Physicians' Occupational Licensing and the Quantity-Quality Trade-off[*]

Juan Pablo Atal   Tomás Larroucau   Pablo Muñoz   Cristóbal Otero

January 31, 2025

PRELIMINARY AND INCOMPLETE
PLEASE DO NOT CIRCULATE

Abstract: Occupational licensing is a quality regulation that ensures service quality but potentially reduces supply. This paper presents a framework to empirically examine this trade-off in the context of health, where quality and access to care are critical concerns. Leveraging rich administrative data and quasi-exogenous variation driven primarily by recent migration patterns of foreign physicians to Chile, we show that both the quantity and quality of physicians are essential inputs in the production of health. Counterfactual simulations, grounded in our theoretical framework and empirical findings, indicate that relaxing licensing policies under current labor market conditions would improve patient outcomes.

1 Introduction

The shortage of physicians is a long-standing and increasingly urgent concern. Providing adequate healthcare access is estimated to require at least an additional 7 million physicians globally, and a deficit of 125 thousand physicians is projected by 2034 for the United States alone (Haakenstad et al., 2022; Association of American Medical Colleges, 2024).

Licensing requirements in the medical profession have been routinely signaled as a key contributor to physician shortages with unclear benefits from increased quality.[1] However, assessing the quality-quantity trade-off embedded in licensing requirements has proven difficult; it requires credible estimates of its effects on the quantity and quality of physicians, and on the resulting impacts to access to healthcare and the quality of care. At the same time, a growing migration of physicians worldwide has made licensing an increasingly relevant and binding policy, potentially changing the nature and magnitude of its associated trade-offs.[2]

In this paper, we provide a novel framework to derive the optimal licensing policy and highlight the key economic fundamentals that govern the quantity-quality trade-off. The theory reveals a set of sufficient statistics needed to evaluate the net benefits of changing the licensing stringency. Informed by the theory, we then provide an empirical framework to evaluate the optimality of the licensing policy currently used to appoint public sector physicians in Chile.

Chile provides an ideal setting to study physician licensing and how its design may depend on changing labor market fundamentals like increased migration. The number of physicians taking the licensing exam increased fivefold between 2009 and 2018, largely fueled by an unprecedented increase in the number of foreign test-takers. In tandem with this surge, passing rates decreased from 92 percent to 58 percent during this period. Second, the Chilean licensing exam has characteristics and purposes similar to those applied in different countries, including the US. Passing the exam is mandatory for working in the public sector and results in the automatic revalidation of a foreign medical degree.[3] Finally, Chile has rich administrative data on patient outcomes, hospital staffing, and physician performance in the licensing exam.

Our theoretical framework highlights the quantity-quality trade-off embedded in the optimal licens-

---

[1] An highly influential criticism to licensing in the medical profession can be found in Friedman (1962). Svorny (2004) provides a more recent literature review.

[2] In the United States, the number of physicians born and educated abroad has increased by 30 percent since 2004 and currently represents 20 percent of the total count (Association of American Medical Colleges, 2023). Similarly, in OECD countries, an increasingly growing share of physicians are foreign-trained physicians and now constitute 30 percent of the physician workforce (Socha-Dietrich and Dumont, 2021). Still, it is estimated that only one-third of employed foreign physicians are on track to practice medicine and that licensing plays a critical role in explaining this fact (Federal Reserve Bank of Minneapolis, 2022).

[3] Similar tests are the USMLE in the US, the NCE in France, and the NMLE in Japan.

ing problem. There is a production function for hospital care, where the quantity and quality of physicians are key inputs. The government has a licensing technology that provides a noisy signal of a physician's quality and grants licenses to those with a signal above a threshold. Changing the licensing threshold affects outcomes depending on the output elasticites with respect to the quantity and quality of labor, and on the elasticity of those inputs with respect to the licensing threshold. These elasticities, in turn, depend on the distribution of quality in the population of test-takers and the precision of the licensing technology (i.e., signal-to-noise ratio). Generally, a higher output elasticity with respect to quantity, a lower output elasticity with respect to quantity, and a lower signal-to-noise ratio decrease the optimal licensing threshold.

In the empirical part of the paper, we begin by estimating physicians' latent quality with a model of scores and retaking behavior. The model enables us to leverage physicians' full test score histories to infer the precision of the licensing test, which we allow to differ by physician's origin. We find that the test has a relatively high signal-to-noise ratio, implying that scores are informative proxies for physicians' latent qualities and that licensing policies can indeed affect the quality distribution across hospitals.

We proceed by estimating the output elasticities with respect to the quantity and quality of physicians. We focus on access and quality of care as key outcomes that jointly determine the value added of the healthcare sector. We measure access primarily by the number of patients seen relative to potential patients (or "service rate") and quality of care by patient mortality. To deal with the common identification challenges related to the endogeneity of inputs in the estimation of production functions, we use an instrumental variables (IV) approach. Specifically, we leverage the increase in the labor supply of physicians brought by the immigration wave towards Chile and also by the expansion of medical schools in the country using a shift-share IV design (Altonji and Card, 1989; Autor et al., 2013; Borusyak et al., 2022).

Our estimates reveal that access to care increases significantly with the number of physicians, with an estimated elasticity of service rate of 1. On the other hand, access does not depend significantly on the quality of physicians. Both the quantity and quality of physicians are important for quality of care. Specifically, holding quality of physicians constant, increasing the number of physicians by 1% decreases the in-hospital death rate by 0.8%; and holding quantity constant, a one point (0.23 SD) increase in the average physicians' quality decreases death rates by 0.04%. We show that this result is not driven by changes in patient composition (i.e., admission of healthier patients) and is robust to an alternative measure of hospital performance that includes out-of-hospital deaths that happened within 28 days after discharge. Consistent with the impacts on mortality, we also find

3

that increasing the number of physicians or their quality leads to a decrease in the percentage of inpatient infections and complications.

To identify the elasticity of quantity and quality of labor with respect to the licensing threshold, we leverage a micro-founded labor matching function between physicians and hospitals. These matching functions depend on physicians' and hospitals' preferences and, therefore, we approximate them as hospital-specific functions of physicians' latent quality and other observables. To accommodate the competitive effects from the changes in the labor supply that occur under counterfactual thresholds, these matching functions also depend on the inflow of physicians to the system.

Equipped with estimates of the sufficient statistics derived from our model, we perform a set counterfactual exercises to evaluate the net benefits of marginally lowering the licensing threshold in the Chilean context. We begin by assessing the short-term effects of lowering the threshold by 2018, at the end of our sample period and 5 years after the onset of the large migration wave. The find that lowering the threshold would have increased access with minimal detrimental effects on quality. On net, population health increases as a consequence. We then investigate how the short-run effects differ across the different phases of the migration wave. Overall, we find a robust positive short-run effect of the policy throughout our sample period, despite the large changes in market fundamentals produced by the migration wave. Still, the policy effects are non-monotonic over time; the largest returns would have occurred in the middle of the migration wave, when the number of marginal physicians was large and the stock of physicians already hired was still relatively low. We finish by estimating the long-run effects of the policy, considering that retaking can dampen the effects of licensing thresholds as scores may improve over attempts. In fact, in our context, more than 80 percent of test-takers who failed their first attempt in 2013 passed by 2018. As a consequence, we find that the effects of lowering the threshold decrease with time. However, there are still net benefits in 2018 of permanently lowering the threshold in 2013.

Our paper contributes to several strands of the literature. We add to a long-standing debate about the desirability of occupational licensing in healthcare (Friedman and Kuznets, 1945; Friedman, 1962; Kleiner, 2014, 2016; Kleiner et al., 2016). Relative to this literature, we provide a framework for understanding the quantity-quality trade-off embedded in licensing and empirically evaluating the stringency of licensing policies in relevant health outcomes. More broadly, our paper adds to the research that studies occupational licensing in other settings, particularly in the public sector (Kleiner and Wang, 2023; Angrist and Guryan, 2008; Kleiner, 2011; Larsen et al., 2020). Close to our work, Kleiner and Soltas (2023) provides a general equilibrium framework to estimate the welfare effect of licensing policies in the US under competitive product and labor markets. In

contrast to their work, we use direct quality measures to derive the social value of licensing instead of relying on wages and equilibrium conditions in the labor market linking them to quality.[4]

We also contribute to the literature studying whether physician quantity matters for health outcomes. Previous work has shown a correlation between location mortality and the number of physicians (Finkelstein et al., 2021) or infant health and the number of primary care physicians (Carrillo and Feres, 2019). Instead, we use quasi-exogenous variation to show that physician quantity matters for health outcomes in tertiary care. Moreover, we complement previous research studying whether physician quality matters. Examples in the literature include the impacts of elite medical training (Doyle et al., 2010), medical school exit exams (Guarin et al., 2021), and physicians' value-added (Fletcher et al., 2014; Ginja et al., 2024). However, to our knowledge, there is no empirical evidence showing that licensing scores correlate with health outcomes. Finally, we also contribute to a scant literature estimating production functions in healthcare (Gaynor et al., 2015; Grieco and McDevitt, 2017). Relative to this literature, we make progress by exploiting exogenous variation in inputs to estimate the elasticity of service rates and quality of care with respect to the quantity and quality of physicians.

2 Background

2.1 Institutional Setting

The Chilean Public Healthcare System: The healthcare system in Chile is divided into public and private insurers and providers. Our focus is on public providers, which mainly serve patients with public insurance. Public insurance covers approximately 80% of the population and is financed through monthly contributions deducted from labor income, cost-sharing mechanisms, and resources from the general government.[5]

The public health insurance provides coverage within the public healthcare system, with varying levels of co-payment determined by the patient's income and family size. Individuals who cannot afford to pay are granted free access to the public system, ensuring nearly universal health coverage in public hospitals. Beneficiaries of the public insurer may also opt for private providers, although the co-payment for such "out-of-network" services is significantly higher than in those in the public

---

[4]Our data and framework allow us to work in a labor-matching market where wages do not clear the market.

[5]Individuals also have the option to redirect their health contributions toward purchasing private insurance, which covers 15% of the population. The remainder 5% is covered through schemes exclusively for the police and armed forces. In practice, public insurance primarily serves the relatively more disadvantaged population, while wealthier, healthier, and younger individuals tend to opt for private insurance (Pardo, 2019).

network.[6]

The network of public hospitals is comprised of 181 hospitals that belong to one of 29 different referral regions called Health Services (*Servicios de Salud*).[7] The referral and counter-referral system also operates at this level. Individuals must register with their local primary healthcare center, and those requiring specialized care are referred to one of the public hospitals within their region. Referrals follow strict and predetermined guidelines based on the location of the primary care center and the patient's diagnosis and demographics.[8] Hospitals are categorized into three levels of complexity, low, medium, and high, depending on their size and the range of medical services they offer.

Physician Licensing: To work in the healthcare sector, physicians are required to pass the EU-NACOM exam (*Examen Único Nacional de Conocimientos de Medicina*).[9] The performance in the EUNACOM is also determinant for accessing training opportunities, as it plays a key role in admissions to residency programs in the country. Importantly, approving the EUNACOM automatically validates the medical degrees of foreign-trained physicians, granting them the same job and training opportunities as locally trained physicians.[10]

The exam was established in 2009 and follows the characteristics and purposes of medical licensing exams in other countries.[11] It comprises a theoretical section and a practical section. The theoretical section corresponds to a multiple-choice test with 180 questions covering various areas of medical knowledge. The score reflects the candidate's absolute performance and is not standardized relative to other test-takers.[12] To pass, candidates must achieve a score of 51 or higher out of 100 points.

Upon passing the theoretical section, the candidates must complete a pass-or-fail practical section, which involves an examination in a real or simulated clinical environment. However, this require-

---

[6]This explains why only around one-fourth of inpatients in private hospitals use public insurance coverage.

[7]The number of 181 hospitals corresponds to hospitals that are present every year throughout our study period 2011-2019.

[8]Patients can also be admitted directly to hospitals through the ER in cases of emergency.

[9]Although the exam is not technically mandatory for physicians in private healthcare organizations, it is effectively binding across all healthcare institutions because passing it is legally required to treat patients covered by public health insurance, regardless of the treatment location. This explains why, *de facto*, most private healthcare institutions only consider applications admissible if candidates include EUNACOM approval.

[10]Countries that allow all prospective physicians hoping to work in a given jurisdiction to take a licensing exam include Canada, Hong Kong, Japan, Korea, the United Arab Emirates, and the United States (Archer et al., 2017).

[11]Comparable exams include the Medical Licensing Examination (USMLE) in the U.S., the Medical Council of Canada Qualifying Examination (MCCQE) in Canada, the National Competency Exam (NCE) in France, the National Medical Licensing Examination (NMLE) in Japan, and the Korean Medical Licensing Examination (KMLE) in Korea. For an in-depth discussion of the exam and evidence of its validity and reliability, see Mena (2021).

[12]A key objective of the EUNACOM designers is to make the difficulty of the exam comparable across years. Empirically, we observe that although the scores are not standardized, the distribution of scores at local universities is very similar across years.

ment is waived for candidates with medical degrees from local universities and for physicians with a medical degree from a select group of countries with which the Ministry of Health has bilateral agreements.[13]. The practical portion of the exam is largely not binding (Kunakov et al., 2018).

Physician's Hiring and Wages in the Public Sector: Physicians can work in both the private and public sectors. In the private sector, employment operates under standard market dynamics, where wages, benefits, and working conditions are negotiated directly between employers and employees. In contrast, in public hospitals, wages are legally regulated and follow a public-sector wage schedule, with annual adjustments based on sector-wide wage revisions.[14] Hiring in public hospitals is decentralized and managed by the regional Health Services within the budgetary constraints established by the Ministry of Health. Health Services determine staffing needs for hospitals, and hospital directors oversee the recruitment process. Although passing the EUNACOM exam is sufficient for eligibility, hospitals reportedly consider the candidate's EUNACOM score an important factor in hiring decisions.

2.2 Changes in Physician's Labor Market Fundamentals

In line with similar trends in other OECD countries, the number of physicians per capita in Chile has increased in the last two decades, driven by a growing number of domestic graduates and greater reliance on foreign-trained physicians (OECD, 2019). Figure 1-A shows that the number of new physicians enrolled in the National Registry of Healthcare Providers rose from around 1,000 in 2001 to over 4,000 in 2019. Figure 1-B shows the number of physicians working in public hospitals and their hourly wages from 2011 to 2019. During this period, the influx of newly registered physicians led to a 60% increase in full-time physicians in public hospitals. Despite this significant growth in the workforce, hourly wages remained flat after accounting for sector-wide remuneration adjustments. This wage rigidity aligns with established wage-setting policies, even as the supply of physicians has increased substantially.

Only one-third of the increase in newly registered physicians shown in Panel A is attributed to the growing number of locally trained physicians, driven by the establishment of new medical schools and the expansion of existing programs. The remaining two-thirds of the increase is explained

---

[13]Argentina, Brazil, Colombia, Ecuador, Spain, the United Kingdom, and Uruguay

[14]This is similar to the pay scales for physicians in the NHS, which are determined nationally through negotiations between government bodies and medical unions.

by an unprecedented inflow of foreign-trained physicians into the system.[15][16] Panels C and D of Figure 1 show that the increase in physicians is strongly reflected in the number of EUNACOM test-takers and their migration status. In 2013, the number of foreign-trained physicians taking the exam was relatively small, with most failing to meet the minimum required score to validate their degrees. By 2018, the number of foreign-trained physicians taking the EUNACOM exam had increased nearly tenfold, and foreign physicians outnumbered local graduates taking the exam by 2.5 to 1. Not only did the mass of foreign-trained physicians increase, but the distribution also shifted to the right, indicating that many were able to validate their medical degrees.[17]

In spite of the large migration influx, physicians remain scarce. As of 2019, Chile had 17.5 physicians per 10,000 inhabitants—nearly half the average of countries with a comparable burden of diseases, injuries, and risk factors (33.4 physicians per 10,000). This figure also falls below the minimum threshold of 20.7 physicians per 10,000 needed to achieve an effective Universal Healthcare coverage score of 80 out of 100 at the global level proposed by Haakenstad et al. (2022). The shortage of physicians is also reflected in the current long waiting lists, with approximately 3 million individuals—15% of the population—awaiting medical attention. An estimated 40,000 individuals die annually while on waiting lists, representing 25% of the country's annual mortality.

3 Data

In this section, we describe the data sources used in the empirical analysis, discuss the construction of aggregate (hospital-year-level) health outcomes, and present summary statistics.

---

[15]While immigrants represented 2% of the population in 2011, this figure rose to 8% by 2022 (INE, 2024), largely driven by the mass migration of Venezuelans to Chile and other countries in the region, which began in late 2015 following the collapse of the Venezuelan economy. The Venezuelan migration to Chile and other countries in the region has been leveraged as a shock to explore the impacts of migration on various outcomes, including crime (Ajzenman et al., 2023), labor markets (Lebow, 2022; Olivieri et al., 2022; Bahar et al., 2024), and discrimination (Groeger et al., 2024), among others.

[16]Differences in remuneration and non-wage amenities are well-documented "push" and "pull" factors in physician migration (OECD, 2019). Various pieces of evidence highlight Chile as an attractive destination for medical professionals in Latin America, driven by significant wage differentials and better working conditions. Physicians in Chile can earn up to eight times more than their counterparts in Argentina, fueling temporary migration among Argentine physicians (Castro, 2024). Similarly, professional insecurity, low salaries, and limited social recognition have prompted physicians to emigrate from Venezuela (Hernández and Ortiz Gómez, 2011).

[17]The share of physicians in Chile who are foreign-trained increased from 15% to 24% between 2014 and 2019, and is currently similar to that of Canada and the US, and higher than the OECD average (OECD, 2015, 2019). Importantly, we do not find evidence of an increase in the share of foreign healthcare workers beyond physicians. We discuss this in greater detail in Appendix B.

Figure 1: Physician's Labor Market

A. Registered Physicians

B. Physician's Wages in Public Hospitals

C. Scores in 2013

D. Scores in 2018
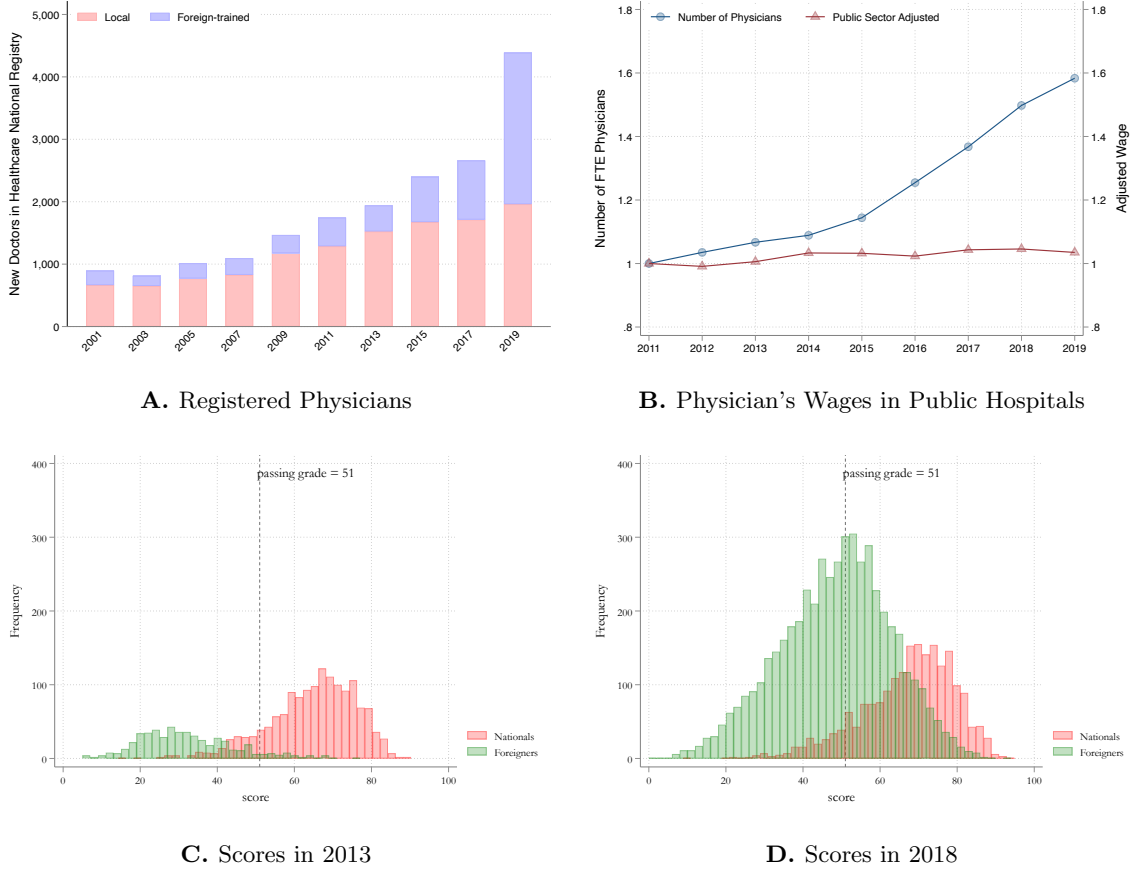
Notes: Panel A shows the number of newly registered physicians in the National Registry of Healthcare Providers between 2001 and 2019. This registry includes all healthcare workers legally allowed to practice in Chile. The bars are divided into locally-trained physicians (light red) and foreign-trained physicians (light blue). Panel B plots the growth of full-time equivalent physicians in public hospitals on the left y-axis and their average wages on the right y-axis, both indexed to 2011 (set equal to 1). Data come from SIRH (2019). Panels C and D display the distribution of test scores for locally-trained (red) and foreign-trained (green) physicians in 2013 and 2018, respectively. The dashed line in both graphs represents the minimum passing score.

3.1 Data Sources

Employer-employee Data for Public Hospitals: We use matched employer-employee administrative records managed by the Ministry of Health. The records cover the universe of physicians employed in all public hospitals in Chile between 2011 and 2019 (SIRH, 2019). These data, called SIRH (*Sistema de Información de Recursos Humanos*), consolidates information from various public health organizations into a unified registry and offers comprehensive data for payroll processing and workforce management and include detailed payroll records, wage information, and employment information. We complement these data with hospital-level characteristics, including size, location and referral area, among others (DEIS, 2020).

National Registry of Public Healthcare Providers: We use a registry of all physicians (and other healthcare professionals) legally authorized to practice in Chile (RNPI, 2024). The registry is managed by Chile's Superintendency of Health (*Superintendencia de Salud*), and provides detailed information on registered healthcare professionals, including their nationality and details of their professional degree including title, date of issuance, and the name and country of the granting institution. For degrees obtained abroad, the registry includes the date of revalidation in Chile.

Licensing Exam Scores: We use confidential score records for all physicians who have taken the national licensing exam, EUNACOM. These data were provided directly by ASOFAMECH (*Asociación de Facultades de Medicina de Chile*), the organization responsible for administering and overseeing the exam (ASOFAMECH, 2019). The dataset includes the date and scores of the theoretical portion of the EUNACOM for every attempt made by the universe of physicians who have taken the exam.

Individual-level Discharge Data: We measure outcomes using administrative records of individual-level inpatient events in all public hospitals in Chile from 2011 to 2019 (DEIS, 2019). The data include diagnoses (ICD-10 codes), discharge or death dates, and patient characteristics such as birth date, gender, residence, and health insurance type. We link these data at the individual level with the universe of death records processed by the Vital Records Office, which includes deaths outside public hospitals.[18]

Waiting Lists: We access individual-level administrative records of waiting lists for surgical procedures and specialist consultations for non-prioritized conditions (SIGTE, 2019), obtained through a Freedom of Information Act (FOIA) request.[19] In Chile, primary care physicians refer patients to hospitals for specialist consultations or surgical procedures, placing them on a waitlist. We observe each patient's entry and exit dates, along with the assigned hospital.

3.2 Data Aggregation and Outcomes

We construct a hospital-by-year panel dataset using all the sources of information described above. We assess hospital performance in terms of access and quality. Our primary measure of access is the yearly hospital service rate, defined as the number of admissions in a given year divided by the eligible population in the hospital's referral region in the corresponding year. We also use surgery rates (inpatient surgeries as a proportion of admissions) and exits from the waitlist as additional

---

[18]We only have access for these records until 2018.

[19]Waitlists are divided into prioritized and non-prioritized conditions. The "Explicit Healthcare Guarantees" program, established by law, prioritizes specific diseases with evidence-based procedures and timelines for diagnosis and treatment. See Menares and Muñoz (2023) for details.

access measures in some robustness exercises.

Our primary measure of quality is the hospital's average death rate, defined as the total count of deaths in a given year divided by the total number of admissions in the corresponding year. We complement this in-hospital death rate with the death rate resulting from counting deaths within 28 days of a patient's admission—regardless of place of death—following Gaynor et al. (2013). In robustness exercises, we replace death rates with complication rates. We construct complication rates as the number of inpatients with with an ICD-10 code related to infections, hemorrhage, or other complications, divided by the number of admissions.

Since EUNACOM was introduced in 2009, we impute scores for physicians who were licensed prior to that.[20] Our imputation procedure assigns scores to these physicians by combining two components: the average score of all physicians at hospital h (the grand mean) and the differential score of physicians from the same region of origin working at hospital h (the group mean). To estimate this differential score, we model licensing scores as a function of region-of-origin fixed effects at the hospital level, imposing sum-to-zero constraints on the fixed effects. We then adjust these estimates using empirical Bayes (Efron and Morris, 1973; Walters, 2024), such that group differentials are shrunk towards zero as estimates' precision decreases. In Appendix A we present additional details on our imputation procedure and conduct robustness checks of the main results under alternative imputation methods.

3.3 Descriptive Stats

Appendix Table A.1 provides descriptive statistics of the data used in the main analysis. Panel A summarizes key patient and hospital characteristics. Patients are predominantly female (57%), and 30% are under the age of 29. As expected, most (97%) of patients in public hospitals are covered by public insurance. The in-hospital death rate averages 3.28%, while the 28-days death rate is higher, at 5.07%. Hospitals serve an average of 5,656 patients annually, corresponding to an average service rate of 2%. The average length of stay is 4.03 days, and complication rates average 11.41%. Hospitals perform an average of 2,018 surgeries annually. They employ on average around 80 full-time equivalent physicians.

Panel B describes the evolution of test scores across multiple years. The number of test takers increased steadily over the period, from 1,389 in 2009 to 7,121 in 2018. Average test scores declined

---

[20]Before the introduction of EUNACOM, Chile implemented a voluntary National Medical Examination (EMN), administered in Chilean medical schools from 2003 to 2008. Prior to the EMN, licensing requirements varied: local medical graduates were required to pass the Medical Surgeon Degree Examination, while foreign-trained physicians had to complete a Foreign Medical Qualification Revalidation Examination.

sharply between 2009 and 2013, dropping from 71.8 to 56.1 and then mostly stabilized to reach 53.9 in 2018. The proportion of test takers achieving a passing score (≥ 51) had an equivalent trend, falling from 92% in 2009 to 58% in 2018. Notably, the number of tests falling within the [40–51] range increased significantly over the years, from 87 in 2009 to 1,552 in 2018, reflecting a growing mass of individuals that could validate their title under a slightly lower passing score.

4 Simple Licensing Problem

We begin with a simple theoretical framework to study optimal licensing policies. We focus on a particular margin, namely the stringency of the requirement as determined by the passing score of the licensing exam.

Consider a licensing technology based on a noisy signal s of quality θ. The output of interest (e.g. hospital production) is a function of labor $L(\underline{s})$ and average quality $\bar{\theta}(\underline{s})$:

$$Y\left(\underline{s}\right) = F\left(L(\underline{s}), \bar{\theta}(\underline{s})\right).$$

The signal s follows a distribution with density $h(s)$ and total mass m. The labor function is given by $L(\underline{s}) = m \int_{\underline{s}}^{\infty} h(s)ds$, which is decreasing in the threshold. The quality index is positively related to the cutoff $\bar{\theta}(\underline{s})' > 0$, generating a quality-quantity trade-off.

A simple parametrization allows us to gain further insights. Let the quality index to be equal to average quality, $\bar{\theta}(\underline{s}) \equiv \mathbb{E}[\bar{\theta}|s > \underline{s}]$, and the production function to have a Cobb-Douglas form $F\left(\bar{\theta}(\underline{s}), L(\underline{s})\right) = L^{\alpha_L} \cdot \exp(\bar{\theta})^{\alpha_{\bar{\theta}}}$. The elasticity of output with respect to the licensing threshold $\underline{s}$ is given by:

$$\frac{1}{Y}\frac{\partial Y}{\partial \underline{s}} = \underbrace{\frac{-m \cdot h(\underline{s})}{L}}_{\eta_L} \cdot \left(\alpha_L - \alpha_{\bar{\theta}} \cdot \underbrace{\left(\mathbb{E}[\bar{\theta}|s > \underline{s}] - \theta(\underline{s})\right)}_{-\tilde{\eta}_{\bar{\theta}}/\eta_L}\right) \tag{1}$$

The first term, $\frac{-m \cdot h(\underline{s})}{L}$, is the elasticity of labor with respect to the threshold, which depends on the mass of workers at the licensing threshold. The second term is the net effect of the marginal physician on health production. The quantity effect of a marginal physician is α_L. However, the inframarginal physician decreases average quality by $\left(\mathbb{E}[\bar{\theta}|s > \underline{s}] - \theta(\underline{s})\right)$, which, multiplied by the return on quality, gives the quality effect of the marginal physician.

The effect of lowering the threshold on average quality depends on the precision of the signal. We

consider a licensing technology based on a noisy test $s = \theta + \epsilon$, with $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\theta \sim N(\mu_\theta, \sigma_\theta^2)$. Defining the signal-to-noise ratio as $\text{SNR} \equiv \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\epsilon^2}$, the elasticity of the output with respect to the licensing threshold is given by:

$$\eta_{Y,\underline{s}} \equiv \frac{1}{Y} \frac{\partial Y}{\partial \underline{s}} = \frac{-m \cdot h(\underline{s})}{L} \cdot (\alpha_L - \alpha_{\bar{\theta}} \cdot SNR \cdot (\mathbb{E}[s|s > \underline{s}] - \underline{s})) \tag{2}$$

Equation 2 highlights the main empirical objects needed to quantify the elasticity of the outcome with respect to the licensing threshold: The net effect of the marginal physician depends on the returns to quantity and quality, the signal-to-noise ratio, and the difference between the average score and the marginal score.

We note that this net effect is positive as long as $\frac{\alpha_L}{\alpha_{\bar{\theta}}} > \text{SNR} \cdot (\mathbb{E}[s|s > \underline{s}] - \underline{s})$. Lowering the licensing threshold increases the output as long as the return to quantity over the return to quality is higher than a threshold. This threshold increases with the signal-to-noise ratio, as a more (less) precise signal implies higher (negative) impacts of quality from lowering the threshold. The threshold also increases with the difference between the average score and the marginal score, which measures the impact of lowering the threshold on the average score.

The optimal threshold takes a simple expression:

$$\underline{s}^* = \mathbb{E}[s|s > \underline{s}^*] - \frac{\alpha_L}{\alpha_{\bar{\theta}}} \cdot \frac{1}{\text{SNR}}.$$

The optimal threshold increases with the mean score. A higher score allows the planner to have more quantity without giving up quality. The optimal threshold decreases with the ratio of the return to quantity over the return to quality. It is also lower when the licensing technology is imprecise, as in such cases a higher threshold does not allow to differentiate between high and low-quality physicians.

Objective Function Let patient i's health in the absence of treatment be equal to m_i and let $\Delta m_i(\bar{\theta}, L)$ be the value added of treatment for individual i (e.g., the decrease in mortality following treatment). Denoting $\mathcal{I}(L, \bar{\theta})$ the set of patients treated, we can write population health as:

$$H(L, \theta) = \int_i m_i di + \underbrace{|\mathcal{I}(\bar{\theta}, L)|}_{\text{service rate}} \times \underbrace{\frac{1}{|\mathcal{I}(L, \bar{\theta})|} \int_{i \in \mathcal{I}(L, \bar{\theta})} \Delta m_i(\bar{\theta}, L) di}_{\text{per-patient treatment value added}} \tag{3}$$

$$= H^0 + Y(L, \bar{\theta}) \tag{4}$$

13

In our empirical analysis below, our main outcome of interest is the effect of the healthcare sector on population mortality, such that $Y(L, \bar{\theta})$ is the impact of the healthcare system in reducing mortality. Therefore, our main measure of per-patient treatment value added is the *decrease* in patient's mortality risk. We can approximate the elasticity of healthcare's impact on mortality with respect to the licensing threshold as the difference between the elasticity of the service rate and the elasticity of per-patient mortality:

$$\eta_{\bar{s}}^Y \simeq \eta_{\bar{s}}^{\text{service rate}} - \eta_{\bar{s}}^{\text{mortality}}. \tag{5}$$

The licensing threshold affects the mortality impact of the healthcare sector in two dimensions. First, it affects access through the service rate. Second, it affects quality by changing patients' mortality. Below, we describe a general framework to estimate the production function for both outcomes as a function of quantity and quality.[21] In the following section, we specify the empirical counterparts of the model used to quantify the elasticity of healthcare's mortality impact with respect to the licensing threshold, $\eta_{\bar{s}}^Y$.

5 Empirical Model

We begin with a model of score determination, with two main goals. First, we aim to infer physicians' latent quality based on the observed vector of scores, which is a key input in the production function. Second, a model of scores allows us to compute score histories under counterfactual policies to evaluate the long-run effects of lowering the threshold. Next, we use an instrumental variables approach to estimate production functions to recover the returns to quantity and quality. Finally, we develop a labor matching model and estimate its empirical analog to determine how the mass of marginal physicians at the licensing threshold matches with each hospital, taking into account potential equilibrium effects of increased crowdout. Taken together, these estimates enable us to measure how sensitive are hospital-specific health outcomes to changes in the licensing threshold.

Scores history determination   Each physician i belongs to type $\tau(i) \in \{N, F\}$ and is characterized by her latent quality θ_i. Physicians can retake the exam and change their test-taking ability over attempts.[22] We denote physician i's test-taking ability in attempt n by Γ_{in}. The score in

---

[21]Formally, $\eta_{\bar{s}}^Y = \eta_{\bar{s}}^{\text{service rate}} - \eta_{\bar{s}}^{\text{mortality}} - \eta_{\bar{s}}^{\text{service rate}} \eta_{\bar{s}}^{\text{mortality}}$. The approximation corresponds to ignoring the cross term. Empirically we will find that $\eta_{\bar{s}}^{\text{mortality}} \simeq 0$.

[22]We assume that quality is constant over time which implies that physicians cannot experience quality gains over attempts. To support this assumption, in Appendix C, we use regression discontinuity designs to show that physicians

attempt n, s_{in}, is a noisy measure of a physician's quality and test-taking ability, with

$$s_{in} = \theta_i + \Gamma_{in} + \epsilon_{in} \tag{6}$$

We assume that test-taking ability improves with exponential decay, such that

$$\Gamma_{in} = \sum_{k=0}^{n_i-1} \gamma_{\tau(i)} \cdot \exp(-\rho_{\tau(i)} \cdot k) \quad n_i \geq 1, \tag{7}$$

where γ and ρ are parameters to be estimated.[23]

Physicians who fail the exam in attempt n, retake it with a probability that is a function of the distance between their average past score and the passing threshold $\bar{s}_{in} - \underline{s}$, the number of attempts, n, physician's type, $\tau(i)$, and the licensing threshold, $\underline{s}$:

$$P(\text{retake}|\bar{s}_{in}, n_i, \tau(i)) = \frac{e^{\beta_{0,\tau(i)} + \beta_{n,\tau(i)} n + \beta_{s,\tau(i)}(\bar{s}_{in} - \underline{s})}}{1 + e^{\beta_{0,\tau(i)} + \beta_{n,\tau(i)} n_i + \beta_{s,\tau(i)}(\bar{s}_{in} - \underline{s})}}. \tag{8}$$

By allowing the retaking probability to depend on both the distance between an individual's average score and the licensing exam threshold, and the number of attempts (a sufficient statistic for predicting future scores and posterior quality), this specification captures the net benefits of retaking. These net benefits arise from two factors: (i) the probability of passing the threshold in subsequent attempts (determined by the gap between expected future scores and the threshold), and (ii) the expected value of the match conditional on passing the exam (which depends on physicians' type and posterior quality). In addition, explicitly incorporating the licensing threshold in the specification ensures that retaking probabilities can vary under counterfactual scenarios.

Since θ and ϵ are normally distributed, the posterior of quality for each physician given type-specific prior means, variances, and their sequences of scores over attempts is equal to:

$$\mathbb{E}[\theta_i \mid s_{i0}, s_{i1}, \ldots, s_{in}] = \mu_{\theta,\tau(i)} + \frac{\sigma^2_{\theta,\tau(i)}}{\sigma^2_{\varepsilon,\tau(i)} + (n+1)\sigma^2_{\theta,\tau(i)}} \left(\sum_{t=0}^{n} (s_{it} \underbrace{-\Gamma_{t,\tau(i)}}_{\text{de-trending}} -\mu_{\theta,\tau(i)}) \right), \tag{9}$$

with

$$\theta_i = \mathbb{E}(\theta_i|s_i) + \nu_i. \tag{10}$$

who retake the exam around the passing threshold do not experience gains on quality proxies.

[23]Appendix Figure A.1 shows that test-taking gains decay over attempts.

15

Equations 9 and 10 are the basis for inferring average quality at the hospital level from its physician types and their vector of scores. Average quality is a key input in the production function, as we specify below.

Production Function We consider a set of public hospitals indexed by $j \in \mathcal{J}$. In period t, each hospital produces an outcome k that we denote by y_{jt}^k, by combining a quantity of physicians L_{jt} with quality index $Q(F_{\theta_{jt}})$, and other determinants of hospital productivity such as capital, patient case-mix, and others collectively represented by A_{jt}. We assume the production function is a Cobb-Douglas, such that:

$$y_{jt} = A_{jt} L_{jt}^{\alpha_L} exp(Q\left(F_{\theta_{jt}}\right))^{\alpha_\theta}, \tag{11}$$

where $Q(F_{\theta_{jt}}) = \int q(\theta) f_{\theta_{jt}}(\theta) \, d\theta$ and $q(\cdot)$ is a *nondecreasing* function of θ. Intuitively, $Q(F_{\theta_{jt}})$ captures how the distribution of quality at hospital j in period t. For our empirical application, we adopt $q(\theta) = \theta$, which implies $Q(\theta_{jt})$ is simply the (unconditional) mean of the quality distribution in hospital j in period t.[24]

Labor Market Matching. Physicians who pass the exam match with hospitals in a decentralized labor market. The main challenge in modeling this matching process is to capture how it changes in counterfactual scenarios with different licensing thresholds. In particular, relaxing the licensing threshold increases the number of (lower-quality) physicians seeking jobs, which may impact the equilibrium matching probabilities due to competition in the labor market. To capture this dimension of the problem, we specify matching probabilities that depend on the inflow of physicians who pass the exam.

Let m_t be the number of test takers in period t and $f_t(s)$ be the distribution of scores in period t. The mass of physicians who pass the exam in year t and have expected quality within a range $r_{\hat{\theta}} \in \mathcal{R}$ is:

$$M_{r_{\hat{\theta}},t} \equiv m_t \int_{s:\ \hat{\theta}(s) \in r_{\hat{\theta}}} f_t(s) ds, \tag{12}$$

Denote by $\kappa_t = \{\kappa_{1t}, ..., \kappa_{|\mathcal{J}|t}\}$, the vector of posted vacancies for physicians in period t. Conditional

---

[24]Reassuringly, as discussed in section 6.2, we obtain qualitatively similar results under an alternative specification that defines the quality index as the share of physicians in the hospital j during period t whose quality falls below the median of the overall distribution. Section 6.2 also discusses the impacts of physician quantity and quality on health outcomes when using a Translog production function that allows for a linear interaction between physicians' quantity and quality.

on passing the licensing exam in period t, each physician $i \in \mathcal{I}$ matches with a hospital $j \in \mathcal{J}$, i.e., $i \in \mu_t(j)$, or opts for their outside option, $\emptyset$, based on a conditional matching probability function denoted by $CMP(\cdot)$, with:

$$CMP_{ijt} = \frac{e^{v(x_{ijt}) + g(M_t, \kappa_t | \tau, r_{\hat{\theta}}(i), j)}}{1 + \sum_{j'} e^{v(x_{ij't}) + g(M_t, \kappa_t | \tau, r_{\hat{\theta}}(i), j')}} \tag{13}$$

where $v(x_{ijt})$ captures the determinants of matching probabilities based on physician i's and hospital j's preferences, such as physician's type, $\tau(i)$, expected quality, $\hat{\theta}_i$, and the distance between hospital j's location and physician i's location and $g(\cdot)$ is a function that captures potential equilibrium effects driven by labor demand and supply channels. We allow the general equilibrium effects for hospital $j \in \mathcal{J}$ to depend on the physician's quality range, their type, the vector of masses of test takers by quality range, and the vector of hospitals' capacities. For our empirical application, we specify this function as a simple linear function of market tightness:

$$g(M_t, \kappa_t | \tau, r_{\hat{\theta}}(i), j) = \boldsymbol{\beta}_j \frac{M_{it}}{\kappa_{jt}} \tag{14}$$

where $M_{it} = [M_{it}^0, M_{it}^+]$ is a vector composed by the mass of physicians in i's quality range, $M_{it}^0 = M_{\tau(i), r_{\hat{\theta}}(i), t}$, and the mass of physicians above i's quality range, $M_{it}^+ = \sum_{r_{\hat{\theta}'} \succ r_{\hat{\theta}}(i)} M_{\tau(i), r_{\hat{\theta}'}, t}$. $\boldsymbol{\beta}_j$ is a vector of hospital-specific coefficients. Our specification of the CMPs captures key determinants of physicians' and hospitals' preferences, as well as market-level conditions that can affect sorting patterns and vary in the counterfactuals.[25] In Appendix F, we generalize and characterize how equilibrium effects can affect our counterfactual exercises, and provide an identification strategy to account for them by exploiting the variations driven by the migration wave.[26]

Labor Supply and Quality. For a physician $i \in \mathcal{I}$ with type $\tau \in \mathcal{T}$, exam score s, and a matrix of physician–hospital characteristics $X \in \mathbb{R}^{K \times |\mathcal{J}|}$ (whose jth column is the vector $\boldsymbol{x}_{ijt} \in \mathbb{R}^K$ for

---

[25]We do not specify the extent to which these conditional matching probabilities are separately influenced by the preferences of physicians and hospitals or by the labor market conditions. However, in Appendix J.5 we provide a micro-foundation for the current model where matching probabilities arise endogenously and are affected by changes in the mass of physicians due to capacity constraints and sorting.

[26]In particular, we assume that equilibrium effects can affect the conditional matching probabilities of physicians with expected quality range $r_{\hat{\theta}}$ only through changes in the distribution of potential physicians in their own range and "upstream" in the order of expected quality ranges. In addition, we assume that equilibrium effects can be approximated by contiguous chained effects, where a change in the mass of physicians in a specific expected quality range can have a direct effect on the CMPs within that range, and can also induce a displacement effect on the expected quality range that immediately precedes it. This displacement effect influences the mass in that range and consequently affects the CMPs indirectly.

hospital $j \in \mathcal{J}$), in a market characterized by (M_t, κ_t) in period t, define:

$$CMP_j\big(X, s \,|\, \tau;\, M_t;\, \kappa_t\big) \;:=\; CMP_{ijt}\Big(\big[x_{i1t}, \ldots, x_{i|\mathcal{J}|t}\big] = X,\; s_{it} = s,\; \tau(i) = \tau,\; M_t,\; \kappa_t\Big).$$

This CMP_j is the conditional probability that such a physician matches with hospital $j \in \mathcal{J}$ in period t. In addition, assume that conditional on type $\tau \in \mathcal{T}$ and period t, observable characteristics (X, s) follow a joint distribution with density $h_t^\tau(X, s)$. Then, given the conditional matching probabilities and the distribution of physicians of type τ's exam scores in period t, $f_t^\tau(s)$, the labor inflow of type τ in hospital j in period t is given by:

$$\Delta L_{jt}^\tau \;=\; m_{\tau,t} \int_{s \geq \underline{s}} \int_X CMP_j\big(X, s \mid \tau;\, M_t;\, \kappa_t\big)\, h_t^\tau(X, s)\, dX\, ds. \tag{15}$$

In addition, the average quality of physicians inflow of type τ in hospital j at time t is given by:

$$\Delta \bar{\theta}_{jt}^\tau \;=\; \frac{\displaystyle\int_{s \geq \underline{s}} \int_X CMP_j\big(X, s \mid \tau;\, M_t;\, \kappa_t\big)\, \hat{\theta}(s, \tau)\, h_t^\tau(X, s)\, dX\, ds}{\displaystyle\int_{s \geq \underline{s}} \int_X CMP_j\big(X, s \mid \tau;\, M_t;\, \kappa_t\big)\, h_t^\tau(X, s)\, dX\, ds.} \tag{16}$$

The total labor inflow in hospital j at time t is given by

$$\Delta L_{jt} \;=\; \sum_{\tau \in \mathcal{T}} L_{jt}^\tau \qquad \Delta \bar{\theta}_{jt} \;=\; \frac{\displaystyle\sum_{\tau \in \mathcal{T}} \Delta L_{jt}^\tau \cdot \Delta \bar{\theta}_{jt}^\tau}{\Delta L_{jt}}. \tag{17}$$

The Licensing Problem. Consider a social planner who decides the licensing threshold $\underline{s}$ to maximize population health outcomes in a given time regime, R:

$$Y_R \equiv \prod_{j \in \mathcal{J}} y_{jR}^{p_j},$$

where p_j are Pareto weights. The elasticity of Y_R with respect to the licensing threshold is given by:

$$\eta_{Y_R, \underline{s}} = \underbrace{\alpha_L \cdot \bar{\eta}_{L_R, \underline{s}}}_{\text{Licensing Quantity Effect}} + \underbrace{\alpha_\theta \cdot \bar{\eta}_{\bar{\theta}_R, \underline{s}}}_{\text{Licensing Quality Effect}},$$

where $\bar{\eta}_{L_R, \underline{s}} = \sum_{j \in \mathcal{J}} p_j \cdot \eta_{L_{jR}, \underline{s}}$, is the average elasticity of labor L_R with respect to $\underline{s}$ in regime R, and $\bar{\eta}_{\bar{\theta}_{jR}, \underline{s}} = \sum_{j \in \mathcal{J}} p_j \cdot \eta_{\bar{\theta}_{jR}, \underline{s}}$, is the average elasticity of quality $\bar{\theta}_R$ with respect to $\underline{s}$ in regime R.

The previous equation, highlights the *licensing quantity-quality trade-off*. On one side, the *Licensing*

Quantity Effect depends on the return to physician quantity, α_L, and the–per-hospital–elasticities of labor, $\eta_{L_{jR},\underline{s}}$. On the other, the *Licensing Quality Effect* depends on the return to quality, α_θ, and the–per-hospital–elasticities of quality, $\eta_{\bar{\theta}_{jR},\underline{s}}$. Both elasticities are functions of physicians' sorting probabilities across hospitals and the density and mass of each type near the licensing threshold.

We define two time regimes for evaluating this trade-off:

(i) *Short-Run Regime* ($R = SR$): Elasticities $\eta_{L_{jSR},\underline{s}}$ and $\eta_{\bar{\theta}_{jSR},\underline{s}}$ reflect the immediate effects of licensing policy changes.

(ii) *Long-Run Regime* ($R = LR$): Elasticities $\eta_{L_{jLR},\underline{s}}$ and $\eta_{\bar{\theta}_{jLR},\underline{s}}$ incorporate steady-state adjustments, including cumulative effects over time, such as exam retakes.

The regime-specific formulation of Y_R and $\eta_{Y_R,\underline{s}}$ allows us to evaluate counterfactual scenarios and compare immediate and long-term impacts of licensing policy changes.

6   Estimation

6.1   Estimation of Latent Quality

We estimate the retaking model in Equation 8 using maximum likelihood. Table 1 shows the results. As expected, the probability of retaking decreases with the number of attempts and with the distance of the score to the cutoff.

Table 1: Retaking Model Estimates

	Foreign	Nationals
n_{it}	-0.231	-0.163
	(0.020)	(0.059)
$\bar{s} - s_{it}$	-0.036	-0.060
	(0.003)	(0.009)
Constant	2.592	1.595
	(0.077)	(0.139)
N	8,221	1,340

Notes: Coefficients are estimated via maximum likelihood. Standard errors in parentheses.

Using the estimated model of retaking probability, we estimate in a second stage the model of score gains via Simulated Method of Moments (SMM).[27] We match the following moments by type

---

[27]This estimation strategy allows us to model the selection behavior in the retaking process.

(national or foreign): the mean over attempts, the mean of gains over attempts, the covariance between attempts, and the variance of the first attempt. Table 2 shows the estimated coefficients by type. We observe that foreigners have larger variances for the estimated quality distribution and exam noise than nationals. In addition, the estimates for score gains predict for foreigners that exam gains decay relatively quickly at an exponential rate of 0.55. On the other hand, for nationals, score gains are imprecisely estimated due to the lack of data on retaking behavior.

With the estimated quality distribution and score gain parameters, and given the history of score realizations for each physician, we use Equation 9 to estimate individual posterior quality means and compute average posterior quality for every hospital in every period.[28]

Table 2: Score Gains and Latent Quality Estimates

	Natives	Foreigners
$\hat{\mu}$	70.385	49.022
	(0.000)	(0.000)
$\hat{\gamma}$	0.044	12.053
	(0.046)	(0.127)
$\hat{\rho}$	5.808	0.556
	(3.514)	(0.018)
$\hat{\sigma}_\theta$	9.130	14.797
	(0.082)	(0.028)
$\hat{\sigma}_\epsilon$	8.173	9.416
	(0.084)	(0.022)

Notes: Coefficients are estimated via Simulated Method of Moments (SMM). Bootstrapped standard errors are in parentheses, computed using the empirical standard deviation of the estimates across 1000 simulations.

6.2 The Impact of Physicians' Quantity and Quality on Hospital-level Outcomes

Building upon the production function presented in Section 5, and using the posterior in Equation 10 to write average quality as a function of average scores and measurement error, the empirical analog of Equation 11 becomes:

$$\ln(y_{jt}^k) = \alpha_L^k \ln(L_{jt}) + \alpha_\theta^k \frac{1}{L_{jt}} \sum_{i \in J_t} E(\theta_i | \mathbf{s}_i) + \gamma_{f(j)t}^k + \rho_j^k + \beta^k X_{jt} + \underbrace{\omega_{jt}^k + \alpha_\theta^k \frac{1}{L_{jt}} \sum_{i \in J_t} \nu_{it} + \varepsilon_{jt}^k}_{\mu_{it}}. \quad (18)$$

Equation 18 is our estimating equation, where the key parameters of interest are the output-

---

[28]Alternatively, we could use only data of initial scores, and leverage our instruments to correct for the associated measurement error, without the need to specify the model of scores (Agostinelli and Wiswall (2016)). However, specifying the model of scores allows us to perform the long-run counterfactuals, and thus we use it for the estimation of the returns to quality for the sake of parsimony.

specific elasticities to quantity (α_L^k) and quality (α_θ^k). This specification implicitly models logged productivity as being influenced by time-invariant unobservable characteristics, such as hospital location and capital stock, which we capture using hospital fixed effects, ρ_j^k. We further account for time-varying changes in capital by including the number of beds per patient in each HRR in the vector X_{jt}. Additionally, following Propper and Van Reenen (2010) and Gaynor et al. (2013), X_{jt} includes demographic controls to account for changes in patient composition that might influence productivity. Specifically, these case-mix controls include the shares of female inpatients, foreign inpatients, and inpatients across eight age bands (0–29, followed by 10-year increments up to 90+), as well as the shares of inpatients with different insurance types categorized by co-payment levels.

To account for potential differential time trends by hospital complexity, the specification also includes a vector of year-fixed effects that vary with hospital complexity, $\gamma_{f(j)t}^k$. Finally, the error term μ_{it} consists of three components: an unobserved productivity shock potentially known before input choices, ω_{jt}^k (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015); measurement error in quality, $\alpha_\theta^k \frac{1}{L_{jt}} \sum_{i \in J_t} \nu_{it}$; and an unobserved productivity shock that occurs after input decisions, ε_{jt}^k.

To address the problem of measurement error and the identification challenge that physicians' quantity and quality could be correlated with unobserved factors affecting hospital outcomes, we implement a two-stage least squares (2SLS) estimation strategy that uses two shift-share (or Bartik) instruments (Altonji and Card, 1989; Autor et al., 2013) to recover the *causal* impact of physician quantity and quality on hospital-level outcomes. Our instruments leverage the increase in the labor supply of physicians brought about by immigration flows from other countries and, to a lesser extent, by the expansion of medical schools in Chile. The instrument for physician quantity, Z_{jt}^L, is constructed by summing the percentage change in the number of physicians clearing the cutoff of the licensing exam from each region of training c (i.e., the shift component), weighted by the percentage of physicians from that region who worked at hospital j the year before (i.e., the share component). Similarly, the instrument for physician quality, Z_{jt}^θ, sums the change in the average quality of eligible test-takers from each training region c, weighted by the percentage of physicians from that region who worked at hospital j the year before.

Results: Table 3 presents the results from our two-stage least squares (2SLS) estimation, which examines the causal impact of physicians' quantity and quality on different hospital-level outcomes of interest. We present point estimates alongside exposure-robust standard errors clustered at the region of origin (i.e., the shock) level in brackets (Adao et al., 2019; Borusyak et al., 2022),[29] and

---

[29]To obtain these standard errors—which are shown to be asymptotically valid (Adao et al., 2019)—we estimate the transformed regression proposed in Borusyak et al. (2022).

we complement our analysis by showing the Anderson and Rubin (1949) p-values (Lee et al., 2022).

Panel A focuses on access, where our main measure is the hospital service rate. Column (1) reports the results. We find that an elasticity of service rate with respect to the number of physicians of 1. Column (2)-(4) show the impacts on additional measures of access. Column (2) focuses on the number of inpatient surgeries, for the subset of hospitals performing surgeries. We find a surgery elasticity of five. We also find that physicians' quality does not affect the service rate or surgeries. Columns (3) and (4) analyze the impact on the completion of wait-listed surgical procedures and specialist consultations, respectively. Column (3) shows that increasing the number of physicians leads to more exits from the surgical waiting list, with no effect from changes in physician quality. Similarly, Column (4) indicates that a higher number of physicians increases the number of patients exiting the specialist waiting list, while physician quality remains inconsequential.

Overall, our findings suggest that the quantity of physicians significantly influences access, consistent with the notion that physician scarcity is the primary binding constraint in expanding hospital capacity and addressing patient demand (e.g., Carrillo and Feres, 2019). Our results also show that the quality of physicians does not impact a hospital's ability to admit more patients, suggesting that the availability of physicians, rather than their expertise, is the primary determinant of hospital utilization rates.

Panel B of Table 3 focuses on quality, where our main measure is the in-hospital death rate. Column (1) shows the results. We find that: i) holding quality constant, increasing the number of physicians by 1% decreases the in-hospital death rate by 0.8%, and ii) holding quantity constant, a one point increase in the average quality of physicians—which corresponds to a 0.23 standard deviations increase—decreases death rates by 0.04%.

To address the concern that a potential correlation between increases in the number of physicians and the admission of healthier patients might be driving our results (i.e., inframarginal, previously non-admitted patients might be healthier), Column (2) shows the impact on the "risk-adjusted mortality rate".[30] Reassuringly, neither physician quantity nor quality significantly affects predicted mortality rates, suggesting that changes in patient composition are not driving our findings.

Columns (3) and (4) show the results for additional quality metrics. As shown by Column (3), the main result of Column (1) is robust to using the 28-day death rate as an outcome instead of in-hospital mortality. Finally, Column (4) focuses on the in-hospital complications rate. Con-

---

[30]We calculate the risk-adjusted mortality rate by fitting a logit model for death outcomes at the inpatient level, controlling for patient demographics and diagnoses group, as per the enhanced Elixhauser comorbidity index (Elixhauser et al., 1998; Quan et al., 2005). The total number of *predicted deaths* at each hospital and year is then divided by the number of admissions.

sistent with the impacts on mortality, we find that: i) holding quality constant, increasing the number of physicians by 1% decreases the in-hospital complications rate by 1%, and ii) holding quantity constant, a one-point increase in the average quality of physicians decreases the in-hospital complications rate by 0.05%.[31]

Identification using shift-share instruments is predicated upon the assumption that either the "shifts" or the "shares" components are as good as random and not correlated with factors that would affect the outcomes of interest (Goldsmith-Pinkham et al., 2020; Borusyak et al., 2022). Identification "from the shifts" can be understood as leveraging a shift-level natural experiment, while identification "from the shares" can be viewed as pooling together multiple difference-in-differences designs leveraging heterogeneous shock exposure (Borusyak et al., 2024). In Appendix E, we present evidence in favor of the exogeneity of shifts and shares components. First, we show that the quantity and quality shocks do not predict predetermined variables related to hospitals' workforce and patients' demographics, as it should be the case if shocks are as-good-as-randomly assigned. Second, to assuage concerns related to potential endogenous mechanisms affecting the composition of physicians from different origins within hospitals and hospital outcomes simultaneously, we show that differential exposures to common shocks (the "shares") are not correlated with changes in our outcomes of interest.

As additional robustness checks, in Appendix D we examine the impact of physicians' quantity and quality on various hospital-level outcomes using variants of our preferred model. Specifically, we consider an alternative quality index and a Translog production function. We obtain qualitatively similar results when the quality index is defined as the share of physicians in the hospital j during period t whose quality falls below the median of the overall distribution (instead of the average quality of the physicians working in the hospital j during period t). The effects of physician quantity and quality on health outcomes remain qualitatively similar using a Translog production function, although the first stage is weaker in this specification.[32]

---

[31]Several mechanisms could explain why quantity and quality affect health outcomes. On the one hand, more physicians can facilitate better patient monitoring and quicker responses to health complications, both of which can improve patient outcomes and reduce mortality. More physicians also mean that inpatients are less likely to experience delays in receiving necessary treatments, which is particularly critical in life-threatening situations. On the other hand, physicians of higher quality may possess greater skill in accurately diagnosing and treating complex cases, making timely decisions, and effectively applying evidence-based practices. This expertise can decrease complications and enhance patient survival by reducing misdiagnosis, treatment errors, and adverse health outcomes.

[32]In this case, we include the interaction between physicians' quantity and quality as an additional endogenous variable and the interaction of our shift-share instruments as an additional instrumental variable.

Table 3: Impact of Physicians' Quantity and Quality on Access and Quality

Panel A: Access

	Ln service rate	Ln inpatient surgeries	Ln exits from waiting list	
			Surgical	Medical
	(1)	(2)	(3)	(4)
Ln Physicians ($\hat{\alpha}_L^{\text{service}}$)	1.01	4.97	3.69	3.00
	[0.25]	[1.99]	[0.71]	[1.04]
Avg. Physicians' Quality ($\hat{\alpha}_\theta^{\text{service}}$)	0.01	0.11	-0.00	0.02
	[0.02]	[0.10]	[0.04]	[0.06]
Observations	1,402	744	738	942
Mean Dep. Var.	0.015	3,803	1,534	8,403
F-stat (First-stage)	21.97	12.16	9.85	15.83
Anderson-Rubin (χ^2)	0.000	0.000	0.000	0.000

Panel B: Quality

	Mortality			In-hospital
	In-Hospital		28-days	Complications
	Ln death rate	Pred. death rate	Ln death rate	Ln complications rate
	(1)	(2)	(3)	(4)
Ln Physicians ($\hat{\alpha}_L^{\text{mortality}}$)	-0.83	0.12	-0.74	-0.98
	[0.19]	[0.05]	[0.21]	[0.19]
Avg. Physicians' Quality ($\hat{\alpha}_\theta^{\text{mortality}}$)	-0.04	-0.00	-0.04	-0.05
	[0.01]	[0.00]	[0.01]	[0.01]
Observations	1,402	1,402	1,402	1,373
Mean Dep. var.	3.28	3.51	5.08	11.65
F-stat (First-stage)	21.97	34.50	21.97	21.15
Anderson-Rubin (χ^2) p-value	0.00	0.03	0.00	0.00

Notes: This table presents the impact of the quantity and quality of physicians on public hospital performance. Panel A focuses on utilization, which we proxy through the service rate, inpatient surgeries, and exits from the waiting list. Panel B focuses on patients' mortality and complications. Estimates come from the two-stage least squares estimation of Equation 18. We present the exposure-robust standard errors clustered at the region of origin (i.e., the shock) level in brackets (Adao et al., 2019; Borusyak et al., 2022).

6.3 Matching Probabilities

We estimate the conditional matching probabilities by maximum likelihood following Equations 13 and 14. We specify:

$$v_{ijt} = \alpha^d \text{Distance}_{ij} + \sum_j \alpha_j^f \text{Foreign}_i + \alpha_j^q \theta_i + \alpha_j^{fq} \theta_i \times \text{Foreign}_i + \alpha_j^s \text{Specialist}_i. \tag{19}$$

We proxy the number of vacancies κ_{jt} in Equation 14 with the ratio of beds and the stock of physicians in the previous period, that is, $\tilde{\kappa}_{jt} = \frac{\text{Beds}_{jt}}{\text{Stock of Physicians}_{j,t-1}}$. We construct quality ranges to construct the vector M_{it} using four quartiles. Finally, we let the vector of coefficients $\boldsymbol{\beta}_j$ to vary across three tiers of hospital quality; which we construct based on the average quality of their physicians.

The results are presented in Table 4, where columns (1)-(3) vary how we incorporate the labor market tightness on the matching probabilities. Column (1) shows a specification where we specify the function $geq(X)$ as depending only on the "direct mass". Column (2) allows the function $geq()$ to depend on the "direct mass" as well as the "above mass". Column (3) allows the effects of the "direct" and "above" mass to vary across different hospital-quality tiers. As a placebo check, in column (4) we show that the "below mass" does not affect matching probabilities.

Table 4: Conditional Matching Probabilities Estimates

	Alternative Models			Placebo
	(1)	(2)	(3)	(4)
Distance$_{ij}$	-0.267	-0.267	-0.267	-0.266
	(0.011)	(0.011)	(0.011)	(0.011)
$(M_{it}^0)/\kappa_{jt}$	-0.636	-0.678	-0.726	
	(0.152)	(0.159)	(0.186)	
$(M_{it}^+)/\kappa_{jt}$		0.018	0.052	
		(0.021)	(0.038)	
$(M_{it}^0)/\kappa_{jt} \times \mathbb{1}[r_j = 2]$			0.004	
			(0.097)	
$(M_{it}^0)/\kappa_{jt} \times \mathbb{1}[r_j = 3]$			0.112	
			(0.122)	
$(M_{it}^+)/\kappa_{jt} \times \mathbb{1}[r_j = 2]$			-0.039	
			(0.045)	
$(M_{it}^+)/\kappa_{jt} \times \mathbb{1}[r_j = 3]$			-0.067	
			(0.056)	
$(M_{it}^-)/\kappa_{jt}$				0.021
				(0.019)
Log likelihood	-15,286.97	-15,286.61	-15,285.41	-15,295.38

Notes: The table shows results from a multinomial logit model for the matching probabilities between physicians and hospital referral regions. The functions $\mathbb{1}[r_j = k]$ indicate the hospital's quality tier. As shown in Equation 19, all specifications include alternative-specific coefficients for the following variables: Foreign indicator, physician quality, and interaction between Foreing indicator and quality, and a Specialist indicator. We display only the estimates for the coefficients that do not depend on the alternative.

7 Counterfactuals

7.1 Short-Run Impacts of Lowering the Licensing Threshold

Two key objects of interest for evaluating the effects of changing the licensing threshold are the elasticity of the quantity and quality of physicians in each hospital with respect to the licensing threshold. We compute these elasticities numerically, by simulating the match under the baseline threshold and under a counterfactual threshold $\underline{s}' = \underline{s} - \Delta\underline{s}$; and the resulting quality and quantity in using the expressions in Equation 17.[33]

Figure 2 summarizes our main results for the short-run elasticies in 2018. For ease of interpretation, we plot the *negative of* the elasticities to quantify the impact of a counterfactual *decrease* in the licensing threshold. The size of the marker is proportional to the number of patients in each hospital. Blue markers denote hospitals with the number of physicians per patient below the median, whereas

---

[33]We provide details on how we perform this calculation in Appendix K.

grey markers denote hospitals with a number of physicians per patient above the median.

Figure 2: Short-run Elasticity of Quantity and Semi-elasticity of Quality by Hospital in 2018
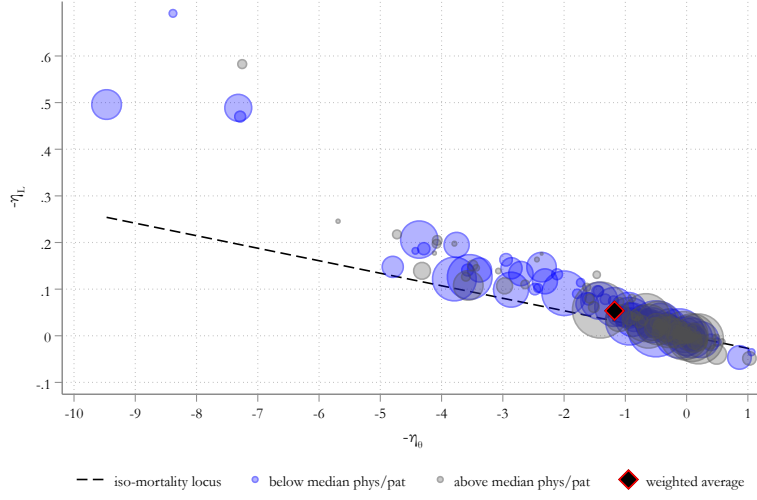
Notes: This figure shows a scatter plot with the estimates for the (negative of) the elasticity of quantity and the semi-elasticity of quality with respect to the licensing threshold. Each dot represents a hospital. The side of the dot is proportional to the number of patients. The dashed line represents the combination of elasticities above with mortality decreases when the threshold is reduced.

There is generally a quantity-quality trade-off of lowering the licensing threshold as $-\eta_{L_j,\underline{s}} > 0$ and $-\eta_{\bar{\theta}_j,\underline{s}} < 0$, although there is ample heterogeneity across hospitals in the magnitudes of the elasticities.

To evaluate the overall effects of lowering the licensing threshold, we contrast these elasticities with the quantity and quality output elasticities noting that, from Equation 5 the elasticity of healthcare's impact with respect to the licensing threshold for hospital j can is equal to:

$$\eta_{Y_j,\underline{s}} = (\alpha_L^{\text{service}} - \alpha_L^{\text{mortality}}) \cdot \eta_{L_j,\underline{s}} + (\alpha_{\bar{\theta}}^{\text{service}} - \alpha_{\bar{\theta}}^{\text{mortality}}) \cdot \tilde{\eta}_{\bar{\theta}_j,\underline{s}}$$

The dashed line in Figure 2 corresponds to the "iso-mortality curve", depicting the minimum quantity elasticity $\eta_{L,\underline{s}}^*(\eta_{\bar{\theta},\underline{s}})$ such that the elasticity of healthcare's value added is non-negative for a given quality semi-elasticity $\eta_{\bar{\theta},\underline{s}}$.[34] For most hospitals we find that $-\eta_{L,\underline{s}} > -\eta_{L,\underline{s}}^*(\eta_{\bar{\theta},\underline{s}})$, which means that the increase in quantity more-than-compensates the decrease in quality when the threshold is reduced. The solid diamond shows the weighted averages of the quantity and quality elasticities; which also lies above the iso-mortality locus. On net, we find that healthcare's

---

[34]That is, the iso-mortality curve is defined by the equation $\eta_{L,\underline{s}}^*(\eta_{\bar{\theta},\underline{s}}) \equiv -\frac{(\alpha_{\bar{\theta}}^{\text{service}} - \alpha_{\bar{\theta}}^{\text{mortality}}) \cdot \tilde{\eta}_{\bar{\theta}_j,\underline{s}}}{(\alpha_L^{\text{service}} - \alpha_L^{\text{mortality}})}$.

value added would increase in the short run with a policy that marginally decreases the licensing threshold.

Time series evolution of short-run impacts   Given the large changes in market fundamentals, mostly resulting from the migration wage, we turn to analyze whether the policy effects found for 2018 —5 years after the onset of the migration wave— would differ from those in earlier years.
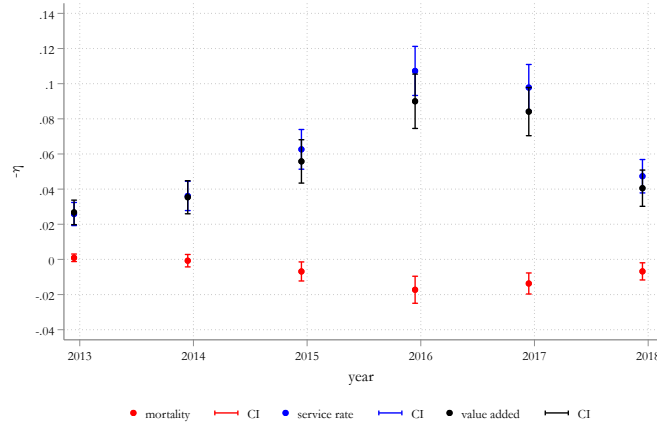
Figure 3-A shows the evolution of the short-run elasticities of quantity and quality over time. For the quantity elasticity (in red), we find an inverse-U pattern. The quantity elasticity is the smallest in 2013 when there are few physicians at the margin that would only pass in the counterfactual scenario with a lower threshold. Over time, migration increases the number of marginal physicians, which increases the quantity elasticity. However, after 2016, the elasticity decreases. Even if the number of marginal physicians increases throughout the sample period, more physicians were already hired when the later cohorts arrived. This decreases the elasticity mechanically as the baseline is higher, but also through the lower capacity in the public sector which increases the share of physicians matching with the outside option. Conversely and by the same arguments, the quality semi-elasticity has a U shape.

Figure 3-B shows the resulting elasticities of mortality, service rate, and healthcare value added, that result from the counterfactual exercises in all years in our sample. The elasticity of mortality stays mostly flat over time, although it becomes slightly negative in 2016. Overall, per-patient mortality increases due to the quality reductions outweighing the quantity gains of lowering the threshold. However, the increase in the service rate more than compensates for this effect, and overall value added increases when the threshold is reduced.

Figure 3: Time Series of Short-Run Elasticities
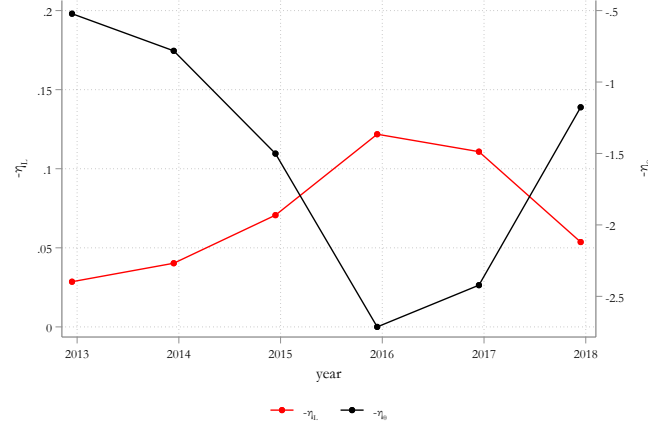
A. Labor Quantity and Quality

B. Mortality, Access and Value Added

Notes: Panel A shows the evolution of the average short-run elasticities of quantity and quality of physicians with respect to the licensing threshold. Panel B shows the resulting average effects on patient's mortality, access, and over value added of the healthcare sector.

7.2 Long-Run Impacts of Lowering the Licensing Threshold

We now turn to analyze the dynamic effects of permanently changing the licensing threshold. The key mechanism we stress in this section is that retaking mitigates the relevance of the licensing threshold over time. For example, in our sample, 83% of test-takers who fail their first attempt in 2013 pass by 2018.

We quantify the dynamic effects of changing the threshold by simulating individual histories for each cohort of test-takers —defined by the year they first take the exam— using the estimated model of scores and retaking from section 5. We use this simulated model to compute yearly

29

elasticities with respect to the threshold that we set permanently lower in 2013.

Table 5 shows the simulated passing rates for cohort 2013, under the status-quo threshold ($\underline{s} = 51$), as well under a counterfactual policy with a threshold set permanently lower at $\underline{s}' = 41$. At baseline, 86% of test takers pass in their first attempt (in 2013), compared to 94% under the counterfactual policy. The 8% gap in passing rates determines the short-run elasticities. However, the magnitude of the gap in passing rates across both scenarios shrinks over time. By 2018, almost 94.7% of physicians from the 2013 cohort pass under the status quo, compared to 98.5% under the counterfactual.

Table 5: Simulated Passing Year for 2013 Cohort

year	$\underline{s} = 51$		$\underline{s} = 41$	
	passing rate	cumulative passing rate	passing rate	cumulative passing rate
2013	86.0	86.0	94.0	94.0
2014	6.8	92.8	3.5	97.5
2015	1.4	94.2	0.7	98.2
2016	0.3	94.6	0.2	98.4
2017	0.1	94.6	0.0	98.4
2018	0.1	94.7	0.1	98.5

Notes: The table shows the simulated passing rates over time for the cohort of physicians that take the exam for the first time in 2013, under a (baseline) scenario where the threshold is kept at 51 points, and under a counterfactual scenario where the threshold is reduced to 41 points.

The ratio between the simulated short-run and long-run elasticities is shown in Figure 4. By definition, the short and long-run elasticities coincide for 2013 when the policy is implemented. For the following years, the long-run elasticities are smaller, as some of the marginal physicians who contribute to the short-run elasticities in any year $t > 2013$ are retakers who pass in a year $t' < t$ when the threshold is set permanently lower in 2013. By 2018, five years after the policy is implemented, the long-run elasticities are approximately 20% of the short-run elasticities. However, lowering the threshold in 2013 still generates positive outcomes by 2018.

Figure 4: Ratio of Simulated Short- and Long-Run Elasticities
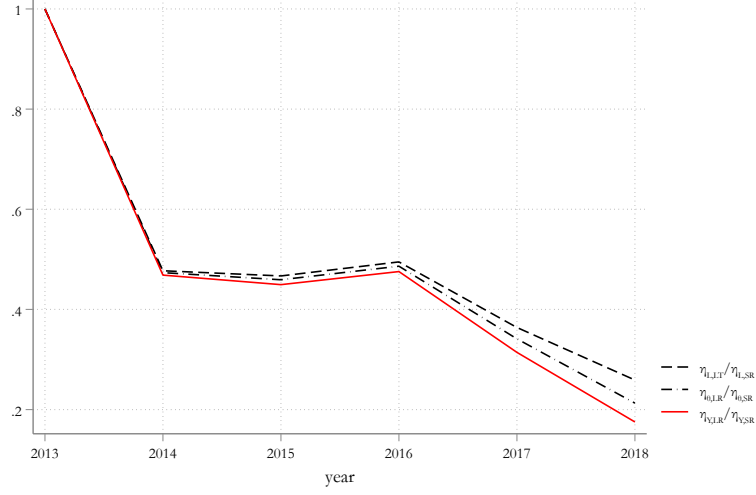
Notes: The figure shows the ratio between the short- and long-run elasticities of mortality, service rate, and value added.

8 Conclusion

Occupational licensing is a widely used regulation in the labor market to ensure a minimum quality standard. However, it comes at the cost of reducing labor supply. This quantity-quality trade-off is particularly relevant in healthcare, where licensing is ubiquitous and could have first-order welfare implications. In this paper, we first analytically characterize this trade-off in an optimal licensing problem. We then propose and estimate an empirical model of physician licensing that allows us to quantify the quantity-quality trade-off in the context of physician licensing in Chile.

Our estimation is built on three parts. First, we infer the latent quality of physicians using individual histories of exam scores. We then aggregate qualities at the hospital level to provide novel estimates of the elasticity of health outcomes with respect to the quantity and quality of physicians. Key to our analysis, we leverage exogenous shocks to these inputs coming from a large migration wave of physicians to Chile. Finally, we introduce a micro-founded labor matching function to estimate the sorting of physicians into hospitals, which allows us to estimate the quantity and quality elasticities of labor with respect to the licensing threshold. These elasticities, paired with the elasticities of health outcomes with respect to the quantity and quality of physicians, allow us to estimate the elasticity of outcomes with respect to the licensing threshold.

Our results indicate that both the quantity and quality of physicians affect health outcomes. As

a consequence, lowering the licensing threshold generally entails a quantity-quality trade-off. Still, on net, population health would improve in the short run. Notably, this result is robust to the large changes in the labor market fundamentals coming from the migration wave. We also investigate the long-run policy effects taking into account exam retaking. While retaking dampens the net effects of lowering the threshold over time, it does not mute it even 5 years after the policy is implemented.

Other policies can also be effective tools to address physician shortages. Expanding nurses' scope of practice, which involves similar trade-offs between increasing access and potentially affecting quality, may be preferable to lowering licensing thresholds for physicians. Additionally, improving access to medical education can help increase the supply of high-quality physicians. Comparing these alternative policies with the reduction in licensing thresholds presents an important avenue for future research.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*(6), 2411–2451.

Adao, R., M. Kolesár, and E. Morales (2019). Shift-share designs: Theory and inference. *The Quarterly Journal of Economics 134*(4), 1949–2010.

Agostinelli, F. and M. Wiswall (2016). Estimating the technology of children's skill formation. Technical report, National Bureau of Economic Research.

Ajzenman, N., P. Dominguez, and R. Undurraga (2023). Immigration, crime, and crime (mis) perceptions. *American Economic Journal: Applied Economics 15*(4), 142–176.

Altonji, J. G. and D. Card (1989, September). The Effects of Immigration on the Labor Market Outcomes of Natives. Working Paper 3123, National Bureau of Economic Research.

Anderson, T. W. and H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of mathematical statistics 20*(1), 46–63.

Angrist, J. D. and J. Guryan (2008). Does Teacher Testing Raise Teacher Quality? Evidence from State Certification Requirements. *Economics of Education Review 27*(5), 483–503.

Archer, J., N. Lynn, L. Coombes, M. Roberts, T. Gale, and S. Regan de Bere (2017). The medical licensing examination debate. *Regulation & Governance 11*(3), 315–322.

ASOFAMECH (2009–2019). EUNACOM. https://www.eunacom.cl. Asociación de Facultades de Medicina de Chile (ASOFAMECH).

Association of American Medical Colleges (2023). 1 in 5 U.S. Physicians Was Born and Educated Abroad: Who Are They and What Do They Contribute? Accessed: 2024-07-30.

Association of American Medical Colleges (2024). The Complexities of Physician Supply and Demand: Projections From 2021 to 2036. Technical report, AAMC, Washington, DC.

Autor, D. H., D. Dorn, and G. H. Hanson (2013). The China Syndrome: Local Labor Market Effects of Import Competition in the United States. *American economic review 103*(6), 2121–2168.

Azevedo, E. M. and J. D. Leshno (2016). A supply and demand framework for two-sided matching markets. *Journal of Political Economy 124*(5), 1235–1268.

Bahar, D., I. Di Tella, and A. Gulek (2024). "formal effects of informal labor and work permits: Evidence from venezuelan refugees in colombia". Working paper.

Borusyak, K., P. Hull, and X. Jaravel (2022). Quasi-Experimental Shift-Share Research Designs. *The Review of economic studies 89*(1), 181–213.

Borusyak, K., P. Hull, and X. Jaravel (2024). A practical guide to shift-share instruments. Technical report, National Bureau of Economic Research.

Borusyak, K. and X. Jaravel (2017). Revisiting event study designs. *Available at SSRN 2826228*.

Carrillo, B. and J. Feres (2019). Provider Supply, Utilization, and Infant Health: Evidence from a Physician Distribution Policy. *American Economic Journal: Economic Policy 11*(3), 156–96.

Castro, M. (2024). Salarios hasta ocho veces más altos: el nuevo fenómeno de los "médicos gaviota" argentinos en chile. *El País*.

DEIS (2019). Egresos Hospitalarios. Available at http://www.deis.cl/estadisticas-egresoshospitalarios/. Accessed: 2020-02-09.

DEIS (2020). Listado de Establecimientos de Salud. Available at https://deis.minsal.cl/#datosabiertos. Accessed: 2020-02-09.

Doyle, J. J., S. M. Ewer, and T. H. Wagner (2010). Returns to Physician Human Capital: Evidence from Patients Randomized to Physician Teams. *Journal of Health Economics 29*(6), 866–882.

Efron, B. and C. Morris (1973). Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association 68*(341), 117–130.

Elixhauser, A., C. Steiner, D. R. Harris, and R. M. Coffey (1998). Comorbidity measures for use with administrative data. *Medical care*, 8–27.

Fack, G., J. Grenet, and Y. He (2019). Beyond truth-telling: Preference estimation with centralized school choice and college admissions. *American Economic Review 109*(4), 1486–1529.

Federal Reserve Bank of Minneapolis (2022). Occupational licensing requirements can limit employment options for immigrants. Accessed: 2024-08-02.

Finkelstein, A., M. Gentzkow, and H. Williams (2021). Place-Based Drivers of Mortality: Evidence from Migration. *American Economic Review 111*(8), 2697–2735.

Fletcher, J. M., L. I. Horwitz, and E. Bradley (2014). Estimating the Value Added of Attending Physicians on Patient Outcomes. Working Paper 20534, National Bureau of Economic Research.

Friedman, M. (1962). *Capitalism and Freedom*. University of Chicago Press.

Friedman, M. and S. Kuznets (1945). *Income from Independent Professional Practice*. NBER.

Gaynor, M., R. Moreno-Serra, and C. Propper (2013, November). Death by market power: Reform, competition, and patient outcomes in the national health service. *American Economic Journal: Economic Policy 5*(4), 134–66.

Gaynor, M. S., S. A. Kleiner, and W. B. Vogt (2015). Analysis of hospital production: An output index approach. *Journal of Applied Econometrics 30*(3), 398–421.

Gilraine, M. and J. Penney (2023, 02). Focused Interventions and Test Score Fade-Out. *The Review of Economics and Statistics*, 1–27.

Ginja, R., J. Riise, B. Willage, and A. Willén (2024). Does Your Doctor Matter? Doctor Quality and Patient Outcomes. *Journal of Political Economy Microeconomics*.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020). Bartik instruments: What, when, why, and how. *American Economic Review 110*(8), 2586–2624.

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

Groeger, A., G. León-Ciliotta, and S. Stillman (2024). Immigration, Labor markets and Discrimination: Evidence from the Venezuelan Exodus in Perú. *World Development 174*, 106437.

Guarin, A., C. Posso, E. Saravia, and J. Tamayo (2021). The Luck of the Draw: The Causal Effect of Physicians on Birth Outcomes. Technical report.

Haakenstad, A., C. M. S. Irvine, M. Knight, C. Bintz, A. Y. Aravkin, P. Zheng, V. Gupta, M. R. Abrigo, A. I. Abushouk, O. M. Adebayo, et al. (2022). Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from 1990 to 2019: a systematic analysis for the global burden of disease study 2019. *The Lancet 399*(10341), 2129–2154.

Haakenstad, A., C. M. S. Irvine, M. Knight, C. Bintz, A. Y. Aravkin, P. Zheng, V. Gupta, M. R. M. Abrigo, A. I. Abushouk, O. M. Adebayo, G. . Agarwal, and R. Lozano (2022). Measuring the Availability of Human Resources for Health and Its Relationship to Universal Health Coverage for 204 Countries and Territories from 1990 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *The Lancet 399*(10341), 2129–2154.

Hernández, T. and Y. Ortiz Gómez (2011). La migración de médicos en venezuela. *Revista Panamericana de Salud Pública 30*(2), 177–181.

INE (2024). Estimación de la población extranjera residente en Chile 2022: Resultados Instituto Nacional de Estadísticas. Accessed: December 11, 2024.

Kleiner, M. M. (2011). Enhancing Quality or Restricting Competition: The Case of Licensing Public School Teachers. *Journal of Law and Public Policy 5*(2), 1–15.

Kleiner, M. M. (2014). Occupational Licensing in Health Care. In A. Culyer (Ed.), *Encyclopedia of Health Economics*. Elsevier.

Kleiner, M. M. (2016). Battling over Jobs: Occupational Licensing in Health Care. *American Economic Review 106*(5), 165–70.

Kleiner, M. M., A. Marier, K. W. Park, and C. Wing (2016, 5). Relaxing Occupational Licensing Requirements: Analyzing Wages and Prices for a Medical Service. *The Journal of Law and Economics 59*(2), 261–291.

Kleiner, M. M. and E. J. Soltas (2023, 02). A Welfare Analysis of Occupational Licensing in U.S. States. *The Review of Economic Studies 90*(5), 2481–2516.

Kleiner, M. M. and W. Wang (2023). The Labor Market Effects of Occupational Licensing in the Public Sector. Working Paper 31213, National Bureau of Economic Research.

Kunakov, N., L. Moraga, and L. Ortiz (2018). Revalidación de títulos médicos extranjeros: eficacia y eficiencia de un examen colaborativo y estandarizado. *Revista médica de Chile 146*(2), 232–240.

Larsen, B., Z. Ju, A. Kapor, and C. Yu (2020). The Effect of Occupational Licensing Stringency on the Teacher Quality Distribution. Working Paper 28158, National Bureau of Economic Research.

Lebow, J. (2022). The labor market effects of venezuelan migration to colombia: Reconciling conflicting results. *IZA Journal of Development and Migration 13*(1), 1–49.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022). Valid t-ratio inference for iv. *American Economic Review 112*(10), 3260–3290.

Levinsohn, J. and A. Petrin (2003). Estimating production functions using inputs to control for unobservables. *The review of economic studies 70*(2), 317–341.

Mena, B. (2021). Validity of the single national examination of medical knowledge (eunacom). *Validity of educational assessments in Chile and Latin America*, 353–369.

Menares, F. and P. Muñoz (2023). The impact of a disease-specific health insurance reform on mortality. Technical report, HEDG, c/o Department of Economics, University of York.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences 116*(44), 22071–22080.

OECD (2015). OECD Health Statistics 2015. Accessed: 2024-06-17.

OECD (2019). *Recent Trends in International Migration of Doctors, Nurses and Medical Students*. Paris: OECD Publishing.

OECD (2019). Recent trends in international migration of doctors, nurses and medical students. Accessed: 2024-06-17.

Olivieri, S., F. Ortega, A. Rivadeneira, and E. Carranza (2022). The labour market effects of venezuelan migration in ecuador. *The Journal of Development Studies 58*(4), 713–729.

Olley, G. S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–1297.

Pardo, C. (2019). Health care reform, adverse selection and health insurance choice. *Journal of health economics 67*, 102221.

Propper, C. and J. Van Reenen (2010). Can pay regulation kill? panel data evidence on the effect of labor markets on hospital performance. *Journal of Political Economy 118*(2), 222–273.

Quan, H., V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali (2005). Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Medical care*, 1130–1139.

RNPI (2024). Registro Nacional de Prestadores Individuales. https://rnpi.superdesalud.gob.cl/. Superintendencia de Salud de Chile.

SIGTE (2015–2019). Sistema de Gestión de Tiempos de Espera. https://sigte.minsal.cl/. Ministerio de Salud de Chile.

SIRH (2011–2019). Sistema de Información de Recursos Humanos. http://sirh.minsal.cl. Ministerio de Salud de Chile.

Socha-Dietrich, K. and J. Dumont (2021). International Migration and Movement of Doctors to and Within OECD Countries - 2000 to 2018: Developments in Countries of Destination and Impact on Countries of Origin. Technical Report 126, OECD Publishing, Paris.

Svorny, S. (2004). Licensing doctors: do economists agree? *Econ Journal Watch 1*(2).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology 58*(1), 267–288.

Walters, C. (2024). Empirical bayes methods in labor economics. In *Handbook of Labor Economics*, Volume 5, pp. 183–260. Elsevier.

ONLINE APPENDIX

Physicians' Occupational Licensing and the Quantity-Quality Trade-off

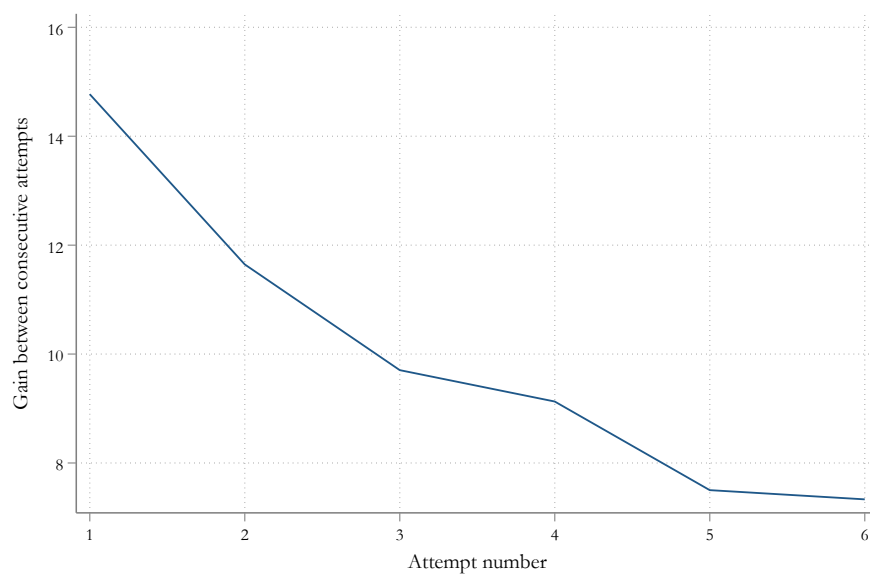Juan Pablo Atal, Tomás Larroucau, Pablo Muñoz, and Cristóbal Otero

List of Figures

List of Tables

Additional Figures

Figure A.1: Score Gains Over Attempts

Notes: This figure presents the score gains after consecutive test-taking attempts by the same individual. It shows that on average score gains diminish with each additional attempt. Data come from ASOFAMECH (2019) and the figure includes the universe of test-takers between 2009 and 2018.

Additional Tables

Table A.1: Descriptive Statistics

Panel A: Health Outcomes				
	Mean	Std. Dev.	Median (p50)	# of Obs.
	(1)	(2)	(3)	(4)
Patient Characteristics:				
% Female	0.57	0.08	0.58	1402
% Age < 29	0.30	0.15	0.31	1402
% Age $\in$ (30,29)	0.10	0.04	0.10	1402
% Age $\in$ (40,49)	0.09	0.03	0.09	1402
% Age $\in$ (50,59)	0.11	0.03	0.11	1402
% Age $\in$ (60,69)	0.12	0.04	0.12	1402
% Age $\in$ (70,79)	0.14	0.06	0.13	1402
% Age $\in$ (80,89)	0.11	0.06	0.10	1402
% Age > 89	0.03	0.02	0.02	1402
% Public Insurance	0.97	0.04	0.98	1402
Hospital Characteristics:				
In-hospital Death Rate	3.28	1.82	2.92	1,402
Death Rate (28-days)	5.07	2.71	4.51	1,402
Patients (# Admissions)	5,656	7,686	1,964	1,402
Service Rate (# Admissions/Beneficiaries)	0.02	0.02	0.01	1,402
Length of Stay	4.03	5.66	3.00	1,402
Complication Rate	11.41	4.25	11.05	1,402
Total Number of Surgeries	2,018	3,332	6.00	1,402
Physicians	77.64	119.64	20.00	1,402

Panel B: Test Scores					
Year	Number Tests Takers	Average score	% Approved (score $\geq$ 51)	Average score if score $\geq$ 51	# Tests $\in [40-51)$
2009	1,389	71.8	92	74.3	87
2010	1,535	65.1	80	72.1	142
2011	1,748	66.6	81	73.3	160
2013	2,003	56.1	66	67.5	231
2014	2,557	55.8	65	67.5	335
2015	3,641	54.7	60	66.5	651
2016	4,999	53.0	54	66.9	1,012
2017	6,014	52.1	55	64.9	1,233
2018	7,121	53.9	58	65.0	1,552

Notes: This table presents descriptive statistics for the data used in the primary analysis. Panel A summarizes patient and hospital characteristics for all public hospitals included in the analysis. These data are derived from individual-level inpatient records reported by DEIS (2019) and restricted administrative records on public hospital employees from SIRH (2019). Panel B provides statistics on EUNACOM scores from 2009 to 2018. For cases where the exam was taken twice in the same year, all data are pooled. The dataset includes all records of test takers and come from ASOFAMECH (2019).

A Score Imputation

The EUNACOM exam was introduced in 2009. Before that, physicians were not required to take a standardized test to practice medicine. To account for the quality of physicians licensed before EUNACOM, we impute their hypothetical scores. This appendix outlines the imputation procedure.

Formally, let y_{ih} denote the licensing score of physician i in hospital h. Our baseline model for scores is:

$$y_{ih} = \mu_h + \alpha^h_{r(i)} + \varepsilon_{ih},$$

where μ_h represents the hospital-level mean, $\alpha^h_{r(i)}$ captures the region-of-origin effect (or "differential score") for physicians from region r working in hospital h, and ε_{ih} is an idiosyncratic error term with mean of zero.

To improve out-of-sample predictions, which is particularly important when hospitals have very few physicians from a particular region, we apply an empirical Bayes shrinkage procedure (Efron and Morris, 1973; Walters, 2024). Specifically, we regress scores on region-of-origin fixed effects for each hospital h using OLS with sum-to-zero constraints (i.e., $\sum_r \alpha^h_r = 0; \forall h$). We then treat each α^h_r and its standard error s_h as coming from a prior distribution centered at 0—reflecting the sum-to-zero constraint—with limiting variance s^2_h. Given the OLS estimate $\widehat{\alpha}^h_r$ and its sampling variance $\widehat{\mathrm{Var}}(\widehat{\alpha}^h_r)$, the empirical Bayes estimate (posterior mean) is a weighted average of the OLS estimate and the prior mean (zero). Specifically:

$$\widetilde{\alpha}^h_r = \underbrace{\frac{\widehat{\mathrm{Var}}(\widehat{\alpha}^h_r)}{\widehat{\mathrm{Var}}(\widehat{\alpha}^h_r) + s^2_h}}_{\text{shrinkage factor}} \widehat{\alpha}^h_r.$$

Since the shrinkage factor ranges between 0 and 1, each estimated effect is pulled ("shrunk") toward zero, with the degree of shrinkage increasing as the precision of $\widehat{\alpha}^h_r$ decreases.

Imputation of Missing Scores: For a physician i at hospital h from region $r(i)$ who does *not* have an observed licensing score, we impute her score using the empirical Bayes-adjusted model:

$$\widehat{y}_{ih} = \widehat{\mu}_h + \widetilde{\alpha}^h_{r(i)},$$

where $\widehat{\mu}_h$ is the average score of physicians working at hospital h. Thus, the imputed score is a combination of the grand mean for hospital h ($\widehat{\mu}_h$), and the "shrunk" region-of-origin differential at hospital h ($\widetilde{\alpha}^h_{r(i)}$).

In very few cases (1.6%), the data does not include a region of origin for the physician. In these cases, we predict the individual score using a LASSO model. This model incorporates hospital indicators as well as controls for physicians' age and gender.

A.1 Alternative Imputation Method:

As an alternative to our Empirical Bayes approach to imputation, we also employ a Least Absolute Shrinkage and Selection Operator (LASSO) model for prediction (Tibshirani, 1996; Murdoch et al.,

2019). Specifically, we estimate a LASSO regression at the physician level to predict EUNACOM scores based on physicians' gender, age, age squared, and a full set of hospital by region-of-origin fixed effects.

Panel A of Figure A.2 plots the mean cross-validation (CV) error against the regularization parameter λ. Each point represents the average error from k-fold CV for a specific value of λ. We highlight the λ that minimizes the CV error, denoted λ_{CV}, and use it to shrink coefficients toward zero, effectively performing variable selection. Panel B of Figure A.2 presents the density of prediction errors from each model. For this comparison, we focus on physicians whose EUNACOM scores we observe and contrast their actual scores with those predicted by both LASSO and our Empirical Bayes approach.

Finally, Table A.2 replicates the analysis presented in Table 3 in the main body of the paper but using the LASSO imputation method. We find quantitative and qualitative similar results, albeit with a weaker first stage.

Figure A.2: Alternative Imputation Methods

A. Cross-validation
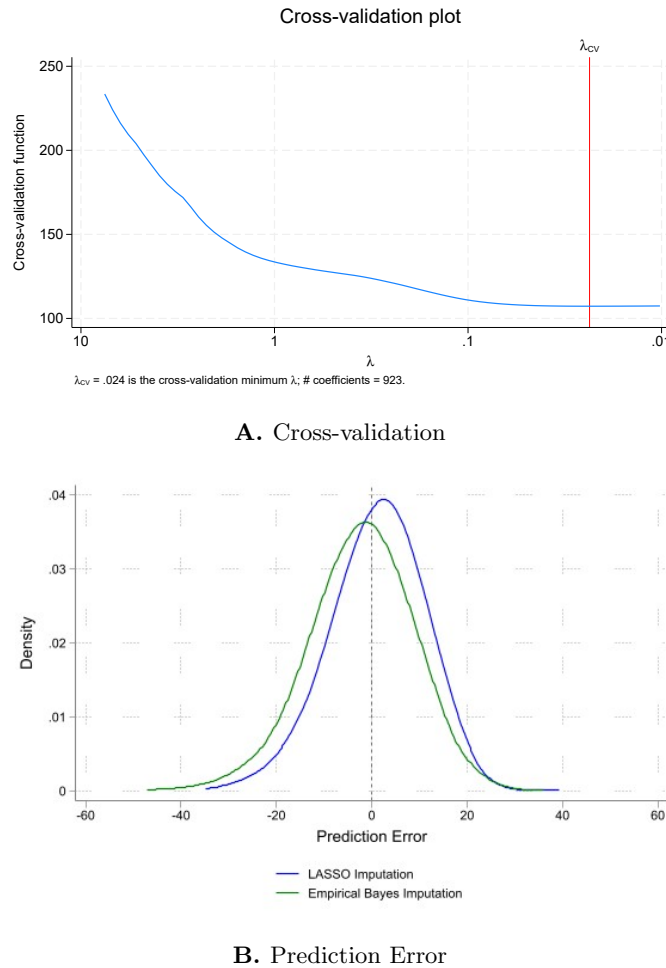
B. Prediction Error

Notes: This figure shows statistics related to the imputation methods. Panel A shows the mean cross-validation (CV) error against the regularization parameter λ. Panel B shows the density of prediction errors from the LASSO and the Empirical Bayes imputation methods.

Table A.2: Impact of Physicians' Quantity and Quality: LASSO Imputation

Panel A: Access

	Ln service rate	Ln inpatient surgeries	Ln exits from waiting list	
			Surgical	Medical
	(1)	(2)	(3)	(4)
Ln Physicians ($\hat{\alpha}_L$)	1.06	6.90	3.62	3.24
	(0.40)	(3.54)	(3.05)	(2.20)
Avg. Physicians' Quality ($\hat{\alpha}_\theta$)	0.01	0.16	-0.01	0.02
	(0.02)	(0.15)	(0.12)	(0.09)
Observations	1,402	744	738	942
Mean Dep. Var.	0.015	3,803	1,534	8,403
F-stat (First-stage)	11.18	3.544	2.979	5.978
Anderson-Rubin (χ^2)	0.00	0.00	0.00	0.00

Panel B: Quality

|	Mortality		28-days	In-hospital Complications
	In-Hospital			
	Ln death rate	Pred. death rate	Ln death rate	Ln complications rate
	(1)	(2)	(3)	(4)
Ln Physicians ($\hat{\alpha}_L$)	-1.21	0.10	-1.11	-1.44
	(0.47)	(0.11)	(0.42)	(0.43)
Avg. Physicians' Quality ($\hat{\alpha}_\theta$)	-0.05	-0.00	-0.04	-0.05
	(0.02)	(0.01)	(0.02)	(0.02)
Observations	1,402	1,402	1,402	1,373
Mean Dep. var.	3.28	3.51	5.08	11.65
F-stat (First-stage)	11.18	19.22	11.18	10.60
Anderson-Rubin (χ^2) p-value	0.00	0.03	0.00	0.00

Notes: This table presents the impact of the quantity and quality of physicians on public hospital performance. Panel A focuses on utilization, which we proxy through the service rate, inpatient surgeries, and exist from the waiting list. Panel B focus on patients' mortality and complications. Estimates come from the two-stage least squares estimation of Equation 18, and we use a quality measure based on the scores imputation using LASSO.

B   Other Labor Inputs

To work in a public hospital, foreign-trained healthcare professionals must validate their degree in Chile. This process is overseen by specific recognized Chilean universities and involves submitting academic credentials, with additional coursework or exams required in some cases. For physicians, passing the EUNACOM is also necessary, as it qualifies them to practice in public healthcare settings.[35] After validation, they must register with the National Registry of Individual Health Providers by providing their validated degree and other required personal documentation.

We observe the impact of the migration wave only in the introduction of foreign-trained physicians. As described in the main text, there is a significant increase in the share of foreign-trained physicians, both among those newly enrolled in the National Registry of Healthcare Providers and those working in public hospitals. However, we do not find similar effects for other healthcare workers. As shown in Panel A of Figure A.3, the share of newly registered healthcare workers in the National Registry of Individual Health Providers remained fairly stable between the 2000s and 2019, except for physicians. Panel B highlights the share of foreign-trained healthcare workers employed in public hospitals, and shows a significant increase for physicians—from 7 percent in 2011 to 18 percent in 2019. In contrast, there is no noticeable change in the share of foreign-trained workers in other healthcare categories.

Figure A.3: Migration Wave: Other Labor Inputs

A. Newly Registered Providers   **B.** Stock in Public Hospitals
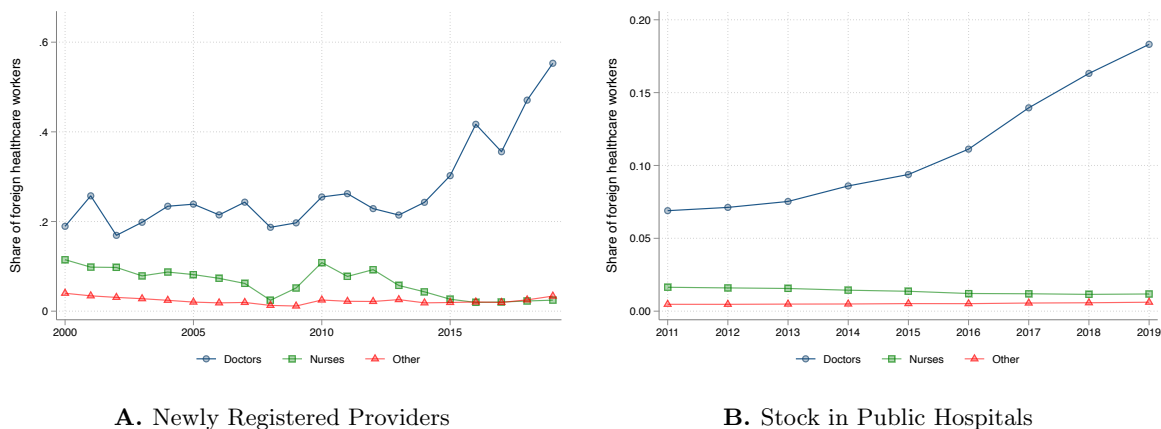
Notes: This figure shows the evolution of the share of foreign-trained healthcare providers in Chile. Panel A examines trends from 2000 to 2019 using data from the National Registry of Healthcare Providers (RNPI, 2024). Panel B focuses on healthcare providers working in public hospitals from 2011 to 2019, based on data from SIRH (2019).

---

[35]Passing the EUNACOM simultaneously qualifies physicians to practice and validates their degree, which reduces the incentive to pursue a separate degree validation process before taking the EUNACOM.

C   Quality Gains

In general, our model could allow for scores to improve over attempts due to increased test-taking ability as well as due to improvements in quality. With data on scores alone, it is not possible to disentangle both mechanisms. In this section, we show that auxiliary data is consistent with no quality improvements over attempts. This justifies the assumption of no quality gains over attempts.

We leverage the discontinuity in retaking probability around the passing threshold to show that retaking does not improve labor-market outcomes that proxy for quality. Figure A.4-A shows the probability of ever retaking the test as a function of the distance to the score in the first attempt to the threshold. It shows that the retaking probability drops substantially around the cutoff. Figure A.4-B shows that the maximum achieved scored changes discontinuously around the threshold. Note that score gains in panel B are a combination of gains in test-taking ability, gains in quality, and selection around cutoff (Gilraine and Penney, 2023).

We investigate how two proxies of quality change around the threshold. Figure A.4-C shows that the probability of being appointed chief general practitioner increases substantially with the score in the first attempt. However, there are no discernible increases around the threshold. Similarly Figure A.4-D shows a (more modest) increase in the hours of contract with the score in the first attempt, with no discernible increases around the threshold. Together, we take this as evidence against the idea that there are quality improvements from retaking.
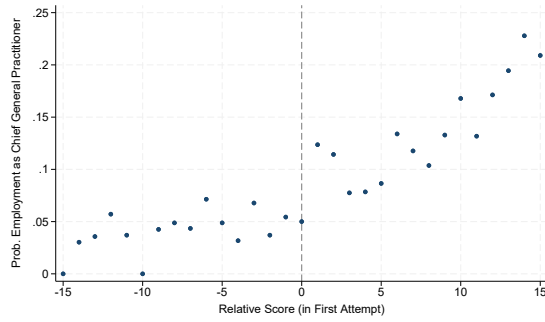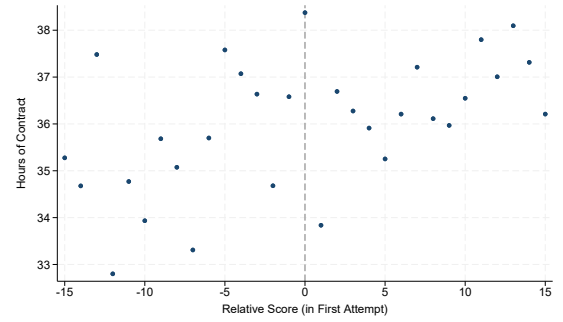
Figure A.4: Testing for Quality Gains
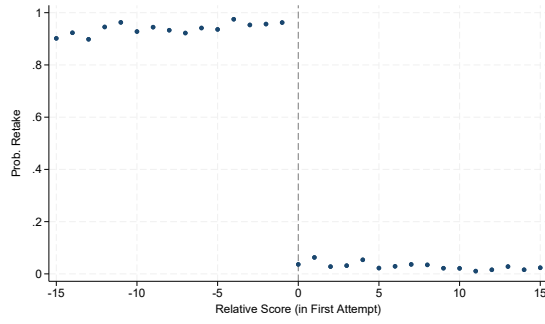
A. Retaking Probability
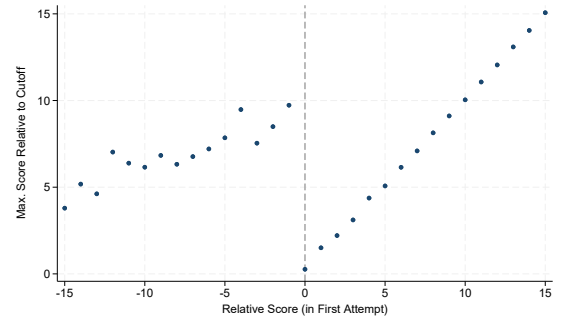
B. Maximum achieved score

C. Appointed Chief General Practitioner

D. Hours of Contract

Notes: This figure illustrates the effect of exam retaking on physician-quality improvements. Panel A depicts the probability of individuals retaking the exam based on whether they scored above the passing threshold on their first attempt. Panel B presents the maximum additional points individuals gained across all their retake attempts, relative to the passing threshold. The line on the right closely follows a 45-degree trajectory, consistent with the fact that individuals who pass the exam do not retake it. In contrast, the line on the left indicates that individuals just below the threshold improve their scores by approximately 10 points on average, with smaller improvements observed for those further below the threshold. Panel C examines the probability of becoming Chief General Practitioner as a proxy for physician quality. There is no visible discontinuity at the threshold on their first test-taking attempt, which is consistent with the absence of quality gains from repeated test-taking. Panel D use weekly working hours as an additional quality measure and similarly there is no visible effect at the threshold. Data on scores come from ASOFAMECH (2019). Panels A and B use the full sample of test-takers, while Panels C and D incorporate data from SIRH (2019), restricting the sample to physicians observed working in public hospitals.

D   Production Function

Translog Production Function: Given its parsimony and tractability, we opt for the Cobb-Douglas specification in our main analysis. Table A.3 below assesses the impacts of physician quantity and quality on health outcomes when using a Translog production function, which allows for a linear interaction between physicians' quantity and quality. The table shows the results from the main model in columns (1) and (3) and the results from the alternative model in columns (2) and (4). At the bottom of the table, we present the overall effect of physicians' quantity and quality on access and mortality. The results are similar, but we obtain a weak first stage for the Translog model, an additional reason in favor of our preferred specification.

Table A.3: Impact of Physicians' Quantity and Quality: Translog Production Function

	Ln Death Rate		Ln Service Rate	
	(1)	(2)	(3)	(4)
Ln Physicians ($\hat{\alpha}_L$)	-0.83	0.43	1.01	1.08
	(0.31)	(1.64)	(0.29)	(1.34)
Avg. Physicians' Quality ($\hat{\alpha}_\theta$)	-0.04	0.07	0.01	0.01
	(0.02)	(0.15)	(0.02)	(0.12)
Interaction ($\hat{\alpha}_{L\theta}$)		-0.02		-0.00
		(0.03)		(0.03)
Observations	1,402	1,402	1,402	1,402
Model	2SLS	2SLS	2SLS	2SLS
Year FE	Yes	Yes	Yes	Yes
Hospital FE	Yes	Yes	Yes	Yes
Mean dep var	3.28	3.28	0.015	0.015
First-stage F-stat	22	2.332	22	2.332

Translog Quantity and Quality Impacts:

Quantity Impact		-1.280		0.987
		(0.460)		(0.309)
Quality Impact		-0.010		0.008
		(0.002)		(0.001)

Notes: This table presents the impact of the quantity and quality of physicians on public hospital performance. Columns (1) and (2) focus on utilization, which we proxy through the service. Columns (3) and (4) focus on health outcomes, which we proxy with in-hospital deaths. Estimates in columns (1) and (3) come from the two-stage least squares estimation of Equation 18. Estimates in columns (2) and (4) come from alternative models that include the interaction between physicians' quantity and quality as an additional endogenous variable and the interaction of our shift-share instruments as an additional instrumental variable.

Alternative Quality Index:

Table A.4 presents estimates of the health production function using an alternative specification. In this specification, the quality index is defined as the share of physicians in the hospital j during

period t whose quality falls below the median of the overall distribution. Encouragingly, the results remain qualitatively similar: quality does not affect access, and worse quality is associated with poorer health outcomes.

Table A.4: Impact of Physicians' Quantity and Quality: Alternative Quality Index

Panel A: Access

	Ln service rate	Ln inpatient surgeries	Ln exits from waiting list	
			Surgical	Medical
	(1)	(2)	(3)	(4)
Ln Physicians ($\hat{\alpha}_L$)	0.98	4.36	3.71	2.94
	(0.25)	(1.29)	(1.32)	(1.19)
% Low Quality Physicians ($\hat{\alpha}_\theta$)	-0.05	-1.00	0.05	-0.17
	(0.18)	(0.80)	(0.78)	(0.82)
Observations	1,376	740	736	934
Mean Dep. Var.	0.0155	3,819	1,537	8,467
F-stat (First-stage)	16.25	10.33	9.29	9.96
Anderson-Rubin (χ^2)	0.00	0.00	0.00	0.00

Panel B: Quality

	Mortality			In-hospital
	In-Hospital		28-days	Complications
	Ln death rate	Pred. death rate	Ln death rate	Ln complications rate
	(1)	(2)	(3)	(4)
Ln Physicians ($\hat{\alpha}_L$)	-0.68	0.13	-0.61	-0.76
	(0.28)	(0.07)	(0.25)	(0.24)
% Low Quality Physicians ($\hat{\alpha}_\theta$)	0.49	0.03	0.48	0.54
	(0.20)	(0.06)	(0.18)	(0.17)
Observations	1,376	1,376	1,376	1,349
Mean Dep. var.	3.30	3.50	5.09	11.65
F-stat (First-stage)	16.25	21.57	16.25	15.46
Anderson-Rubin (χ^2) p-value	0.00	0.03	0.00	0.00

Notes: This table presents the impact of the quantity and quality of physicians on public hospital performance. Panel A focuses on utilization, which we proxy through the service rate, inpatient surgeries, and exist from the waiting list. Panel B focus on patients' mortality and complications. Estimates come from the two-stage least squares estimation of Equation 18 when we use the share of physicians in each hospital and time with quality below the median of the entire distribution as quality index.

E Shift-share Instruments

To assess the robustness of our instrumental variables approach, we build on a recent econometric literature which suggests two distinct paths to identification. One path, developed by Borusyak and Jaravel (2017) and Adao et al. (2019), leverages many exogenous shifts while making no assumption on the exogeneity of the shares. The second path, proposed by Goldsmith-Pinkham et al. (2020), instead focuses on share exogeneity. As pointed out in Borusyak et al. (2024), identification "from the shifts" can be understood as leveraging a shift-level natural experiment, while identification "from the shares" can be viewed as pooling together multiple difference-in-differences designs leveraging heterogeneous shock exposure. In this appendix, we present different robustness checks assessing the exogeneity of our shifts and shares.

E.1 Identification "from the shifts"

Shift-based identification stems from the observation that a share-weighted average of random shifts is itself as-good-as-random (Borusyak et al., 2024). This is true even if the shares are econometrically endogenous, in the sense that units with different shares may have systematically different unobservables. Indeed, Borusyak et al. (2022) shows that classical shift-share IV regression coefficients are numerically equivalent to those obtained from a regression where the outcome and treatment variables are first averaged, using exposure shares as weights, and the shocks are directly used as instrument for the aggregated treatment.

The fact that shift-share estimates can be equivalently obtained by a shock-level IV procedure suggests ways to establishing their consistency. Following Borusyak et al. (2022), we begin by showing statistics of the shocks. Table A.5 shows the distribution of the quantity and quality shocks. As shown by columns (1) and (3), the distribution of the quantity shocks has an average of 3.0, a standard deviation of 7.6, and an interquartile range of 1.8; while the distribution of the quality shocks has an average of -0.5, a standard deviation of 2.2, and an interquartile range of 3.9. Columns (2) and (4) show that there is residual shock variation even conditional on period fixed effects. The inverse HHI of the exposure shares is 28.8 across region by period cells. The largest shock weights are 6% across region-by-periods.
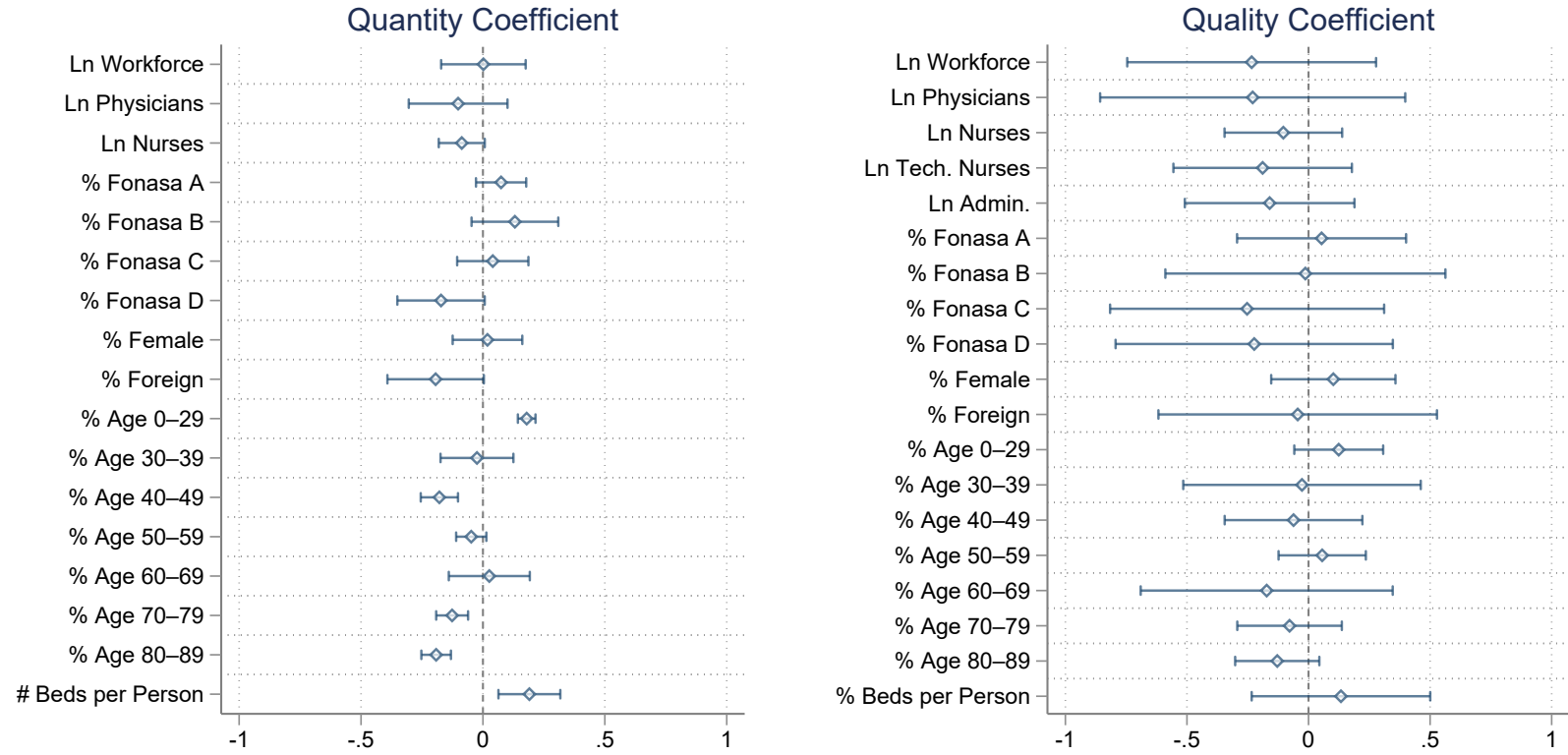
To corroborate the plausibility of the conditional quasi-random shock assignment we perform a shock balance test. If the migration shocks are as-good-as-randomly assigned, we expect them to not predict predetermined variables related to hospitals' workforce and patients' demographics. Figure A.5 reports the results of our balance tests on potential confounders. Panel A focuses on the quantity shocks and Panel B on the quality shocks. Reassuringly, we find that there is no statistically significant relationship between most variables and the shocks. There is some evidence of unbalance between the quantity shock and a few patient and hospital characteristics. Nonetheless, these are variables we control for in our main analysis.

Table A.5: Shock Summary Statistics

	Quantity Shock		Quality Shock	
	(1)	(2)	(3)	(4)
Mean	3.0	0.0	-0.5	0.0
Standard deviation	7.6	7.5	2.2	1.3
Interquartile range	1.8	1.2	3.9	1.3
Residualizing on period FE	No	Yes	No	Yes
Effective sample size ($1/HHI$ of s_{nt} weights)	28.83	28.83	28.83	28.83
Largest s_{nt} weight	0.06	0.06	0.06	0.06
No. of regions-period shocks	126	126	126	126
No. of regions	16	16	16	16

Notes: This table summarizes the distribution of migration shocks across regions of origin and periods. Quantity shocks are measured as the percentage change in the number of physicians from each region of origin clearing the cutoff of the licensing exam. Similarly, quality shocks are measured as the change in the average quality of eligible test-takers from different regions of origin. All statistics are weighted by the region of origin exposure shares. Columns (1) and (2) consider the quantity shocks and columns (3) and (4) consider the quality shocks. Columns (2) and (4) residualize the migration shocks on period indicators. As in Borusyak et al. (2022), we also report the effective sample size (the inverse re-normalized Herfindahl index of the weights) as well as the largest shares.

Figure A.5: Shocks Balance Test

Notes: These figures assess the plausibility of the conditional quasi-random assignment of shocks. Panels A and B show the point estimate and confidence interval obtained from separate regressions of predetermined variables (as of 2012) on the quantity and quality shocks, respectively. All variables are standardized to have a mean of zero and a standard deviation of one.

E.2   Identification "from the shares"

Share-based identification stems from the work by Goldsmith-Pinkham et al. (2020) showing that the Bartik 2SLS estimator can be decomposed into a weighted sum of the just-identified instrumental variable estimators that use each entity-specific share as a separate instrument.

The fact that the exogeneity of the shift-share instruments might also rely on the exogeneity of the shares implies that for our empirical strategy to be valid, we require the differential exposures to common immigration shocks (the "shares") to be independent of differential changes in our outcome of interest during the pre-period. In our case, this would not be the case if there are endogenous mechanisms affecting both the composition of immigrants within hospitals and patients' outcomes.

To assess the plausibility of the identification assumption we follow Goldsmith-Pinkham et al. (2020) and proceed in two steps. First, we compute the Rotemberg weights for each country-specific instrument. Rotemberg weights indicate which country-specific exposure gets a larger weight in the overall Bartik-2SLS estimate, and thus which nationality-share effects are more important for testing. In our data, Ecuador has the highest weight (0.585), followed by Colombia (0.262), mid and low tier universities in the central region of Chile (0.22), and Cuba (0.182). Second, we test for "parallel trends". Table A.6 presents our results on the relationship between differential exposures to common immigration shocks (the "shares") and differential changes in our outcome of interest during the pre-period. Specifically, we assess "pre-trends" by regressing the outcome of interest against the region of origin-shares during the (pre) period 2012-2015, controlling by hospital and year fixed effects as well our main battery of control variables. Generally, we find that the differential exposure to immigrants from different regions do not statistically or economically predict differential utilization or health outcomes. An exception is the share of European physicians, which seem to be statistically associated with surgeries and infection rates. However, the Rotemberg weight of Europe is among the smallest (0.007). Figure A.6 complements our analysis by visually showing the point estimate and confidence interval associated to each share, for all our outcomes of interest.

Table A.6: Rotemberg Weights and Pre-trends

						Region of Origin						
Other S. America	Europe	Venezuela	Bolivia	Colombia	Cuba	Ecuador	Chile (north)	Chile (south)	Chile (tier 1)	Chile (tier 2)	Chile (tier 3)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)

Ln Service Rate

-0.002	-0.007	0.007	0.007	0.006	-0.004	0.002	-0.001	-0.000	-0.000	0.001	0.000
(0.008)	(0.009)	(0.006)	(0.018)	(0.008)	(0.007)	(0.004)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)

Ln Surgery

0.018	0.057	0.040	-0.012	-0.021	0.029	-0.007	0.002	-0.000	-0.001	-0.002	-0.006
(0.023)	(0.024)	(0.017)	(0.048)	(0.022)	(0.019)	(0.009)	(0.006)	(0.002)	(0.002)	(0.003)	(0.004)

Ln Exit Surgical Wailist

0.031	0.091	0.004	0.023	-0.019	0.064	0.011	0.002	-0.005	0.004	-0.002	-0.006
(0.033)	(0.036)	(0.024)	(0.071)	(0.032)	(0.028)	(0.014)	(0.009)	(0.003)	(0.004)	(0.005)	(0.006)

Ln Exit Medical Wailist

0.021	0.007	-0.005	0.034	0.011	0.009	0.023	0.014	-0.007	0.006	-0.002	-0.004
(0.037)	(0.039)	(0.027)	(0.078)	(0.035)	(0.031)	(0.015)	(0.010)	(0.004)	(0.004)	(0.005)	(0.007)

Ln Death Rate

-0.010	0.022	0.011	0.013	0.017	0.012	0.006	0.011	-0.005	-0.000	-0.002	0.000
(0.012)	(0.013)	(0.009)	(0.026)	(0.012)	(0.011)	(0.005)	(0.003)	(0.001)	(0.001)	(0.002)	(0.002)

Ln Death Rate (28-days)

-0.012	0.007	0.011	0.004	0.009	0.008	0.004	0.006	-0.004	0.001	-0.002	0.001
(0.010)	(0.011)	(0.007)	(0.021)	(0.010)	(0.009)	(0.004)	(0.003)	(0.001)	(0.001)	(0.001)	(0.002)

Ln Infection Rate

0.010	0.050	0.002	-0.024	-0.010	0.005	-0.000	0.001	0.001	0.001	-0.003	-0.000
(0.009)	(0.009)	(0.006)	(0.018)	(0.008)	(0.007)	(0.004)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)

Notes: This table presents robustness checks in the spirit of Goldsmith-Pinkham et al. (2020). Panel A shows the Rotemberg weights associated to each region of origin. Panel B assesses pre-trends by regressing each of the outcomes of interest against the region of origin shares in the (pre) period 2012-2015. Point estimates reflect the differential effect of region-specific shares on the dependent variable.

Figure A.6: Share Balance Test
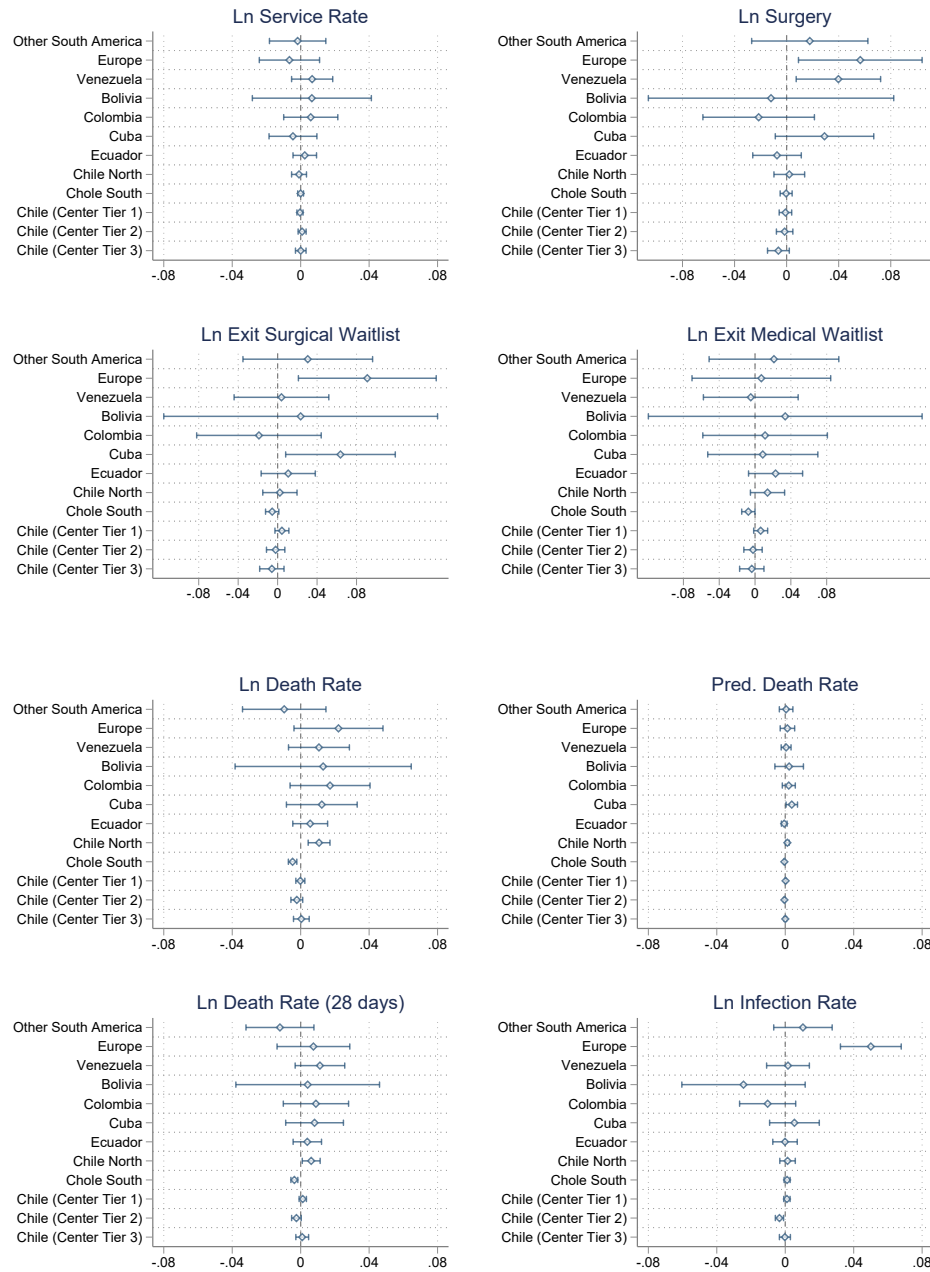
Notes: These figures show the results from regressing the outcome of interest against the region of origin-shares during the (pre) period 2012-2015, controlling by hospital and year fixed effects as well our main battery of control variables.

F   Equilibrium Effects

In this section, we generalize how equilibrium effects can affect the conditional matching probabilities. We choose the following linear function:

$$g_{eq}(M|\tau, r_{\hat{\theta}}, j) = \sum_{\tau' \in \mathcal{T}} \sum_{r'_{\hat{\theta}}} M_{\tau', r'_{\hat{\theta}}} \beta^{\tau, r_{\hat{\theta}}}_{\tau', r'_{\hat{\theta}}, j}, \tag{A.1}$$

where $\beta^{\tau, r_{\hat{\theta}}}_{\tau', r'_{\hat{\theta}}, j} \in \mathbb{R}$ is a parameter capturing the marginal effect of a change in the mass of physicians of type τ' in range $r'_{\hat{\theta}}$, on the conditional match probability of physicians of type τ in range $r_{\hat{\theta}}$ to match with hospital j.

We introduce some definitions and assumptions to characterize how equilibrium effects can affect our counterfactual exercises. We define an order, denoted as $\succ$, over the expected quality ranges in $\mathcal{R}$. Specifically, for any expected quality ranges $r_{\hat{\theta}}$ and $r'_{\hat{\theta}}$ in $\mathcal{R}$, we say that $r'_{\hat{\theta}} \succ r_{\hat{\theta}}$ if the expected qualities in the range $r'_{\hat{\theta}}$ are strictly higher than those in $r_{\hat{\theta}}$. Furthermore, we index quality ranges in increasing order by $r_k \in \mathcal{R}$ $\forall k \in \{1, ..., K\}$, with $K = |\mathcal{R}|$. We introduce the following simplifying assumptions:

Assumption 1.

$$\beta^{\tau, r_k}_{\tau', r_l, j} = 0, \quad \forall r_k, r_l \in \mathcal{R} : k > l, \tau \in \mathcal{T}, j \in \mathcal{J} \tag{A.2}$$

Under Assumption 1, equilibrium effects can affect the conditional matching probabilities of physicians in expected quality range r_k only through changes in the distribution of potential physicians in their own range and "upstream" in the order of expected quality ranges.

We further assume that equilibrium effects can be approximated by contiguous chained effects, where a change in the mass of physicians in a specific expected quality range can have a direct effect on the CMPs within that range, and can also induce a displacement effect on the expected quality range that immediately precedes it. This displacement effect influences the mass in that range and consequently affects the CMPs indirectly. The effect on the predecessor range's CMP is given by the product of its direct effect and the indirect effects from each succeeding range in the chain, thereby chaining these effects downstream and reducing the number of parameters that need to be identified and estimated. We formalize the structure of chain effects in the following assumption:

Assumption 2.

$$\beta^{\tau, r_k}_{\tau', r_l, j} = \begin{cases} \beta^{\tau, r_k}_{\tau', j} & \text{if } l = k \\ \beta^{\tau, r_k}_{N, j} \left(\delta^N_{\tau', r_l} \prod_{i=k+1}^{l-1} \delta^N_{N, r_i} \right), & \text{if } K \geq l > k, \\ 0 & o.w. \end{cases} \tag{A.3}$$

Where $\delta^N_{\tau', r_l} \in \mathbb{R}$, $\forall \tau' \in \mathcal{T}, K \geq l > k$, are parameters capturing the contiguous displacement effect of a change in the mass of physicians of type τ' in range r_l into an equivalent mass of Nationals in the immediate range below, i.e., r_{l-1}. Notice that under Assumption 2, the number of parameters that specify how changes in the distribution of potential physicians, M, affect the conditional matching probabilities, is vastly reduced. Indeed, we only need to identify the direct effects, β, and the contiguous displacement effects, δ, to identify the effects of any change in the distribution.[36] The previous assumptions lead to the following corollary:

---

[36] We expect the contiguous displacement effect parameters, δ, to be non-negative, and the direct effect parameters,

Corollary 1. *Under Assumptions 1 and 2, the general equilibrium effects are given by:*

$$g(M|\tau, r_{\hat{\theta}} = r_k, j) = \beta_{N,j}^{\tau,r_k} \tilde{M}_{N,r_k} + \beta_{F,j}^{\tau,r_k} M_{F,r_k}, \quad \forall k \in \{1, ..., K\}, \tag{A.4}$$

with

$$\tilde{M}_{N,r_k} = \begin{cases} M_{N,r_k} + \delta_{N,r_{k+1}}^N \tilde{M}_{N,r_{k+1}} + \delta_{F,r_{k+1}}^N M_{F,r_{k+1}} & \text{if } k < K \\ M_{N,k} & \text{if } k = K. \end{cases} \tag{A.5}$$

We can now introduce the propositions that will allow us to characterize the equilibrium effects in the counterfactual exercises, i.e., with marginal changes of the physician migration wave, m_F, and marginal changes of the licensing threshold, $\underline{s}$.

Proposition 1. *Under Assumptions 1 and 2, the equilibrium effects of a marginal change on the physician migration wave, m_F, on the conditional matching probabilities, is given by:*

$$\frac{\partial}{\partial m_F} CMP_j = CMP_j \left(\frac{\partial}{\partial m_F} g_{eq}(M|\tau, r_{\hat{\theta}}, j) - \sum_{j'} CMP_{j'} \frac{\partial}{\partial m_F} g_{eq}(M|\tau, r_{\hat{\theta}}, j') \right) \tag{A.6}$$

where

$$\frac{\partial}{\partial m_F} g_{eq}(M|\tau, r_{\hat{\theta}} = r_k, j) = \begin{cases} \underbrace{\beta_{F,j}^{\tau,r_k} \int_{\hat{\theta}(s,F) \in r_k} f^F(s)ds}_{\text{Direct effect}} + \underbrace{\sum_{r_l \succ r_k} \left[\beta_{N,j}^{\tau,r_k} \left(\delta_{\tau',r_l}^N \prod_{i=k+1}^{l-1} \delta_{N,r_i}^N \right) \int_{\hat{\theta}(s,F) \in r_l} f^F(s)ds \right]}_{\text{Indirect Displacement effect}} & \text{if } k < K \\ \underbrace{\beta_{F,j}^{\tau,r_k} \int_{\hat{\theta}(s,F) \in r_k} f^F(s)ds}_{\text{Direct effect}} & o.w. \end{cases} \tag{A.7}$$

Proposition 1 gives us intuition on how the physicians' migration wave might affect the matching probabilities of different hospitals. If the direct effect parameters, β, are non-positive, and the displacement effect parameters, δ, are non-negative, a marginal increase in the physicians' migration wave will lower the conditional matching probability of matching with a hospital j if the sum of its direct and indirect effects, i.e., $\frac{\partial}{\partial m_F} g_{eq}(M|\tau, r_{\hat{\theta}} = r_k, j)$, are larger in magnitude than the weighted average of these effects over every hospital, weighted by their conditional matching probabilities. Finally, Proposition 1 also states that the physicians' migration wave will be negligible for physicians of a given type at hospitals that are not already exposed to that type, i.e., if $CMP_j \approx 0$.

Proposition 2. *Under Assumption 1, the equilibrium effects of a marginal change on the licensing threshold, $\underline{s}$, on the conditional matching probabilities, is given by:*

$$\frac{\partial}{\partial \underline{s}} CMP_j = CMP_j \left(\frac{\partial}{\partial \underline{s}} g_{eq}(M|\tau, r_{\hat{\theta}}, j) - \sum_{j'} CMP_{j'} \frac{\partial}{\partial \underline{s}} g_{eq}(M|\tau, r_{\hat{\theta}}, j') \right) \tag{A.8}$$

β, to be non-positive.

where

$$
\frac{\partial}{\partial \underline{s}} g_{eq}(M|\tau, r_{\hat{\theta}} = r_k, j) = \begin{cases} \underbrace{- \sum_{\tau \in \mathcal{T}} \beta_{\tau,j}^{\tau, r_k} m_\tau f^\tau(\underline{s})}_{Direct \ effect} & if \ k = 1 \\ \\ 0 & o.w. \end{cases} \tag{A.9}
$$

Proposition 2 shows that under Assumption 1, a marginal increase in the licensing threshold has only a direct effect on matching probabilities for physicians at the bottom of the expected quality ranges. In addition, the overall effect on each hospital depends again on the relative magnitude of the direct effects–which are non-negative if the direct effect parameters β are non-positive–and the relative weights given by the conditional matching probabilities across hospitals. Finally, if physicians at the lowest quality range have negligible matching probabilities with respect to highly competitive hospitals, i.e., $CMP_j \approx 0$, a change in the licensing threshold will have no effects on their matching probabilities to these hospitals, revealing a strong vertical sorting of physicians across hospitals.

F.1   Identification strategy

To show identification of the the parameters governing the CMPs, we can take log-odds over $y_j := CMP_j/CMP_\emptyset$. Then, fixing a market, i.e., fixing time t, cross-sectional variation identifies the parameters up to a fixed effect of the combination of $(\tau, r_{\hat{\theta}}, j)$. To separately identify β_j^τ from the parameters within the function g, we need to exploit variation over markets. The mean effect for a given type τ and hospital j will be captured by β_j^τ, because it is fixed over markets, and variation over M across time identifies the parameters within g.

To formalize this identification strategy, we can write in matrix form the direct and indirect effects from the accumulated displacement effects by type of mass. For simplicity of exposition, we omit the indices τ, j, and t. Let M_N and M_F be the vectors of masses for nationals and foreigners respectively from range r_k upwards. These vectors can be represented as:

$$
M_{\tau'} = \begin{bmatrix} M_{\tau', r_K} \\ M_{\tau', r_{K-1}} \\ M_{\tau', r_{K-2}} \\ \vdots \\ M_{\tau', r_k} \end{bmatrix}, \quad \forall \tau' \in \mathcal{T}
$$

and let B_N and B_F be diagonal matrices of the direct effect parameters for nationals and foreigners respectively:

$$
B_{\tau'} = \begin{bmatrix} \beta_{\tau'}^{r_K} & 0 & 0 & \cdots & 0 \\ 0 & \beta_{\tau'}^{r_{K-1}} & 0 & \cdots & 0 \\ 0 & 0 & \beta_{\tau'}^{r_{K-2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \beta_{\tau'}^{r_k} \end{bmatrix}, \quad \forall \tau' \in \mathcal{T}
$$

In addition, let by D_N the matrix containing the displacement effects from masses of nationals:

$$D_N = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \delta^N_{N,r_K} & 0 & 0 & \cdots & 0 \\ \delta^N_{N,r_K}\delta^N_{N,r_{K-1}} & \delta^N_{N,r_{K-1}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta^N_{N,r_K}\prod^{K-1}_{i=k+1}\delta^N_{N,r_i} & \delta^N_{N,r_{K-1}}\prod^{K-2}_{i=k+1}\delta^N_{N,r_i} & \cdots & \delta^N_{N,r_{k+1}} & 0 \end{bmatrix}$$

and let by D^F the matrix containing the displacement effects from masses of foreigners:

$$D_F = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \delta^N_{F,r_K} & 0 & 0 & \cdots & 0 \\ \delta^N_{F,r_K}\delta^N_{N,r_{K-1}} & \delta^N_{F,r_{K-1}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta^N_{F,r_K}\prod^{K-1}_{i=k+1}\delta^N_{N,r_i} & \delta^N_{F,r_{K-1}}\prod^{K-2}_{i=k+1}\delta^N_{N,r_i} & \cdots & \delta^N_{F,r_{k+1}} & 0 \end{bmatrix}$$

The total direct and indirect effects on Y from both nationals and foreigners can be represented as the sum of the effects from each group. This is given by:

$$Y = \underbrace{B_N M_N + B_F M_F}_{\text{Direct effects}} + \underbrace{B_N(D_N M_N + D_F M_F)}_{\text{Indirect effects}}$$

Due to the cumulative and diagonal structure of the displacement (chain) effects, it is clear that variation of masses at the top, M_{τ',r_K}, will identify the direct effect at the top, $\beta^{r_K}_{\tau'}$, and variation on the mass at the immediate predecessor range, $M_{\tau',r_{K-1}}$, together with variation at the successor range, M_{τ',r_K}, helps us to separately identify both the direct effect from range r_{K-1}, $\beta^{r_K-1}_{\tau'}$, from the displacement effect of its immediate successor, δ^N_{τ',r_K-1}. This argument can be applied by moving downward in the order of ranges until we reach the bottom range r_1. To see this more clearly, we can exploit the structure of our Corollary 1:

$$g(M|\tau, r_{\hat\theta} = r_k, j) = \beta^{\tau,r_k}_{N,j}\tilde M_{N,r_k} + \beta^{\tau,r_k}_{F,j}M_{F,r_k}, \quad \forall k \in \{1,...,K\},$$

with

$$\tilde M_{N,r_k} = \begin{cases} M_{N,r_k} + \delta^N_{N,r_{k+1}}\tilde M_{N,r_{k+1}} + \delta^N_{F,r_{k+1}}M_{F,r_{k+1}} & \text{if } k < K \\ M_{N,k} & \text{if } k = K. \end{cases}$$

thus, for $k < K$ we get:

$$g(M|\tau, r_{\hat\theta} = r_k, j) = \beta^{\tau,r_k}_{N,j}M_{N,r_k} + \beta^{\tau,r_k}_{N,j}\delta^N_{N,r_{k+1}}\tilde M_{N,r_{k+1}} + \beta^{\tau,r_k}_{N,j}\delta^N_{F,r_{k+1}}M_{F,r_{k+1}} + \beta^{\tau,r_k}_{F,j}M_{F,r_k}, \quad \forall k \in \{1,...,K\}$$
$$\text{(A.10)}$$

Notice that conditional on having identified the upstream displacement parameters, variation over the displaced mass of equivalent nationals over range $k+1$, $\tilde M_{\tau,r_{k+1}}$, together with variation over the masses from k, M_{τ,r_k}, identifies each of the parameters up to range r_k.

F.2 Likelihood

The previous identification strategy suggests that we can estimate the CMPs by Maximum likelihood, where the individual log-likelihood is given in a recursive manner:

CMPs Log-likelihood. For all $i \in \mathcal{I}$ with $\hat{\theta}(s, \tau, t) \in r_k$, let the individual log-likelihood to be defined recursively by:

$$l\left(\beta, \delta | i \in \mu(j); \tau, s, X, M, t\right) = \log\left(\frac{e^{\beta_j^{\tau} + X_{j\cdot,t}\beta_x^{\tau} + g(M_t|\tau, r_k, j) + \beta_{\hat{\theta}}^{\tau}\hat{\theta}(s,\tau,t)}}{1 + \sum_{j'} e^{\beta_{j'}^{\tau} + X_{j'\cdot,t}\beta_x^{\tau} + g(M_t|\tau, r_k, j') + \beta_{\hat{\theta}}^{\tau}\hat{\theta}(s,\tau,t)}}\right), \forall j \in \mathcal{J} \quad \text{(A.11)}$$

with

$$g(M|\tau, r_k, j) = \beta_{N,j}^{\tau, r_k} \tilde{M}_{N, r_k} + \beta_{F,j}^{\tau, r_k} M_{F, r_k}, \quad \forall k \in \{1, ..., K\},$$

with

$$\tilde{M}_{N, r_k} = \begin{cases} M_{N, r_k} + \delta_{N, r_{k+1}}^N \tilde{M}_{N, r_{k+1}} + \delta_{F, r_{k+1}}^N M_{F, r_{k+1}} & \text{if } k < K \\ M_{N, k} & \text{if } k = K. \end{cases}$$

G   Proof of corollary 1

By the matrix definitions of the total effects we have:

$$Y = B_N M_N + B_F M_F + B_N(D_N M_N + D_F M_F)$$

Calculating each of the terms we have:

$$B_N \times M_N = \begin{bmatrix} \beta_N^{r_K} M_{N,r_K} \\ \beta_N^{r_{K-1}} M_{N,r_{K-1}} \\ \vdots \\ \beta_N^{r_k} M_{N,r_k} \end{bmatrix}$$

$$B_F \times M_F = \begin{bmatrix} \beta_F^{r_K} M_{F,r_K} \\ \beta_F^{r_{K-1}} M_{F,r_{K-1}} \\ \vdots \\ \beta_F^{r_k} M_{F,r_k} \end{bmatrix}$$

$$D_N \times M_N = \begin{bmatrix} 0 \\ \delta_{N,r_K}^N M_{N,r_K} \\ \delta_{N,r_K}^N \delta_{N,r_{K-1}}^N M_{N,r_K} + \delta_{N,r_{K-1}}^N M_{N,r_{K-1}} \\ \vdots \\ \delta_{N,r_K}^N \prod_{i=k+1}^{K-1} \delta_{N,r_i}^N M_{N,r_K} + \delta_{N,r_{K-1}}^N \prod_{i=k+1}^{K-2} \delta_{N,r_i}^N M_{N,r_{K-1}} \cdots + \delta_{N,r_{k+1}}^N M_{N,r_{k+1}} + 0 \end{bmatrix}$$

$$D_F \times M_F = \begin{bmatrix} 0 \\ \delta_{F,r_K}^N M_{F,r_K} \\ \delta_{F,r_K}^N \delta_{N,r_{K-1}}^N M_{F,r_K} + \delta_{F,r_{K-1}}^N M_{F,r_{K-1}} \\ \vdots \\ \delta_{F,r_K}^N \prod_{i=k+1}^{K-1} \delta_{N,r_i}^N M_{F,r_K} + \delta_{F,r_{K-1}}^N \prod_{i=k+1}^{K-2} \delta_{N,r_i}^N M_{F,r_{K-1}} \cdots + \delta_{F,r_{k+1}}^N M_{F,r_{k+1}} + 0 \end{bmatrix}$$

$$B_N(D_N M_N + D_F M_F) = \begin{bmatrix} 0 \\ \beta_N^{r_{K-1}}(\delta_{N,r_K}^N M_{N,r_K} + \delta_{F,r_K}^N M_{F,r_K}) \\ \beta_{N,j}^{r_{K-2}}\left(\delta_{N,r_K}^N \delta_{N,r_{K-1}}^N M_{N,r_K} + \delta_{N,r_{K-1}}^N M_{N,r_{K-1}} + \delta_{F,r_K}^N \delta_{N,r_{K-1}}^N M_{F,r_K} + \delta_{F,r_{K-1}}^N M_{F,r_{K-1}}\right) \\ \vdots \end{bmatrix}$$

With this we can write the effect when $k = K$ as:

$$g(M|\tau, r_K, j) = \beta_{N,j}^{\tau, r_K} M_{N,r_K} + \beta_{F,j}^{\tau, r_K} M_{F,r_K}$$

Now for $K - 1$ we have:

$$g(M|\tau, r_{K-1}, j) = \beta_{N,j}^{\tau, r_{K-1}} M_{N,r_{K-1}} + \beta_{F,j}^{\tau, r_{K-1}} M_{F,r_{K-1}} + \beta_{N,j}^{r_{K-1}} M_{N,r_{K-1}}(\delta_{N,r_K}^N M_{N,r_K} + \delta_{F,r_K}^N M_{F,r_K})$$
$$= \beta_{N,j}^{\tau, r_{K-1}} \left(M_{N,r_{K-1}} + \delta_{N,r_K}^N M_{N,r_K} + \delta_{F,r_K}^N M_{F,r_K}\right) + \beta_{F,j}^{\tau, r_{K-1}} M_{F,r_{K-1}}$$

For $K-2$ we have:

$$g(M|\tau, r_{K-2}, j) = \beta_{N,j}^{\tau, r_{K-2}} M_{N,r_{K-2}} + \beta_{F,j}^{\tau, r_{K-2}} M_{F,r_{K-2}}$$
$$+ \beta_{N,j}^{r_{K-2}} \left(\delta_{N,r_K}^N \delta_{N,r_{K-1}}^N M_{N,r_K} + \delta_{N,r_{K-1}}^N M_{N,r_{K-1}} + \delta_{F,r_K}^N \delta_{N,r_{K-1}}^N M_{F,r_K} + \delta_{F,r_{K-1}}^N M_{F,r_{K-1}} \right)$$
$$= \beta_{N,j}^{\tau, r_{K-2}} \left(M_{N,r_{K-2}} + \delta_{N,r_K}^N \delta_{N,r_{K-1}}^N M_{N,r_K} + \delta_{N,r_{K-1}}^N M_{N,r_{K-1}} + \delta_{F,r_K}^N \delta_{N,r_{K-1}}^N M_{F,r_K} + \delta_{F,r_{K-1}}^N M_{F,r_{K-1}} \right)$$
$$+ \beta_{F,j}^{\tau, r_{K-2}} M_{F,r_{K-2}}$$

Defining

$$\tilde{M}_{N,r_k} = \begin{cases} M_{N,r_k} + \delta_{N,r_{k+1}}^N \tilde{M}_{N,r_{k+1}} + \delta_{F,r_{k+1}}^N M_{F,r_{k+1}} & \text{if } k < K \\ M_{N,k} & \text{if } k = K. \end{cases}$$

we can write the case $K-1$ as:

$$g(M|\tau, r_{K-1}, j) = \beta_{N,j}^{\tau, r_{K-1}} \tilde{M}_{N,r_{K-1}} + \beta_{F,j}^{\tau, r_{K-1}} M_{F,r_{K-1}}$$

Because of the definition we have that:

$$\tilde{M}_{N,r_{K-1}} = M_{N,r_{K-1}} + \delta_{N,r_K}^N M_{N,r_K} + \delta_{F,r_K}^N M_{F,r_K}$$

Wich is equal to the original expression:

$$g(M|\tau, r_{K-1}, j) = \beta_{N,j}^{\tau, r_{K-1}} \left(M_{N,r_{K-1}} + \delta_{N,r_K}^N M_{N,r_K} + \delta_{F,r_K}^N M_{F,r_K} \right) + \beta_{F,j}^{\tau, r_{K-1}} M_{F,r_{K-1}}$$

Now for the case of $K-2$ we can write it as:

$$g(M|\tau, r_{K-2}, j) = \beta_{N,j}^{\tau, r_{K-2}} \tilde{M}_{N,r_{K-2}} + \beta_{F,j}^{\tau, r_{K-2}} M_{F,r_{K-2}}$$

Now we have:

$$\tilde{M}_{N,r_{K-2}} = M_{N,r_{K-2}} + \delta_{N,r_{K-1}}^N \tilde{M}_{N,r_{K-1}} + \delta_{F,r_{K-1}}^N M_{F,r_{K-1}}$$

From before we have that:

$$\tilde{M}_{N,r_{K-1}} = M_{N,r_{K-1}} + \delta_{N,r_K}^N M_{N,r_K} + \delta_{F,r_K}^N M_{F,r_K}$$

Now combining the two previous expressions we have:

$$\tilde{M}_{N,r_{K-2}} = M_{N,r_{K-2}} + \delta_{N,r_{K-1}}^N \left(M_{N,r_{K-1}} + \delta_{N,r_K}^N M_{N,r_K} + \delta_{F,r_K}^N M_{F,r_K} \right) + \delta_{F,r_{K-1}}^N M_{F,r_{K-1}}$$

Replacing in the original expression for $K-2$, yields the same result:

$$g(M|\tau, r_{K-2}, j) = \beta_{N,j}^{\tau, r_{K-2}} M_{N,r_{K-2}} + \beta_{F,j}^{\tau, r_{K-2}} M_{F,r_{K-2}}$$
$$+ \beta_{N,j}^{r_{K-2}} \left(\delta_{N,r_K}^N \delta_{N,r_{K-1}}^N M_{N,r_K} + \delta_{N,r_{K-1}}^N M_{N,r_{K-1}} + \delta_{F,r_K}^N \delta_{N,r_{K-1}}^N M_{F,r_K} + \delta_{F,r_{K-1}}^N M_{F,r_{K-1}} \right)$$

In conclusion the general equilibrium effect is defined recursively by:

$$g(M|\tau, r_k, j) = \beta_{N,j}^{\tau, r_k} \tilde{M}_{N,r_k} + \beta_{F,j}^{\tau, r_k} M_{F,r_k}, \quad \forall k \in \{1, ..., K\},$$

with

$$\tilde{M}_{N,r_k} = \begin{cases} M_{N,r_k} + \delta^N_{N,r_{k+1}} \tilde{M}_{N,r_{k+1}} + \delta^N_{F,r_{k+1}} M_{F,r_{k+1}} & \text{if } k < K \\ M_{N,k} & \text{if } k = K. \end{cases}$$

$\square$

H   Proof of Proposition 1

Differentiating **??** with respect to m_F, and noting that we have a quotient rule, and that m_F has effect through equation A.1

$$\frac{\partial}{\partial m_F} CMP_j(\tau, s, X, M) =$$

$$\frac{e^{\beta^\tau_j + X_j.\beta^\tau_x + g_{\text{eq}}(M|\tau,r_{\hat\theta},j) + \beta^\tau_{\hat\theta}\hat\theta(s,\tau)} \cdot \frac{\partial}{\partial m_F} g_{\text{eq}}(M|\tau,r_{\hat\theta},j) \cdot \left(1 + \sum_{j'} e^{\beta^\tau_{j'} + X_{j'}.\beta^\tau_x + g_{\text{eq}}(M|\tau,r_{\hat\theta},j') + \beta^\tau_{\hat\theta}\hat\theta(s,\tau)}\right)}{\left(1 + \sum_{j'} e^{\beta^\tau_{j'} + X_{j'}.\beta^\tau_x + g_{\text{eq}}(M|\tau,r_{\hat\theta},j') + \beta^\tau_{\hat\theta}\hat\theta(s,\tau)}\right)^2}$$

$$- \frac{e^{\beta^\tau_j + X_j.\beta^\tau_x + g_{\text{eq}}(M|\tau,r_{\hat\theta},j) + \beta^\tau_{\hat\theta}\hat\theta(s,\tau)} \cdot \left(\sum_{j'} e^{\beta^\tau_{j'} + X_{j'}.\beta^\tau_x + g_{\text{eq}}(M|\tau,r_{\hat\theta},j') + \beta^\tau_{\hat\theta}\hat\theta(s,\tau)} \cdot \frac{\partial}{\partial m_F} g_{\text{eq}}(M|\tau,r_{\hat\theta},j')\right)}{\left(1 + \sum_{j'} e^{\beta^\tau_{j'} + X_{j'}.\beta^\tau_x + g_{\text{eq}}(M|\tau,r_{\hat\theta},j') + \beta^\tau_{\hat\theta}\hat\theta(s,\tau)}\right)^2}$$

This can be rearranged, note that we recover the CMP_j terms

$$\frac{\partial}{\partial m_F} CMP_j(\tau, s, X, M) = CMP_j \frac{\partial}{\partial m_F} g_{\text{eq}}(M|\tau,r_{\hat\theta},j) - CMP_j \cdot \frac{\sum_{j'} e^{\beta^\tau_{j'} + X_{j'}.\beta^\tau_x + g_{\text{eq}}(M|\tau,r_{\hat\theta},j') + \beta^\tau_{\hat\theta}\hat\theta(s,\tau)} \cdot \frac{\partial}{\partial m_F} g_{\text{eq}}(M|\tau,r_{\hat\theta},j')}{1 + \sum_{j'} e^{\beta^\tau_{j'} + X_{j'}.\beta^\tau_x + g_{\text{eq}}(M|\tau,r_{\hat\theta},j') + \beta^\tau_{\hat\theta}\hat\theta(s,\tau)}}$$

$$= CMP_j \left(\frac{\partial}{\partial m_F} g_{\text{eq}}(M|\tau,r_{\hat\theta},j) - \sum_{j'} CMP_{j'} \frac{\partial}{\partial m_F} g_{\text{eq}}(M|\tau,r_{\hat\theta},j') \right)$$

Note that the sign will depend on the sign of the parenthesis. Now we must calculate the effect of m_F through the channel of the general equilibrium effect (equation A.1) note that because of assumption 1, we will have different cases (if we are in the final range, theres only a direct effect). When $r_{\hat\theta} = r_{\bar{\hat\theta}}$

$$g(M|\tau, r_k, j) = \sum_{\tau' \in \mathcal{T}} \sum_{r_l} M_{\tau',r_l} \beta^{\tau,r_k}_{\tau',r_l,j}$$

$$= \sum_{\tau' \in \mathcal{T}} M_{\tau,r_l} \beta^{\tau',r_k}_{\tau,j}$$

Recall that $M_{\tau,r_{\hat\theta}}$ is defined by equation 12, so the previous expression becomes:

$$g(M|\tau, r_k, j) = \beta^{\tau,r_k}_{F,j} m_F \int_{s:\,\hat\theta(s,F)\in r_k} f^F(s)ds + \beta^{\tau,r_k}_{N,j} m_N \int_{s:\,\hat\theta(s,N)\in r_k} f^N(s)ds$$

Differentiating with respect to m_F, we have:

$$\frac{\partial}{\partial m_F} g_{\text{eq}}(M|\tau, r_k, j) = \beta_{F,j}^{\tau, r_k} \int_{s:\ \hat{\theta}(s,F)\in r} f^F(s)ds$$

Now for the case that $r \neq \bar{r}$, the expression becomes:

$$g(M|\tau, r_{\hat{\theta}}, j) = \sum_{\tau'\in\mathcal{T}} \sum_{r_l} M_{\tau', r_l} \beta_{\tau', r_l, j}^{\tau, r_k}$$

$$= \sum_{\tau'\in\mathcal{T}} \sum_{r_l} m_{\tau'} \int_{s:\ \hat{\theta}(s,\tau')\in r_l} f^{\tau'}(s)ds \beta_{\tau', r_l, j}^{\tau, r_k}$$

$$= \sum_{\tau'\in\mathcal{T}} m_{\tau'} \sum_{r_l} \int_{s:\ \hat{\theta}(s,\tau')\in r_l} f^{\tau'}(s)ds \beta_{\tau', r_l, j}^{\tau, r_k}$$

Differentiating with respect to m_F, and taking in consideration the definition of $\beta_{\tau', r_l, j}^{\tau, r_k}$

$$\frac{\partial}{\partial m_F} g_{\text{eq}}(M|\tau, r_{\hat{\theta}}, j) = \sum_{r_l} \int_{s:\ \hat{\theta}(s,F)\in r_l} f^F(s)ds \beta_{F, r_l, j}^{\tau, r_k}$$

$$= \int_{s:\ \hat{\theta}(s,F)\in r_k} f^F(s)ds \beta_{F,j}^{\tau, r_k} + \sum_{r_l \succ r_k} \int_{s:\ \hat{\theta}(s,F)\in r_l} f^F(s)ds \beta_{F,j}^{\tau, r_k} \left(\delta_{\tau', r_l}^N \prod_{i=k+1}^{l-1} \delta_{N, r_i}^N \right)$$

∎

I Proof of Proposition 2

Now in this case we must differentiate $CMP_j(\underline{s})$ with respect to $\underline{s}$. Just as before, the CMP_j has an effect from $\underline{s}$ through general equilibrium effect A.1 so the derivative is the same as before, the only change is the derivative of Equation A.1.

$$\frac{\partial}{\partial \underline{s}} CMP_j\left(\tau, s, X, M\right) =$$

$$\frac{CMP_j}{1 + \sum_{j'} e^{\beta_{j'}^\tau + X_{j'}.\beta_x^\tau + g_{\text{eq}}(M|\tau, r_{\hat{\theta}}, j') + \beta_{\hat{\theta}}^\tau \hat{\theta}(s,\tau)}}$$

$$\cdot \left(\frac{\partial}{\partial \underline{s}} g_{\text{eq}}(M|\tau, r_{\hat{\theta}}, j) + \sum_{j'} e^{\beta_{j'}^\tau + X_{j'}.\beta_x^\tau + g_{\text{eq}}(M|\tau, r_{\hat{\theta}}, j') + \beta_{\hat{\theta}}^\tau \hat{\theta}(s,\tau)} \cdot \left(\frac{\partial}{\partial \underline{s}} g_{\text{eq}}(M|\tau, r_{\hat{\theta}}, j) - \frac{\partial}{\partial \underline{s}} g_{\text{eq}}(M|\tau, r_{\hat{\theta}}, j') \right) \right)$$

is equal to:

$$\frac{\partial}{\partial \underline{s}} CMP_j = CMP_j \left(\frac{\partial}{\partial \underline{s}} g_{\text{eq}}(M|\tau, r_{\hat{\theta}}, j) - \sum_{j'} CMP_{j'} \frac{\partial}{\partial \underline{s}} g_{\text{eq}}(M|\tau, r_{\hat{\theta}}, j') \right) \qquad \text{(A.12)}$$

Now the derivative of Equation A.1 is (please note that the threshold only affects the first range $r_k = r_1$), so we have:

$$g(M|\tau, r_k, j) = \sum_{\tau' \in \mathcal{T}} \sum_{r_l} M_{\tau', r_l} \beta^{\tau, r_k}_{\tau', r_l, j}$$

$$= \sum_{\tau' \in \mathcal{T}} M_{\tau', r_k} \beta^{\tau, r_k}_{\tau', j}$$

$$= \sum_{\tau' \in \mathcal{T}} m_{\tau'} \int_{s:\, \hat{\theta}(s, \tau') \in r_k} f^{\tau'}(s) ds \, \beta^{\tau, r_k}_{\tau', j}$$

Differentiating, and taking in consideration leibnitz rule of integration, we have:(note that the integrand and the upper bound are both independent of $\underline{s}$, but the lower bound is $\underline{s}$

$$\frac{\partial}{\partial \underline{s}} g_{\text{eq}}(M|\tau, r_{\hat{\theta}}, j) = -\sum_{\tau' \in \mathcal{T}} m_{\tau'} f^{\tau'}(\underline{s}) \beta^{\tau, r_k}_{\tau', j}$$

J   Microfoundation of CMPs

We now provide a micro-foundation for the CMPs in a simplified environment.

J.1   Physicians' preferences

Consider a continuum of physicians indexed by i. Each physician i belongs to a type $\tau(i) \in \{N, F\}$, with total mass m_τ for each type τ. The physician's indirect utility from matching with hospital j is given by

$$u_{ij} = X_{\tau(i)j}\beta + \gamma_j s_i + \varepsilon_{ij}, \tag{A.13}$$

where $X_{\tau(i)j}$ represents observable characteristics (e.g., hospital attributes), s_i is physician i's licensing score, γ_j is a hospital-specific coefficient on that score, and ε_{ij} are i.i.d. Type-I Extreme Value shocks.

J.2   Hospitals' preferences

Let $\mathcal{J}$ be a finite set of hospitals, each indexed by j, and suppose hospital j is endowed with κ_j vacancies, where $\kappa_j \in \mathcal{K}$. Hospitals maximize an objective function that is strictly increasing in the expected quality $\hat{\theta}$ of the physicians they match with. Strict monotonicity in $\hat{\theta}$ implies that if hospital j can replace a physician of lower expected quality with one of higher expected quality, it will strictly prefer to do so.

J.3   Equilibrium

An equilibrium allocation satisfies pairwise stability: no hospital–physician pair can profitably deviate and match with each other, given the matches of all other agents. Following Azevedo and Leshno (2016), the equilibrium can be characterized by a system of expected quality cutoffs. In particular, each hospital j admits any physician i with $\hat{\theta}_i$ above the relevant cutoff $\bar{\theta}_j$, while each

physician i chooses the hospital that maximizes her indirect utility among those that would admit her.

J.4 Example with two vertically sorted hospitals.

Assume there are two hospitals, i.e., $\mathcal{J} = \{U, R\}$ with capacities κ_U and κ_R respectively. Then, the equilibrium expected quality cutoffs are characterized by the following set of equations:

Assuming the Urban hospital is more selective than the Rural hospital, we get that the Urban hospital cutoff is characterized by:

$$\kappa_U = \sum_\tau m_\tau \int_{\hat{\theta}(s) \geq \hat{\underline{\theta}}_U} \frac{e^{X_{\tau U}\beta + \gamma_U s}}{1 + \sum_j e^{X_{\tau j}\beta + \gamma_j s}} f^\tau(s) \, ds \tag{A.14}$$

The Rural hospital cutoff must then satisfy the following equation:

$$\kappa_R = \sum_\tau m_\tau \int_{\hat{\theta}(s) \geq \hat{\underline{\theta}}_R} \frac{e^{X_{\tau R}\beta + \gamma_R s}}{1 + \sum_j e^{X_{\tau j}\beta + \gamma_j s}} f^\tau(s) \, ds + \sum_\tau m_\tau \int_{\hat{\underline{\theta}}_R \leq \hat{\theta}(s) \leq \hat{\underline{\theta}}_U} \left(\frac{e^{X_{\tau U}\beta + \gamma_U s}}{1 + \sum_j e^{X_{\tau j}\beta + \gamma_j s}} \right) \cdot \left(\frac{e^{X_{\tau R}\beta + \gamma_R s}}{1 + e^{X_{\tau R}\beta + \gamma_R s}} \right) f^\tau(s) \, ds, \tag{A.15}$$

where the terms inside the integrals represent the probability of an individual of type τ with score s choosing a particular hospital, following an exploded logit model.

J.5 Micro-founded CMPs

Following Fack et al. (2019), in a continuum model the micro-founded Conditional Matching Probability (CMP) for a physician i of type $\tau(i)$ and hospital j is:

$$CMP^m \big(i \in \mu(j) \big) \;=\; \frac{\exp\big(X_{\tau(i)j}\beta + \gamma_j s_i \big) \, \mathbf{1}\{\hat{\theta}(s_i) \geq \hat{\underline{\theta}}_j\}}{1 \;+\; \sum_{j'} \exp\big(X_{\tau(i)j'}\beta + \gamma_{j'} s_i \big) \, \mathbf{1}\{\hat{\theta}(s_i) \geq \hat{\underline{\theta}}_{j'}\}}. \tag{A.16}$$

The indicator $\mathbf{1}\{\hat{\theta}(s_i) \geq \hat{\underline{\theta}}_j\}$ enforces that j only matches with physicians whose expected quality $\hat{\theta}(s_i)$ exceeds the relevant threshold.

By contrast, our reduced-form CMPs for this simplified model, would be given by:

$$CMP^r \big(i \in \mu(j) \big) \;=\; \frac{\exp\big(X_{\tau(i)j}\tilde{\beta} + \tilde{\gamma}_j s_i - geq_j(M_{\hat{\theta}_i}) \big)}{1 \;+\; \sum_{j'} \exp\big(X_{\tau(i)j'}\tilde{\beta} + \tilde{\gamma}_{j'} s_i - geq_{j'}(M_{\hat{\theta}_i}) \big)}. \tag{A.17}$$

Equivalence between (A.17) and the micro-founded CMP arises if:

$$\exp\Big(-\big[geq_j(M_{\hat{\theta}_i}) - X_{\tau(i)j}\,\beta_{eq} + \gamma_{eqj}\,s_i \big] \Big) \;\approx\; \mathbf{1}\{\hat{\theta}(s_i) \geq \hat{\underline{\theta}}_j\} \quad \forall j \in \mathcal{J}, \tag{A.18}$$

where $\beta_{eq} \equiv \tilde{\beta} - \beta$ and $\gamma_{eqj} \equiv \tilde{\gamma}_j - \gamma_j$.

A smooth approximation of the indicator can be obtained using the Gompertz function,

$$e^{-e^{-\lambda\left(\hat{\theta}_i - \hat{\underline{\theta}}_j\right)}},$$

which converges to an indicator as $\lambda \to \infty$. Thus, under a suitable parameterization of $geq_j(\cdot)$, our reduced-form CMPs can approximate the micro-founded counterparts if

$$geq_j\left(M_{\hat{\theta}_i}\right) - X_{\tau(i)j}\,\beta_{eq} + \gamma_{eqj}\,s_i \approx e^{-\lambda\left(\hat{\theta}_i - \hat{\underline{\theta}}_j\right)}. \tag{A.19}$$

Finally, notice that the function $geq_j(\cdot)$ depends in principle on the vector mass of physicians by quality (which can vary in the counterfactuals). We make this modeling choice such that we can capture changes in equilibrium quality cutoffs, which are themselves a function of preferences, the mass of physicians by quality, and hospital vacancies. To see this, in the next Subsection, we derive expressions for the two-hospital example that make explicit this dependency.

J.5.1 Approximating the two-hospital choice set

For simplicity, assume that physicians' preferences are independent of their scores, i.e., $\gamma_j = 0, \quad \forall j \in \mathcal{J}$. Under this assumption, the equilibrium cutoff for the Urban hospital is given by:

$$1 = \sum_{\tau} \frac{e^{X_{\tau U}\beta}}{1 + \sum_j e^{X_{\tau j}\beta}} \left[\frac{m_{\tau}\left(1 - F^{\tau}(\hat{\underline{\theta}}_U)\right)}{\kappa_U} \right] \tag{A.20}$$

The first term shows the dependency of the equilibrium cutoff on physicians' preferences while the second term makes it explicit the dependency on the stratified market tightness, i.e., the mass of test takers by type who are above a given quality $\hat{\theta}$, divided by the number of posted vacancies.

To get a closed-form expression for the Urban hospital cutoff, we can further assume for simplicity that there is one type τ. Under this assumption, the previous equation gets to:

$$\hat{\underline{\theta}}_U = F^{-1}\left(1 - \frac{\kappa_U}{m\frac{e^{X_U\beta}}{1+\sum_j e^{X_j\beta}}}\right). \tag{A.21}$$

Using the logistic approximation to the inverse CDF, the previous expression becomes:

$$\hat{\underline{\theta}}_U \approx \hat{\mu}_{\theta} + \hat{\sigma}_{\theta}\left(\frac{\pi}{\sqrt{3}}\ln\left(\frac{m\frac{e^{X_U\beta}}{1+\sum_j e^{X_j\beta}} - \kappa_U}{\kappa_U}\right)\right), \tag{A.22}$$

thus, we can write the Gompertz fcn as:

$$e^{-e^{-\lambda(\hat{\theta}_i - \underline{\hat{\theta}}_U)}} = \exp\left(-\exp\left(-\lambda(\hat{\theta}_i - \hat{\mu}_\theta)\right)\left(\underbrace{\frac{m}{\kappa_U}\left(\frac{e^{X_U\beta}}{1+\sum_j e^{X_j\beta}}\right) - 1}_{\text{Preference-adjusted market tightness}}\right)^{\lambda\hat{\sigma}_\theta \frac{\pi}{\sqrt{3}}}\right)$$

$$= \exp\left(-\exp\left(-\lambda(\hat{\theta}_i - \hat{\mu}_\theta)\right)\left(\alpha_U\left(\frac{M}{\kappa_U}\right) - 1\right)^{\lambda\hat{\sigma}_\theta \frac{\pi}{\sqrt{3}}}\right)$$

The previous equation suggests that the matching probability for a hospital at the top of the quality distribution, depends on the share of physicians of high quality that prefer this hospital as their top preference, divided by the number of vacancies for this hospital.

To approximate the Rural hospital choice set, we can further assume that physicians have common preferences over hospitals. Under this assumption, the original equation for the Rural equilibrium cutoff can be written as:

$$\kappa_R = m\left(\frac{e^{X_U\beta}}{1+\sum_j e^{X_j\beta}}\right) \cdot \left(\frac{e^{X_R\beta}}{1+e^{X_R\beta}}\right)\left(F(\underline{\hat{\theta}}_U) - F(\underline{\hat{\theta}}_R))\right) \tag{A.23}$$

Thus, the quality cutoff gets:

$$\underline{\hat{\theta}}_R = F^{-1}\left(\underbrace{\left(1 - \frac{\kappa_U}{m\frac{e^{X_U\beta}}{1+\sum_j e^{X_j\beta}}}\right)}_{P_U} - \underbrace{\frac{\kappa_R}{m\left(\frac{e^{X_U\beta}}{1+\sum_j e^{X_j\beta}}\right) \cdot \left(\frac{e^{X_R\beta}}{1+e^{X_R\beta}}\right)}}_{P_R}\right) \tag{A.24}$$

Doing similar steps than before, we can approximate the Gompertz function with the following expression:

$$e^{-e^{-\lambda(\hat{\theta}_i - \underline{\hat{\theta}}_R)}} = \exp\left(-\exp\left(-\lambda(\hat{\theta}_i - \hat{\mu}_\theta)\right) \cdot \left(\frac{P}{1-P}\right)^{\lambda\hat{\sigma}_\theta \frac{\pi}{\sqrt{3}}}\right)$$

where $P = P_U - P_R$ which is a function of the stratified preference-adjusted market tightness for each of the two hospitals.

The previous expression makes explicit that for the Rural hospital, matching probabilities can be affected by a direct effect, through the term P_R, and an indirect (displacement) effect, through the

term P_U.

The derivations from our simplified micro-founded model, suggest that under reasonable assumptions, the true CMPs can be approximated with multinomial logit probabilities, where the equilibrium effects are a function of physicians' preferences and the stratified market tightness. Therefore, in our empirical application, we allow equilibrium effects to be affected by the market tightness of physicians (of each type) in their own quality range, and the market tightness of physicians ranked above their own quality range.

K   Computation of numerical elasticities

To compute the elasticity of labor and quality, we simulate $R = 100$ draws for the EV1 shocks in the estimated CMPs, and compute the resulting matches for physicians who take the test in year t and have a score above the corresponding threshold. If $\Delta^r L_{jt}(\underline{s})$ physicians of average quality $\Delta^r \theta_{jt}(\underline{s})$ match with hospital j in simulation r; we compute the resulting quantity and quality of each hospital by adding the marginal physicians that match with hospital j in year t to the stock of physicians of the hospital j in year $t-1$. That is, the resulting simulated physicians is $L^r_{jt}(\underline{s}) = L_{jt-1} + \Delta^r L_{jt}(\underline{s})$ and the resulting simulated average quality is $\bar{\theta}^r_{jt}(\underline{s}) = \frac{(\bar{\theta}_{jt-1}L_{jt-1} + \Delta^r L_{jt}\Delta^r \theta_{jt}(\underline{s}))}{L^r_{jt}(\underline{s})}$.

We compute the quantity elasticity numerically as

$$\eta^r_{L_{jt},\underline{s}} \approx -\frac{\left(L^r_{jt}(\underline{s} + \Delta\underline{s}) - L^r_{jt}(\underline{s})\right)}{\Delta\underline{s}} \cdot \frac{\underline{s}}{L^r_{jt}(\underline{s})}$$

and the quality semi-elasticity as

$$\tilde{\eta}^r_{\bar{\theta}_{jt},\underline{s}} \approx -\frac{\left(\bar{\theta}^r_{jt}(\underline{s} + \Delta\underline{s}) - \bar{\theta}^r_{jt}(\underline{s})\right)}{\Delta\underline{s}} \cdot \underline{s}$$

.

We then compute the estimated elasticities by averaging out across simulations.