Regulating Hospital Quality: Input Regulation, Factor Market Structure, and Misallocation in Healthcare

Chandni Raja *

September 25, 2026

Abstract

This paper estimates a multi-input production model for hospital quality using identifying variation from the 1999 California nurse staffing mandate. I use the recovered production primitives to study the within- and across-hospital factor misallocation from quality regulation targeting production inputs. I find that nurses and physicians are highly complementary inputs (near Leontief) and regulation targeting only nurses increases annual costs by 0.5 percent ($10 million) holding quality constant. Misallocation is doubled under second-best where the mandate exacerbates understaffing of physicians at low revenue, government, and rural hospitals. I find no evidence of across-hospital misallocation of nurses to low productivity hospitals due to the regulation – low staffing hospitals are as productive as their high staffing neighbors. However, relative to a mandate that sets the same input floor for all hospitals, I find aggregate efficiency gains can be made by reallocating nurses to higher severity hospitals where labor is more valuable.

Keywords: hospital quality, production models, healthcare staffing

*Antitrust and Competition Economics Practice, Charles River Associates (CRA) and California Center for Population Research (CCPR) (email: chandni.raja@gmail.com). The views expressed in this article are the views of the author only and do not necessarily represent the views of CRA, its affiliates, its employees, or its clients. I am grateful for the encouragement and support of my advisors: John Asker, Martin Hackmann, and Will Rafey. I thank Maya Ayoub, Jesper Böjeryd, David Henning, Michael Rubens, and seminar participants at UCLA, DePaul University, Caltech, University College London, and the Antitrust Division at the U.S. Department of Justice for helpful comments. I thank Harry Dhami at HCAI for answering my questions regarding the financial reporting data. This work benefited from secure data storage facilities and support provided by CCPR (P2CHD041022). The content does not necessarily represent the views of any organizations listed. All errors are mine.

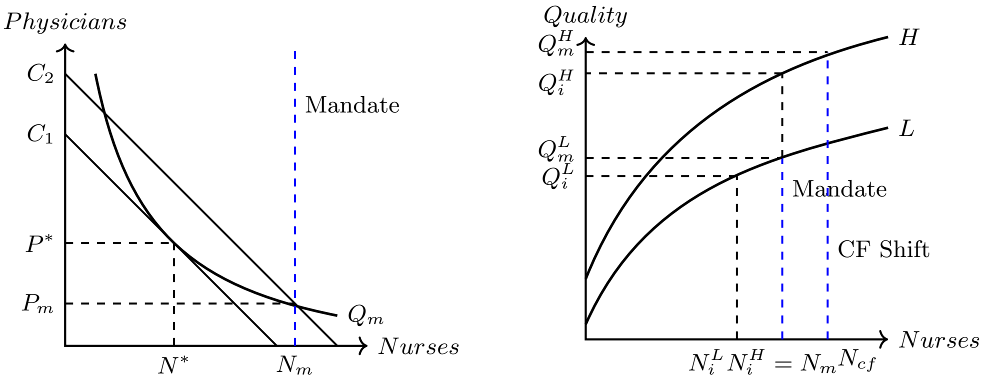
Low quality hospitals are a source of regulatory concern. Given a myriad of options for regulating low performers, the regulator's choice of design has important implications for the productive efficiency of the healthcare industry and the macroeconomy: healthcare constitutes 17 percent of U.S. GDP and employs nearly 10 percent of the U.S. workforce.[1]

In this paper, I use a production framework to evaluate the productive efficiency implications of minimum nurse-to-patient ratios which are widely considered in the U.S. and globally as an option for regulating low performing hospitals.[2] Minimum staffing ratios and other input regulations aim to indirectly regulate a firm's output by targeting the firm's input usage. Regulating inputs can be appealing when direct output regulation leads to unintended responses by firms or when monitoring costs for output are high. However, regulating inputs is in theory misallocative because hospitals use multiple inputs in production (regulation of a single input leads to misallocation across inputs within the hospital) and hospitals admit different types of patients and have heterogeneous productivities (regulation leads to misallocation of the regulated input across hospitals). This regulation directly affects the allocation of resources and output across hospitals. These two dimensions are illustrated in Figure 1. Their quantification depends on the production primitives.

Despite the labor intensive nature of healthcare, we have a limited knowledge of the structural primitives for hospital quality production and the returns to labor. This is for several reasons: prior work on returns to labor in healthcare employs largely reduced-form methods; data on hospital inputs and quality are difficult to obtain; and healthcare market frictions render standard assumptions in the identification of production functions implausible. I fill this gap in the literature by estimating a structural value-added production model of hospital quality with two important features: first, the model includes multiple inputs and flexible elasticities of substitution between inputs which allows me to study the within-hospital misallocation across inputs; second, labor productivity varies with both observed patient type and unobserved hospital productivity which allows me to study the across-hospital misallocation of nurses to low marginal product hospitals. By recovering the underlying primitives from the production rather than cost function, I am able to consider departures from perfectly

[1] See CMS (2024) and BLS (2024), respectively, for the 17 percent and 10 percent figures. Performance pay, cost-based reimbursement, public reporting requirements, labor licensing restrictions, and labor staffing ratios are examples of options aimed at regulating low performers.

[2] Minimum nurse-to-patient ratios have received significant attention over the past five years. Ratios are in consideration at the federal and state levels in the U.S. with active bills S.1567 (U.S. Senate), SB240 (Pennsylvania Senate), and S6855 (New York Senate).



(a) Within-Hospital Misallocation

(b) Across-Hospital Misallocation

Figure 1: Misallocation Within and Across Hospitals

Notes: In panel (a), I show hypothetical isoquant and isocost curves associated with the quality Q_m produced under the mandate. N_m and P_m represent the nurse- and physician-to-patient ratios used to produce Q_m. However, Q_m can be produced at lower cost under isocost curve C_1 using the cost-minimizing input vector $\{N^*, P^*\}$ with the difference in costs $(C_2 - C_1)$ representing the within-hospital misallocation. In panel (b), I show a hypothetical quality production function for two hospitals in close proximity to one another: H and L. L is affected by the mandate and shifts nurse staffing from N_i^L to N_m while hospital H has an initial level of N_m and is unaffected. If we consider moving the mandate-induced number of nurses from hospital L to H ("CF Shift"), the production differential due to the heterogeneous marginal products is $(Q_m^H - Q_i^H) - (Q_m^L - Q_i^L)$.

competitive labor markets and hospital cost-minimization.

I use California hospitals between 1995 and 2008 as an empirical setting to estimate the model. This setting is important for several reasons. First, its sheer size – California hospitals deliver medical care to nearly 3 million patients each year and employ over 150,000 licensed nurses. Second, I have access to rich, administrative patient-level discharge data to construct a measure of healthcare quality (30-day non-readmission rate)[3] and use risk-adjustment to address observable endogenous patient selection into hospitals which is a common issue in healthcare research. Third, California hospitals experienced a large shift in nurse staffing levels in the early 2000s due to the 1999 California nurse staffing mandate – one of the first and to-date few pieces of comprehensive legislation on nurse staffing levels in hospitals. I use a difference-in-differences design to show that the mandate led to a 12 percent increase in the nurse-to-patient ratio and a 0.7 percent increase in the non-readmission rate among treated hospitals within one year of implementation.

[3]The hospital's 30-day non-readmission rate is the share of patients discharged alive from the hospital who were not readmitted to any inpatient hospital for any cause within 30 days of discharge from the initial inpatient stay. Readmission within 30 days is an indicator of poor quality of care during the first inpatient stay and non-readmission an indicator of good quality of care.

My identification strategy addresses the endogeneity of the firm's input choices to its unobserved productivity including unobservable patient selection: I use the quasi-experimental variation from the mandate and lagged input variables as instruments to address the endogeneity of productivity using multiple estimation strategies: the IV2SLS estimator and the dynamic panel estimator.

Next, I use the recovered production primitives in a cost-minimization counterfactual to quantify the within-hospital misallocation shown in Figure 1. My production function estimates indicate that nurses and physicians are highly complementary (near Leontief) in production and that the efficient ratio of nurses to physicians engaged in patient care is between 1.35 and 2.05.[4] I find that labor regulation that targets nurses alone leads to misallocation of 0.5 percent of healthcare labor costs equaling $48,553 for the average hospital and $10 million in total costs (roughly 20 percent of the total healthcare labor costs that hospitals incurred due to the mandate). Given evidence that healthcare labor markets are imperfectly competitive, I estimate the misallocation under second-best. First, I assume hospitals' pre-regulatory input choices are optimal and calibrate reduced-form wedges to factor markets that reconcile these choices. Next, I use these wedges in the post-regulatory period to solve for misallocation under second-best and find that misallocation is doubled ($104,834 for the average hospital and $21 million across hospitals) because the mandate exacerbates pre-existing understaffing of physicians at low revenue, government, and rural hospitals.

Finally, I evaluate the across-hospital misallocation of nurses to low marginal product hospitals due to the regulation. I estimate the variance in the marginal products of labor across hospitals within California geographic markets for hospital services before and after the reallocation. I find that the mandate led to improvements in productive efficiency in the majority of geographic markets – treated hospitals have comparable productivities to control hospitals. However, I find more efficient labor allocation can be made for specific pairs of hospitals located in densely populated counties where untreated hospitals admit higher severity patients. Given my finding that labor productivity increases with the severity of the patient population, allocating resources based on patient severity is more efficient.

My results have three important policy implications. First, my findings highlight that knowledge of the firm's production technology is central to efficient regulatory

[4] I find elasticities of substitution between zero and 0.2 under the main estimator and up to 0.6 under the dynamic panel estimator. The elasticity of substitution between nurses and physicians is the percent change in the nurse to physician ratio divided by the percent change in the physician to nurse marginal product ratio. A low elasticity of substitution indicates that even large changes in the relative marginal products do not lead to large changes in labor allocations.

design – whether in healthcare or other markets (e.g. class sizes in education or technology standards in environmental regulation). Second, I show that the size of misallocation depends on factor market structure: when inputs are complements, a relatively inelastic market for the unregulated input exacerbates the misallocation from regulation. In my setting, substantial dispersion in the nurse marginal product across hospitals can be attributed to a (mis)allocation of physicians. This is consistent with longstanding policy discussion over inelastic supply of physicians and their geographic allocation (Rosenthal et al., 2005, Rittenhouse et al., 2021). Taking these two findings together, it is unsurprising that recent healthcare workforce policy changes have rolled back licensing restrictions on nurses to allow them to operate more independently of physicians for at least a subset of less severe patients. Third, my findings show that allocation of resources based on severity is more efficient than not.

This paper makes three contributions. First, I document the role of labor as a driver of the well documented productivity differentials across healthcare providers (Chandra et al., 2016a, Einav et al., 2022, Fisher et al., 2009) and examine how labor regulation affects these differentials. Labor use has been understudied from an efficiency perspective despite the labor intensive nature of healthcare production and healthcare labor shortages which amplify the need for efficient allocation (KFF, 2023).[5] IO tools have been used in a few papers to study healthcare quality (Romley and Goldman, 2011, Grieco and McDevitt, 2017, Gertler and Waldman, 1992).[6] My work is complementary to prior papers in that I focus on hospitals rather than dialysis care (Grieco and McDevitt, 2017) or nursing homes (Gertler and Waldman, 1992) and my model captures to greater extent the organizational complexity of the firm: I observe multiple labor inputs and I use a flexible functional form to study their interactions.[7] By doing so, my work highlights how the organizational structure of the hospital affects the returns to labor regulation.

Second, I build upon the body of empirical work that studies the misallocation of resources across firms in the economy (Hsieh and Klenow, 2009, Asker et al., 2019). Under a first-best where low productivity firms produce lower output, input regulation

[5]Healthcare is labor intensive and labor costs are estimated to make up two-thirds of total healthcare expenditures (WHO, 2000). Prior work has focused on the choice of treatment as a driver of healthcare productivity (Chandra et al., 2023, Skinner and Staiger, 2015, Chandra and Skinner, 2012, Garber and Skinner, 2008).

[6]Lee et al. (2013) do not consider hospital quality but estimate a hospital revenue production function from which they consider the marginal products of information technology inputs. Chandra and Staiger (2020) allow the hospital's treatment use to vary as a function of its productivity. Notably a few papers study how the returns to physicians vary based on the hospital's productivity and the implications for the allocation of physicians across hospitals (Mourot, Mourot, Huckman and Pisano, 2006).

[7]Health services research has highlighted the coordination and interaction between nurses and physicians in providing care (Havens et al., 2010).

shifts resources and output towards low productivity firms. However, this first-best is not obviously the case in healthcare: government intervention, imperfectly competitive product and factor markets, healthcare access concerns, and unique market features outlined in Arrow (1963) dampen the linkage between productivity and output. I assess the theoretical implication by estimating the underlying primitives and using the recovered primitives to assess the misallocation relative to an efficient benchmark. Similar to my paper is Chandra and Staiger (2020) who assess the role of hospital productivity in the variation in treatment rates.

The third contribution of this paper is methodological. Monopsony and bargaining power between physician practice groups and hospitals; nurse unions and hospitals; and insurers and hospitals; are commonly discussed features of healthcare markets yet the identification methods for IO production models rely heavily on conduct assumptions in the product and factor markets are unlikely to apply to hospitals. I rely on quasi-experimental variation from the mandate and restrictions on the productivity process for identification, in the process showing that input market frictions are significant.

The remainder of the paper proceeds as follows. Section 1 introduces the empirical framework and objects to be estimated. Section 2 introduces the data, measurement, and institutional context. Section 3 highlights key industry facts that inform the structural model. Section 4 presents the structural model and Section 5 discusses the model identification. Section 6 presents the parameter estimates and recovers the production primitives. Section 7 presents the estimates of the within- and across-hospital misallocation objects from Section 1. Section 8 concludes.

I. Empirical Framework

In this section, I discuss the empirical framework that allows me to estimate the misallocation due to the input regulation. I take a production approach: I recover the underlying production primitives and use them to estimate within-hospital misallocation under first-best (perfect competition in input markets) and second-best (using observed input choices to recover the implied wedges under imperfect competition and estimating the misallocation given the wedges).

A. Producer behavior

Consider hospitals indexed by $h = 1,..., N$ and period t which produce quality of care Q_{ht}. Hospitals are heterogeneous in productivity ω and the patients they admit

H and otherwise use a common production technology. In each period t, hospital h minimizes the contemporaneous cost of production given the following technology:

$$Q(N_{ht}, P_{ht}, H_{ht}, \mathbf{M}_{ht}, \omega_{ht}, \epsilon_{ht}) = e^{\omega_{ht}+\epsilon_{ht}} F(N_{ht}, P_{ht}, H_{ht}, \mathbf{M}_{ht})$$

where N is nurse labor, P is physician labor, H is patient type, $\mathbf{M} = (M^1,..., M^J)$ represents the set of other inputs in production (including capital, intermediate inputs, materials), and ω is the hospital and year-specific Hicks-neutral productivity term. ϵ is measurement error. I consider the cost-minimization problem below with indexes h and t omitted for brevity:

$$\begin{aligned} \min_{N,P}\{\mu^N w^N N + \mu^P w^P P\} \\ \text{s.t.} e^{\omega+\epsilon} F(N, P, H, \mathbf{M}) \geq Q \end{aligned} \tag{1}$$

where w^N, w^P are the prices of nurse and physician labor, respectively, and μ^N, μ^P are wedges signifying market frictions – for example, imperfectly competitive factor markets[8] or hospital objectives which deviate from healthcare quality. Q is a scalar. The cost of materials $\mathbf{M}$ is not included in Equation (1) due to assumptions made in Section 4 but this conceptual exercise does not rely on these assumptions.

The associated Lagrangian:

$$\mathcal{L} = \mu^N w^N N + \mu^P w^P P + \lambda\{Q - e^{\omega+\epsilon} F(N, P, H, \mathbf{M})\}$$

The first-order conditions:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial N} &= \mu^N w^N - \lambda e^{\omega+\epsilon} \frac{\partial F}{\partial N} = 0 \\ \frac{\partial \mathcal{L}}{\partial P} &= \mu^P w^P - \lambda e^{\omega+\epsilon} \frac{\partial F}{\partial P} = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= Q - e^{\omega+\epsilon} F(N, P, H, \mathbf{M}) = 0 \end{aligned}$$

Suppose the constraint of Equation (1) represents a mandate that directly regulates hospital quality. λ^*_{dreg} represents the marginal cost of quality at (N^*_{dreg}, P^*_{dreg}) and the total cost $C_{dreg} = (\mu^N w^N N^*_{dreg} + \mu^P w^P P^*_{dreg})$.

[8] In this case, the wedges would represent constant markups as the ratio of marginal revenue product to marginal cost in the corresponding labor market ($\mu_L = \frac{\lambda \frac{\partial F}{\partial L}}{w^L}$). In the product market power context the multiplicative markup is defined as $\mu = \frac{P}{MC}$. Note that my notation reflects the reciprocal of the "markdown" parameter used in the monopsony and labor market power literature $\theta^L = \frac{w^L}{\lambda \frac{\partial F}{\partial L}}$.

B. Within-hospital input misallocation from first-best

Under the assumption of no market frictions ($\mu^N = \mu^P = 1$), the staffing regulation introduces the following misallocation across factors of production:

$$
\begin{aligned}
&\min_{N,P}\{\mu^N w^N N + \mu^P w^P P\} \\
&\text{s.t. } e^{\omega+\epsilon}F(N,P,H,\mathbf{M}) \geq Q \\
&N \geq N_{min}
\end{aligned}
\tag{2}
$$

The associated Lagrangian:

$$\mathcal{L} = w^N N + w^P P + \lambda_1\{Q - e^{\omega+\epsilon}F(N,P,H,\mathbf{M})\} + \lambda_2\{N_{min} - N\}$$

The first-order conditions:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial N} &= w^N - \lambda_1 e^{\omega+\epsilon}\frac{\partial F}{\partial N} - \lambda_2 = 0 \\
\frac{\partial \mathcal{L}}{\partial P} &= w^P - \lambda_1 e^{\omega+\epsilon}\frac{\partial F}{\partial P} = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda_1} &= Q - e^{\omega+\epsilon}F(N,P,H,\mathbf{M}) \\
\frac{\partial \mathcal{L}}{\partial \lambda_2} &= N_{min} - N
\end{aligned}
$$

Suppose the constraint of Equation (2) represents a nurse staffing mandate. The solution to (2) is $(\lambda^*_{sreg}, N^*_{sreg}, P^*_{sreg})$. Under the standard assumption of $\frac{\partial F}{\partial^2 N} < 0$ and a binding mandate $\lambda_2 > 0$, the regulation leads to overuse of the regulated input relative to first-best. I define the misallocation as the difference in total costs between the equilibria under the staffing regulation and the direct quality regulation: $\Delta C = C^*_{sreg} - C^*_{dreg}$ where $C^*_{sreg} = (w^N N^*_{sreg} + w^P P^*_{sreg})$. The misallocation is a function of production primitives ($\frac{\partial F}{\partial N}$, $\frac{\partial F}{\partial P}$, λ_1, λ_2) and factor prices (w^N, w^P).

C. Within-hospital input misallocation from second-best

Hospitals are unlikely to be price-takers in the labor markets. Physicians are not directly employed by hospitals with the exception of special cases[9] and engage in complex negotiations with hospitals over payment (Burns and Muller, 2008). Nurses are directly employed by hospitals but roughly 50 percent of hospital nurses in California

[9] Hospitals are prohibited from directly employing physicians in California under the Corporate Practice of Medicine Ban (CPOM). County hospitals and teaching hospitals are exempt from CPOM.

are unionized and collectively bargain over wages (Raja, 2023). Prior work identifies market power in healthcare labor markets (Prager and Schmitt, 2021).

The first-order conditions absent regulation can be rewritten to indicate these distortions:

$$w^N = \frac{\lambda e^{\omega+\epsilon}\frac{\partial F}{\partial N}}{\mu^N}, \qquad w^P = \frac{\lambda e^{\omega+\epsilon}\frac{\partial F}{\partial P}}{\mu^P} \tag{3}$$

The regulation enters as a force that could correct or exacerbate the pre-existing distortion. For example, under frictions $\mu^N > 1$ and $\mu^P = 1$ (e.g. nurse wage markdown and perfectly competitive physician labor market, respectively) the regulation plays a corrective role and addresses distortionary under-hiring. On the other hand, under frictions $\mu^N < 1$ and $\mu^P = 1$ (e.g. nurse wage markup) the regulation exacerbates pre-existing over-hiring.

$$w^N = \frac{\lambda_1 e^{\omega+\epsilon}\frac{\partial F}{\partial N} + \lambda_2}{\mu^N}, \qquad w^P = \frac{\lambda_1 e^{\omega+\epsilon}\frac{\partial F}{\partial P}}{\mu^P}$$

The measure of misallocation ΔC under second-best is analogous under first-best except μ^N and μ^P are not assumed to be equal to one under second-best.

D. Across-hospital input misallocation

As indicated by the nurse labor first-order condition with respect to Equation (1):

$$\mu^N w^N = \lambda e^{\omega+\epsilon}\frac{\partial F}{\partial N}$$

In the absence of market frictions, μ^N, and with the same choice of hospital quality Q cost-minimizing hospitals would choose their nurse-to-patient ratio to equalize their marginal quality products of nurse labor. High productivity hospitals would hire fewer nurses per patient than low productivity hospitals facing the same market wage.

On the other hand, under the presence of market frictions including variation in quality choice or imperfectly competitive markets, there could be variance in the marginal products of labor across hospitals prior to the regulation. Depending on whether low staffing hospitals are low or high marginal product hospitals, the regulation could reduce or increase productive efficiency in the market, respectively.

I define the across-hospital misallocation from the mandate as the change in the variance of the marginal quality products of nurse labor at substitutable hospitals i in geographic market m.

$$\Delta \text{var}(\text{MPL}_m) = \text{var}(\text{MPL}_{i \in m})_{reg} - \text{var}(\text{MPL}_{i \in m})_{noreg}$$

The variance in marginal products across firms has been used as a measure of misallocation in prior work (Hsieh and Klenow, 2009) with the equalization of the marginal revenue products reflecting an efficient allocation of resources. I focus on marginal quality product rather than marginal revenue product. [10]

The marginal product of labor is a function of the hospital's input mix $(N, P, \mathbf{M})$, patient mix (H), and unobserved productivity ω. The across-hospital misallocation depends on how nurses are allocated across hospitals with heterogeneous productivities and patient types. The market m is defined by the Dartmouth Atlas of Healthcare's Hospital Service Area (HSA). A tight definition of the geographic market addresses healthcare access concerns: we may be indifferent between purchasing steel produced in factory A or B in far apart locations but we are not indifferent between receiving emergency medical care from hospital A or B in far apart locations.

II. Data and Implementation

I construct measures of hospital quality (Q), wages (w^N, w^P), patient types (H), and inputs $(N, P, \mathbf{M})$ from a data set that I constructed linking hospital financial reporting data and administrative patient level hospital records covering all admissions at California hospitals between 1995 and 2008.

A. Hospitals and their Inputs (N, P, and M) and Patient Type (H)

In this paper, I focus on general acute care (GAC) hospitals in California which admit inpatients (patients who stay at least overnight). [11]

I observe hospitals in financial reporting data from the California Department of Healthcare Access and Information (HCAI)'s Hospital Annual Financial Disclosure Reports. These data are desk-audited and notable for their granularity and detail. The data report patient volumes, capacity, revenues, nurse and administrative labor hours,

[10] I do not consider the misallocation from allocating the nurses outside of the hospital setting because nurses who work in hospitals form a different labor market relative to nurses who work in nursing homes, home health, or physician practice groups. See Raja (2023) for more details.

[11] General acute care hospitals provide short-term acute medical care. Other types of hospitals are pediatric, psychiatric, and long-term care. Outpatients are admitted and discharged in the same day.

and expenditures on labor, materials, and capital for each inpatient hospital unit (e.g. Medical/Surgical Acute Care) of reporting hospitals. Hospital characteristics including ownership type, medical staff numbers and specialties, and services inventories are reported at the hospital level.

Hospitals organize medical care delivery for inpatients into hospital units. See the Supplemental Appendix for details. I focus on the relationship between the hospital's inputs in the Medical/Surgical Acute Care unit (hereafter "acute care") and quality of care at the hospital level. Quality of care is not observable at the unit level. I used a balanced panel of 208 hospitals that report non-zero inpatient days in the acute care unit for each of the thirteen years in my sample between 1996-2008.[12]

In Table 1, I display the descriptive statistics for hospitals in my sample for 1996-2002, the period prior to the implementation of the California nurse staffing mandate, according to position in the nurse-to-patient ratio distribution. The average hospital in my sample reports roughly 56,000 inpatient days and 10,000 inpatient discharges annually with significant dispersion in size.

I measure nurse labor N in terms of hours of clinical nursing time (nurses per 1,000 patient days). [13] The average hospital-year between 1996-2008 has 2.88 nurses per 1,000 patient days implying a nurse-to-patient ratio of 0.250 (one nurse per four patients). This average was far lower in the period prior to the mandate's implementation (0.222) compared to the period after implementation (0.287) but the dispersion across hospitals remained stable indicating that the hospital's input demand increased independently of the mandate on the upper end of the staffing distribution. The interquartile range was 0.200-0.273 in 2002 and increased to 0.231-0.307 in 2004.

I measure physician labor in terms of the number of "active medical staff" affiliated with the hospital at the end of each reporting period. I use the revenue share in acute care to allocate the number of physicians per patient at the hospital level to the unit level. I multiply the number of physicians for each hospital and year by the acute care share of total inpatient revenue, divide the resulting number by the number of acute care patient days, and then multiply by 1,000 days to obtain the physicians per 1,000 patient days. [15] To standardize nurse and physician labor hours to time spent

[12]The year 1995 is used solely as a one-year lookback period to conduct risk-adjustment for patients admitted in 1996. Federal hospitals administered by the Veterans Administration, Department of Defense, or Public Health Service are exempt from California state reporting requirements for patient discharges by because they are not subject to state licensure. Kaiser hospitals were not required to submit hospital financial reporting data separately for their separate facilities until 2022. Therefore both federal and Kaiser hospitals are excluded from this analysis.

[13]I aggregate the number of nurse hours to the number of nurses under the assumption that one nurse works 40 hours per week for 52 weeks of the year.[14] I then divide the number of acute care nurses by the number of acute care patient days and multiply by 1,000 days to obtain the number of nurses per 1,000 patient days.

[15]The HCAI hospital financial accounting manual uses the revenue share of the hospital unit to allocate costs

Table 1: Descriptive Statistics for California Hospitals from 1996-2002

	Nurse-to-Patient Ratio Distribution			
	Bottom 25	25-50	50-75	Top 25
Hospitals	52	52	52	52
Annual discharges	9,367	10,368	10,993	9,433
Annual inpatient revenue ($)	60,769,720	78,174,402	77,836,713	77,893,051
Acute share of revenue	0.366	0.401	0.417	0.448
Case Mix Index	1.03	1.08	1.10	1.14
Hospital-wide discharges				
Hospital-wide 30-day non-readmission rate	0.902	0.897	0.887	0.897
Hospital-wide risk-residualized rate	0.970	0.968	0.966	0.973
Hospital-wide length of stay	3.407	3.510	3.496	3.544
Inputs in acute care				
Nurses per 1,000 patient days	2.196	2.443	2.725	3.230
Physicians per 1,000 patient days	1.091	1.289	1.295	1.233
Materials expenditures per 1,000 patient days ($)	4,403	3,531	3,872	4,120
Capital expenditures per 1,000 patient days ($)	433,019	468,360	541,930	580,107
Patient care costs per 1,000 patient days ($)	399,171	473,644	547,243	626,019
Hospital characteristics				
Share church or non-profit	0.654	0.596	0.692	0.731
Share investor-owned	0.115	0.192	0.154	0.096
Share teaching hospitals	0.038	0.096	0.115	0.154
Share small/rural hospitals	0.173	0.115	0.135	0.212

Notes: This table includes the 208 hospitals in my balanced panel sample from 1996-2008. Hospitals are grouped into quartiles of the nurse-to-patient ratio distribution based on their average values from 2000-2002 (prior to the implementation of the California nurse staffing mandate in 2003).

on patient care, I assume that 56 percent of physicians' time is spent on patient care. See Supplemental Appendix for details. The numbers are reported in Table 1.

The average number of FTE physicians in direct patient care per 1,000 patient days (hereafter "physicians per patient") in a hospital-year in the sample is 1.05 with an interquartile range of 0.67-1.58. This implies a physician-to-patient ratio of 0.091 (approximately one physician per ten patients). An average of 85 physicians are assigned to acute care with an interquartile range of 41-162.

Finally, I measure H as is standard in the literature using the Case Mix Index which is an hospital level index accounting for the diagnoses, co-morbidities, treatments administered, and demographics of the hospital's patient population. A higher CMI reflects a case mix that requires greater resource intensity to improve the patient's health. Further details on the measurement of nurses, physicians, and case mix are included in the Supplemental Appendix.

B. Hospital Quality Q

This paper contributes to healthcare quality measurement by constructing standardized, risk-adjusted hospital level quality measures for California hospitals between 1996-2008 using administrative discharge records.

The administrative discharge records include data on patient characteristics, date of admission and discharge, and primary and secondary diagnoses and procedure codes for the universe of discharges at California hospitals between 1995-2008. Importantly, I do not restrict to the Medicare population given that hospitals treat different shares of Medicare patients and performance among Medicare patients may not be representative of overall performance. As far as I am aware, hospital level quality measures are publicly available beginning in the mid-2010s (after the mandate was implemented) and are calculated for the Medicare population only.

I construct two measures of hospital quality Q: the 30-day non-readmission rate and the 30-day survival rate. A hospital's 30-day non-readmission rate is the proportion of patients admitted to the given hospital that are not readmitted to any hospital within 30 days of discharge from their initial inpatient stay. I study non-readmission as the outcome of the production model and I study survival for robustness: as constructed in the literature the denominator of the non-readmission rate includes patients that die outside of the hospital and studying the 30-day survival rate in conjunction ensures

associated with physicians from the hospital level to the unit level. Specifically, the costs listed in the columns associated with Page 18, line 255 are then allocated to the according to the statistic on Page 19, Column 11 (gross patient revenue).

that patients who are not readmitted remain alive. In the Supplemental Appendix, I discuss the measurement details and reasons for focusing on non-readmission and survival specifically.

C. Wages w^N, w^P

I observe the average hourly nurse wage in the financial reporting data. However, I do not observe physician wages in my data. [16]. I measure physician wages using data from Gottlieb et al. (2023) who use IRS tax records to report summary statistics on physicians' individual total income, inclusive of business income, at the commuting zone level. The commuting zone level averages for 2017 wages are shown in Figure E.5 of the Online Appendix to Gottlieb et al. (2023). [17]

D. Implementation

I use these data to recover ω and F. For identification, I use quasi-experimental variation from two data sources: the Dartmouth Atlas of Healthcare's publicly available hospital tracking file with hospital closure and merger and acquisition events and participation in CMS's Critical Access Hospital (CAH) program; and data from the California Department of Health on the 1999 California nurse staffing mandate.

Using my estimates of $\widehat{\omega}$ and $\widehat{F}$, I solve the problem in Equation 1 for each hospital to produce its observed quality in the post-staffing mandate period and recover $(N^*_{dreg}, P^*_{dreg}, \lambda^*_{dreg})$ under the assumption that $\mu^N = \mu^P = 1$. Next, I solve the problem in Equation 2 for each hospital to produce its observed quality in the post staffing mandate period under a staffing constraint and recover $(N^*_{sreg}, P^*_{sreg}, \lambda^*_{sreg})$. Misallocation under first-best is the difference in total costs under these two scenarios.

To quantify the misallocation from second-best, I use the estimate of λ^*_{dreg} recovered under a perfectly competitive factor markets assumption in addition to observed values of (N, P) during the pre staffing mandate period and Equation 3 to calibrate μ_N and μ_P for each hospital assuming that each hospital's behavior in the pre staffing

[16] Most physicians own or are employed by physician practice groups that contract with hospitals. Physicians' salaries are therefore not reported in the financial reporting data the way that direct employees' salaries are reported. Their financial contracts with hospitals and any payments from hospitals to physicians under these contracts may or may not need to be reported in the form. Furthermore, the nature of these contracts vary so greatly across and within hospitals (across hospital units) that one cannot be sure that reported payments, if they are observed, are made for physician labor as opposed to hospital reimbursements for expenditures made by physicians who in some cases supply their own inputs to the unit they staff.

[17] I adjust these 2017 wages to 2005 wages (denominated in 2017 USD) at the commuting zone level using the growth rate in physician salaries implied by Figure E.3(A). I then inflation-adjust the 2005 wages to be denominated in 2005 USD. Given the assumption that the average medical staff physician is 0.4 FTEs (discussed in Section 2), I multiply the wage by 0.4 to obtain the wage of an affiliated physician.

mandate period is optimal given $(\widehat{\mu_N}, \widehat{\mu_P})$. Then I solve Equations 1 and 2 using the calibrated $(\widehat{\mu_N}, \widehat{\mu_P})$ calculate ΔC with the total costs calculated under second-best.

To quantify the across-hospital misallocation, I define the pairs of hospitals in each market m and estimate the MPL for each hospital i from the recovered estimates of $\widehat{\omega}$ and $\widehat{F}$ and observed values of (N, P) before and after the staffing mandate.

III. Key Industry Facts

A. Differences in patient mix drive resource allocation

A well-documented fact in the hospital cost and production literature is that hospitals with more complex cases use more resources for each patient day (Tatchell, 1983, Lave and Lave, 1970, Lee and Wallace, 1973). Higher resource use among higher CMI hospitals is also expected from the regulator's perspective.[18]

In Figure 2, I plot a binned scatter of the patient care costs per patient day and CMI. Figure 2 shows the positive, statistically signficant relationship between costs per patient day and case mix in my sample and provides suggestive evidence that case mix drives resource allocation. This relationship is further observed within hospitals: across units which provide medical care to patients of varying severity, hospitals allocate more resources per patient day to intensive compared to acute care.

If this observed resource allocation is efficient, then a mandate which does not account for case mix risks resource misallocation towards low severity hospitals. I observe and identify case mix, H, in the model to determine whether the mandate efficiently allocates nurses across hospitals based on patient type.

B. Licensing restrictions imply low substitutability across labor types

How substitutable the labor types are in the production of hospital quality depends on licensing (whether each type is legally allowed to independently perform the range of tasks that are required to improve patient health) and skill (whether each type is able to perform equally on the same range of tasks).

"Scope of practice" regulations for nurses limit the tasks that nurses are allowed to perform in the hospital setting. Nurses were not allowed to practice independently of

[18] The California Health Department states that a "higher CMI indicates a more complex and resource-intensive case load" (HCAI, 2025).

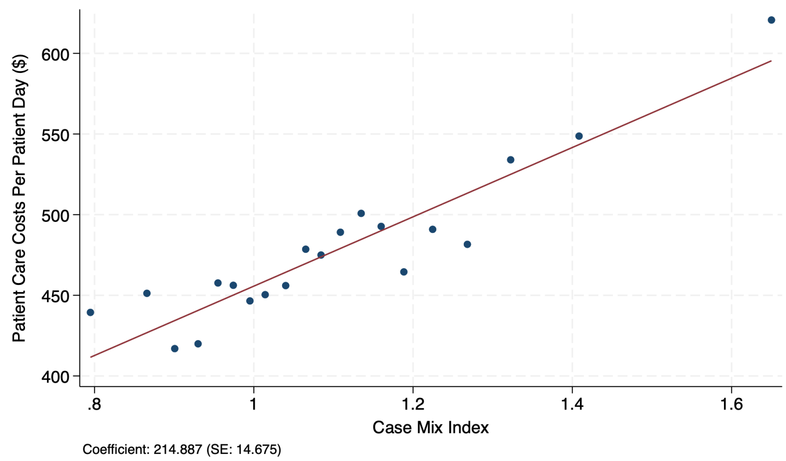


Figure 2: Case Mix and Costs Per Patient Day

Notes: This figure shows a binned scatter plot of the costs per patient day by Case Mix Index for hospitals in the sample. Costs per patient day increase in Case Mix indicating that more complex and unhealthy patients require more resource use. This relationship remains statistically significant if we consider only cross-sectional variation.

physicians in the inpatient hospital setting during my sample period.[19] The California Nursing Practice Act states that RNs require physicians' orders to perform dependent activities including administering medications and therapeutic agents and require written authorization at the provider-level to perform interdependent functions that overlap with medical practice. These interdependent functions which are termed "beyond the usual scope of nursing practice" include diagnosing disease, prescribing medication or treatment, and penetrating or severing tissue (California Nursing Practice Act, 2012). Prior work suggests that nurses yield lower quality returns than physicians on tasks that they are both licensed to perform (Chan and Chen, 2022). If the estimated production relationships confirm what the institutional context suggests, a mandate nurses may misallocate resources within the hospital unless hospitals make complementary investments in physicians.

C. Market structure indicates inelasticity of physician labor supply

As indicated in Section I.C., hospitals are unlikely to be price-takers in the labor markets. Restrictions on aggregate physician labor supply are notable features of the

[19]State-level changes over the past decade including in California have loosened the practice restrictions on high-skilled Registered Nurses with a Nurse Practitioner (NP) license and allowed them to practice as independent practitioners. The proliferation of NPs is more widespread in the primary care setting than in the inpatient setting.

market and have a long history of study (Kessel, 1958).

D. Mandate causally increases nurses and quality

The 1999 California nurse staffing mandate imposed minimum nurse-to-patient ratios in the acute care units of GAC hospitals. Intensive care units were already required to maintain legislated ratios beginning in the 1970s. Ratios were announced on January 2002 for initial implementation dates January 2004 for the Medical/Surgical Acute Care unit to reach an intermediate ratio of 0.16 and January 2005 to reach a final ratio of 0.2 (Raja, 2023). See Raja (2023) for details on the institutional context and reduced-form effects of the mandate.

I follow Raja (2023) in defining hospitals as treated by the mandate if they have an average nurse-to-patient ratio of below 0.25 between 2000-2002. [20] In Figure 3, I plot the distribution of nurses and physicians for hospitals in 2000 (in red) and in 2008 (in purple) among treated hospitals. The figure illustrates the large shift in nurses among treated hospitals with no discernible change in physicians.[21]

To estimate the causal effect, I use the following event-study specification for y_{ht} as the log of nurses per 1,000 patient days and the logs of the risk-adjusted non-readmission and survival rates:

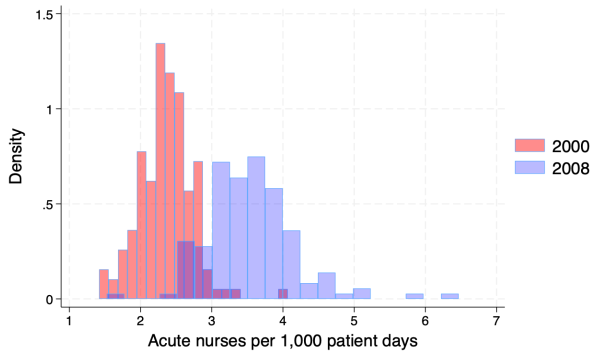
$$y_{ht} = \beta_0 + \sum_{t \neq 2003} \beta_t \{YEAR_t = t\} * BELOW_h + \gamma_h + \xi_t + \epsilon_{ht} \tag{4}$$

where $BELOW_h$ is an indicator variable for the treatment that takes on a value of one if the hospital had an average nurse-to-patient ratio of below 0.25 in 2000-2002, and β_t are the treatment effects of interest for years following the excluded year 2003, and γ_h and ξ_t are hospital and year fixed effects.

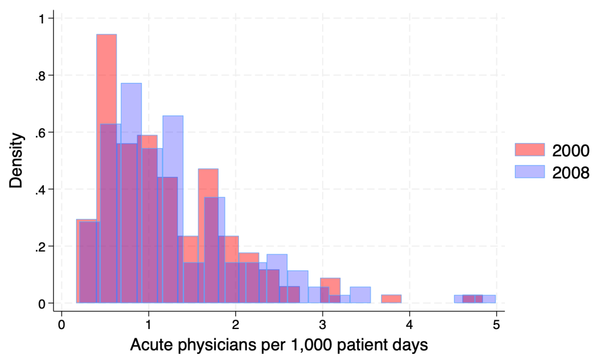
In the Supplemental Appendix, I plot the raw means of nurses for the treated and control hospitals and the estimated treatment effects β_t from the estimation of Equation (4) with nurses per patient as the outcome variable. I show that the mandate led to a 12 percent increase in acute care nurses per patient at treated hospitals within one year of implementation. In Figure 4 and a corresponding figure in the Supplemental Appendix, I plot the raw means and treatment effects for the risk-adjusted non-readmission rate and mortality rate. My findings indicate that the mandate had a statistically significant effect on the non-readmission rate at treated

[20] Raja (2023) finds that the effects of the mandate are estimated to be similar if research design uses a 0.2 threshold to assign treatment instead.

[21] See figure in Supplemental Appendix for the same distributions for control hospitals.



(a) Nurses Per Patient in 2000 vs. 2008



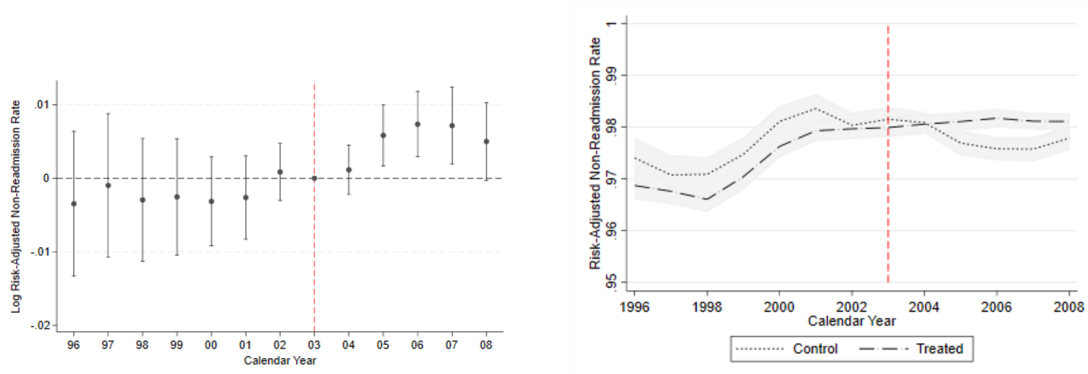
(b) Physicians Per Patient in 2000 vs. 2008

Figure 3: Input Use in 2000 vs. 2008 for Hospitals Treated by the Mandate

Notes: In panel (a), this figure shows the histogram of hospitals treated by the mandate according to the number of nurses per 1,000 patient days prior to the mandate in 2000 (red) and after the mandate in 2008 (blue). In panel (b), I do the same for physicians per 1,000 patient days which represent the physician FTEs constructed using patient care time. Taken together, the figures indicate that the mandate led to a large shift in the ratio of nurses to physicians.

hospitals with a magnitude of 0.7 percent within one year of implementation with no effect on patient mortality. Since the mandate did not have an effect on 30-day survival rates, we can rule out the possibility that the non-readmission rate increased due to a decrease in survival. [22]

[22] To benchmark the magnitude of these findings to the literature, Gupta (2021) finds that the HRRP, Medicare's value-based purchasing program which links reimbursements to readmission rates, led to a 5 percent decline in the readmission rate with approximately 40-50 percent of the decline in the probability of readmission explained by more stringent readmission policies for patients returning within 30 days.



(a) Event-Study Estimates

(b) Raw Means

Figure 4: Event-Study Estimates of the Effect of the Mandate on Log Non-Readmission

Notes: In panel (a), this figure plots coefficients β_t and 95 percent confidence intervals from Equation 1 with the log of the risk-adjusted non-readmission rate as the outcome variable. Standard errors are clustered at the hospital level. The mandate leads to an average quality effect of 0.7 to 1 percent between 2005 and 2008. In Panel (b), this figure plots average values and standard error bands of the rate by group.

IV. Structural Model of Hospital Quality Production

The reduced-form effects of the mandate indicate that an increase in nurse staffing led to improvement in hospital quality. However, we require a model to recover the production primitives and evaluate the regulation relative to direct quality regulation (to recover the within-hospital misallocation) and relative to alternate allocations of nurses across hospitals (to recover the across-hospital misallocation).

A. Hospital as multi-product firm

I consider the hospital to be a multi-product firm that produces healthcare for patients with a myriad of conditions. Early work has highlighted the challenge in studying hospital production-cost functions under output heterogeneity: hospitals vary in the diagnoses they treat, the services they produce, and the severity of the patients they admit and these differences are linked to the costs of production (Tatchell, 1983, Lave and Lave, 1970, Lee and Wallace, 1973).

Patient i is admitted to hospital h with initial health x_i^h and is assigned x_i^m materials, x_i^k capital, x_i^n nurse labor, and x_i^p physician labor per day. Patient i is not readmitted within 30 days if they received a quality of care $X - x_i^h$ during their inpatient stay and is readmitted if they receive quality of care less than this amount. Specifically:

$$Q_i = \mathbb{1}\{(X - x_i^h) \geq Y\} = f(x_i^n, x_i^p, x_i^m, x_i^k, x_i^h, \epsilon)$$

Given this differentiation in x_i^h, a key challenge is that I do not observe input allocations x_i^n and x_i^p at the patient level to estimate a patient-level production function. [23] However, I argue that the unique features of this setting satisfy the existence conditions for a hospital-level aggregate production function for $\overline{Q_h}$.[24]

$$\overline{Q_h} = \frac{\sum_i Q_i}{N} = F\left(\frac{\sum_i x_i^n}{N}, \frac{\sum_i x_i^p}{N}, \frac{\sum_i x_i^m}{N}, \frac{\sum_i x_i^k}{N}, \frac{\sum_i x_i^h}{N}\right)$$

Fisher (1969) shows that if all inputs are physically homogeneous and mobile between firms then the aggregate production function $Q^{**} = F(K, L)$ over firm-specific production functions $Y(v) = f^v(K(v), L(v))$ exists if we assume the efficient allocation of inputs across firms so as to maximize output. Q^{**} represents the maximized output. This is a stringent condition when we consider the aggregation of production across firms but it is reasonable to assume that the allocation of inputs across patients within the hospital is efficient from the perspective of maximizing average quality. In the Appendix, I show that the aggregate hospital-level production function exists under the condition that inputs are allocated efficiently across patients within the hospital.

The issue in this setting – analogous to the capital aggregation issue in Fisher (1969) – is that physicians are highly specialized and therefore in many cases immobile across patients. Surgeons, for example, produce coronary artery bypass grafting procedures for patients with coronary artery diseases. Oncologists produce chemotherapy and other treatments for cancer patients. For many treatments we can imagine that the types of physicians are not homogeneous. I therefore impose the following:

Assumption 1 – Nurse labor x^n, materials x^m, and capital x^k are homogeneous and mobile across patients and are efficiently allocated across patients to maximize the hospital's average quality across patients Q_{ht}.

[23]The inclusion of x_i^h in the model follows prior empirical work on hospital cost-production functions. For example, Lave and Lave (1970) state: "The multi-product nature of output can be taken into account either by deriving a composite output measure which is a weighted sum of the various outputs or by including measures of the proportions of each output as explanatory variables."

[24]Prior work on production has taken one of two approaches to address unobserved input allocation across product lines. The first is to estimate a joint production function for the vector of outputs, e.g. Grieco and McDevitt (2017) model a joint production function for quality and patient volumes with separability. The second is to assume the technology that produces the vector of outputs is nonjoint and use structural assumptions or the identifying variation from single-product firms. Examples include Diewert (1978) who assumes perfectly competitive factor markets, Valmari (2023) who assumes profit-maximizing firms, and De Loecker (2011) who rely on single-product firms for identification. Unfortunately there are no single-product hospitals and structural assumptions on the factor markets and profit-maximizing behavior would preclude institutional features of the market.

Assumption 2 – Physician labor x_i^p is a fixed factor which is immobile across patients. For any two types of physicians $i = 1, 2$ and for every patient v, the marginal rate of transformation between $x_1^p(v)$ and $x_2^p(v)$ is independent of $x^n(v)$.

Assumption 3 – For the physician aggregate x^p and for every patient v:

$$\frac{f^v_{x^p x^n}}{f^v_{x^p} f^v_{x^n x^n}} = g(f^v_{x^n})$$

Assumption 2 guarantees the existence of a physician aggregate at the patient level. Under Assumption 2, an additional nurse (or additional unit of any other input) affects the marginal product of one type of physician proportionally to the marginal product of any other type of physician. Assumption 3 guarantees the aggregate of x^p across patients v. Under constant quality returns to scale, Assumption 3 implies that the average product of nurse labor is equal across patients when nurse labor is optimally allocated. Assumptions 1-3 together guarantee the existence of an aggregate production function for each hospital and year:

$$Q^* = F(x^n, x^p, x^h, x^m)$$

B. Structural value-added in capital and materials

Nurse and physician labor are measured in physical quantities while capital and materials are measured in expenditures. I assume capital and materials cannot be substituted with either labor or the patient's initial health and I assume that capital and materials per patient are always used in optimal proportion to labor per patient and the patient's initial health. Let the production function be given by Equation (5):

$$Q_{ht} = g(x^n_{ht}, x^p_{ht}, x^h_{ht}, x^m_{ht}) e^{\omega_{ht} + \epsilon_{ht}} \tag{5}$$

Assumption 4 – Capital and materials cannot be substituted with either labor the patient's initial health.

$$Q_{ht} = \min[F(x^n_{ht}, x^p_{ht}, x^h_{ht}), s(x^m_{ht})] e^{\omega_{ht} + \epsilon_{ht}}$$

Assumption 5 – Hospitals choose per patient capital and materials optimally given their choices of labor per patient and given the patient's initial health.

$$F(x^n_{ht}, x^p_{ht}, x^h_{ht}) = s(x^m_{ht})$$

Assumptions 4-5 underlie the structural value-added production function (Diewert, 1978, Gandhi et al., 2017). I argue that these assumptions are realistic in the hospital industry. Nurses and physicians play a diagnostic, prescriptive, and monitoring role in producing quality of care which requires the use of capital to test and diagnose patients, to communicate with one another, and to administer treatment using pharmaceuticals and medical instruments.[25] Table 1 indicates that capital and materials per patient increase nearly proportionally with nurses per patient in the cross-section.

Further evidence can be seen from the effects of regulatory interventions. We should expect that if the mandate led to increases in average quality per patient and Assumptions 4-5 hold then the mandate must have led to increases in capital and materials per patient as well. Raja (2023) shows that this is indeed the case. These results are furthermore consistent with Gupta (2021)'s finding that the improvement in quality under Medicare's HRRP was associated with increases in the use of materials and diagnostic imaging in addition to an increase in physician time.

In the presence of labor adjustment costs, Assumption 5 is unlikely to hold for each hospital and year absent imposing greater structure on the function $s(\cdot)$ in Assumption 6 (Ackerberg et al., 2015, Gandhi et al., 2017).

Assumption 6 – The function s is linear in x_{ht}^m.

Assumptions 4-6 allow me to rewrite the gross output production function as the structural value-added production function below:

$$Q_{ht} = F(x_{ht}^n, x_{ht}^p, x_{ht}^h)e^{\omega_{ht}+\epsilon_{ht}}$$

C. Translog parameterization

I parameterize the production function using a translog form. Figure 1 illustrates how the elasticity of substitution between the factors is a key primitive driving the degree of misallocation from the mandate. I therefore argue that a flexible functional form is required to recover the elasticity of substitution rather than assume it. Common functional forms such as Leontief, Cobb-Douglas, and linear production functions impose elasticities of substitution of zero, one, and infinity, respectively, on the factors and the constant elasticity of substitution (CES) form imposes that the elasticity is invariant to the input levels. The latter could be an issue if we imagine the first doctor

[25] Capital-labor substitutability has become more common with the recent use of labor-replacing capital equipment in healthcare – for example, artificial intelligence as a diagnostic tool in healthcare. However, IT capital expenditures during my sample period were concentrated in hospital billing, provider monitoring, and clinical decision support which perform distinct tasks from clinical workforce (Lee et al., 2013). Recent developments may require that production function estimates based on current data make different assumptions on the relationships between labor and capital.

and first nurse treating a patient are more substitutable with one another than the third or fourth, etc... Instead, I use the translog functional form which allows me to recover the elasticities of substitution that vary with input levels.

$$Q_{ht} \equiv e^{\omega_{ht}+\epsilon_{ht}} \prod_{i\in\{h,n,p\}} (x^i_{ht})^{\beta_i} \prod_{i\in\{h,n,p\}} (x^i_{ht})^{\frac{1}{2}\left(\sum_{j\in\{h,n,p\}} \beta_{ij} \ln(x^j_{ht})\right)}.$$

I take the logs of both sides to obtain the linear-in-parameters estimating equation:

$$\ln(Q_{ht}) = \sum_{i\in\{h,n,p\}} \beta_i \ln(x^i_{ht}) + \frac{1}{2} \sum_{i\in\{h,n,p\}} \sum_{j\in\{h,n,p\}} \beta_{ij} \ln(x^i_{ht}) \ln(x^j_{ht}) + \omega_{ht} + \epsilon_{ht}. \tag{6}$$

D. Factor markets

I assume nurses and physicians are variable inputs that are chosen in the same period as they are used. While the markets for healthcare professionals are notably rigid relative to other labor markets, the average "time-to-hire" between the initial search date and the contract date in both nurse and physician markets falls below the one-year mark: the average time-to-hire for hospital nurses is three months and for physicians is four months with the specialties with longest time-to-hire (urology, neurology) taking slightly under one year (Gradney, 2023). See the Supplemental Appendix for more details. Both nurses and physicians are assumed to be dynamic inputs and allow for adjustment costs in labor consistent with prior work in healthcare production (Lee et al., 2013, Grieco and McDevitt, 2017).

Nurse and physician labor markets do not need to be perfectly competitive. I introduce wedges in the cost-minimization problem of the hospital, which I calibrate in the counterfactual exercises, to illustrate how quantification of the misallocation from the mandate varies between frictionless and imperfectly competitive markets.

E. Productivity

Productivity is assumed to be Hicks-neutral and assumed to be the sum of a hospital fixed effect ω_h, year fixed effect γ_t, and hospital-year specific shock ξ_{ht}:

$$\omega_{ht} = \omega_h + \gamma_t + \xi_{ht} \tag{7}$$

Assumption 8 – ξ_{ht} is serially uncorrelated.

In this model, productivity represents any element of the hospital's quality unattributed to the physical quantities of labor inputs or to the case mix. Productivity includes management practices, quality of labor inputs, health shocks, and economies of scale in patient volumes. Production is multiplicatively separable or Hicks-neutral in productivity which forecloses the possibility of factor-biased technological change that improves nurse labor productivity differentially from physician labor productivity.[26]

I argue that the dynamics of the productivity process are not of first-order concern in the hospital industry – the large component of hospital quality is constant over time with hospital fixed effects capturing 50 percent of the variation in the 30-day non-readmission rate and year fixed effects capturing two percent. Assumption 8 imposes that the remaining hospital-year variation is serially uncorrelated.

I conduct several robustness checks. I test this assumption using the Sargan test of overidentifying restrictions whose null hypothesis would be rejected if the hospital-year shock were serially correlated. In the Appendix, I estimate a robustness which imposes an alternate productivity restriction that ω_{ht} follows an AR(1) process. The results are discussed in Section 6.

F. Entry and exit

I abstract from entry and exit given that the majority of care is provided by long-lived hospitals. The 208 hospitals in my balanced panel comprise 77 percent of the patient days in acute care over the sample period. [27]

V. Identification and Estimation

The identification of production models has received significant attention in the industrial organization literature (Marschak and Andrews, 1944, Ackerberg et al., 2015). Two endogeneity concerns arise from the estimation of Equation (6): the hospital's unobserved productivity ω_{ht} determines whether and when the hospital enters or exits the market (selection) and it determines the hospital's choice of inputs when it operates (simultaneity). [28] Approaches to address these endogeneity may

[26]Other work including Gandhi et al. (2017) have proposed the use of first-order conditions to identify multiple dimensions of heterogeneity but this paper refrains from imposing behavioral assumptions on the hospital's objectives which forecloses the use of standard tools for identification.

[27]This is far larger coverage than prior work that has focused on selection bias. For example, Olley and Pakes (1996) find that restricting to a balanced panel in their setting uses only 35 percent of the full set of observations.

[28]The simultaneity issue applies to patient health in addition to labor inputs because patients can observe the hospital's productivity in a given period and choose the hospital leading to endogenous patient selection. Roughly fifty percent of hospital admissions are elective.

combine: timing assumptions over the factors of production, use of instrumental variables for endogenous factors, statistical restrictions over the productivity process, or the construction of a control function for the unobserved productivity term, among others. The appropriateness of the approach depends on the empirical features of production in the industry.

A. Identification

I impose the statistical restriction on the productivity process in Assumption 8 in conjunction with timing assumptions over the factors and internal and external instruments to identify the production model in Equation (6). Hospitals choose labor x^n, x^p and patients with case mix x^h select hospitals after the productivity shock in period t is realized meaning we require instruments for all three variables.

I use lagged nurse labor x^n_{ht-1}, lagged physician labor x^p_{ht-1}, lagged patient case mix x^h_{ht-1}, the mandate M_{ht}, and interactions between these variables in the instrument set. Additionally I include a second external instrument C_{ht} described below which shifts the patient case mix within hospitals. Assumption 8 ensures that the endogeneity of input choices to productivity in period $t-1$: $x^i_{ht-1} = f(\xi_{ht-1})$ does not imply the endogeneity of input choices in period t to productivity in period t conditional on input choices in period $t-1$. The adjustment cost model implies persistence in the input variables needed to satisfy the relevance criterion of the instrumental variables model even if the productivity shocks themselves are uncorrelated (Bond and Soderbom, 2005). The following exclusion restrictions follow:

$$E[\xi_{ht}|(x^n_{ht-1}, x^p_{ht-1}, x^h_{ht-1}, M_{ht}, C_{ht})]_{t\in[2,....T]} = 0$$

I include two external instruments: the first is a kinked treatment indicator for the nurse staffing mandate, M_{ht}, which takes on a value equal to the difference between 0.25 and the average nurse-to-patient ratio of the hospital in 2000-2002 if the hospital had an average below 0.25 (a measure of the incidence of the mandate) and zero otherwise. M_{ht} is interacted with an indicator for $t > 2003$.

The second external instrument captures whether the hospital converted to Critical Access Hospital status in the early 2000s. The CAH program was established in 1997 as part of the Balanced Budget Act to support the financial health of rural hospitals: it enabled small and rural hospitals to be reimbursed on cost-basis rather than through Prospective Payment System (PPS). CAH participant hospitals were required to maintain fewer than 25 acute care beds and an average length of stay of less than 96 hours per patient which led to reductions in length of stay and changes in the DRGs of admitted patients towards less severe patients following conversion (Schoenman and Sutton, 2008). The conversion event shifted x^h downwards for participant hospitals. I discuss the event and threats to identification from the external instruments in greater detail in the Supplemental Appendix.

This identification strategy places a strong statistical restriction on the productivity process and relies on instrument validity in lieu of strong economic restrictions on product market structure or on the optimality of observed input choices. The construction of a control function for unobserved productivity assumes that observed input use for at least one input increases monotonically in unobserved productivity (Pakes, 1991) which I argue to be unlikely given heterogeneity in product and insurer market structures that drives heterogeneity in hospital quality choice.

B. Estimation

I use IV2SLS to estimate Equation (6) with hospital and year fixed effects included in both stages of the estimation procedure. This estimation procedure is equivalent to an IV2SLS estimation of Equation (6) with demeaned variables and demeaned instruments. [29]

In the Appendix, I conduct a robustness which employs an alternate, AR(1) assumption on the productivity process and the System GMM dynamic panel estimator of Blundell and Bond (1998). I discuss the results in Section 6.

VI. Production Model Estimates

In this section, I present the results from the estimation of Equation (6) and use the estimated primitives to recover the elasticities of substitution between the inputs and the marginal product of nurse labor.

In Table 2, I report the production function estimates for the Cobb-Douglas and translog models estimated using OLS and IVFE. Columns (2)-(4) utilize fixed effects and instrumental variables to address the endogeneity of input choices to hospital

[29]The estimation of an IV2SLS model with demeaned variables and demeaned lagged input variables as instruments introduces a mechanical correlation between the demeaned lagged input variables and the error term similar to the problems noted in Nickell (1981) over the inconsistency of a fixed effects model in which there is an underlying dynamic process. The bias is significant in large N, small T panels and reduces as $T \to \infty$ with the bias of the estimates bounded to order T^{-1}. Given that my sample consists of 12 years of data the bias is minimal and undetected by the Sargan test of overidentifying restrictions.

productivity. Column (4) shows my preferred specification – a partially restricted translog which excludes the two terms that were found to be statistically insignificant in the full translog specification shown in Column (3).

Regarding instrument relevance, I report the first-stage estimates and F-statistics for the joint significance of the instruments. My findings indicate that the instruments are relevant and my main model is not weakly identified. In response to prior work on weak or underidentification in models with multiple endogenous variables (Sanderson and Windmeijer, 2016), I report the conditional F-statistics of Sanderson-Windmeijer and compare them to the corresponding Stock-Yogo weak identification critical values and find that of the nine first-stage regressions that I estimate, I can reject that the variable is weakly identified in all but three cases. In these three cases, the F-statistic is larger than lowest critical value (representing 5 percent maximal IV relative bias). Regarding bias, I find that in all of my models I cannot reject the null hypothesis of the Sargan test that the overidentifying restrictions are valid.

Notably, my findings reject the Cobb-Douglas model: I employ an F-test of joint significance for the squared and interaction terms in Column (3) to assess this assumption formally and I reject the null hypothesis of this test. Said plainly, the use of the Cobb-Douglas parameterization would yield biased estimates of misallocation.

In the Appendix, I present the results from the dynamic panel estimation of Equation (6) under the assumption that ω_{ht} follows an AR(1) process. In Column (1), I present results using Arellano and Bond (1991)'s Difference GMM and in Columns (2)-(3), results using Blundell and Bond (1998)'s System GMM.

A. Nurses and physicians are highly complementary

I use the estimated production parameters in Table 2, I derive the structural objects of interest in this paper. I estimate the "direct" elasticity of substitution between nurses and physicians as in (J.D. Sargan, 1971) (see Supplemental Appendix for the algebraic derivation). In Table 3, I present the percentiles of the elasticity. The 95 percent confidence intervals based on a parametric bootstrap are shown in brackets. For each hospital-year observation in the data, I compute the elasticity of substitution implied at each of the levels of nurse staffing indicated in the first column of Table 3 taking the levels of physicians and patient health as data.

I find nurses and physicians to be highly complementary. The estimated elasticities for the 50th percentile of the distribution for all nurse staffing levels shown lie around zero (Leontief or perfect complements) with the 95 percent confidence intervals falling

Table 2: Production Model for Hospital 30-Day Non-Readmission

	OLS	IVFE		
	(1) Translog	(2) Cobb-Douglas	(3) Translog	(4) Translog
β_n	0.026	-0.002	-0.028	–
	(0.020)	(0.006)	(0.055)	–
β_p	-0.004	0.001	0.110	0.103
	(0.005)	(0.002)	(0.042)	(0.032)
β_h	0.054	-0.018	0.227	0.197
	(0.020)	(0.012)	(0.065)	(0.053)
β_{nn}	-0.029	–	-0.075	-0.105
	(0.021)	–	(0.044)	(0.021)
β_{pp}	-0.001	–	-0.111	-0.107
	(0.003)	–	(0.035)	(0.026)
β_{hh}	-0.025	–	-0.207	–
	(0.031)	–	(0.291)	–
β_{np}	0.009	–	0.053	0.056
	(0.004)	–	(0.016)	(0.011)
β_{nh}	-0.037	–	-0.122	-0.100
	(0.017)	–	(0.049)	(0.027)
β_{ph}	-0.006	–	-0.060	-0.047
	(0.006)	–	(0.022)	(0.018)
N	2,704	2,496	2,496	2,496

Notes: This table reports the production function estimates for Cobb-Douglas and translog models estimated using OLS and IVFE. Analytic standard errors are shown in parentheses. Column 4 shows the preferred specification.

below one (Cobb-Douglas). The estimated elasticities for the 10th and 25th percentiles are negative which is a byproduct of the flexible parameterization. It is a commonly known issue that translog and other flexible models do not often satisfy regularity conditions globally. However, locally, at the median of the data (2.88 nurses per 1,000 patient days) most estimates are positive. These findings are robust to alternate assumptions on the productivity process.

In the Supplemental Appendix, I produce a version of the table that relies on the dynamic panel estimation. I find that the largest elasticity of substitution across all levels of nurse staffing is 0.62. While the dynamic panel approach does not allow one to

B. Complementarity increases with staffing level and patient complexity

Table 3 indicates that the complementarity increases as nurse staffing increases (down each column of the table). For example, the confidence intervals range from 0.20 to -0.23 at the lower bound and 1.52 to 1.04 at the upper bound for the 75th percentile. Under low staffing, nurse and physician roles are more fluid and substitutable.

However, there is an upper bound to the substitutability: the 90th percentile always falls below 0.23 and the upper bound of its confidence interval in all but one case fall below 3. This is sensible with respect to the institutional context – licensing restrictions do not vary with staffing levels.

The complementarity also increases with patient complexity. For each level of nurses, I regress the elasticity on hospital fixed effects, level of physicians, and level of patient health. I find that the patient health is positively and statistically significantly correlated with the elasticity of substitution at each level of nurse staffing. Nurses and physicians are more substitutable when patients are healthy whether due to licensing (tasks for healthy patients more likely to overlap between nurses and physicians) or skill (nurses are skilled enough to handle the overlapping tasks) consistent with prior work (Chan and Chen, 2022)

Table 3: Elasticities of Substitution - Nurses and Physicians

	Percentiles of Distribution				
Nurses	10th	25th	50th	75th	90th
1	−0.54	−0.35	−0.13	0.07	0.18
	[−1.84, 0.35]	[−0.43, 0.64]	[0.06, 1.00]	[0.20, 1.52]	[0.27, 2.86]
1.5	−0.29	−0.14	−0.00	0.12	0.21
	[−1.79, 0.23]	[−0.46, 0.56]	[0.00, 0.85]	[0.17, 1.40]	[0.24, 2.75]
2	−0.25	−0.10	0.02	0.12	0.22
	[−2.20, 0.16]	[−0.77, 0.53]	[−0.18, 0.80]	[0.09, 1.29]	[0.21, 2.56]
2.5	−0.26	−0.10	0.01	0.11	0.20
	[−3.24, 0.08]	[−1.35, 0.40]	[−0.47, 0.75]	[−0.03, 1.16]	[0.18, 2.54]
3	−0.25	−0.12	0.00	0.10	0.16
	[−4.80, 0.08]	[−1.91, 0.30]	[−0.72, 0.67]	[−0.13, 1.07]	[0.15, 2.43]
3.5	−0.25	−0.13	0.00	0.10	0.17
	[−6.21, 0.17]	[−2.44, 0.29]	[−0.91, 0.57]	[−0.23, 1.04]	[0.12, 4.77]

Notes: In this table, I present the percentiles of the distribution of elasticities of substitution for each hospital-year in my sample. The near zero elasticities of substitution indicate strong complementarities in quality production between nurses and physicians. Nurses and physicians are more substitutable at low levels of the two inputs i.e. when hospitals are relatively understaffed.

C. Marginal product of nurse labor

In Figure 5, I plot the nurse marginal product. For each hospital-year observation, I compute the marginal product for each level of nurse staffing from a grid. I take the levels of physicians and patient health as data and the productivity as estimated for each hospital-year. This figure indicates wide dispersion in the returns from regulation. As with the negative elasticities, the range of negative marginal product indicates that regularity conditions are not satisfied globally or that there are true negative effects from higher staffing stemming from nursing handoffs or an increase in the span of control for leaders that supervise and monitor nursing teams.

I plot the estimated marginal products against productivities and observed case mixes in Figure 6. Note that the standard deviation of hospital productivity that I estimate is 0.029 points where the 90-10 range is 0.070 implying that a patient at the 90th percentile hospital is 1.07 times more likely to not be readmitted than a patient treated at the 10th percentile hospital ($e^{0.07} = 1.07$).[30] Figure 6a indicates that the dispersion in productivity explains little of the dispersion in marginal products.

Figure 6b shows that the dispersion in case mix, on the other hand, explains much of the dispersion in marginal products. High marginal product hospitals are just as likely to be low or high productivity but they are much more likely to be high severity.

VII. Misallocation Estimates

A. Within-hospital input misallocation from first-best

I present the within-hospital misallocation results under the assumption of perfectly competitive factor markets: $\Delta C = C^*_{sreg} - C^*_{dreg}$ where $C^*_{sreg} = (w^N N^*_{sreg} + w^P P^*_{sreg})$ and $C^*_{dreg} = (w^N N^*_{dreg} + w^P P^*_{dreg})$. The average annual costs per hospital due to the mandate relative to direct quality regulation are \$48,553 and the total annual costs across hospitals (ΔC) are \$10,001,908. I plot the cost distribution in Figure 7. Given annual U.S. hospital revenues of \$1.6 trillion (CMS, 2024), 5.2% average operating margins (KFF, 2024), and 6,093 hospitals in the U.S. (KFF, 2025), this puts the costs equal to 0.4 percent of annual profits.

[30]This productivity dispersion is notably smaller than an estimate of the productivity dispersion for heart attack survival in other work over a similar time period (Chandra et al., 2016b).

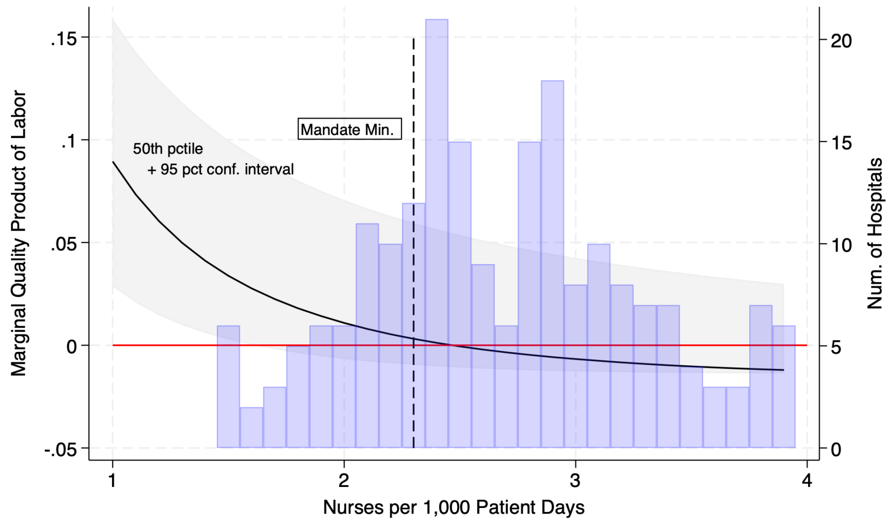


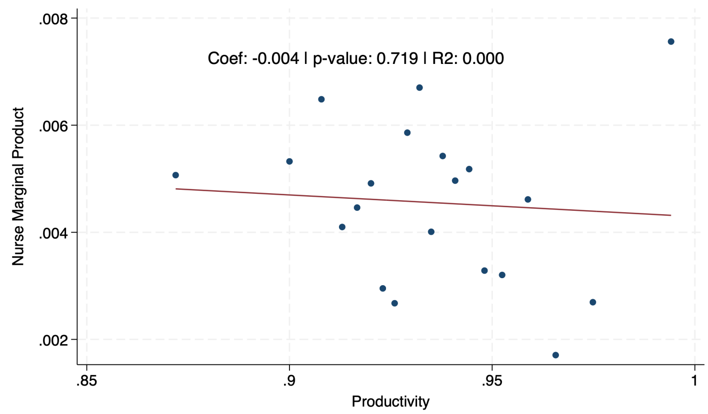
Figure 5: Median Nurse Marginal Product and the Mandate

Notes: In this figure, I plot the percentiles of the marginal product of nurse labor distribution. I compute the marginal product for each point on the nurse grid taking the physicians, patient health, and productivity as data and plot the percentiles of the distribution. A marginal product of 0.04 indicates that a one unit increase in nurses per 1,000 patient days at the specified level leads to a 0.04 point increase in the non-readmission rate (corresponding to approximately 4 percent).

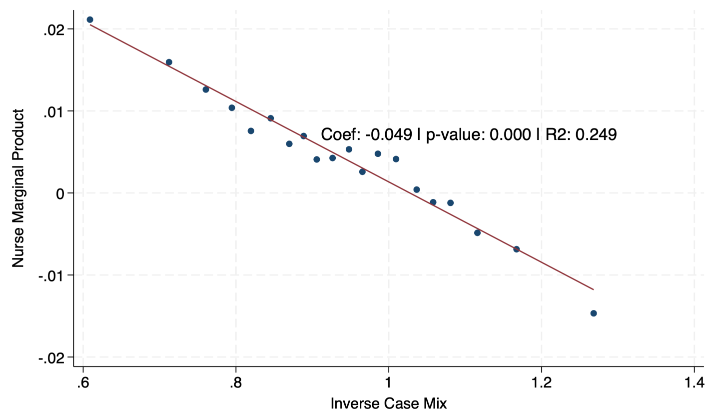
B. Within-hospital input misallocation from second-best

I present the within-hospital misallocation results under imperfectly competitive factor markets where μ^N and μ^P are wedges calibrated based on input choices in the pre-regulatory period: $\Delta C = C^*_{sreg} - C^*_{dreg}$ where $C^*_{sreg} = (\mu^N w^N N^*_{sreg} + \mu^P w^P P^*_{sreg})$ and $C^*_{dreg} = (\mu^N w^N N^*_{dreg} + \mu^P w^P P^*_{dreg})$. The average annual costs per hospital due to the mandate relative to direct quality regulation are \$104,834 and the total annual costs across hospitals are \$21,386,210. The costs under second-best are double the costs under first-best an amount to 0.8 percent of annual profits. Prior to the regulation, many hospitals were inefficiently understaffed with respect to physicians and the mandate exacerbated these pre-existing distortions. The costs under second-best account for the pre-existing distortions by way of the calibrated wedges.

The range of efficient ratios of nurses to physicians under first-best is between 1.35 and 2.05 nurses per physician. However, the 25-75 range of nurse to physician ratios prior to the mandate was 0.94 to 2.35. The hospitals which experience large misallocation costs after the mandate are understaffed with respect to physicians prior to regulation. In Table 4, I show the descriptive statistics of hospitals whose



(a) Marginal Product by Productivity Level



(b) Marginal Product by Patient Health Level

Figure 6: Heterogeneity in Marginal Product Across Hospitals

Notes: In panel (a) this figure plots the marginal product of nurse labor from the data against the recovered productivities. In panel (b), I show the marginal product of nurse labor and the inverse Case Mix. The coefficients, p-values, and R^2 for each regression are reported in the chart. The differences in patient health are more significant drivers of observed differences in marginal product.

nurse-physician ratio falls below 1.35 ("physicians overstaffed"), between 1.35 and 2.05 ("efficient"), and above 2.05 ("physicians understaffed"). Hospitals with physicians understaffed are more likely to be small and rural, government owned, and/or hospitals that serve a disproportionate number of Medicare and/or MediCal patients and less likely to be teaching hospitals.[31] The average revenue per inpatient day in acute care at these hospitals is $1,215 per patient day which is lower than both efficient ($1,349) and overstaffed ($1,633) hospitals. My findings corroborate recent health services work highlighting the low share of physicians out of total practitioners at critical access, non-urban, and small hospitals (Gettel et al., 2025).

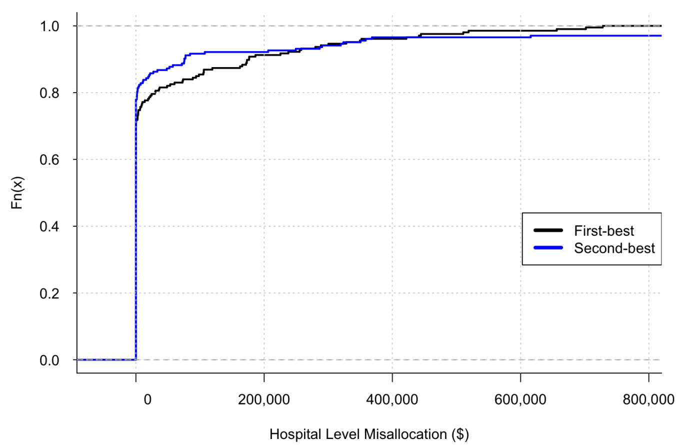


Figure 7: Empirical CDFs of Within-Hospital Misallocation

Notes: This figure shows the empirical CDFs for the within-hospital misallocation calculated for each hospital under first-best (perfect competition) and second-best (market power). Hospital-level misallocation reflects annual costs from the mandate relative to direct quality regulation.

C. Across-hospital input misallocation

To evaluate the across-hospital misallocation of nurses to low marginal product hospitals due to the mandate, I study the 150 markets for hospital services in California as defined by Hospital Service Area (HSA) by the Dartmouth Atlas Project. The HSA is defined by assigning each zip code to the hospital area where the largest proportion of the zip code's Medicare residents were hospitalized (Dartmouth Atlas, 2024).

[31]These descriptives are suggestive of several possible frictions – teaching hospitals value educating physicians in addition to providing quality healthcare therefore they "overstaff" from a quality perspective; if demand is not linked to hospital quality then low revenue hospitals may not make the requisite investments in physicians.

Table 4: Descriptive Statistics by Physician Staffing

	Nurse-Phys. Ratio	Nurses per 1,000 p.d.	Phys. per 1,000 p.d.	Share Teach	Share Small/ Rural	Share Govt. Owned	Share DSH	Acute Rev. p.p.d. ($)
Efficient	1.66	2.84	1.72	0.17	0.17	0.12	0.26	1,349
Physicians overstaffed	0.89	2.62	3.29	0.12	0.10	0.14	0.28	1,633
Physicians understaffed	2.99	2.83	1.01	0.05	0.24	0.27	0.36	1,215

Notes: In this table, I present descriptive statistics by the hospital's nurse to physician ratio prior to regulation. "Efficient" refers to hospitals with a ratio between 1.35 and 2.05. "Physicians overstaffed" refers to hospitals with a ratio below 1.35. "Physicians understaffed" refers to hospitals with a ratio above 2.05. "DSH" refers to Disproportionate Share Hospitals that receive more than 50 percent of their revenue from Medicare and/or MediCal.

Of these 150 geographic markets, only 29 include more than one hospital and only 8 include at least one treated hospital and one control hospital. For each of the 29 markets with more than one hospital, I estimate the change in the variance of the nurse marginal product before and after the mandate $\Delta\text{var}(\widehat{\text{MPL}_m}) = \text{var}(\widehat{\text{MPL}_{i\in m}})_{reg} - \text{var}(\widehat{\text{MPL}_{i\in m}})_{noreg}$. Figure 8 shows the distribution of $\Delta\text{var}(\widehat{\text{MPL}_m})$ across the 29 markets. A negative value for the change in variance indicates an improvement in productive efficiency due to reallocation that took place between 2002 and 2006. Notably, the majority of markets experienced an improvement in productive efficiency. As a robustness, in the Supplemental Appendix, I plot separate distributions of the same metric for markets with and without treated hospitals and show that the increase in productive efficiency is driven by treated hospitals. In other words, treated hospitals had high marginal products relative to their peers.

Table 5 presents the market-level results underlying Figure 8. This table shows several markets with treated hospitals which did not see improvements (Anaheim, Stockton, Burbank, San Gabriel, Vacaville, Glendora, Orange). By analyzing these markets and assessing the optimal allocation of nurses added between 2002 and 2006 within geographic markets, one finds that efficiency gains can be made by reallocating nurses to higher severity hospitals within each market where labor is more valuable.

VIII. Conclusion

Many regulatory options exist to address low quality hospitals. Labor regulation is an increasingly popular choice for several reasons. In settings with strong healthcare labor unions, staffing mandates are politically feasible. Relative to the regulation of

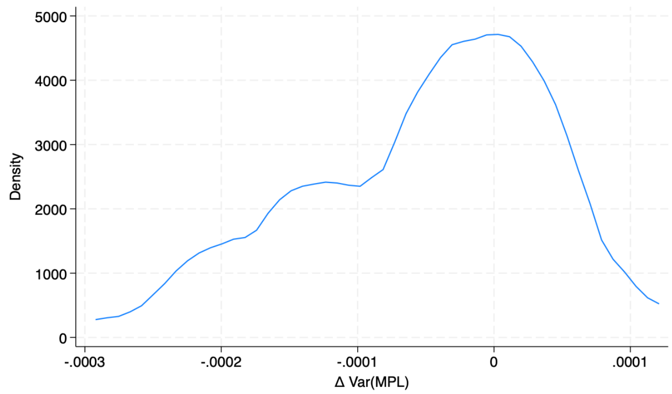


Figure 8: Distribution of Δ Var(MPL) Across Markets, All Hospitals

Notes: This figure plots the distribution of the Δ Var(MPL) across geographic markets. The majority of 29 markets shown here experienced a reduction in the variance of marginal products across hospitals in their markets due to the mandate.

healthcare outcomes, staffing mandates have relatively low enforcement costs. Finally, in a setting with proven inadvertent consequences of healthcare outcomes regulation, the behavioral costs of direct quality regulation can be significant.[32] However, theory indicates that input regulation is misallocative both within and across firms.

This paper considers the productive efficiency implications of minimum nurse-to-patient ratios in California's hospitals. To evaluate the misallocation within and across hospitals, this paper estimates a production model for hospital quality using input data from hospital reporting forms and healthcare outcomes constructed from administrative patient discharge data. I use the 1999 California nurse staffing mandate and structural assumptions on productivity for identification which allows me to consider departures from perfectly competitive labor markets and hospital cost-minimization.

The main empirical findings indicate that labor regulation targeting a single input can be substantially misallocative in the healthcare setting where substitution across types is limited by licensing and skill. Regulation targeting only nurses increased annual costs by 0.5 percent ($10 million) holding quality constant. Misallocation is doubled under second-best where labor regulation exacerbates pre-existing understaffing of physicians at low revenue, government, and rural hospitals. Interestingly, the regulation was misallocative across hospitals primarily due to its one-size-fits-all approach across severity levels. Low staffing hospitals were high marginal product hospitals indicating pre-existing (mis)allocation of labor across hospitals.

[32] Gupta (2021) finds that nearly half of the average reduction in readmission due to CMS's HRRP was due to selective readmission.

Table 5: Productivity Implications of Nurse Allocation from Mandate, by Market

		Num. Hospitals		Max. MPL		Ave. Nurses (Trtd.)		Ave. Nurses (Ctrl.)	
Market (HSA)	Δ Var(MPL)	Treated	Control	2002	2006	2002	2006	2002	2006
Mission Hills	−0.000 26	2	0	0.016	0.002	2.31	3.31	–	–
Glendale	−0.000 19	3	0	0.014	−0.001	2.06	3.08	–	–
Victorville	−0.000 18	1	1	0.015	−0.005	1.62	3.16	2.90	3.74
Lancaster	−0.000 18	2	0	0.009	−0.002	2.33	2.81	–	–
Bakersfield	−0.000 16	3	1	0.023	0.009	2.45	2.79	3.17	3.56
Torrance	−0.000 12	3	0	0.006	0.005	2.39	2.83	–	–
Los Angeles	−0.000 10	8	3	0.026	0.013	2.27	3.22	3.07	3.68
Long Beach	−0.000 10	5	0	0.012	0.003	2.80	3.17	–	–
Ventura	−0.000 09	2	0	0.012	0.001	2.52	3.48	–	–
Oakland	−0.000 09	1	1	0.008	−0.004	2.58	4.52	3.61	4.87
Fresno	−0.000 06	3	0	0.031	0.021	2.01	2.57	–	–
San Jose	−0.000 05	3	0	0.015	0.001	2.63	3.55	–	–
Riverside	−0.000 02	2	1	0.002	−0.002	2.52	3.15	2.96	3.19
San Francisco	−0.000 02	4	4	0.006	0.001	2.62	3.59	3.52	4.05
San Bernardino	−0.000 02	2	0	0.007	−0.008	1.92	3.74	–	–
Sacramento	0.000 00	0	3	0.006	−0.002	2.55	–	3.73	3.55
Anaheim	0.000 00	4	0	0.012	0.015	2.53	3.14	–	–
San Diego	0.000 00	0	2	0.005	−0.001	–	–	2.98	3.69
Stockton	0.000 00	3	0	0.001	0.000	2.21	2.93	–	–
Santa Rosa	0.000 00	0	2	−0.008	−0.005	–	–	3.72	4.05
Santa Barbara	0.000 00	0	2	−0.003	−0.002	–	–	3.61	3.85
Burbank	0.000 01	2	0	−0.016	−0.015	2.25	3.39	–	–
Santa Monica	0.000 02	0	2	0.009	0.010	–	–	3.21	3.85
San Gabriel	0.000 03	2	0	0.018	0.001	1.89	2.88	–	–
Vacaville	0.000 04	1	1	0.006	0.001	2.94	3.20	3.24	4.19
Glendora	0.000 08	1	1	−0.010	0.005	2.44	2.82	3.40	3.74
Orange	0.000 09	3	0	0.016	0.002	2.38	3.32	–	–

Notes: This table shows the market-level changes in the variance in marginal products of labor across hospitals within each of the geographic markets listed in the first column. A negative change in variance indicates a reduction in the dispersion of marginal products of nurse labor between 2002 and 2006.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification Properties of Recent Production Function Estimators. *Econometrica 83*(6), 2411–2451.

Arellano, M. and S. Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies 58*(2), 277–297.

Arrow, K. J. (1963, December). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review*.

Asker, J., A. Collard-Wexler, and J. De Loecker (2019, April). (Mis)Allocation, Market Power, and Global Oil Extraction. *American Economic Review 109*(4), 1568–1615.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*.

Bond, S. and M. Soderbom (2005, January). Adjustment costs and the identification of Cobb Douglas production functions. Publication Title: Institute for Fiscal Studies.

Burns, L. R. and R. W. Muller (2008). Hospital-Physician Collaboration: Landscape of Economic Integration and Impact on Clinical Integration. *The Milbank Quarterly 86*(3), 375–434.

Chan, David C., J. and Y. Chen (2022, October). The Productivity of Professions: Evidence from the Emergency Department. Series: Working Paper Series Type: Working Paper.

Chandra, A., C. Colla, and J. Skinner (2023, August). Productivity Variation and Input Misallocation: Evidence from Hospitals. Technical Report w31569, National Bureau of Economic Research, Cambridge, MA.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016a). Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector. *The American Economic Review 106*(8), 2110–2144.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016b, May). Productivity Dispersion in Medicine and Manufacturing. *American Economic Review 106*(5), 99–103.

Chandra, A. and J. Skinner (2012, July). Technology Growth and Expenditure Growth in Health Care. *Journal of Economic Literature 50*(3), 645–680.

Chandra, A. and D. O. Staiger (2020, May). Identifying Sources of Inefficiency in Healthcare*. *The Quarterly Journal of Economics 135*(2), 785–843.

De Loecker, J. (2011). PRODUCT DIFFERENTIATION, MULTIPRODUCT FIRMS, AND ESTIMATING THE IMPACT OF TRADE LIBERALIZATION ON PRODUCTIVITY. *Econometrica 79*(5), 1407–1451.

Diewert, W. E. (1978, January). Hicks' Aggregation Theorem and the Existence of a Real Value-Added Function*. In M. Fuss and D. McFADDEN (Eds.), *Contributions to Economic Analysis*, Volume 2 of *Applications of the Theory of Production*, pp. 17–51. Elsevier.

Einav, L., A. Finkelstein, and N. Mahoney (2022, August). Producing Health: Measuring Value Added of Nursing Homes. Technical Report w30228, National Bureau of Economic Research, Cambridge, MA.

Fisher, E., D. Goodman, J. Skinner, and K. Bronner (2009). Health Care Spending, Quality, and Outcomes. Technical report, Dartmouth Atlas Project.

Fisher, F. M. (1969). The Existence of Aggregate Production Functions. *Econometrica 37*(4), 553–577.

Gandhi, A., S. Navarro, and D. Rivers (2017). How Heterogeneous is Productivity? A Comparison of Gross Output and Value Added. *University of Western Ontario, Centre for Human Capital and Productivity (CHCP) Working Papers*.

Garber, A. M. and J. Skinner (2008, December). Is American Health Care Uniquely Inefficient? *Journal of Economic Perspectives 22*(4), 27–50.

Gertler, P. J. and D. M. Waldman (1992). Quality-adjusted Cost Functions and Policy Evaluation in the Nursing Home Industry. *Journal of Political Economy 100*(6), 1232–1256.

Gettel, C. J., Z. Lin, C. Rothenberg, Z. Lin, T. Lagu, K. Goodrich, J. S. Ross, and A. K. Venkatesh (2025, November). Physician Composition of Hospitals' Workforce And Mortality Across U.S. Hospitals.

Gottlieb, J. D., M. Polyakova, K. Rinz, H. Shiplett, and V. Udalova (2023, July). Who Values Human Capitalists' Human Capital? The Earnings and Labor Supply of U.S. Physicians. Series: Working Paper Series Type: Working Paper.

Gradney, A. (2023, October). AAPPR Report Shows Demand for Physicians Reaching a New High as Staffing Shortage Continues. Publication Title: AAPPR - Association for Advancing Physician and Provider Recruitment.

Grieco, P. L. E. and R. C. McDevitt (2017, July). Productivity and Quality in Health Care: Evidence from the Dialysis Industry. *The Review of Economic Studies 84*(3), 1071–1105.

Gupta, A. (2021, April). Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program. *American Economic Review 111*(4), 1241–1283.

Havens, D. S., J. Vasey, J. H. Gitl, and W.-T. Lin (2010). Relational coordination among nurses and other providers: impact on the quality of patient care. *Journal of Nursing Management 18*(8), 926–937.

Hsieh, C.-T. and P. J. Klenow (2009, November). Misallocation and Manufacturing TFP in China and India*. *The Quarterly Journal of Economics 124*(4), 1403–1448.

Huckman, R. S. and G. P. Pisano (2006). The firm specificity of individual performance: Evidence from cardiac surgery. *Management Science 52*(4), 473–488.

J.D. Sargan (1971). Production Functions. In *Qualified manpower and economic performance : an inter-plant study in the electrical engineering industry*. London : Allen Lane The Penguin Press.

Kessel, R. A. (1958). The A.M.A. and the Supply of Physicians. *Law and Contemporary Problems*.

Lave, J. R. and L. B. Lave (1970). Hospital Cost Functions. *The American Economic Review 60*(3), 379–395.

Lee, J., J. S. McCullough, and R. J. Town (2013). The impact of health information technology on hospital productivity. *The RAND Journal of Economics 44*(3), 545–568.

Lee, M. L. and R. L. Wallace (1973). Problems in Estimating Multiproduct Cost Functions:an Application to Hospitals. *Economic Inquiry 11*(3), 350–363.

Marschak, J. and W. H. Andrews (1944). Random Simultaneous Equations and the Theory of Production. *Econometrica 12*(3/4), 143–205.

Mourot, P. Should Top Surgeons Practice at Top Hospitals?

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica 49*(6), 1417–1426.

Olley, G. S. and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica 64*(6), 1263–1297.

Pakes, A. (1991, July). Dynamic Structural Models: Problems and Prospects. Mixed Continuous Discrete Controls and Market Interactions. *Cowles Foundation Discussion Papers*.

Prager, E. and M. Schmitt (2021, February). Employer Consolidation and Wages: Evidence from Hospitals. *American Economic Review 111*(2), 397–427.

Raja, C. (2023, December). How do hospitals respond to input regulation? Evidence from the California nurse staffing mandate. *Journal of Health Economics 92*, 102826.

Rittenhouse, D., A. Ament, J. Genevro, and K. Contreary (2021, April). Health Workforce Strategies for California: A Review of the Evidence.

Romley, J. A. and D. P. Goldman (2011). How Costly is Hospital Quality? A Revealed-Preference Approach. *The Journal of industrial economics 59*(4), 578–608.

Rosenthal, M. B., A. Zaslavsky, and J. P. Newhouse (2005, December). The Geographic Distribution of Physicians Revisited. *Health Services Research 40*(6 Pt 1), 1931–1952.

Sanderson, E. and F. Windmeijer (2016, February). A weak instrument F\textlessmath\textgreater\textlessmi is="true"\textgreaterF\textless/mi\textgreater\textless/math\textgreater-test in linear IV models with multiple endogenous variables. *Journal of Econometrics 190*(2), 212–221.

Schoenman, J. A. and J. P. Sutton (2008). Impact of CAH Conversion on Hospital Finances and Mix of Inpatient Services. Technical report, NORC.

Skinner, J. and D. Staiger (2015, December). Technology Diffusion and Productivity Growth in Health Care. *The review of economics and statistics 97*(5), 951–964.

Tatchell, M. (1983, January). Measuring hospital output: A review of the service mix and case mix approaches. *Social Science & Medicine 17*(13), 871–883.

Valmari, N. (2023, November). Estimating Production Functions of Multiproduct Firms. *The Review of Economic Studies 90*(6), 3315–3342.

IX. Appendix

A. Existence of an aggregate production function

For ease of exposition assume a three-factor production function for patient i:

$$Q_i = e^{\epsilon_i} \prod_{v \in \{h,n,p\}} (x_i^v)^{\beta_v} \prod_{v \in \{h,n,p\}} (x_i^v)^{\frac{1}{2}\left(\sum_{j \in \{h,n,p\}} \beta_{vj} \ln(x_i^j)\right)}.$$

Now assume that the hospital solves the following problem to allocate x^n and x^p across patients with x_i^h within the hospital:

$$\begin{aligned} \max_{x_1^n, x_2^n, x_1^p, x_2^p} Q_1 + Q_2 \\ \text{s.t.} x_1^n + x_2^n = x^n \\ x_1^p + x_2^p = x^p \end{aligned}$$

The Lagrangian:

$$\mathcal{L} = Q_1 + Q_2 - \lambda_1 \left(x^n - x_1^n - x_2^n\right) - \lambda_2 \left(x^p - x_1^p - x_2^p\right)$$

The first-order conditions:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x_1^n} = \frac{\partial Q_1}{\partial x_1^n} - \lambda_1 = 0 \\ \frac{\partial \mathcal{L}}{\partial x_2^n} = \frac{\partial Q_2}{\partial x_2^n} - \lambda_1 = 0 \\ \frac{\partial \mathcal{L}}{\partial x_1^p} = \frac{\partial Q_1}{\partial x_1^p} - \lambda_2 = 0 \\ \frac{\partial \mathcal{L}}{\partial x_2^p} = \frac{\partial Q_2}{\partial x_2^p} - \lambda_2 = 0 \\ x^n - x_1^n - x_2^n = 0 \\ x^p - x_1^p - x_2^p = 0 \end{aligned}$$

Combining (1)-(2) and (3)-(4):

$$\frac{\partial Q_1}{\partial x_1^n} = \frac{\partial Q_2}{\partial x_2^n}, \qquad \frac{\partial Q_1}{\partial x_1^p} = \frac{\partial Q_2}{\partial x_2^p}$$

The first-order conditions indicate that nurses and physicians will be allocated across patients to equalize their marginal products. In this model, the marginal products vary across patients due to their initial health x_i^h but no other productivity differences across patients. Now, note one solution to the problem exists where $x_1^h = x_2^h$:

$$x_1^n = x_2^n = \frac{x^n}{2}, \qquad x_1^p = x_2^p = \frac{x^p}{2}$$

We can derive the aggregate production function across two patients:

$$F = 2\left(\frac{x^n}{2}\right)^{\beta_n}\left(\frac{x^p}{2}\right)^{\beta_p}$$
$$\left(x^h\right)^{\beta_h}\left(\frac{x^n}{2}\right)^{\frac{1}{2}\left(\sum_{j\in\{h,n,p\}}\beta_{nj}\ln(x^j)\right)}$$
$$\left(\frac{x^p}{2}\right)^{\frac{1}{2}\left(\sum_{j\in\{h,n,p\}}\beta_{pj}\ln(x^j)\right)}\left(x^h\right)^{\frac{1}{2}\left(\sum_{j\in\{h,n,p\}}\beta_{hj}\ln(x^j)\right)}$$

And it is apparent that for the aggregate production function to take the form:

$$Q_{ht} = \prod_{v\in\{h,n,p\}}(x_{ht}^v)^{\beta_v}\prod_{v\in\{h,n,p\}}(x_{ht}^v)^{\frac{1}{2}\left(\sum_{j\in\{h,n,p\}}\beta_{vj}\ln(x_{ht}^j)\right)}$$

We require the following:

$$\beta_n + \beta_p + \frac{1}{2}\left(\sum_{j\in\{h,n,p\}}\beta_{nj}\ln(x^j)\right) + \frac{1}{2}\left(\sum_{j\in\{h,n,p\}}\beta_{pj}\ln(x^j)\right) = 1 \tag{8}$$

In other words, the patient-level production function should have constant returns to scale in quality after we account for the differences in health across patients. Said differently, for a given patient for whom the hospital produces $(X - x^h)$ prior to discharge, we should find that doubling the number of nurses per patient and physicians per patient should yield double the non-readmission rate. [33]

[33] We can evaluate this assumption using estimates of $\widehat{\beta}$ from the hospital-level production function. I evaluate the LHS of Equation 8 based on my parameter estimates to be equal to 0.93 which is reasonably close to 1.

B. AR(1) productivity and dynamic panel estimation

In Equation (7) and Assumption 8 of the text, I impose a structure on the productivity process and assume the hospital- and year-specific shock, ξ_{ht}, is serially uncorrelated.

In this section, I identify and estimate the production function over an alternate assumption that productivity ω_{ht} follows an AR(1) process and σ_{ht} represents the hospital- and year-specific productivity innovation.

$$\omega_{ht} = \rho\omega_{ht-1} + \sigma_{ht}$$

We can ρ-difference the production model in Equation (6) to obtain

$$\begin{aligned}
&\ln(Q_{ht}) - \rho\ln(Q_{ht-1}) = \\
&\quad \beta_n(\ln(x^n_{ht}) - \rho\ln(x^n_{ht-1})) + \beta_h(\ln(x^h_{ht}) - \rho\ln(x^h_{ht-1})) + \beta_p(\ln(x^p_{ht}) - \rho\ln(x^p_{ht-1})) \\
&+ \frac{1}{2}\beta_{nn}((\ln(x^n_{ht}))^2 - \rho(\ln(x^n_{ht-1}))^2) + \frac{1}{2}\beta_{hh}((\ln(x^h_{ht}))^2 - \rho(\ln(x^h_{ht-1}))^2) + \frac{1}{2}\beta_{pp}((\ln(x^p_{ht}))^2 \\
&\quad - \rho(\ln(x^p_{ht-1}))^2) + \beta_{np}(\ln(x^n_{ht})\ln(x^p_{ht}) - \rho\ln(x^n_{ht-1})\ln(x^p_{ht-1})) + \beta_{nh}(\ln(x^n_{ht})\ln(x^h_{ht}) \\
&- \rho\ln(x^n_{ht-1})\ln(x^h_{ht-1})) + \beta_{ph}(\ln(x^p_{ht})\ln(x^h_{ht}) - \rho\ln(x^p_{ht-1})\ln(x^h_{ht-1})) + \sigma_{ht} + \epsilon_{ht} - \rho\epsilon_{ht-1}
\end{aligned}$$

From the ρ-differenced equation, we can derive two sets of moment conditions. First, the difference equation using lagged levels as instruments

$$(9)\quad E[(\sigma_{ht} + \epsilon_{ht} - \rho\epsilon_{ht}) \otimes (\ln(x^n_{ht-s}), \ln(x^p_{ht-s}), \ln(x^h_{ht-s}), Z_{ht}] = 0$$

[34] And second, the levels equation using lagged differences as instruments

$$(10)\quad E[(\omega_{ht} + \epsilon_{ht}) \otimes (\Delta\ln(x^n_{ht-1}), \Delta\ln(x^p_{ht-1}), \Delta\ln(x^h_{ht-1}), Z_{ht}] = 0$$

[34] Equivalently:

$$\begin{aligned}
&E[(ln(Q_{ht}) - \rho ln(Q_{ht-1}) - \beta_n(\ln(x^n_{ht}) - \rho\ln(x^n_{ht-1})) - \beta_h(\ln(x^h_{ht}) - \rho\ln(x^h_{ht-1})) - \\
&\quad \beta_p(\ln(x^p_{ht}) - \rho\ln(x^p_{ht-1})) - \frac{1}{2}\beta_{nn}((\ln(x^n_{ht}))^2 - \rho(\ln(x^n_{ht-1}))^2) - \frac{1}{2}\beta_{hh}((\ln(x^h_{ht}))^2 - \rho(\ln(x^h_{ht-1}))^2) - \\
&\quad \frac{1}{2}\beta_{pp}((\ln(x^p_{ht}))^2 - \rho(\ln(x^p_{ht-1}))^2) - \beta_{np}(\ln(x^n_{ht})\ln(x^p_{ht}) - \rho\ln(x^n_{ht-1})\ln(x^p_{ht-1})) - \\
&\quad \beta_{nh}(\ln(x^n_{ht})\ln(x^h_{ht}) - \rho\ln(x^n_{ht-1})\ln(x^h_{ht-1})) - \beta_{ph}(\ln(x^p_{ht})\ln(x^h_{ht}) - \rho\ln(x^p_{ht-1})\ln(x^h_{ht-1})) \otimes \\
&\qquad (\ln(x^n_{ht-s}), \ln(x^p_{ht-s}), \ln(x^h_{ht-s}), Z_{ht}] = 0
\end{aligned}$$

[35] The moment conditions in Equation (9) are used in Difference GMM estimation (Arellano and Bond, 1991) and in Equations (9) and (10) are used in System GMM estimation (Blundell and Bond, 1998). The parameter estimates are shown in Table 6.

Table 6: Production Model Robustness - Dynamic Panel GMM Estimation

	Arellano-Bond	Blundell-Bond	
	(1)	(2)	(3)
β_n	0.012	0.032	0.022
	(0.004)	(0.002)	(0.002)
β_p	-0.048	-0.067	-0.061
	(0.002)	(0.001)	(0.001)
β_h	-0.106	-0.006	–
	(0.007)	(0.005)	–
β_{nn}	-0.044	-0.078	-0.068
	(0.005)	(0.003)	(0.003)
β_{pp}	0.026	0.019	0.016
	(0.001)	(0.001)	(0.001)
β_{hh}	0.005	-0.088	-0.026
	(0.014)	(0.007)	(0.011)
β_{np}	0.020	0.040	0.038
	(0.002)	(0.001)	(0.001)
β_{nh}	0.053	-0.030	-0.026
	(0.004)	(0.003)	(0.001)
β_{ph}	0.012	0.013	0.011
	(0.002)	(0.002)	(0.001)
ρ	0.831	0.754	0.756
	0.004	0.001	0.001
N	2,496	2,496	2,496

Notes: This table reports the production function estimates under alternate assumptions on the productivity process and estimated using the dynamic panel. Column 3 shows the preferred specification.

[35] Equivalently:

$$E[(\ln Q_{ht} - \beta_n \ln(x^n_{ht}) - \beta_h \ln(x^h_{ht}) - \beta_p \ln(x^p_{ht}) - \frac{1}{2}\beta_{nn}(\ln(x^n_{ht}))^2 - \frac{1}{2}\beta_{hh}(\ln(x^h_{ht}))^2 - \frac{1}{2}\beta_{pp}(\ln(x^p_{ht}))^2 - \beta_{np}\ln(x^n_{ht})\ln(x^p_{ht}) - \beta_{nh}\ln(x^n_{ht})\ln(x^h_{ht}) - \beta_{ph}\ln(x^p_{ht})\ln(x^h_{ht})) \otimes (\Delta\ln(x^n_{ht-1}), \Delta\ln(x^p_{ht-1}), \Delta\ln(x^h_{ht-1}), Z_{ht}] = 0$$

Table 7: Elasticities of Substitution - Nurses and Physicians (Dynamic Panel GMM Estimation)

	Percentiles of Distribution				
Nurses	10th	25th	50th	75th	90th
1	0.22	0.31	0.36	0.38	0.40
	[0.19, 0.24]	[0.29, 0.32]	[0.34, 0.38]	[0.36, 0.40]	[0.38, 0.42]
1.5	0.07	0.19	0.25	0.28	0.30
	[0.02, 0.10]	[0.16, 0.21]	[0.24, 0.27]	[0.26, 0.30]	[0.28, 0.32]
2	–0.30	0.01	0.13	0.18	0.21
	[–0.39, –0.21]	[–0.00, 0.03]	[0.12, 0.15]	[0.17, 0.19]	[0.21, 0.23]
2.5	–0.58	–0.12	0.06	0.62	1.01
	[–0.62, –0.53]	[–0.13, –0.10]	[0.05, 0.06]	[0.55, 0.65]	[0.94, 1.08]
3	–0.40	–0.00	0.22	0.42	0.62
	[–0.66, –0.13]	[–0.02, 0.01]	[0.20, 0.24]	[0.40, 0.44]	[0.59, 0.66]
3.5	–0.13	–0.07	0.08	0.26	0.42
	[–0.17, –0.10]	[–0.10, –0.04]	[0.05, 0.11]	[0.23, 0.29]	[0.38, 0.46]

Notes: In this table, I present the percentiles of the distribution of elasticities of substitution for each hospital-year in my sample using the dynamic panel estimates in Column (3) of Table 6. The results in this table confirm the high complementarity and that the complementarity increases with nurse staffing levels.