLICA (2011), "Dialysis: High Costs and Hidden Perils of a Treatment Guaranteed to All" http://www.propublica.org/series/dialysis.
RAMANARAYANAN, S. and 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. and DAROUICHE, R. O. (2007), 'Healthcare Costs Associated with Hemodialysis Catheter-Related Infections: A Single-Center Experience", *Infection Control and Hospital Epitemiology*, **28**, 606–609.
ROBINSON, P. (1988), "Root-n-consistent Semiparametric Regression", *Econometrica*, **56**, 931–954.
ROMLEY, J. A. and GOLDMAN, D. P. (2011), "How Costly is Hospital Quality? a Revealed-preference Approach", *The Journal of Industrial Economics*, **59**, 578–608.
SCOTT, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley series in probability and mathematical statistics, (New York: John Wiley and Sons, Ltd.).
SILVERMAN, B. W. (1986), *Density Estimaton for Statistics and Data Analysis*, Monographs on statistics and applied probability, (London: Chapman and Hall).
SLOAN, F. (2000), "Not-for-profit Ownership and Hospital Behavior", *in* A. Culyer and 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**, 326–65.
TENTORI, F., ZHANG, J., LI, Y., KARABOYAS, A., KERR, P., SARAN, R., BOMMER, J., PORT, F., AKIBA, T., PISONI, R. and ROBINSON, B. (2012), "Longer Dialysis Session Length is Associated with Beter Intermediate Outcomes and Survival Among Patients on In-centre Three Times Per Week Hemodialysis: Results from the Dialysis Outcomes and Practice Patterns Study", *Nephrology Dialysis Transplantation*, **27**, 4180–4188.
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. and STASON, W. B. (1977), "Foundations of Cost-Effectiveness Analysis for Health and Medical Practices", *The New England Journal of Medicine*, **296**, 716–721.
ZHANG, H. (2014), "Biased Technology and Contribution of Technological Change to Economic Growth: Firm-level Evidence from China" (University of Hong Kong).