IEPS
Instituto de Estudos
para Políticas de Saúde

WORKING PAPER SERIES

IEPS WP No. 1

How to Attract Physicians to Underserved Areas? Policy Recommendations from a Structural Model

Francisco Costa
Letícia Nunes
Fábio Sanches

December, 2019

Instituto de Estudos para Políticas de Saúde
Texto para Discussão nº 1
Dezembro de 2019

Rua Itapeva 286 | 81-84
Bela Vista, São Paulo – SP
01332-000 - Brasil

www.ieps.org.br
+55 11 4550-2556
contato@ieps.org.br

How to Attract Physicians to Underserved Areas? Policy Recommendations from a Structural Model

Francisco Costa   Letícia Nunes   Fábio Sanches

July 4, 2019

Abstract

The lack of physicians in poorer areas is a matter of concern in developed and developing countries. This paper exploits location choices and individual characteristics of all generalist physicians who graduated in Brazil between 2001 and 2013 to study policies that aim at increasing the supply of physicians in underserved areas. We estimate physicians' locational preferences using a random coefficients discrete choice model. We find that physicians have substantial utility gains if they work close to the region they were born or from where they graduated. We show that wages and health infrastructure, though relevant, are not the main drivers of physicians' location choices. Simulations from the model indicate that quotas in medical schools for students born in underserved areas and the opening of vacancies in medical schools in deprived areas improve the spatial distribution of physicians at lower costs than financial incentives or investments in health infrastructure.

We thank, without implicating, Kym Ardison, Samuel Bazzi, Luis Braido, Bladimir Carrillo, Claudio Ferraz, Sergio Firpo, Jason Garred, Igal Hendel, Francisco Lima, Claudio Lucinda, Cecilia Machado, Renata Narita, Cezar Santos, Rodrigo Soares, André Trindade, Mar Reguant, Leo Rezende, Rudi Rocha, Marcelo Sant'Anna, Manisha Shah, Gabriel Ulyssea, Guo Xu, and seminar participants at Northwestern, EPGE, USP, Anpec 2015, SBE 2017, and 4th SBE/REAP meeting. We thank the Brazilian Federal Council of Medicine for sharing important data. Costa and Nunes gratefully acknowledge financial support from FAPERJ and from Rede de Pesquisa Aplicada FGV. This study was financed in part by CAPES/Brasil; Grant #001.
F. Costa, FGV EPGE, e-mail: francisco.costa@fgv.br;
L. Nunes, FGV EPGE, e-mail: leticia.nunes@fgvmail.br;
F. Sanches, Insper, e-mail: fabioams1@insper.edu.br.

1 Introduction

The delivery of basic services, such as healthcare, hinges on a key resource: human capital. The lack of qualified health professionals in rural and underdeveloped areas forms a barrier to the improvement of health outcomes of those living in these places.[1] As a result, the imbalances in the geographic distribution of physicians, who tend to be more concentrated in metropolitan areas, have been a matter of concern in developing and developed countries (WHO, 2010). Many governments have resorted to the use of financial and non-financial incentives to recruit specialized professionals to needy regions.[2] However, as the thriving literature on recruiting of health and frontline providers indicates,[3] attracting qualified personnel to poorer locations has been proven challenging.

This paper exploits practice location choices of all generalist physicians graduated in Brazil between 2001 and 2013 to estimate physicians' locational preferences and to study counterfactual policies that aim at reducing imbalances in the geographic distribution of these professionals.[4] We estimate a random coefficients discrete choice model to understand how wages, local living and working conditions, as well as place of birth and graduation influence physicians' locational choice just after graduation. The richness of our microdata permits us to accommodate flexible forms of heterogeneity in individuals' preferences, including the quality of physicians' education. Our estimates suggest that physicians utility increases substantially if they work close to the place they were born or completed medical school. Wages and quality of local health infrastructure are relevant but less important than these two types of geographic (home and graduation place) biases. Our counterfactual experiments indicate that investments in medical schools in underserved areas and the adoption

[1] Some papers have associated the availability and quality of human resources with better health outcomes, e.g., Anand and Barnighausen (2004); Banerjee et al. (2004); WHO (2006); Bjorkman and Svensson (2009).

[2] E.g., the United States (Holmes, 2005), Canada (Bolduc et al., 1996), India (Rao et al., 2012), Ghana (Kruk et al., 2010), and Brazil (Carrillo and Feres, 2019).

[3] In particular, Dal Bo et al. (2013); Dunne et al. (2013); Ashraf et al. (2016); Finan et al. (2017).

[4] We focus on generalists because these are the health professionals directly linked with basic healthcare. We do not study medical specialists, such as surgeons or oncologists, because these tend to be concentrated in specialized health centers and hospitals.

of affirmative action policies – such as quotas in medical schools for students born in poorer regions – are more cost-effective in improving the geographic distribution of physicians than policies based on financial incentives or investment in health infrastructure.

Brazil presents a compelling setting to study physicians' locational preferences. First, there is a sizable imbalance in the presence of doctors across regions within the country. Figure 1 shows the number of registered physicians per thousand inhabitants in each state's capital and countryside in 2015.[5] While the number of physicians per thousand people ranges from 11.9 to 1.42 across state capitals,[6] the supply of doctors outside the capitals is substantially lower, ranging from 2 to 0.1 doctors per thousand people. The poorest states, mostly located in the Northern regions, have fewer doctors than the richest states. Second, the imbalance of physicians across the country comes associated with lower access to preventive care and worse health outcomes. Figure 2 shows that basic health indicators such as access to doctors, exams and infant mortality are worse in the countryside than in metropolitan areas.

We assembled a unique dataset following physicians after graduation. Our primary database comprises all 60,563 generalist physicians that received a medical degree in Brazil between 2001 and 2013. We merge the registries of all new graduates with the official records of all active physicians. We exploit this data to track physicians from birth, through medical school and the first years of their professional lives. A descriptive analysis of the data reveals that more than 50 percent of the physicians in our sample choose to work in the same region as they were born or completed medical school. Metropolitan areas in the richest regions of the country are the main destinations of physicians that decided to migrate to a different region. Curiously, real wages in these metropolitan areas are relatively lower than in other regions. Yet, these areas have better amenities and health infrastructure. Together, these empirical patterns may indicate that wages are not the main local characteristic explaining

[5]In Brazil, in contrast to some countries in North America and Europe, the state capital is the largest and most developed metropolitan region in each state.

[6]E.g., the United States had 2.6 per thousand people in 2013 and Germany had a ratio of 4.1 in 2014.

physicians location choices, and that home biases seem to be particularly relevant.

To better understand how physicians preferences depend on their own characteristics and choice attributes, we estimate a discrete choice model with random coefficients (Berry et al., 2004). This model has the advantage of accommodating spatial correlation across locations and flexible forms of heterogeneity across individuals. We address potential endogeneity of wages through a control function approach (Petrin and Train, 2010). Physicians' choice set is composed by the pairs metropolitan region-state and countryside-state for all Brazil. Guided by our descriptive study, we model physicians' location choices as a function of local (i) expected real wages, (ii) amenities, (iii) health infrastructure at work, (iv) stock of physicians, (v) penetration of private health insurance, as well as physicians' (vi) age, (vii) gender, (viii) birthplace, (ix) graduation place, and (x) quality of the school where they graduated.

Our estimates show that physicians' supply function is inelastic, with mean and median wage elasticity ranging around 0.41 and 0.70 in metropolitan areas and in the countryside, respectively. Our results also suggest that health infrastructure and amenities impact positively physicians' utility. Importantly, we find that both home bias and place of graduation bias play a central role in the choice of job location. Physicians derive great utility for working close to their place of birth and for staying in the same region from where they graduated. As in other papers in the literature (Dal Bo et al., 2013; Agarwal, 2015; Diamond, 2016), local characteristics appear to be as important to explain locational choices of skilled workers as wages. In particular, the low wage elasticities may explain why financial incentives in Brazil have not been sufficient to attract more physicians to underserved areas.

Finally, we find that preferences are heterogeneous according to the rank of the medical school from which the physicians graduated. Physicians graduated from better medical schools value more local amenities, are more inelastic to wages, and derive lower value for returning to their region of birth. Those graduated in top ranked schools in metropolitan regions are the most inclined for staying in their locale of graduation. Because better schools are located in the richest areas of the country, this finding suggests that taste heterogeneity may contribute to regional inequality in the quality of physicians.

In the last step of our study, we use the structural model to implement two types of counterfactual exercises. First, we perform a heuristic analysis to assess how characteristics of individuals and choices affect the geographic distribution of physicians. We homogenize local characteristics across regions and quantify its impact on physicians' spatial distribution. Our benchmark distribution is such that the physicians to population ratio is the same in all regions. We find that if health infrastructure was geographically uniform, the distribution of physicians would improve by 24 percent towards our benchmark. Analogously, if local amenities were homogeneous everywhere, the distribution of doctors would improve by only 4 percent. These results indicate that health infrastructure seems to be more important in explaining the shortage of health professionals in poorer regions than local amenities. Our simulations also point that if medical schools were distributed across regions in proportion to the local population, the distribution of doctors across the country would improve by 39 percent. Surprisingly, if we homogenize the place of origin (birthplace) of medical students, the geographic imbalance in the distribution of physicians' would reduce by 52 percent.

We then provide back-of-the-envelope estimates of the cost-effectiveness of actionable policies. We find that policies exploiting physicians' geographic (home and graduation place) biases are the most cost-effective. First, quotas on student enrollment in medical schools for students born in underserved areas considerably improve the geographic distribution of physicians at little costs. Second, the opening of vacancies in medical schools in areas lacking generalist physicians seems to make the geographic distribution of physicians more even, but at higher costs compared to quotas. Third, offering a large wage premium for doctors in needy areas is also effective but the cost of this policy is higher compared to the first two alternatives. Last, investing in health infrastructure is less effective and substantially more costly than these other options.

Our paper directly relates to the literature that studies labor supply of qualified professionals in rural and poorer areas (Bolduc et al., 1996; Holmes, 2005; Dunne et al., 2013; Agarwal, 2015).[7] More broadly, this paper contributes to the development literature on recruitment of workers for service delivery (Dal Bo et al., 2013; Ashraf et al., 2016; Bau and Das, 2017; Alva et al., 2017; Leon, 2018; Deserranno, 2019), and the growing literature that applies discrete choice models to study the determinants of migration decisions, demand for neighborhoods and labor sorting (Bayer et al., 2007; Albouy, 2009; Bishop and Murphy, 2011; Kennan and Walker, 2011; Albouy and Stuart, 2014; Bayer et al., 2016; Diamond, 2016). Our study contributes to this body of work by shedding light on new dimensions relevant to locational choices of qualified professionals and by providing cost-effectiveness analyses of different policies. We show that the lack of basic working conditions can be more important than local living conditions for physicians. Further, our estimates indicate that geographic biases are decisive to explain locational decisions of qualified health professionals. Importantly, our counterfactuals clarify that the cost-effectiveness of policies acting on home bias and on graduation place bias are substantially different.[8]

Last, we complement this literature by providing evidence of preferences heterogeneity according to the quality of the education of high skill professionals. While higher wages may attract more job candidates, it may also select individuals with weaker pro-social motivation (Francois, 2000) and affect retention and performance (Ashraf et al., 2016; Deserranno, 2019). Our estimates suggest that policies based on wages may also affect the composition of recruited health professionals by attracting individuals more responsive to financial incentives, in our setting, those graduated from lower ranked universities.

This paper is organized as follows: in Section 2, we provide an overview of the Brazilian

[7] Studies in a developing country context use stated-preference surveys (see de Bekker-Grob et al., 2012).

[8] Two recent papers (Kulka and McWeeny, 2018; Falcettoni, 2018) study the shortage of doctors in rural areas in the United States. Like ours, both find that physicians prefer to practice close to their residency location. Differently from these papers, we can distinguish physicians' preferences for working close to their home region (birthplace) or close to the school from where they graduated. As we showed above, this differentiation is important because policies exploiting the two types of geographic bias – home or graduation place – produce different results. Another difference is that we consider physicians' preferences for health infrastructure, which tend to be substantially lower in underdeveloped areas, particularly, in the developing world.

health system and of physicians labor market. We then present a description of our data and some descriptive evidences in Section 3. Section 4 describes the theoretical model of physicians' supply and demand. The estimation results are presented and discussed in Section 5. We present the counterfactual exercises in Section 6 and discuss the policy implications of our results. Section 7 concludes.

2 Empirical Context

This section provides background information on the Brazilian context. First, we describe the labor market for physicians in Brazil. Both public and private sectors are important providers of health services in the country and physicians can work freely in both sectors. Next, we characterize the geographic distribution of physicians and discuss recently implemented public policies to mitigate regional imbalances in the distribution of physicians. Our description suggests that policies based mainly on financial incentives were not sufficient to reduce regional imbalances and that excess demand for physicians still prevails in most parts of the country. We close the section showing correlations between access to basic healthcare and health outcomes of the Brazilian population. These correlations suggest that access to basic health care improves health outcomes, which serve to justify why the presence of physicians in underserved regions may have positive implications for populations living in these areas.

2.1 The Labor Market for Physicians

The Brazilian market for physicians is dichotomous; public and private sectors are important providers of health services. The public health system, known as the Unified Health System (*Sistema Único de Saúde*, SUS), was inspired by the National Health Services in the United Kingdom and is now one of the largest in the world in terms of coverage. Free health care is provided both in public establishments and private health facilities associated with SUS. The services offered range from simple outpatient care to organ transplantation. System management is decentralized with responsibilities being transferred by the Federal government to states and cities. The purpose of this decentralization is to make local health systems more aligned with local interests needs (Elias and Cohn, 2003). Important to our study, state and municipal governments have autonomy to hire physicians and health workers. While the main admission process to jobs in the public health system is through public exams, most entry level jobs for doctors in the public sector are temporary contracts which are more flexible and offered through standard selection process.

The private sector also plays an important role in the Brazilian health system. The country has the second largest private insurance market by number of users in the world, covering 25 percent of the population. This is a regulated market, where most private health plans are either small or medium sized local firms. These plans often operate via contractual arrangements and their medical care is provided through partnership with private clinics and hospitals (Elias and Cohn, 2003).

The labor market for physicians in Brazil is very flexible. Physicians can freely choose where to work and contractual arrangements. It is very common that physicians work in more than one health facility and under a variety of employment relationships at the same time.[9] A survey conducted by the Federal Council of Medicine (Scheffer et al., 2015) in 2014 with 2,400 doctors shows that around 21.6 percent of physicians work exclusively in the public sector, 26.9 percent work exclusively in the private sector and 51.5 percent have joint appointments in the public and private sectors.[10] Because every work arrangement requires the doctors to be registered at the regional CRM, we will observe all of them in our data.

This completes the description of physician market in Brazil. Next we describe the

[9]The constitution limits physicians to hold at most two public jobs, adding to a maximum of 60 hours per week. A decree from the Ministry of Health limits to five the number of private employments per physician, although the municipal government can authorize physicians to accumulate more jobs (Ministry of Heath, Decree 134/2011).

[10]Physicians that are in the public sector mainly work in hospitals (51.5%) and in primary health care (23%). Among those in the private sector, 40.1% own their practices, 38.1% work in private hospitals and 31.1% in clinics or private ambulatories. These percentages do not sum 100% because physicians usually work in several health facilities.

geographic distribution of physicians in the country.

2.2 Geographic Distribution of Physicians

In 1980 Brazil had 1.15 physicians per 1,000 inhabitants. This number increased substantially in the past years followed by the inauguration of several medical schools. By the end of 2014, there were 251 medical schools in the country, but 74 of them opened after 2008 and are not included in our sample since we look at physicians that graduated between 2001 and 2013.[11] In 2015 the number of physicians per 1,000 inhabitants jumped to about 2.11 (Scheffer et al., 2015).

Despite the growth in the number of physicians, some regions are still severely underserved. As Figure 1 shows, physicians are mainly concentrated in state capitals. Concentration is especially high in the richest regions (South and Southeast). The physicians to population ratio ranges from as high as 11.9 in Espirito Santo's capital to 1.27 in its countryside. Across regions there is also a considerable variation: while the countryside of Rio de Janeiro has 2.11 physicians per 1,000 people, the countryside of Piauí, one of the poorest states, has a ratio equal to 0.01. Another way of looking at these imbalances: Brazilians living in cities with populations smaller than 50,000 people, which represents 32.6 percent of the total population, can rely on only 31,500 thousand doctors, a ratio below 0.5 physician per 1,000 inhabitants (Scheffer et al., 2015).

Federal and local governments implemented a series of programs to mitigate the undersupply of physicians in disadvantaged areas. The two most important are the Primary Care Professional Valorization Program (*Provab*) and the More Physicians Program (*Mais Médicos*). The first was created in 2011 and intended to attract newly trained physicians to needy regions using two incentives: competitive and tax-free wages and a 10 percent increase

[11] Around 47% of the medical schools are public, financed by the Federal or State Ministry of Education and tuition free – SIGRAS (System of Indicators of Undergraduate Health Courses), 2014. The other 53% are private schools with tuition that may vary from around $11,971-$41,879 dollars annually – https://www.escolasmedicas.com.br/mensalidades.php; conversion from Brazilian Reais to US Dollars based on the exchange rate in 31/01/2019 (3.65 BRL/USD).

in the final grade in admission exams to medical specialization programs. The second was implemented in the middle of 2013 and involved three strategies: (i) expansion and building of new primary health care units in needy areas; (ii) increasing in the number of medical schools and medical residency programs in areas suffering from undersupply; and (iii) opening of primary health care jobs with good wages in underserved areas (Carrillo and Feres, 2019).

Despite all the effort, the government could not fill new positions. Figure 3 presents the number of vacancies created by the More Physicians Program and the number of physicians graduated in Brazil that filled the vacancies (per 1,000 people) between July 2013 to July 2014. As is evident, most of the vacancies remained unfilled, especially those in the countryside and in poorer states. Given the vast excess demand for physicians, the government started to source foreign doctors, especially from Cuba. This suggests that the main hurdle to overcome regional imbalance of health professionals is not the lack of positions with good wages, but some other aspect behind physicians' locational preferences.

2.3 Health Provision and Health Outcomes

The inadequate supply of physicians has immediate implications for local access to healthcare and, possibly, for the health status of the population. Figure 2 displays some measures of access to healthcare and health outcomes in rural and urban areas in Brazil. The data comes from the 2013 National Health Survey (*Pesquisa Nacional de Saúde*) and the Mortality Information System (SIM-Datasus). The picture is clear: the population in rural areas tend to have lower access to healthcare and worse related health outcomes than in urban areas. For example, those living in the countryside are about 38 percent less likely to have been to a doctor's appointment in the last 12 months than those living in urban areas. Infants in rural areas are about 20 percent less likely to visit the doctor in the first 30 days of life, and about 50 percent more likely to not have gone through the recommended number of prenatal care visits than infants in urban areas. The same pattern can be seen in measures related to men's health and diabetes.[12] Perhaps as a consequence, infant mortality is 18% higher in rural areas.

This gap between rural and urban seems to be driven in part by the lower number of physicians per capita in rural areas. Table A5 presents the regression estimates of three different measures of infants' and women's health statuses on the number of physicians per capita across Brazilian municipalities between 2005 and 2016. We find a positive and strong correlation between physicians per capita in a given municipality-year and infants' health outcomes. These correlations are virtually unchanged when we add a dummy for municipalities in the countryside and do not loose statistical significance when we add state fixed effects and compare municipalities within states. The relationship gets weaker, however, when we add municipality fixed effects. This may be driven by other local characteristics relevant to infants' health but also by small variation in the number of physicians within underserved municipalities over the decade.

In summary, the descriptive evidences shown in this section suggests that:[13] (i) there is a large regional imbalance in the distribution of doctors in Brazil, both across different states and across rural or urban areas within states; (ii) health outcomes in underserved areas tend to be worse. This helps understanding the amount of attention given to this matter by national and international institutions. Next, we describe our dataset and show preliminary evidence on the determinants of the geographic distribution of physicians in Brazil.

3 Data and Descriptive Analysis

This section describes the data sources, our sample and variables. Alongside, we present the summary statistics of our data in Table 1. Appendix E provides more details on our dataset. We also present descriptive statistics on practice location choices of physicians in our sample.

[12]The Brazilian Ministry of Health recommends a minimum of six prenatal care visits and that the first visit to the doctor be in the first week after birth.

[13]Please, see Appendix B for a more detailed discussion.

Our objective is to obtain preliminary evidence on the importance of different local attributes to physicians choices. This analysis will guide the formulation of the structural model in the next sections.

3.1 Data

Sample and aggregation level. Our sample consists of all physicians graduated between 2001 and 2013. For technical reasons, we aggregate geographical units into pairs metropolitan region/state and countryside/state, amounting to 52 possible choices.[14] All explanatory variables are averaged at this geographical level for the year prior to graduation – i.e., ranging from 2000 to 2012 – because we assume physicians make their decision based on the local characteristics one year before graduating.[15]

Physicians and practice location. Our primary data source are the records of all physicians in Brazil, active or not, maintained by the Federal Council of Medicine (CFM). This registry contains physicians' names, cities of birth, medical schools attended and the beginning and conclusion dates of physicians' training. We have 149,637 doctors who graduated between 2001 and 2013, our period of study.[16] To restrict attention to generalist physicians, we merge this database through name and registration number with the National Commission of Medical Residency (CNRM), which contains all physicians who applied for specialist training.[17] We have that 40% of recent graduates did not pursue a residency program, leaving us with 60,563 generalist physicians, which are the focus of our research.

We link all generalists with the Annual Report of Social Information (RAIS/MTE) from

[14] Using finer spatial units leads to higher computational costs and, more importantly, a large number of cells (region-year) not being chosen by any physician. Brazil has 26 states and the Federal District. This would represent 53 decision choices according to our approach. However, the state of Santa Catarina is composed primarily of metropolitan regions.

[15] Physicians graduating and making locational choice at period t observe local characteristics at $t-1$.

[16] We consider only physicians that were born and completed medical school in Brazil. This figure is similar to the 149,002 physicians graduated in the same period according SIGRAS/INEP.

[17] Medical residency is the golden standard of medical specialization in Brazil. Other forms of graduate training are recognized by the Brazilian Medical Association (AMB), but they are very heterogeneous in their quality, curriculum, and duration. Thus, our definition of specialist comprise only the doctors with a medical residency degree.

2001 to 2015 and the National Register of Health Establishments (CNES) from 2005 to 2016, using their full name. The first database is maintained by the Ministry of Labor and Employment in Brazil and currently covers around 97% of the formal sector in the country. For each establishment we observe detailed information about its employees, which makes it possible to see in which cities physicians with formal job contracts are working, as well as their wages, working hours, age and gender. The second database keeps records of all active physicians working in the public and private sectors, including the ones self-employed. With these two datasets we know the workplace of all active physicians – i.e., those acting as paid workers or associates in any private or public health facilities in the country.[18] We find the workplace of 78% of all generalist physicians within their first three years from graduation, leaving us with a final sample of 46,989 physicians.[19] Reassuringly, as the descriptive statistics in Table A1 show, physicians that made it into our final sample do not look systematically different from the ones we lose.

Table 1 column 1 shows physicians' decision of practice location. It depicts a similar scenario to the one in Figure 1. Physicians choose to work more in capitals and metropolitan areas, and more than 50 percent of all physicians took a job in the Southeast region. Only 26.6% of the generalists who graduated in the period started their careers in the North and Northeast regions, which includes more than 36% of the Brazilian population.[20]

Birth and medical school location. We create dummies equal to one for physicians that choose to work in their birth place or in the same location where they completed medical school. As Table 1 column 2 shows, 55% of doctors graduated in the Southeast. On the other extreme, the countrysides of the North, Northeast, and Midwest graduated only 1.14% of all physicians. A similar, but less extreme, pattern can be seen in physicians' birth location.

Quality of medical schools. We proxy the quality of Brazilian medical schools using

[18] If a physician works in more than one region, we consider only the one with higher working hours.

[19] We actually lose only 5.5% of generalist physicians in the merge with RAIS and CNES. We find the remaining 16.5% after more than three years from graduation. We do not consider these physicians because we cannot assure these regions were their first location choice after medical school.

[20] We also show that most physicians still work in the same region of their first job after graduation five years later (see Table A3). This supports our focus on the first job after medical school.

the ranks of university courses published by *Folha de São Paulo* newspaper in 2013. Brazil had 214 medical schools in our time frame, and 44% of medical schools in Brazil are in the Southeast region (INEP). The top 25 medical schools graduated 17% of the doctors in our sample, and these universities are mostly concentrated in metropolitan regions of the South, Southeast, and Midwest. Around 38% of generalists graduated in universities ranked outside the top 100. The mean and median rank of physicians in our sample are 83 and 71, respectively.

Physicians per 1,000 people. We compute the number of physicians per 1,000 people from CNES. This variable captures both competition effects (that is, more saturated local markets), and peer effects (more physicians could bring greater learning opportunities and networking). The numbers in Table 1 column 4 again resemble the physicians' concentration pattern shown in Figure 1.

Health infrastructure and private market. The lack of equipment, supplies and appropriate health units may affect physicians' decisions of where to work. To assess how physicians perceive the working environment in each region, we measured the availability of essential medical equipment to diagnose and treat patients using CNES data from 2005 to 2015 on the per capita number of ultrasound and x-ray machines, mammographs, computed tomography (CT) and magnetic resonance imaging (MRI) scanners across Brazilian regions. With these ratios, we developed a normalized index of health infrastructure, the same way as Kling et al. (2007). Table 1 column 5 shows that health infrastructure is the worst exactly in regions with the biggest deficiency in physicians, the countryside of the North and Northeast. The regions with the best health infrastructure are the metropolitan areas from the South, Southeast and Midwest.

Private health insurance coverage. As a proxy for opportunities in the private market, we use the coverage of private health insurance in each region, obtained from the National Regulatory Agency for Private Health Insurance and Plans (ANS). Column 6 shows, again, that the richest regions are also the ones with a higher percentage of the population covered by health insurance – 37.5% and 25.5% in the metropolitan regions of the Southeast and South against 1.8% and 2.4% in the countryside of the North and Northeast.

Amenities. Local amenities play an important role in labor sorting of skilled workers (e.g. Diamond, 2016). To summarize an area's observable quality of life,we computed a local amenity index which includes six features: (i) education, comprising the scores in a national exam of local elementary schools from first to fifth and sixth to ninth grades (INEP); (ii) entertainment, quantified by the number of cinemas, hotels, restaurants and recreation firms per capita – RAIS and the National Cinema Regulatory Agency (ANCINE); (iii) transportation – RAIS and the National Traffic Department (DENATRAN); (iv) violence measured by the number of violent deaths per capita – Mortality Information System (SIM/DATASUS); (v) local GDP per capita and; (vi) public investment by the state and municipal governments – National Treasure.[21] We combine all these variables into one amenity index as Kling et al. (2007).Table 1 column 7 presents the statistics of this amenity index by region. The imbalance follows the same pattern documented in the health infrastructure index with the countryside of the North and the Northeast, again, having the worst amenities.

Wages. We also construct a measure of physicians' expected compensation. For each of the 52 decision units we calculate the average wage per hour of recently graduated (below 35 years old)physicians using RAIS. Given that this data only covers formal jobs, this measure underestimates total physician income as it does not account for off-book earnings. However, the average hourly wage in the formal labor market is correlated with the earnings opportunities in the informal markets,[22] so this measure captures a meaningful variation in physicians' compensation over time and across regions. We adjust the expected wages for local purchase power. We calculate local living costs using the values of real estate rentals from the National Household Sample Survey (PNAD) and the 2010 Census.[23]

[21] The choice of these features was based on the amenity index developed by Diamond (2016), and adapted to the Brazilian data availability and context.

[22] It is difficult to rationalize that areas with higher formal wages tend to have lower informal wages.

[23] We regress the rental value against many property's characteristics and a set of dummies that identifies in which region the property is located (see Summers, 1973; Seabra and Azzoni, 2015). Potential sample

The last column in Table 1 shows the average wage per hour earned by recently graduated generalists (in 2010 BRL) adjusted by local living costs. Differently from the other variables, we see that physicians' real hourly wages are considerably higher in the countryside than in the corresponding metropolitan regions. Likewise, generalists in the poorest regions tend to earn more in real terms than those in the most developed areas. This indicates that less developed areas already pay a premium to physicians in order to compensate worse amenities and working conditions. These financial incentives, however, do not suffice to correct the large imbalance in the number of generalist physicians per capita across the country.

3.2 Descriptive Analysis

We present descriptive evidence that illustrates how the set of local attributes characterized in the last subsection may influence physicians practice location choices. Panel A in Table 2 describes practice location choices of physicians born in different regions of the country. Each cell (i, j) in the table has the fraction of physicians born in region i (row) that decided to work in region j (column). Analogously, the rows in Panel B indicate the region physicians completed medical school and columns their practice location choices.[24] The analysis of these two panels reveals interesting patterns.

First, as the main diagonal of both panels illustrates, most physicians prefer to stay in the same region they were born or completed medical school. Overall, 62.2 percent(60.6%) of the physicians in our sample stays in the same region they were born (completed medical school). This effect is even stronger for physicians that were born (or completed medical school) in the most developed areas of the country (South and Southeast). In particular, 74.1 percent(72.2%) of the physicians born (completed medical school) in the metropolitan areas of the Southeast continues to work in the same area after graduation. Similar patterns

selection issues related to the fact that rented households may differ in many ways from those not rented are addressed using Heckman (1979). For more details on how this measure is constructed, see E.

[24]This analysis does not consider the 52 alternatives in our setting, but the aggregation of them into ten bigger regions as displayed in the Table 1 - N, NE, SE, S and MW divided in metropolitan regions and countryside.

are observed for physicians that were born or completed medical school in the metropolitan areas of the South region. These numbers suggest that home and graduation place biases are very strong for physicians.

Second, the Southeast, the richest area of the country, attracts a relatively large number of physicians that were born (or completed medical school) in the other regions of the country. For example, 13.2 percent(8.3%) of the physicians that were born in metropolitan areas (countryside) of the North migrated to have their first job in the Southeast region. In total, approximately 10 percent of physicians born in the North move to the Southeast after graduation. Similar patterns are observed when we look at panel B. This fact is interesting because, as shown in Table 1, real wages in the Southeast region are relatively lower than in the poorest regions. On the other hand, the Southeast has better health infrastructure and amenities than the other regions.

The descriptive analysis appears to indicate that, after graduation, physicians tend to stay close to the place they were born or completed medical school. Physicians that migrate to a geographic region that is different from the geographic region they were born (or completed medical school) prefer regions with better health infrastructure and amenities but that pay relatively lower wages. Jointly, these evidence may indicate that wages are not as important as other local attributes to explain physicians locational preferences. In the next section, we analyze physicians' locational preferences using a structural model of supply and demand of physicians in Brazil.

4 Empirical Framework

This section characterizes our structural model and estimation procedure. The framework is intended to capture key characteristics of supply and demand of physicians in Brazil. The labor supply is a discrete choice model that describes physicians' practice location choices. Berry et al. (2004) show how detailed information on individual characteristics and choices can help researchers to identify individual preferences. As our data contains a wide set of information on physicians' characteristics and practice location choices, Berry et al. (2004) is the natural benchmark for our supply model. Next we develop a simple model of demand for physicians, and discuss its main assumptions and implications. We close the section by describing strategies that we employ to estimate the model.

4.1 Supply Side

We assume that right after graduating from medical school at year t physician i chooses a practice location j among $J \geq 1$ different practice locations. We define location as the pair $(state, area)$, where state is one of the 26 Brazilian states plus the Federal District; and area is either state capital including metropolitan region or countryside. Physician i's indirect latent utility from choosing location j is given by:

$$u_{ijt} = \sum_k x_{jtk}\tilde{\beta}_{ik} + \xi_j + \tilde{\xi}_j \cdot t + \tilde{\xi}_{jt} + \varepsilon_{ij}, \tag{1}$$

with,

$$\tilde{\beta}_{ik} = \beta^c_k + \sum_r z_{ir}\beta^o_{kr} + \beta^u_k v_{ik}. \tag{2}$$

In this model, the variables x_{jtk} represent observed attributes of location j at year t, such as local health infrastructure, indexes that capture quality of local amenities, etc. Physicians' real wages are included in the vector of observed attributes of each location. Analogously, $\tilde{\xi}_{jt}$ condenses local characteristics that are not in our data (e.g. quality of local restaurants, quality of cultural life, etc) and is left as an error term. We also include in the model a location fixed-effect, ξ_j, capturing unobserved attributes of location j that are constant over time (e.g. natural attributes) and $\tilde{\xi}_j \cdot t$ is a location specific time trend. In practice, as in Nevo (2000), ξ_j is modeled as location specific dummies and the term $\tilde{\xi}_j \cdot t$ is modeled as an interaction between a time trend and the location dummies. The remaining error term, ε_{ij}, represents an idiosyncratic preference that physician i has over location j and $\tilde{\beta}_{ik}$ represents the effect of a given observed attribute of location j at year t, say x_{jtk}, on physician i's indirect utility.[25]

The terms $\tilde{\beta}_{ik}$ are decomposed into a choice specific constant, β^c_k, observed physicians characteristics, z_{ir}, and unobserved physicians characteristics, v_{ik}. In other words, physician i's "tastes" for each observed attribute of location j at year t are allowed to vary according to their observed and unobserved characteristics. The components β^o_{kr} and β^u_k capture, respectively, the effects of observed and unobserved physicians characteristics on $\tilde{\beta}_{ik}$. Sometimes we use β^c, β^o and β^u to denote the vectors $\{\beta^c_k\}_k$, $\{\beta^o_{kr}\}_{kr}$ and $\{\beta^u_k\}_k$, respectively. The variables z_{ir} contain physicians attributes that are present in our data, such as age, gender and the quality of the medical school physician i graduated from. The variables v_{ik} contain physicians' characteristics that we do not observe (e.g. marriage status, number of children, etc).

Substituting equation (2) into equation (1) we obtain a model that governs physicians' practice location choice:

$$u_{ijt} = \delta_{jt} + \sum_{k,r} x_{jtk} z_{ir} \beta^o_{kr} + \sum_k x_{jtk} v_{ik} \beta^u_k + \varepsilon_{ij}, \tag{3}$$

where,

$$\delta_{jt} = \sum_k x_{jtk} \beta^c_k + \xi_j + \tilde{\xi}_j \cdot t + \tilde{\xi}_{jt}. \tag{4}$$

The formulation above captures two important features of our framework. First, substitution patterns across different locations are allowed to depend on observed and unobserved physicians' attributes. Physicians with different observed and/or unobserved characteristics give different weight for the same observed choice attribute. That is, it can be the case that local health infrastructure is more important for graduates from better schools than

[25] We are suppressing the time index t for all variables that are already indexed by i. We are doing this because each individual is observed only at the year they graduate from medical school, i.e., the index i also represents year of graduation. By using the index t together with the index i we would give to the reader the erroneous impression that each individual is observed at different points in time. However, keep in mind that observed choice attributes can be different for individuals graduating in different cohorts.

for graduates from worse schools, or that physicians that have children are more concerned with the quality of local amenities than physicians that do not have children. In practice, the inclusion of these interactions produces a model with very flexible substitution patterns (see Berry et al., 1995, 2004; Nevo, 2000). Second, the variables x_{jtk} summarize a finite set of attributes of location j that are relevant for i's decision process. However, as the list of relevant local aspects can be quite large and/or partially unobserved by the econometrician, not all the relevant characteristics of location j are included in our x_{jtk}. The role of $\tilde{\xi}_{jt}$, of the time invariant and time varying location fixed-effects – ξ_j and $\tilde{\xi}_j \cdot t$, respectively – is to account for all the relevant characteristics of location j affecting i's decision that are not included in x_{jtk}.

Physicians are assumed to choose a single location, $j \in \{1, 2, ..., J\}$ in order to maximize their utility – expressed by equations (3) and (4). This defines a set of unobserved individual/location attributes that is associated with the choice of each location. Under distributional assumptions on unobservables – which will be specified later in this section – we can integrate this distribution over the set of individual and local characteristics leading to each choice and obtain the probability of any given physician i choosing any given location j, $s_{ijt}(\mathbf{x_t}, \mathbf{z_i}; \theta)$, as a function of preference parameters, θ, observed individual characteristics, $\mathbf{z_i}$, and observed location characteristics, $\mathbf{x_t}$. We will precisely characterize these probabilities later in this section. Now we turn to the demand model.

4.2 Demand Side

The demand of physicians is highly dependent on the public sector. This makes the task of modeling the demand side more involved as the objective function of public administrators may be influenced by a myriad of idiosyncratic factors. On the other hand, evidences shown in Section 2 suggest that there exists considerable excess demand for physicians in the Brazilian market. This is consistent with a model where wages are not set in order to clear the market. Based on these observations, our baseline model assumes that the demand for physicians is inelastic and depends only on observed and unobserved characteristics of each location.

More specifically, we assume that wages, wjt, are a linear function of the other observed location attributes (except wages) including location and year fixed effects, $\tilde{\mathbf{x}}_{\mathbf{jt}}$, an instrumental variable (that will be specified later), h_{jt}, and an error term, η_{jt}:

$$w_{jt} = \tilde{\mathbf{x}}_{\mathbf{jt}}\gamma + h_{jt}\lambda + \eta_{jt}. \tag{5}$$

The instrumental variable, h_{jt}, does not enter utility directly but affects wages. The vector (γ, λ) contains the parameters of the wage equation. We further assume that η_{jt} and $\tilde{\xi}_{jt}$ are uncorrelated with $\tilde{\mathbf{x}}_{\mathbf{jt}}$ and h_{jt} but are not independent of each other. In other words, wages in location j depend on local observed attributes and an idiosyncratic term, η_{jt}, that may be correlated with unobserved local attributes, $\tilde{\xi}_{jt}$. The correlation between η_{jt} and $\tilde{\xi}_{jt}$ is captured by the following process:

$$\tilde{\xi}_{jt} = \eta_{jt}\psi_1 + \tilde{\eta}_{jt}\psi_2, \tag{6}$$

where, ψ_1 and ψ_2 are parameters to be estimated and $\tilde{\eta}_{jt}$ is an error term.

We also consider two different formulations for the demand model: one where local governments are oligopsonists and set wages to maximize a welfare function (without any constraint) and one where they are subject to a budget constraint which is always binding. We describe and estimate these models in Appendix C. The fitting to the data and the estimates of the budget constraint model are very similar to the baseline model, the oligopsony model fits worse to the data than the other two models.

4.3 Estimation

The main issue behind the estimation of the supply model is that wages at any given location are likely to depend on unobserved location attributes, $\tilde{\xi}_{jt}$ – which is assumed to be known by physicians and public administrators but not by the econometrician – see equations (5) and (6). To deal with this issue, we apply the control function approach pioneered by Heckman and Robb (1985) and adapted by Petrin and Train (2010) for the estimation of Random Utility Models.[26]

We define the instruments for wages in location j, period t as the average value of observed attributes of other locations except location j that are in the same geographic region[27] as location j, period t – i.e., the average of variables x_{jtk} in equation (1), except wages, across all $j' \neq j$ that are at the same geographic region as location j (see Berry et al., 1995). The intuition for this instrument is that health facilities in a given region may also take into consideration the characteristics of neighboring regions when setting wages. For example, if neighboring region "B" has better health infrastructure or amenities than region "A", health facilities at region "A" may have to increase their wages in order to attract physicians. This approach will be valid if observed location attributes are determined exogenously (see Nevo, 2000). As in other papers in the literature, the main pitfall behind these instruments is that observed local attributes can be correlated with unobserved local attributes. We believe, however, that the inclusion of location and year fixed effects may help to mitigate this problem.

We next define physicians' location choice probabilities based on equations (3), (4), (5) and (6). Given the instrument, h_{jt}, we can recover the variable η_{jt} via OLS from equation

[26]Another popular method to solve endogeneity problems of this type is the one developed in Berry et al. (1995). This method is not suitable for our application because the observed probabilities of some individuals choosing some locations are zero. In this case the contraction mapping used to recover ξ_{jt} does not converge and we cannot form the moment conditions necessary to recover the parameters of interest. According to Petrin and Train (2010) both procedures produce very similar results.

[27]Brazil is divided in five geographic regions (North, Northeast, Midwest, Southeast and South). These regions have similar geographic and economic characteristics. To calculate the instrument we consider the pairs Region-Metropolitan Area and Region-Countryside.

(5), so we proceed as if η_{jt} and (γ, λ) are known. This term is our control function. It captures the correlation between wages and unobserved local attributes. Therefore, the observed variables of the model are x_{jtk} and z_{ir}, the unobserved variables are v_{ik}, $\tilde{\eta}_{jt}$ and ε_{ij}, and the parameters are $\theta = \left(\psi_1, \psi_2, \beta^c, \beta^o, \beta^u, \{\xi_j\}_j, \left\{\tilde{\xi}_j\right\}_j\right)$.

To obtain physicians' choice probabilities, we still have to specify the joint distribution of the unobserved variables, v_{ik}, $\tilde{\eta}_{jt}$ and ε_{ij}. Following Petrin and Train (2010) and Berry et al. (2004) we assume that: *(i)* ε_{ij} is iid across i and j with Extreme Value distribution; *(ii)* the unobserved individual characteristics, v_{ik}, are iid across i and k with a standard normal distribution; and *(iii)* the term $\tilde{\eta}_{jt}$ is iid across j and t with standard normal distribution. Based on these assumptions, the probability of physician i graduating at year t choosing practice location j as a function of the the vector of parameters θ and the observed individual and location characteristics can be expressed as:

$$s_{ijt}\left(\mathbf{x_t}, \mathbf{z_i}; \theta\right) = \int \frac{exp\left(\delta_{jt} + \sum_{k,r} x_{jtk} z_{ir} \beta^o_{kr} + \sum_k x_{jtk} v_{ik} \beta^u_k\right)}{\sum_q exp\left(\delta_{qt} + \sum_{k,r} x_{qtk} z_{ir} \beta^o_{kr} + \sum_k x_{qtk} v_{ik} \beta^u_k\right)} dF_{\mathbf{v}} dF_{\tilde{\eta}}, \tag{7}$$

where, $\delta_{jt} = \sum_k x_{jtk}\beta^c_k + \xi_j + \tilde{\xi}_j \cdot t + \eta_{jt}\psi_1 + \tilde{\eta}_{jt}\psi_2$, $F_{\mathbf{v}}$ is the cumulative distribution of unobserved individual tastes v_{ik}, and $F_{\tilde{\eta}}$ is the cumulative distribution of the error terms, $\tilde{\eta}_{jt}$.

We first estimate equation (5) by OLS and recover the error term η_{jt}. This term, along with the observed variables, is plugged into the integral above. The integral is approximated via simulation. The terms v_{ik} and $\tilde{\eta}_{jt}$ are drawn from a standard normal distribution. For each draw, the logit equation inside the integral is calculated. This process is repeated 150 times – i.e., for each individual in our sample we draw a sequence of 150 $(\mathbf{v_i}, \tilde{\eta})$ vectors from $F_{\mathbf{v}}$ and $F_{\tilde{\eta}}$. We calculate the integral in (7) as the average across draws of the logit formula. We estimate the vector of parameters, θ, via Simulated Maximum Likelihood.

Having estimated $s_{ijt}\left(\mathbf{x_t}, \mathbf{z_i}; \theta\right)$, the aggregate supply of physicians at each location-year can be calculated taking draws – we take in total 300 draws for each individual in our sample – from the distribution (7) and aggregating the outcomes.

5 Labor Supply Estimates

This section presents the estimates of physicians' supply function and shows the fitting of our model.

5.1 Labor Supply Estimates

We estimate four different versions of the supply model. Tables 4 and 5 show the estimates of the parameters β^c and β^o, and Table 6 shows the estimates of β^u – see equation (2).[28] The first two columns in Table 4 illustrates the estimates of the Logit model – i.e. the version of the full model where β^u is restricted to zero. The last two columns have the Random Coefficients Logit estimates. In both models, we present estimates without (columns 1 and 3) and with (columns 2 and 4) correction for endogeneity of wages – that is, using the control function.[29]All equations include location fixed effects and an interaction between location fixed effects and year. The observed local attributes used in all specifications were described in Section 3. All variables were normalized to be between 0 and 1, thus the magnitudes of the coefficients are comparable across different variables. As a robustness, we estimate the model restricting physicians' choice set according to the historical placement of each medical school.[30]

[28]Table 6 has only two columns, one for each of the two versions of the Random Coefficients Logit.

[29]The inclusion of the control function in equation 7 biases maximum likelihood standard-errors. We attempted to correct our standard-errors using bootstrap. However, as the random coefficients model takes on average 3 days to run, the bootstrap method for the random coefficients model showed to be computationally unfeasible. We computed bootstrap standard-errors for the logit models only. We observed that the differences between bootstrapped standard-errors and maximum likelihood standard-errors are minimal. We take this as evidence that the bias in the standard-errors are of second order.

[30]Precisely, we identify the set of locations graduates from each medical school chose over our time frame. We, then, estimate a version of the model restricting the choice set of all students in each medical school to this set locations. As Appendix Tables A8, A7 and A9 show, results are quantitatively similar.

Overall, the sign of our estimates are as expected in both the Logit and Random Coefficient models and in consonance with the descriptive evidence in Section 3. Health infrastructure and amenities seem to increase physicians' utility and are statistically significant, suggesting that physicians value working and living conditions. The dummies of place of birth and local where the physician graduated are also positive and significant, meaning that home bias and moving costs are taken into account in the choice of work location. On the other hand, the number of physicians per capita and the coverage of health insurance are not statistically significant.

The main difference between both models is the sign of wages. In the models we do not use control function (columns 1 and 3), wages have a negative but small sign. However, when we use the control function (columns 2 and 4), the coefficients attached to wages are positive and statistically significant. We report the first stage estimates – parameters (γ, λ) in equation (2) – in Table 3. The coefficient attached to η_{jt} – from equation (6) displayed in the last row of second column – is also negative and statistically significant, suggesting that wages and unobserved local attributes are negatively correlated. This result is also expected: health facilities in areas where the value of unobserved local attributes is higher may be able to attract physicians paying lower wages. Similar patterns are also observed in studies of demand for differentiated products.[31]

Table 6 reports the coefficients attached to the interactions between local attributes and unobserved physicians' characteristics. In both models only the interactions with place of birth and place where the physician completed medical school are significant. This suggests that unobserved physicians' characteristics may help to explain practice location choices.

[31] This type of endogeneity induces a positive bias in the coefficient attached to prices in studies of demand for differentiated products, as prices are expected to be positively correlated with unobserved product attributes (see Berry et al., 1995; Nevo, 2000).

5.1.1 Wage elasticity

Table 7 reports the wage elasticities implied by the Random Coefficients Logit with control function. Each row corresponds to one Brazilian state. The first two columns show own elasticities and the minimum of cross wage elasticities for the state capital and its metropolitan area. The other two columns have the same numbers for the countryside of each state.[32] These results indicate that physicians' supply function is inelastic. The highest own wage elasticity across capitals and metropolitan areas is around 0.86 and the lowest is around 0.13. Elasticities in the countryside are relatively higher than in capitals and metropolitan areas for most states. These numbers are in line with previous estimates found in the health literature.[33] They suggest that policies targeting (only) pecuniary incentives may not have a decisive effect on the redistribution of physicians across the country. In the next section, we quantify the implication of this type of policy.

5.1.2 Heterogeneous preferences

Finally, in all specifications a number of interactions between local attributes and physicians' observed characteristics is statistically significant. Our estimates suggest that while men value amenities relatively more than women, they derive lower value for staying in the locale of their graduation and those born in metropolitan regions seem to be less inclined to work close their birth place. We interpret this as suggestive evidence that the fixed cost of migration is lower for men than for woman.[34]

[32] The Federal District (DF), by definition, has no countryside. For Santa Catarina (SC), we aggregated capital and countryside because SC is a small state and capital (plus metropolitan area) covers a large fraction of the population living in the state.

[33] Baltagi et al. (2005) use a dynamic panel of 1303 male physicians in Norway to estimate a standard supply function (hours worked on wages) and find a wage elasticity of 0.33. Andreassen et al. (2013) studies how wages affect physicians choices over 10 different combinations of jobs packages that differ with respect to working load, place of work (hospitals or primary care) and type of institution (public and private) in Norway. Wage elasticities obtained in this study are around 0.04. We refer the reader to these papers for a detailed survey on the topic. The general conclusion of these studies is that the supply of physicians is inelastic.

[34] It is difficult to interpret the coefficients of the interaction with physician age because this variable has small variance (see Section 3).

More importantly, physicians' taste seem to be meaningfully different according to the rank of the course each physician graduated from. Because the university rank is closely related to physician's qualifications, this heterogeneity in taste may have important policy implication. Note that the the quality of the medical school decreases with university rank, for instance the top medical school in Brazil has a university rank index equal to zero, and the universities with the worse evaluations have a rank index close to one. We see that those who graduated from better medical schools derive higher utility from local amenities, have lower wage elasticity, and derive lower utility from returning to their region of birth. Those who graduated from better quality schools in metropolitan regions derive the greatest utility from staying in their local of graduation.[35]

5.2 Model Fitting

Table 8 shows the fitting of our model. For each Brazilian geographic region – North (N), Northeast (NE), Southeast (SE), South (S) and Midwest (MW). The first column shows the distribution of physicians as observed in the data. The second column reports the number of physicians in each geographic region/area as predicted by the model. The third column has the frequency of physician choices that were correctly predicted by the model;[36] we assume that physician i has their choice correctly predicted by the model if the probability of going to location j is higher than the probability of going to any location $j' \neq j$ and, in the data, we observed physician i choosing location j. Column four (five) is the ratio between column two (three) and one.

In general, the model overestimates the number of physicians in capitals and metropolitan areas and underestimates in the country side. Columns 4 and 5 suggest that on average the model predicts correctly the locational choice of 58.3 percent of the physicians in our sample.

[35]Although these results may indicate that – observed – heterogeneity in physicians' tastes for different local attributes must be taken into consideration in our counterfactuals, an exercise in the next section suggests that it is not a prominent factor underlying physicians' distribution.

[36]Although the results are shown using ten aggregated regions, the correctly predicted frequencies were calculated by assessing the 52 alternatives.

Notice that as choice sets have 52 alternatives, a model that randomly allocates physicians to all possible locations would predict correctly the choice of approximately only 1.92 percent (1/52) of the physicians in our sample.

6 Counterfactuals and Policy Analysis

This section reports the results of a series of counterfactual studies and discusses how these exercises can inform the policy debate. Our main counterfactual question is: what types of policies are more effective to reduce the regional imbalances in the geographic distribution of physicians in Brazil? To answer this question we compare the effects of four policies on the geographic distribution of physicians: financial incentives (increases in wages in poorer areas), improvement of health infrastructure in underdeveloped regions, redistribution of medical schools across the country, and affirmative action (quotas) according to place of birth. We also quantify the importance of heterogeneity in physicians' tastes – according to the quality of medical schools – for different local attributes. We close the section discussing policy implications of our results and calculating the cost-effectiveness of these different policies.

6.1 Counterfactuals

To quantify the relative importance of each local or individual characteristic to the observed geographic imbalance of physicians, we assess physicians choices when we homogenize each characteristic across regions – i.e., we set the value of each variable to zero in all regions.[37] We evaluate these different counterfactual distributions against the population share in each region, shown in Table 9 column 1. We chose the population benchmark following the WHO recommendation that a minimum of 2.3 health workers per thousand people are necessary for the provision of basic health services (WHO, 2006).

[37]Note that in discrete choice models only differences in utility matter, so homogenizing one characteristic in all choices to zero or one would entail the same results.

Column 2 shows the predicted distribution of physicians according to our random coefficient model with control function. The quadratic error of 0.59 depicts the baseline distribution imbalance originated from our model relative to the population distribution.

Table 9 columns 3 to 7 present the geographical distribution generated by our counterfactual scenarios. We first simulate physicians' choice if the share of physicians place of birth in each medical school was proportional to the population of each region. Column 3 shows that this scenario would improve the distribution of physicians by more than 51 percent by increasing the share of physicians choosing to work in the countryside, especially the Northeast countryside. Column 4 shows the counterfactual distribution if medical schools were distributed across regions in proportion to the local population. The distribution of doctors across the country would improve by 39 percent.

In the following counterfactuals, we level health infrastructure, amenities and wages across the country. Table 9 column 5 shows that if health infrastructure was geographically uniform, the distribution of physicians would improve by 24 percent. Analogously, column 6 shows that if local amenities were similar everywhere, doctors' distribution would improve by only 4 percent. These results suggest that while the quality of local amenities affects physicians' locational choice, health infrastructure seems to be a barrier to the recruitment of physicians in poorer areas.

These counterfactuals also help to explain why the wage premium in impoverished areas needs to be larger to compensate the other less desirable local traits or home biases. We find that wage differentials contribute to improve the distribution of physicians. If real wages were homogeneous countrywide, as in column 7, the distribution of physicians would be 13 percent worse than in the baseline distribution. In this scenario, the share of physicians in the countryside, places that already pay a wage premium, would be even smaller than the baseline distribution, and the concentration in state capitals of the South and Southeast regions would be even larger.

Finally, to quantify the importance of (observable) taste heterogeneity on the geographical distribution of physicians, we make two counterfactuals. We simulate the geographic distribution of physicians assuming that they have the same preferences as physicians graduated from (i) the first and (ii) the last medical school in the universities ranking. The mean quadratic error considering the tastes of graduates from the last ranked school is 0.61, whereas the quadratic error considering the tastes of the graduates from the first school is 0.64. This result seems to indicate that increasing/decreasing the quality of physicians' training would not produce a major change in the distribution of physicians across the country.

6.2 Policy Analysis

We now discuss the effect and cost-effectiveness of four different policy counterfactuals. Despite the challenge involved in obtaining the monetary costs for these counterfactual policies, we try to provide meaningful back-of-the-envelope calculations using the best information available. The goal of this exercise is to have an internally consistent way to compare the cost-effectiveness of policies acting on different margins of physicians' preferences.

6.2.1 Medical school quotas based on place of birth

Home bias seems to be quite important to explain physicians' location decisions in Brazil. Based on this evidence we first consider an affirmative action policy that sets quotas in medical schools according to the fraction of the population living in each area of the country. For example, if region A has 30% of the population and region B has 70%, then 30% (70%) of the vacancies in all Brazilian medical schools would be allocated to students born in region "A" ("B"). This policy could be implemented, for example, through a centralized admission system resembling the Brazilian Unified Selection System (SISU) already in place.[38]

[38] Machado and Szerman (2016) find that SISU enabled universities to attract more students from different states, with no effect on registration or dropouts. In the United States, Fitzpatrick and Jones (2016) find that merit aid programs targeted at state-born individuals increase the likelihood that residents live in their home state after graduation.

Note that as this policy does not change the location of medical schools, only the composition of their students, it is relatively cheap to be implemented.[39]

According to our counterfactual, this policy has a powerful result: implementing this quota system would improve the distribution of physicians in the country by 51.5 percent – see Table 9 column 3.

6.2.2 Targeted creation of new vacancies in medical schools

Now, we analyze how the distribution of physicians across the country would be affected by the redistribution of medical schools towards regions lacking generalist physicians. As we estimate sizable migration costs, such policy could help to keep physicians in a specific area. We provide the cost-effectiveness of this policy when the policy is implemented through expansion of vacancies in public universities or vouchers for private universities.

According to SIGRAS/INEP, there were on average 12,174 vacancies available each year between 1996-2008, the period the physicians in our sample started medical school. The federal government expanded the number of vacancies in public universities and extended the number of government-funded scholarships for private colleges (FIES). As a result, from 2009-2016, there were on average 22,515 vacancies available in medical schools each year, an increase of 10,341 vacancies/year relative to the previous period. Considering that around 40.49 percent of physicians remain generalist, we conjecture in our model that 4,187 vacancies/year were opened in the eight years subsequent the years we study. We create a scenario where these 4,187 vacancies/year are opened in places with the lowest student-population ratio. We add vacancies to each region in an interactive way as to guarantee that the regions with the lowest student-population ratio receive new vacancies first.[40]

[39] The main hidden cost of this type of policy is a potential loss in efficiency of the educational system caused by a mismatch between school and student quality. Evidence on the effect of affirmative actions on graduation rates and earnings are weak – Holzer and Neumark (2000) and Arcidiacono and Lovenheim (2016) review the literature. Estevan et al. (2018) show that an affirmative action policy targeting underprivileged students in Brazil did not distort the incentives of targeted applicants while still in high-school, as captured by test performance or application decision.

[40] When we add a vacancy to one of the 52 alternatives, we randomly duplicate physicians graduated in the region. If no physician graduated in the region (some areas have no medical school) we randomly select

Table 10 column 3 present the counterfactual placement of physicians across regions under this policy. It increases the share of physicians in the North and Northeast countryside by more than 70 percent, attracting mostly physicians from the Northeast metropolitan areas and from the Southeast countryside. As a result, we find that opening new university vacancies in needy areas improve the distribution of physicians by 52.5 percent – it halves the quadratic error between the difference of population and physicians distributions.

Considering the average cost of $13,796 for each undergraduate student in a public university,[41] the total cost of creating 10,341 federal vacancies would be $143 million.[42] This cost likely underestimates the real cost as it is based on the cost per student across all courses, and medical school is one of the most expensive programs. To produce an upper bound to the cost of such program, we can compute what would be the cost of the program if supplied through vouchers to private universities. The average yearly tuition of private medical schools in Brazil is around $32,295 per year.[43] The cost of implementing this policy using a voucher policy would be $334 million per year.

Thus, using the targeted expansion of vacancies in medical schools to improve the spatial allocation of physicians by 1 percentage point would cost between 2.7 and 6.4 million per year.

6.2.3 Improving health infrastructure in N/NE countryside

Our estimates suggest that physicians weigh the local health infrastructure when choosing their work place. Analogous to the previous exercise, we consider a policy that improves local health infrastructure, such as investing in X-ray machines, MRAs and other medical and hospital equipment. To implement this counterfactual exercise we improve the health infrastructure index by 50 percent in the countryside of the North and Northeast. The

a physician from the same region/CS or region/MR to duplicate (region here comprehends N, NE, SE, S, and MW), changing their place of graduation to the one where we are adding a vacancy.

[41] Ministry of Education, MEC Technical Note No. 4/2018, page 13.

[42] We deflate all costs to BRL in 2010, and convert to USD using the exchange rate 1.779 BRL to USD from January 2010.

[43] https://www.escolasmedicas.com.br/mensalidades.php accessed on February 21, 2019.

counterfactual distribution of physicians resulting from this policy – shown in Table 10 column 4 – suggests that better infrastructure helps to attract 14 percent more physicians to the targeted areas, improving the regional imbalance by 8.6 percent only.

Calculating the cost of such policy is challenging. Our strategy is to obtain the cost of a 0.1 increase in the health infrastructure index. We focus on the Federal District (DF) to avoid double counting in the way municipalities and states inform their public spending. We quantify the improvement in the health infrastructure index over 2005-2012 and calculate the total amount invested in health, excluding the wage bill.[44] This accounts for fixed cost investments in health infrastructures. However, improving infrastructures also increases the variable operational cost of the public health system – new equipment needs additional maintenance, and better service provision may increase demand. To capture these costs we consider the amount spent in health investment as the fixed cost component, and the health operational costs (i.e., the total cost minus expenses with personnel and investments) the variable operational cost. Between 2005 and 2012, the Federal District's health infrastructure index increased 0.1855 for an investment of $248.1 million in health infrastructure (fixed cost). During the same period health operational costs increased $95 million per year. Splitting the fixed cost over this 7-year period, we calculate that the cost of improving the health infrastructure by 0.1 in the period was $70.3 million per year.[45]

Based on these numbers, we calculate that this policy – i.e. an increase of 50 percent in the infrastructure index of the North and Northeast regions – would cost $600 million per year. Therefore, using investments in health infrastructure to improve one percentage point in the distribution of physicians would cost $69.8 million per year, of which $50.8 million are due to increased operational costs and $19 million due to fixed costs. Note that attracting more and better physicians is not the only consequence from improving health infrastructure. Investing in clinics and hospitals infrastructure directly improves the quality

[44] Health expenditure data from the Information System on Public Budgets in Health (SIOPS).

[45] This calculation overestimates the annual cost of the policy by effectively assuming that the new infrastructure fully depreciates in this 7-year period.

of health provision and probably health outcomes and, as a side effect, help recruiting doctors to work in the region.

6.2.4 Increasing wages in N/NE countryside

Last, we consider a 50 percent bonus on wages for physicians that begin their first job in the countryside of the North or Northeast. This is likely the most studied policy used to recruit qualified personnel to underserved areas.[46] The first two columns of Table 10 show the population distribution (our benchmark) and the share of physicians who chose to work in each region according to our original model. Column 5 shows the counterfactual placement of physicians under the wage bonus scheme. Consistent with the own wage elasticities shown in Table 7, a 50 percent wage increase attracts about 26 percent more physicians to these regions, and therefore improves the national allocation of generalists by around 13.4 percent. Note that such policy would reduce the number of physicians in other underserved areas.

The status quo wage bill of physicians in these areas is around $2,490 million per year – considering 40 hours per week work contracts.[47] The 50 percent bonus would increase the wage bill on $144 million per year. This means that such wage incentive scheme would would cost about $10.7 million for each percentage point improvement in the distribution of physicians.

7 Conclusion

We exploit revealed preferences of all generalist physicians graduated in Brazil between 2001 and 2013 to estimate physicians' practice location preferences using a random coefficients discrete choice model. We have detailed information on physicians characteristics – including the quality of the school from which they graduated –, practice location choices and attributes of these choices. Our estimates indicate that physicians tend to stay in the same

[46] E.g., Dal Bo et al. (2013); Andreassen et al. (2013); Kennan and Walker (2011); Ashraf et al. (2016); Finan et al. (2017).

[47] Data from RAIS, considering generalist physicians below 35 years old.

area as where they were born or completed medical school. These two types of geographic biases are much more important to explain physicians' location choice than wages, quality of health infrastructure and amenities. In particular, our estimates indicate that physicians are inelastic to wages. This result may explain why programs purely based on financial incentives often struggle to attract physicians to underserved areas.

We take advantage of the structural model to simulate the effects of different policies on the geographic distribution of physicians in Brazil. Our back-of-the-envelope cost calculations suggest that policies exploiting physicians' home bias and place of graduation bias are the most cost-effective. Affirmative action policies in the form of quotas on student enrollment aimed at increasing the proportion of students born in underserved areas in medical schools appear to greatly improve the geographic distribution of physicians at little cost. The opening of new vacancies in medical schools in areas lacking generalist physicians is also cost-effective. Offering even larger wage premiums for doctors in needy areas also appears to be effective, but at a much higher annual cost than the two previous policies.

References

Agarwal, N. (2015). An Empirical Model of the Medical Match. *American Economic Review*, 105(7):1939–78.

Albouy, D. (2009). The Unequal Geographic Burden of Federal Taxation. *Journal of Political Economy*, 117(4):635–667.

Albouy, D. and Stuart, B. (2014). Urban Population and Amenities: The Neoclassical Model of Location. NBER Working Papers 19919.

Alva, C., Bobba, M., Neilson, C., Leon, G., and Nieddu, M. (2017). Teacher Wages, Student Achievement and the Recruitment of Talent in Rural Peru.

Anand, S. and Barnighausen, T. (2004). Human resources and health outcomes: Cross-country econometric study. *Lancet*, 364(9445):1603–1609.

Andreassen, L., Di Tommaso, M. L., and Strøm, S. (2013). Do Medical Doctors Respond to Economic Incentives? *Journal of Health Economics*, 32(2):392–409.

Arcidiacono, P. and Lovenheim, M. (2016). Affirmative Action and the Quality-Fit Trade-off. *Journal of Economic Literature*, 54(1):3–51.

Ashraf, N., Bandiera, O., and Lee, S. S. (2016). Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the delivery of Public Services.

Baltagi, B. H., Bratberg, E., and Holmås, T. H. (2005). A Panel Data Study of Physicians' Labor Supply: the Case of Norway. *Health Economics*, 14(10):1035–1045.

Banerjee, A., Deaton, A., and Duflo, E. (2004). Wealth, Health, and Health Services in Rural Rajasthan. *AER Papers and Proceedings*, 94(2):326–331.

Bau, N. and Das, J. (2017). *The Misallocation of Pay and Productivity in the Public Sector: Evidence from the Labor Market for Teachers*. The World Bank.

Bayer, P., Ferreira, F., and McMillan, R. (2007). A Unified Framework for Measuring Preferences for Schools and Neighborhoods. *Journal of Political Economy*, 115(4):588–638.

Bayer, P., McMillan, R., Murphy, A., and Timmins, C. (2016). A Dynamic Model of Demand for Houses and Neighborhoods. *Econometrica*, 84(3):893–942.

Berry, S., Levinsohn, J., and Pakes, A. (2004). Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market. *Journal of Political Economy*, 112(1):68–105.

Berry, S. T., Levinsohn, J., and Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, 63(4):841–890.

Bishop, K. C. and Murphy, A. D. (2011). Estimating the Willingness to Pay to Avoid Violent Crime: A Dynamic Approach. *American Economic Review*, 101(3):625–29.

Bjorkman, M. and Svensson, J. (2009). Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda. *Quarterly Journal of Economics*, 124(May):735–769.

Bolduc, D., Fortin, B., and Fournier, M.-A. (1996). The Effect of Incentive Policies on the Practice Location of Doctors: A Multinomial Probit Analysis. *Journal of Labor Economics*, 14(4):703.

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

Dal Bo, E., Finan, F., and Rossi, M. (2013). Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service. *Quarterly Journal of Economics*, 128(3):1169–1218.

de Bekker-Grob, E. W., Ryan, M., and Gerard, K. (2012). Discrete Choice Experiments in Health Economics: a Review of the Literature. *Health economics*, 21(2):145–172.

Deserranno, E. (2019). Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda. *American Economic Journal: Applied Economics*, 11(1):277–317.

Diamond, R. (2016). The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000. *The American Economic Review*, 106(3):479–524.

Dunne, T., Klimek, S. D., Roberts, M. J., and Xu, D. Y. (2013). Entry, Exit, and the Determinants of Market Structure. *The RAND Journal of Economics*, 44(3):462–487.

Elias, P. E. M. and Cohn, A. (2003). Health reform in Brazil: Lessons to consider. *American Journal of Public Health*, 93(1):44–48.

Estevan, F., Gall, T., and Morin, L.-P. (2018). Redistribution without Distortion: Evidence from An Affirmative Action Programme At a Large Brazilian University. *The Economic Journal*, 129(619):1182–1220.

Falcettoni, E. (2018). The Determinants of Physicians' Location Choice: Understanding the Rural Shortage.

Finan, F., Olken, B., and Pande, R. (2017). The Personnel Economics of the Developing State. *Handbook of Economic Field Experiments*, 2:467–514.

Fitzpatrick, M. D. and Jones, D. (2016). Post-Baccalaureate Migration and Merit-Based Scholarships. *Economics of education review*, 54:155–172.

Francois, P. (2000). 'Public Service Motivation' as an Argument for Government Provision. *Journal of Public Economics*, 78(3):275–299.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Heckman, J. J. and Robb, R. (1985). Alternative Methods for Evaluating the Impact of Interventions. *Journal of Econometrics*, 30(1-2):239–267.

Holmes, G. M. (2005). Increasing Physician Supply in Medically Underserved Areas. *Labour Economics*, 12(5):697–725.

Holzer, H. and Neumark, D. (2000). Assessing affirmative action. *Journal of Economic Literature*, 38(3):483–568.

Kennan, J. and Walker, J. R. (2011). The Effect of Expected Income on Individual Migration Decisions. *Econometrica*, 79(1):211–251.

Kling, J., Liebman, J., and Katz, L. (2007). Experimental Analysis of Neighborhood Effects. *Econometrica*, 75(1):83–119.

Kruk, M. E., Johnson, J. C., Gyakobo, M., Agyei-Baffour, P., Asabir, K., Kotha, S. R., Kwansah, J., Nakua, E., Snow, R. C., and Dzodzomenyo, M. (2010). Rural Practice Preferences Among Medical Students in Ghana: A Discrete Choice Experiment. *Bulletin of the World Health Organization*, 88(5):333–341.

Kulka, A. and McWeeny, D. (2018). Rural Physician Shortages and Policy Intervention.

Leon, M. P. (2018). Inducing Teacher Retention in Remote Locations: Evidence from Peru.

Machado, C. and Szerman, C. (2016). Centralized Admission and the Student-College Match. IZA Discussion Papers 10251.

Nevo, A. (2000). A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand. *Journal of Economics & Management Strategy*, 9(4):513–548.

Petrin, A. and Train, K. (2010). A Control Function Approach to Endogeneity in Consumer Choice Models. *Journal of Marketing Research*, 47(1):3–13.

Rao, K. D., Zubin, S., Sudha, R., Khandpur, N., Seema, M., Indrajit, H., Mandy, R., Peter, B., and Marko, V. (2012). *How to Attract Health Workers to Rural Areas? Findings from a Discrete Choice Experiment in India*, volume 6. The World Bank, Washington, D.C.

Scheffer, M., Cassenote, A., and Aureliano, B. (2015). Demografia Médica No Brasil 2015. Technical report, Conselho Regional de Medicina do Estado de São Paulo.

Seabra, D. and Azzoni, C. (2015). Custo de Vida Comparativo para os Distritos das 100 Maiores Cidades Brasileiras. *Temas de Economia Aplicada*, pages 12–24.

Summers, R. (1973). International Price Comparisons Based Upon Incomplete Data. *Review of Income and Wealth*, 19(1):1–16.

WHO (2006). World Health Report 2006: Working Together for Health. Technical report, World Health Organization.

WHO (2010). Increasing access to health workers in remote and rural areas through improved retention. Technical Report February, World Health Organization.

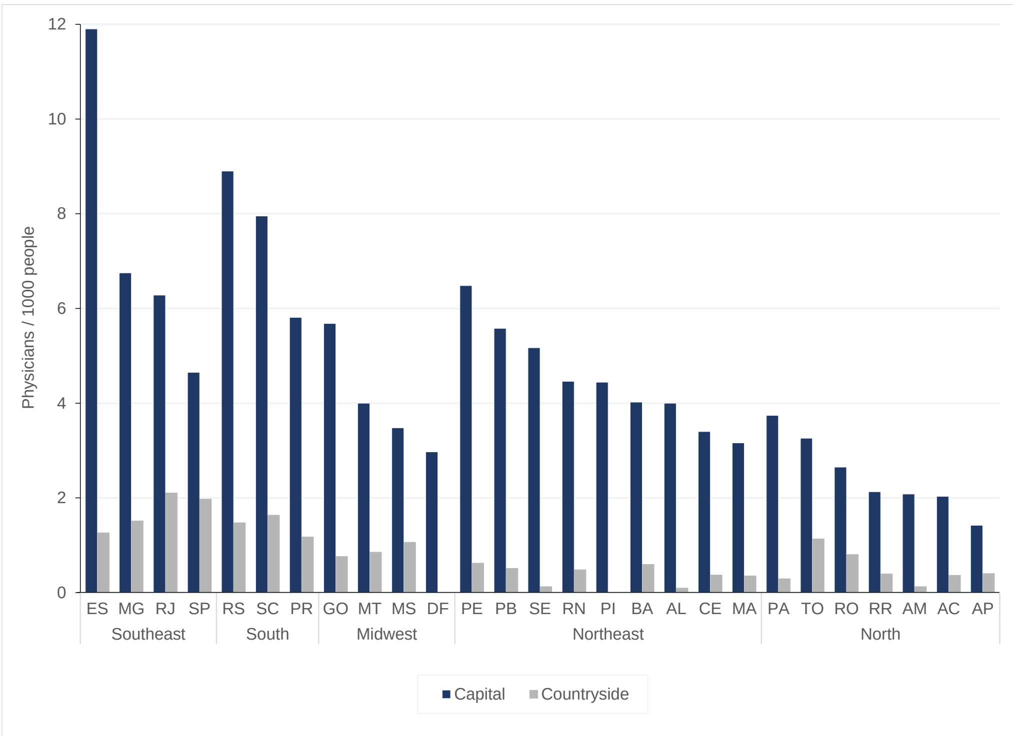
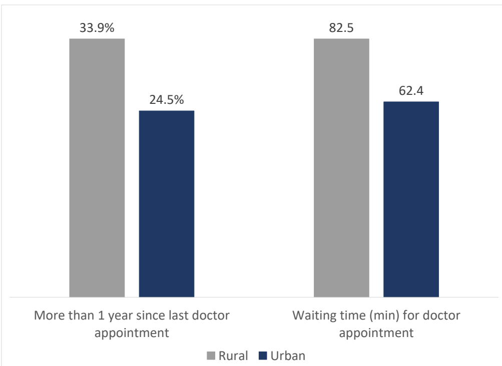
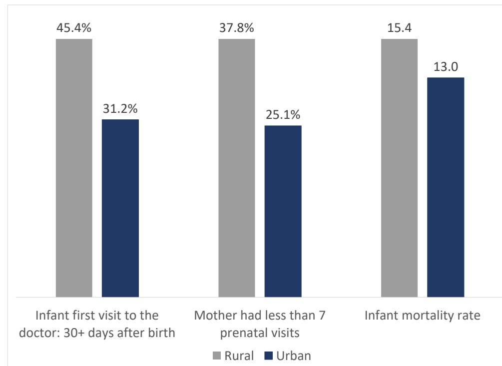


Figure 1: Physicians in Capitals and Countryside by State (per 1,000 people)

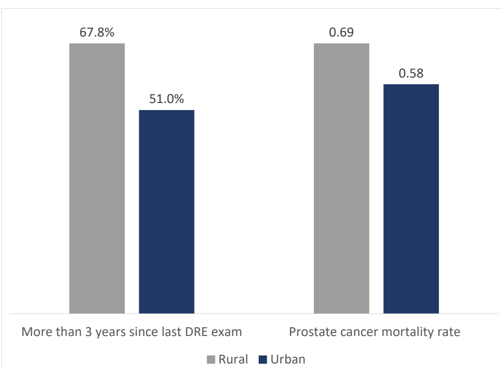
This graph presents the number of physicians per 1,000 people in 2014 in capitals (in blue) and countryside (in gray) by state. Data is from the Federal Council of Medicine-CFM (Scheffer et al., 2015) and the Brazilian Institute of Geography and Statistics (IBGE).



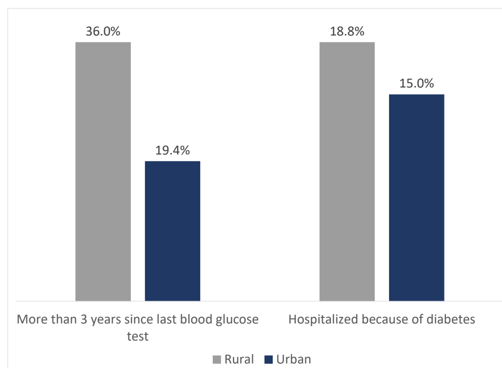
(a) Access to Doctors



(b) Infants' Health



(c) Men's Health



(d) Diabetes

Figure 2: Access to Healthcare and Health Outcomes

This figure shows health access indicators and health outcomes in the rural and urban areas of Brazil. Figure (a) has information on access to doctors, showing the percentage of inhabitants that did not have any appointment in the past year and the time (in minutes) patients have to wait in line to be examined by a doctor. Figure (b) depicts information related to infants' health, showing first the proportion that had their first visit to the doctor more than 30 days after birth, then the percentage of mothers that had less than 7 prenatal care visits and, last the infant mortality. Figure (c) shows the percentage of men with more than 50 years that did not do a Digital Rectal Exam and the prostate cancer mortality rate. Figure (d) depicts the percentage of the population that did not do a blood glucose test in the past three years and the proportion of the population that was hospitalized because of diabetes. Data related to mortality comes from Datasus. The others are from the 2013 National Health Survey (PNS).

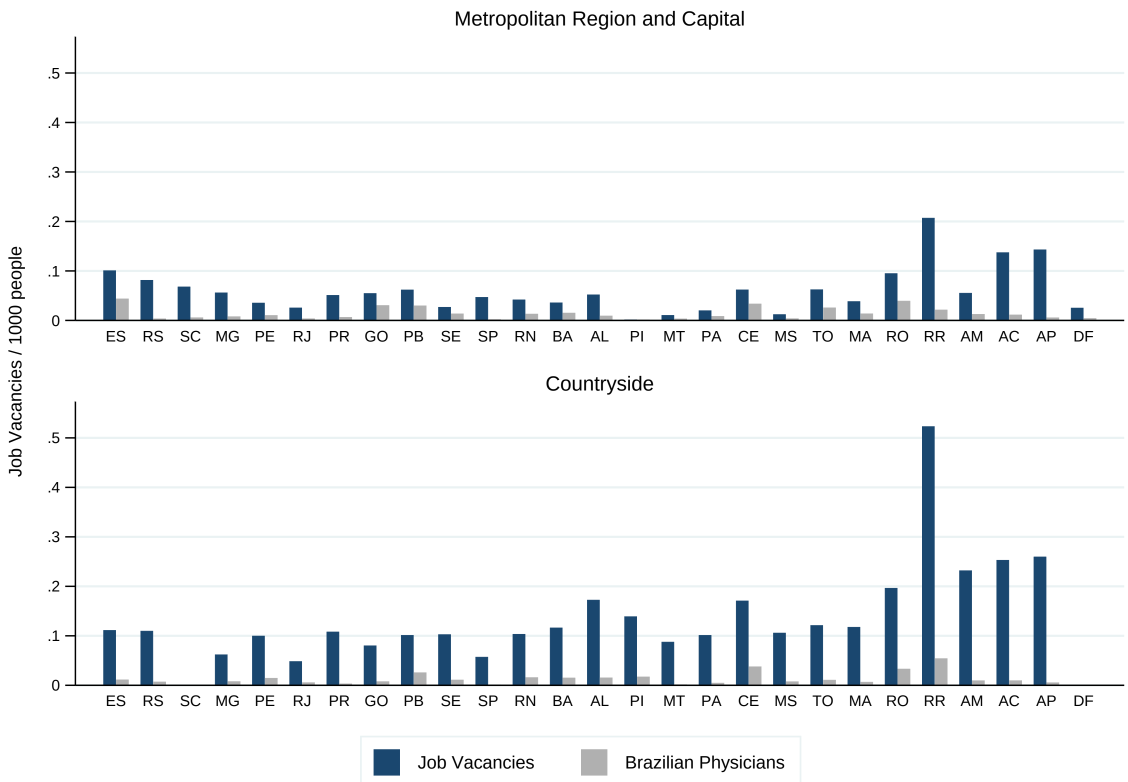


Figure 3: Job Vacancies from *Mais Médicos* Program and Positions Filled by Brazilians (per 1,000 people)

This graph presents the number of job vacancies created by the *Mais Médicos* Program (in blue) from the Federal Government and the number of Brazilian physicians that filled the open positions (in gray) per 1,000 people. The upper panel shows these figures in metropolitan regions across states, and the bottom panel shows the numbers in the countryside across state. Note that, as we explain in the text, SC and DF only have metropolitan region. Data from Ministry of Health, July 2013 to July 2014.

Table 1: Descriptive Statistics

	(%) Generalists			Regions' Attributes (2001-2012)				
	Work Region	Medschool Region	Birth Region	Physicians' Ratio	Health Infra Index	Health Insurance (%)	Amenity Index	Generalists' Avg. Wage/hr
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
North								
Metro. Reg	3.30	8.05	3.73	1.11	-0.10	11.54	0.05	10.14
				(0.24)	(0.65)	(5.29)	(0.49)	(6.01)
Countryside	2.92	0.14	1.75	0.40	-1.01	1.76	-0.36	18.41
				(0.12)	(0.42)	(1.55)	(0.35)	(15.06)
Northeast								
Metro. Reg	12.26	17.93	13.20	1.46	0.35	16.86	-0.36	18.14
				(0.41)	(0.67)	(6.28)	(0.38)	(8.31)
Countryside	8.20	0.72	7.38	0.48	-0.96	2.44	-0.61	26.31
				(0.10)	(0.28)	(1.36)	(0.25)	(9.69)
Southeast								
Metro. Reg	30.44	27.19	27.14	2.02	0.84	37.53	0.39	9.28
				(0.28)	(0.53)	(6.44)	(0.47)	(4.76)
Countryside	21.59	27.92	24.75	1.37	0.52	18.68	0.67	18.84
				(0.32)	(0.69)	(6.56)	(0.45)	(8.67)
South								
Metro. Reg	9.97	10.49	7.96	1.71	0.65	25.49	0.81	10.25
				(0.35)	(0.45)	(6.65)	(0.39)	(3.59)
Countryside	3.93	2.86	5.87	1.01	0.31	9.61	0.52	19.80
				(0.22)	(0.46)	(2.47)	(0.31)	(4.64)
Midwest								
Metro. Reg	3.99	4.42	4.64	1.97	1.41	21.54	0.59	14.04
				(0.44)	(0.77)	(4.42)	(0.52)	(8.95)
Countryside	3.41	0.28	3.58	0.75	-0.10	7.51	0.28	27.57
				(0.06)	(0.27)	(3.11)	(0.42)	(17.58)
Brazil								
				1.13	0.00	13.39	0.00	17.72
				(0.62)	(0.93)	(11.27)	(0.62)	(11.47)

This table shows the summary statistics of the main variables and data used in this study. Our sample consists of 46,989 generalist physicians that graduated from 2001 and 2013. Column 1 presents their decision of practice location right after graduating. Columns 2 and 3 display where they finished the medical school and where they were born. The average age upon graduation is 25.7 with a standard deviation of 3. 52.3% are men. Medical schools were ranked in 2013 by the newspaper Folha de São Paulo (range from 1 to 183). Columns 4 through 9 show the regions' attributes, with indexes constructed using the KKL method. We detail the variables in Section 3. Column 4 and 5 show the ratio of physicians per 1,000 people and the health infrastructure index. Columns 6 shows the percentage of the population that has health insurance. Column 7 show the amenity index, which encompasses education, entertainment, transportation, violence, local GDP per capita, and public investment. Column 8 shows the average hourly wage generalists up to 35 years old receive in each region, multiplied by a living cost index that ranges from 0 to 1.

Table 2: Physicians choice given place of birth and medical school region

		Metropolitan Regions					Countryside				
		N (1)	NE (2)	SE (3)	S (4)	MW (5)	N (6)	NE (7)	SE (8)	S (9)	MW (10)
Panel A. Birth Region											
Metropolitan Regions	N	47.32	4.90	13.23	2.00	1.82	21.84	3.42	2.85	0.80	1.82
	NE	1.56	64.91	6.32	0.74	1.16	1.37	21.46	1.77	0.16	0.53
	SE	0.95	1.84	74.10	2.80	1.00	0.87	1.73	15.26	0.67	0.78
	S	0.64	0.91	5.78	74.06	0.86	0.96	1.02	3.48	10.83	1.47
	MW	3.53	3.12	11.46	3.71	45.16	3.71	2.84	7.93	0.92	17.61
Countryside	N	19.61	4.14	8.28	3.65	4.51	44.21	4.26	5.85	1.71	3.78
	NE	1.27	32.02	5.80	0.46	1.01	1.87	53.01	3.78	0.12	0.66
	SE	0.83	0.81	26.91	2.19	1.53	1.05	1.60	61.94	0.95	2.19
	S	1.38	1.45	5.94	36.59	1.52	1.49	1.52	4.75	41.81	3.55
	MW	3.81	1.90	11.85	4.94	19.88	5.18	2.44	13.10	1.67	35.24
Panel B. Medical School Region											
Metropolitan Regions	N	34.36	4.97	10.62	1.64	4.70	24.63	7.58	3.83	1.22	6.45
	NE	0.80	59.94	3.61	0.78	0.62	1.72	29.98	1.09	0.53	0.94
	SE	0.47	1.21	72.18	1.33	0.73	0.59	2.47	19.55	0.45	1.02
	S	0.49	1.05	4.60	72.62	0.65	0.73	1.74	2.41	14.12	1.58
	MW	1.06	2.17	6.55	2.02	53.25	2.17	2.65	5.01	1.98	23.13
Countryside	N	4.55	3.03	3.03	1.52	4.55	65.15	1.52	1.52	1.52	13.64
	NE	0.30	25.52	1.78	0.89	0.00	0.00	69.44	1.19	0.00	0.89
	SE	0.54	1.33	30.05	2.50	2.72	0.70	2.62	54.40	1.42	3.71
	S	0.30	0.45	3.94	31.75	1.34	0.37	0.37	2.53	56.73	2.23
	MW	0.00	0.00	7.52	3.01	27.07	0.75	0.00	7.52	6.77	47.37

This table describes practice location choices of physicians born and graduated in different regions of the country. In *Panel A*, each cell (i, j) in the table has the fraction of physicians born in region i (row) – metropolitan areas (including capitals) and countryside for the 5 Brazilian geographic regions – that decided to work in region j (column). Analogously, in *Panel B* rows indicate the region physicians did medical school (rows) and their practice location choices (columns). Rows sum to 100.

Table 3: Control Function First Stage

	Estimates (1)
Constant	0.016
	(0.030)
Health Infrastructure	-0.060
	(0.052)
Physicians Ratio	0.018
	(0.102)
Health Insurance	0.013
	(0.072)
Amenity Index	0.073
	(0.047)
Health Infra Instrument	0.212**
	(0.095)
Physicians Ratio Instrument	-0.479***
	(0.139)
Health Insurance Instrument	0.062
	(0.094)
Amenity Index Instrument	0.302*
	(0.157)
Observations	676
F-Statistics	78.97

This table displays the control function first stage. Regression includes region and year dummies. Robust standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Preference Estimates – Regions' Characteristics

	Multinomial Logit	Multinomial Logit with Control Function	Random Coefficients	Random Coefficients with Control Function
	(1)	(2)	(3)	(4)
Physicians Ratio	**0.341**	**0.522**	**0.734**	**1.042**
	(0.536)	**(0.540)**	**(0.706)**	**(0.717)**
× Male	-0.303	-0.306	-0.301	-0.306
	(0.188)	(0.188)	(0.247)	(0.247)
× Age	-0.368	-0.356	-1.549**	-1.542**
	(0.580)	(0.580)	(0.759)	(0.759)
× Medschool Rank	-0.107	-0.105	-0.379	-0.363
	(0.322)	(0.322)	(0.416)	(0.416)
Health Infrastructure	**2.291*** **	**2.483*** **	**2.878*** **	**3.081*** **
	(0.500)	**(0.504)**	**(0.670)**	**(0.675)**
× Male	-0.093	-0.090	-0.105	-0.100
	(0.182)	(0.182)	(0.248)	(0.248)
× Age	-0.845	-0.850	0.145	0.141
	(0.547)	(0.547)	(0.736)	(0.736)
× Medschool Rank	-0.314	-0.312	-0.291	-0.304
	(0.310)	(0.310)	(0.415)	(0.415)
Health Insurance	**0.174**	**0.022**	**0.249**	**0.058**
	(0.371)	**(0.375)**	**(0.490)**	**(0.496)**
× Male	-0.247**	-0.245**	-0.300**	-0.298**
	(0.102)	(0.102)	(0.127)	(0.127)
× Age	-1.621***	-1.630***	-1.529***	-1.533***
	(0.323)	(0.323)	(0.404)	(0.404)
× Medschool Rank	1.126***	1.124***	1.986***	1.978***
	(0.175)	(0.175)	(0.215)	(0.215)
Amenity Index	**0.783*** **	**0.654** **	**1.311*** **	**1.100*** **
	(0.271)	**(0.275)**	**(0.356)**	**(0.366)**
× Male	0.139	0.138	0.246**	0.244**
	(0.091)	(0.091)	(0.120)	(0.120)
× Age	0.067	0.068	0.005	0.007
	(0.273)	(0.273)	(0.357)	(0.357)
× Medschool Rank	-0.466***	-0.463***	-1.045***	-1.039***
	(0.151)	(0.151)	(0.199)	(0.199)
Avg Hourly Wage	**-0.394* **	**2.673** **	**-0.784*** **	**2.788* **
	(0.231)	**(1.147)**	**(0.291)**	**(1.485)**
× Male	0.216	0.215	0.438**	0.436**
	(0.137)	(0.137)	(0.173)	(0.173)
× Age	-0.898**	-0.899**	-0.969**	-0.965**
	(0.393)	(0.393)	(0.483)	(0.483)
× Medschool Rank	0.381*	0.384*	0.567**	0.562**
	(0.225)	(0.225)	(0.282)	(0.282)
Region Unobs		**-2.327*** **		**-2.740** **
		(0.853)		**(1.099)**

This table displays the preference estimates for a standard and random coefficients logit, both with and without a control function. Sample size: 46,989. Respective log likelihoods: -82399.53, -82395.80, -78976.27 and -78972.99. Standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. Point estimates using 150 simulation draws. All columns include alternative-specific dummies and region-specific year trends.

Table 5: Preference Estimates – Physicians' Place of Birth and Medical School Region

	Multinomial Logit	Multinomial Logit with Control Function	Random Coefficients	Random Coefficients with Control Function
	(1)	(2)	(3)	(4)
Birth Metrop Region	**1.752*****	**1.751*****	**2.534*****	**2.534*****
	(0.051)	**(0.051)**	**(0.099)**	**(0.099)**
× Male	-0.084**	-0.083**	-0.260***	-0.260***
	(0.040)	(0.040)	(0.075)	(0.075)
× Age	0.290**	0.292**	0.137	0.141
	(0.127)	(0.127)	(0.233)	(0.233)
× Medschool Rank	1.306***	1.307***	1.168***	1.169***
	(0.066)	(0.066)	(0.123)	(0.123)
Birth Countryside Region	**3.100*****	**3.100*****	**2.842*****	**2.841*****
	(0.050)	**(0.050)**	**(0.133)**	**(0.133)**
× Male	0.083**	0.083**	0.197**	0.196**
	(0.038)	(0.038)	(0.099)	(0.099)
× Age	-0.263**	-0.262**	-1.122***	-1.118***
	(0.110)	(0.110)	(0.299)	(0.299)
× Medschool Rank	0.036	0.036	0.813***	0.815***
	(0.063)	(0.063)	(0.170)	(0.171)
Medschool Metrop Region	**3.723*****	**3.723*****	**4.997*****	**4.998*****
	(0.046)	**(0.046)**	**(0.084)**	**(0.085)**
× Male	-0.169***	-0.169***	-0.455***	-0.456***
	(0.036)	(0.036)	(0.064)	(0.064)
× Age	0.625***	0.624***	0.156	0.155
	(0.107)	(0.107)	(0.184)	(0.184)
× Medschool Rank	-0.425***	-0.425***	-0.548***	-0.548***
	(0.061)	(0.061)	(0.108)	(0.108)
Medschool Countryside Region	**1.726*****	**1.727*****	**2.823*****	**2.824*****
	(0.074)	**(0.074)**	**(0.096)**	**(0.096)**
× Male	-0.143***	-0.143***	-0.162**	-0.161**
	(0.052)	(0.052)	(0.064)	(0.064)
× Age	1.152***	1.151***	0.769***	0.766***
	(0.161)	(0.161)	(0.198)	(0.198)
× Medschool Rank	-0.008	-0.009	0.123	0.125
	(0.098)	(0.098)	(0.125)	(0.125)

This table displays the preference estimates for a standard and random coefficients logit, both with and without a control function. Sample size: 46,989. Respective log likelihoods: -82399.53, -82395.80, -78976.27 and -78972.99. Standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. Point estimates using 150 simulation draws. All columns include alternative-specific dummies and region-specific year trends.

Table 6: Preference Estimates – Interaction Term β^u

	Random Coefficients	Random Coefficients with Control Function
	(1)	(2)
Birth Metrop Region	3.121***	3.119***
	(0.081)	(0.081)
Birth Countryside Region	4.827***	4.831***
	(0.146)	(0.147)
Medschool Metrop Region	2.3***	2.305***
	(0.102)	(0.102)
Medschool Countryside Region	0.437	0.424
	(0.304)	(0.314)
Physicians Ratio	0.067	0.063
	(0.139)	(0.139)
Health Infrastructure	0.023	0.024
	(0.126)	(0.126)
Health Insurance	0.046	0.043
	(0.106)	(0.106)
Amenity Index	0.012	0.012
	(0.075)	(0.075)
Avg Hourly Wage	0.258	0.207
	(0.396)	(0.388)
Region Unobs		0.507
		(0.711)

This table displays the preference estimates for a random coefficients logit with and without a control function. Sample size: 46,989. Respective log likelihoods: -78976.27 and -78972.99. Standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. Point estimates using 150 simulation draws. All columns include alternative-specific dummies and region-specific year trends.

Table 7: Wage Elasticities

States	Metrop. Regions		Countryside	
	Own (1)	Cross (Min) (2)	Own (3)	Cross (Min) (4)
RO	0.127	-0.016	0.656	-0.083
AC	0.294	-0.036	0.259	-0.031
AM	0.348	-0.046	0.914	-0.127
RR	0.427	-0.053	0.465	-0.053
PA	0.241	-0.030	0.771	-0.095
AP	0.167	-0.019	0.184	-0.022
TO	0.559	-0.068	0.616	-0.073
MA	0.480	-0.063	0.554	-0.067
PI	0.860	-0.110	0.756	-0.098
CE	0.616	-0.077	1.234	-0.156
RN	0.380	-0.047	0.687	-0.085
PB	0.569	-0.068	0.946	-0.114
PE	0.508	-0.061	0.656	-0.079
AL	0.546	-0.065	0.622	-0.077
SE	0.689	-0.078	0.866	-0.109
BA	0.424	-0.055	0.914	-0.115
MG	0.448	-0.058	0.722	-0.097
ES	0.334	-0.042	0.701	-0.087
RJ	0.134	-0.017	0.241	-0.028
SP	0.166	-0.019	0.620	-0.080
PR	0.402	-0.051	0.705	-0.089
SC	0.306	-0.039	-	-
RS	0.214	-0.027	0.500	-0.062
MS	0.846	-0.103	0.643	-0.081
MT	0.310	-0.038	0.720	-0.088
GO	0.423	-0.052	1.447	-0.162
DF	0.186	-0.023	-	-

This table shows own and the minimum cross wage elasticity for each alternative in our model. The lowest cross elasticities are associated to the metropolitan regions of RJ and SP and the countryside of MG.

Table 8: Model Fit – Actual and Predicted Frequencies

		Actual Frequency (1)	Predicted Frequency (2)	Correctly Predicted (3)	% Correctly Predicted (3/1) (4)	% Predictions Correct (3/2) (5)
Metropolitan Regions	N	1,552	1,626	748	48.2	46.0
	NE	5,759	6,449	3,857	67.0	59.8
	SE	14,303	15,460	10,060	70.3	65.1
	S	4,683	5,363	3,355	71.6	62.6
	MW	1,874	1,810	872	46.5	48.2
Countryside	N	1,374	691	283	20.6	41.0
	NE	3,854	2,689	1,321	34.3	49.1
	SE	10,143	9,889	5,705	56.2	57.7
	S	1,845	1,651	771	41.8	46.7
	MW	1,602	1,361	437	27.3	32.1
Total		46,989	46,989	27,409	58.3	58.3

This table shows prediction tests using the estimated parameters. The predictions are obtained by assuming each physician chooses the region that has the highest predicted probability. Although the results are shown using ten aggregated regions, the correctly predicted frequencies were calculated by assessing the 52 alternatives. Specification used: random coefficients logit with control function. Sample size: 46,989. 10,000 draws were taken from the normal distributions estimated to evaluate the choice of each physician.

Table 9: Counterfactual Physicians' Distribution

		Population Distrib. (%) (1)	Predicted Distrib. (%) (2)	Counterfactuals (%) Birth Region (3)	Medschool Region (4)	Homog. Health Infra (5)	Homog. Amenity (6)	Homog. Wage (7)
Metropolitan Regions	N	3.55	3.29	2.50	2.21	3.38	3.35	3.56
	NE	12.09	12.04	10.28	9.72	11.50	12.72	11.98
	SE	24.98	30.84	31.95	31.88	26.79	30.01	33.75
	S	8.54	9.79	9.99	8.90	8.69	9.26	10.40
	MW	3.32	3.96	3.36	3.69	2.72	3.81	4.23
Countryside	N	4.51	2.94	3.97	3.75	5.83	3.44	2.60
	NE	15.77	7.87	11.07	11.38	13.19	9.25	6.32
	SE	17.39	21.49	19.04	19.25	19.74	20.59	20.52
	S	5.99	4.10	4.05	4.78	3.90	3.96	3.89
	MW	3.87	3.66	3.78	4.45	4.27	3.60	2.74
Quadratic Error			0.59	0.29	0.36	0.45	0.57	0.67
% Reduction in Imbalance				51.54	38.62	23.62	4.11	-13.02

This table displays the counterfactuals performed to assess how each factor contributes to physicians' decision and to reduce their imbalance across the country. The model used was the random coefficients with a control function, and 150 draws were taken from the normal distributions estimated to evaluate physicians' choices in each situation. Column (1) shows the population distribution (the optimum benchmark to be achieved) and Column (2) where physicians chose to work according to our original model. The following two columns show the counterfactual distribution of generalists if: (3) each medical school implemented an affirmative policy in which their admissions would also consider candidates birthplace and try to match it to the population distribution; and (4) the medical schools were redistributed such that the proportion of physicians graduating in each region mirrored the population distribution. The last three columns detail how the distribution of generalists would be if: (5) health infrastructure, (6) amenities, and (7) wages were homogeneous across the whole country. The quadratic error indicates how far the distributions originated from the policy simulations are from the population one. Bellow, there is the percentage reduction in imbalance each counterfactual would produce relative to the predicted distribution quadratic error.

Table 10: Cost-Effectiveness Analysis

		Population Distrib. (%) (1)	Predicted Distrib. (%) (2)	Counterfactuals (%) New Medschool Vacancies (3)	Infra x1.5 (N/NE CS) (4)	Wage x1.5 (N/NE CS) (5)
Metropolitan Regions	N	3.55	3.29	2.17	3.22	3.15
	NE	12.09	12.04	9.89	11.81	11.61
	SE	24.98	30.84	29.27	30.39	29.93
	S	8.54	9.79	8.82	9.64	9.49
	MW	3.32	3.96	3.94	3.88	3.81
Countryside	N	4.51	2.94	6.38	3.34	3.70
	NE	15.77	7.87	12.14	9.01	10.14
	SE	17.39	21.49	17.63	21.14	20.78
	S	5.99	4.10	4.01	4.02	3.94
	MW	3.87	3.66	5.76	3.55	3.44
Quadratic Error			0.59	0.28	0.54	0.51
% Reduction in Imbalance				52.47	8.56	13.42
Cost (1,000 USD) per % Reduction in Imbalance				[2,718 ; 6,362]	69,811	10,724

This table displays the cost-benefit of different policy simulations. The first two columns show the population distribution (the optimum benchmark to be achieved) and where physicians chose to work according to our original model. Column (3) shows how the distribution of physicians would be if the 4,187 new medical school vacancies created after 2009 had been created in a way to approximate the distribution of students to the population distribution. The last two counterfactual columns detail how the distribution of physicians would be if (4) wages and (5) health infrastructure in the North and Northeast countrysides increased by 50%. The quadratic error indicates how far the distributions originated from the policy simulations are from the population one. Bellow, there is the percentage reduction in imbalance each counterfactual would produce relative to the predicted distribution quadratic error. The last row shows the cost incurred in each policy for a 1% reduction in imbalance. The lower bound cost of opening new medical school vacancies were calculated considering the average cost of $13,796 for each undergraduate student in a public university. To produce an upper bound we used the average yearly tuition of private medical schools in Brazil, which was around around $32,295 per year. To calculate the health infrastructure cost we compared how much the Federal District of Brazil spent in investment and operational costs (excluding expenses with personnel) with the improvement in its health infrastructure index over 2005-2012 using data from the National Treasury. This same ratio of expenses per increase in health infrastructure was used to estimate the cost a 50% improvement in the N and NE countryside would entail. The cost of an increase in wages is straightforward. All costs are in 2010 USD.

Appendix (for online publication only)

Appendix for "How to Attract Physicians to Underserved Areas? Policy Recommendations from a Structural Model" (Costa, Nunes and Sanches, June 2019)

- Section A contains supplementary material discussed in Section 2 in the text.
- Section B discusses correlation between health provision and health outcomes mentioned in Section 2.3.
- Section C presents the two alternative demand models to the baseline model presented in Section 4.2.
- Section D presents the estimates using a restricted choice set for physicians, as mentioned in Section 5.
- Section E describes the data sources and data cleaning process in detail.

A Supplementary Tables and Descriptive Evidence

This appendix provides supplementary tables and evidence that complement Section 3. Table A1 shows that physicians in our sample (matched) are similar to those physicians we lose when matching the different data sources (not matched). Figure A1 and Table A2 display the living costs we use (see Appendix E for details). Table A3 presents the transition matrix of physicians' practice location choice just after graduation and practice location 5 years later.

We estimate some correlations between local attributes and the geographic distribution of physicians in Brazil by regressing

$$y_{jt} = \alpha + \beta X_{jt-1} + \gamma_j + \delta_j \times t + \varepsilon_{jt} \tag{8}$$

where y_{jt} is the number of new generalist physicians (per 1,000 people) that chose their first job in state-region j (i.e., metropolitan areas or countryside) in year t, X_{jt} is a vector of local covariates in year $t-1$, γ_j is state-region fixed effects, and δ_j are state-region specific trends. We estimate this equation with and without population weights. To account for potential endogeneity of wages and filled job positions, we also instrument local wages using local characteristics in the neighboring regions – we describe the instrument in greater detail in section 4.3.[48]

The first five columns of Table A4 present the OLS estimates. We see that higher wages are positively associated with the number of physicians starting to work in the area. The estimates also show that other local characteristics influence the number of physicians choosing to go to the area. As columns 3 and 4 show, better local amenities, health infrastructure and the percent of medical school generalist students graduating in that area contributes to attract more new graduates to that region – both across and within state-region.[49] The results from the specification without population weights (column 5) is qualitatively similar.

Our instrumental variables estimates – shown in Table A4 columns 6 to 10 – tell a similar story. Again, local characteristics play an important role in physicians choice. While the specification with no local controls suggests that higher wages attract fewer physicians, this relationship changes when we account for local fixed characteristics (column 7) or when we control for other local time-varying characteristics (column 8 to 10). The coefficients associated with these local characteristics have roughly similar magnitude in the instrumented and OLS specifications. The coefficient attached to wages, instead, increases substantially

[48] These are the same type of instruments as proposed in Berry et al. (1995). See also Nevo (2000) and Petrin and Train (2010).

[49] As discussed above, the negative coefficient of the stock of physicians per capita suggest that physicians avoid areas with greater competition.

in comparison to the OLS coefficients. This finding suggests that wages are negatively correlated with unobserved local attributes: places with better unobservable characteristics can attract more physicians paying relatively lower wages.

Differently from the linear regression shown above, the structural model in the paper accounts for both local and individual characteristics, preference heterogeneity and spatial correlation across locations. As it will be shown in the following sections, these elements are important determinants of physicians' locational choices.

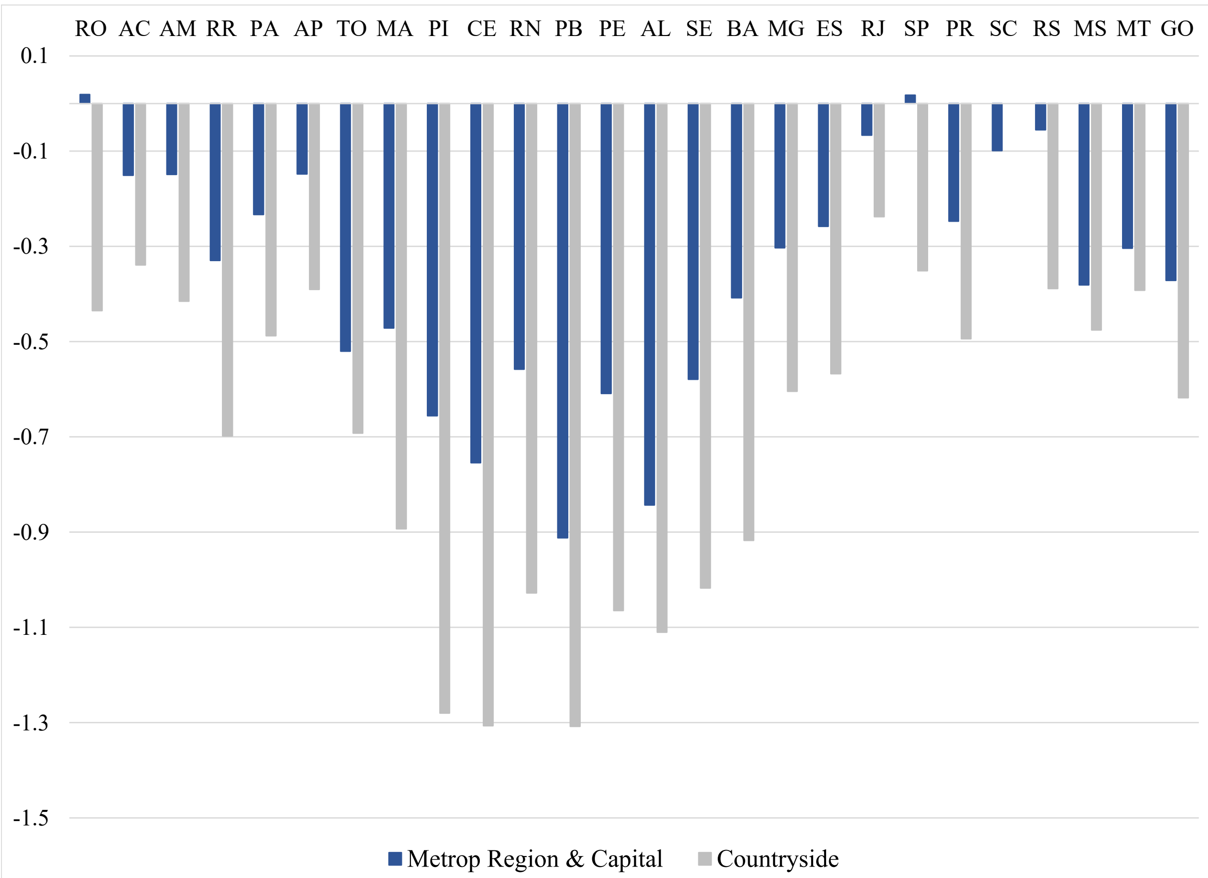


Figure A1: 2010 Living Cost Index

This graph presents our 2010 living cost estimates for each state/(countryside or metropolitan region & capital) using the 2010 Census. We chose the Federal District as the omitted state in the regressions, so its index is zero.

Table A1: Physicians Lost During Merge process - Mean Difference

	Matched (1)	Not Matched (2)	Difference p-value (3)
Birth Region (%)			
MR N	3.73	3.77	0.000
MR NE	13.20	11.56	0.016
MR SE	27.14	36.16	-0.090
MR S	7.96	9.81	-0.019
MR MW	4.64	5.57	-0.009
CS N	1.75	1.12	0.006
CS NE	7.38	3.67	0.037
CS SE	24.75	19.57	0.052
CS S	5.87	6.11	-0.002
MR MW	3.58	2.67	0.009
Medschool Region (%)			
MR N	8.05	5.32	0.000
MR NE	17.93	15.89	0.020
MR SE	27.19	32.16	-0.050
MR S	10.49	12.22	-0.017
MR MW	4.42	4.91	-0.005
CS N	0.14	0.01	0.001
CS NE	0.72	0.16	0.006
CS SE	27.92	26.34	0.016
CS S	2.86	2.90	0.000
MR MW	0.28	0.10	0.002
Graduation Year (%)			
2001	2.61	14.18	0.000
2002	4.38	10.01	-0.056
2003	5.06	10.79	-0.057
2004	5.70	10.93	-0.052
2005	6.06	8.42	-0.024
2006	6.17	7.79	-0.016
2007	6.18	7.17	-0.010
2008	6.38	6.06	0.003
2009	7.45	4.91	0.025
2010	8.67	4.77	0.039
2011	11.48	4.81	0.067
2012	13.52	4.71	0.088
2013	16.34	5.44	0.109
Age	25.74	24.89	0.000
% Male	52.26	34.02	0.000
Medschool Rank	83.49	70.44	0.000
Number of Obs	46,989	13,574	

Table A2: Living Cost Index

		Years														
		2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Metropolitan Region	RO	0.037	0.039	0.032	0.030	0.027	0.024	0.021	0.023	0.028	0.019	0.021	0.024	0.028	0.031	0.031
	AC	-0.096	-0.115	-0.059	-0.096	-0.148	-0.141	0.002	-0.118	-0.105	-0.151	-0.193	-0.195	-0.211	-0.290	-0.259
	AM	-0.542	-0.361	-0.285	-0.235	-0.300	-0.298	-0.291	-0.128	-0.104	-0.149	-0.147	-0.211	-0.204	-0.203	-0.259
	RR	-0.334	-0.259	-0.320	-0.408	-0.284	-0.294	-0.326	-0.390	-0.297	-0.330	-0.387	-0.403	-0.410	-0.460	-0.398
	PA	-0.227	-0.232	-0.248	-0.273	-0.234	-0.260	-0.239	-0.247	-0.265	-0.233	-0.245	-0.263	-0.238	-0.335	-0.340
	AP	0.078	-0.107	0.021	-0.015	-0.104	-0.067	-0.141	-0.167	-0.152	-0.148	-0.097	-0.195	-0.178	-0.200	-0.252
	TO	-0.722	-0.665	-0.631	-0.595	-0.598	-0.553	-0.600	-0.566	-0.599	-0.520	-0.599	-0.557	-0.572	-0.580	-0.588
	MA	-0.650	-0.616	-0.513	-0.471	-0.524	-0.508	-0.438	-0.497	-0.531	-0.471	-0.483	-0.582	-0.408	-0.475	-0.519
	PI	-0.625	-0.699	-0.630	-0.670	-0.728	-0.655	-0.592	-0.639	-0.657	-0.656	-0.672	-0.661	-0.620	-0.620	-0.633
	CE	-0.777	-0.779	-0.747	-0.760	-0.751	-0.755	-0.766	-0.797	-0.768	-0.754	-0.762	-0.767	-0.676	-0.685	-0.668
	RN	-0.633	-0.666	-0.636	-0.639	-0.633	-0.586	-0.597	-0.575	-0.527	-0.558	-0.587	-0.636	-0.615	-0.665	-0.640
	PB	-1.153	-1.042	-0.988	-0.983	-0.973	-0.965	-1.040	-0.944	-0.908	-0.912	-0.942	-0.907	-0.873	-0.908	-0.901
	PE	-0.536	-0.586	-0.566	-0.594	-0.572	-0.590	-0.624	-0.624	-0.641	-0.609	-0.592	-0.540	-0.533	-0.524	-0.520
	AL	-0.875	-0.897	-0.875	-0.908	-0.873	-0.839	-0.870	-0.845	-0.876	-0.843	-0.886	-0.904	-0.847	-0.850	-0.812
	SE	-0.725	-0.699	-0.611	-0.653	-0.670	-0.640	-0.678	-0.638	-0.647	-0.579	-0.579	-0.594	-0.537	-0.570	-0.582
	BA	-0.462	-0.471	-0.436	-0.430	-0.416	-0.390	-0.410	-0.419	-0.422	-0.408	-0.439	-0.450	-0.435	-0.452	-0.435
	MG	-0.395	-0.372	-0.331	-0.356	-0.345	-0.332	-0.339	-0.344	-0.346	-0.303	-0.311	-0.314	-0.297	-0.304	-0.312
	ES	-0.501	-0.432	-0.391	-0.385	-0.364	-0.349	-0.299	-0.319	-0.298	-0.258	-0.278	-0.302	-0.261	-0.309	-0.284
	RJ	-0.007	0.008	0.049	0.028	0.008	-0.002	-0.043	-0.043	-0.066	-0.066	-0.089	-0.096	-0.075	-0.083	-0.063
	SP	0.013	0.018	0.012	0.014	0.021	0.021	0.028	0.029	0.027	0.018	0.019	0.023	0.012	0.016	0.014
	PR	-0.384	-0.367	-0.311	-0.318	-0.309	-0.271	-0.269	-0.266	-0.262	-0.247	-0.261	-0.248	-0.239	-0.241	-0.223
	SC	-0.330	-0.297	-0.147	-0.161	-0.103	-0.034	-0.008	-0.076	-0.061	-0.076	-0.104	-0.169	-0.090	-0.158	-0.135
	RS	-0.077	-0.063	-0.054	-0.054	-0.046	-0.038	-0.049	-0.062	-0.064	-0.055	-0.062	-0.065	-0.063	-0.067	-0.061
	MS	-0.609	-0.576	-0.537	-0.508	-0.435	-0.399	-0.430	-0.416	-0.457	-0.381	-0.383	-0.415	-0.353	-0.411	-0.393
	MT	-0.442	-0.398	-0.354	-0.278	-0.216	-0.201	-0.297	-0.322	-0.345	-0.304	-0.349	-0.390	-0.304	-0.311	-0.316
	GO	-0.603	-0.540	-0.472	-0.460	-0.454	-0.430	-0.411	-0.438	-0.428	-0.372	-0.378	-0.395	-0.353	-0.374	-0.357
	DF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Countryside	RO	-0.831	-0.872	-0.720	-0.686	-0.601	-0.539	-0.481	-0.526	-0.622	-0.435	-0.481	-0.547	-0.622	-0.693	-0.696
	AC	-0.216	-0.258	-0.133	-0.216	-0.334	-0.317	0.004	-0.266	-0.236	-0.339	-0.434	-0.439	-0.475	-0.651	-0.582
	AM	-1.512	-1.008	-0.795	-0.655	-0.837	-0.831	-0.811	-0.356	-0.290	-0.415	-0.410	-0.588	-0.568	-0.567	-0.722
	RR	-0.708	-0.548	-0.677	-0.863	-0.601	-0.624	-0.691	-0.825	-0.629	-0.698	-0.819	-0.853	-0.869	-0.975	-0.844
	PA	-0.476	-0.486	-0.519	-0.572	-0.489	-0.544	-0.500	-0.516	-0.554	-0.487	-0.513	-0.550	-0.499	-0.700	-0.711
	AP	0.205	-0.281	0.055	-0.039	-0.274	-0.177	-0.373	-0.439	-0.400	-0.390	-0.255	-0.515	-0.470	-0.527	-0.666
	TO	-0.960	-0.884	-0.839	-0.792	-0.796	-0.736	-0.798	-0.753	-0.798	-0.692	-0.797	-0.741	-0.761	-0.772	-0.783
	MA	-1.232	-1.168	-0.972	-0.892	-0.992	-0.964	-0.829	-0.941	-1.007	-0.893	-0.915	-1.103	-0.774	-0.901	-0.983
	PI	-1.220	-1.365	-1.230	-1.308	-1.421	-1.279	-1.155	-1.248	-1.283	-1.280	-1.311	-1.290	-1.211	-1.210	-1.237
	CE	-1.347	-1.350	-1.293	-1.317	-1.301	-1.308	-1.326	-1.380	-1.330	-1.306	-1.320	-1.328	-1.172	-1.186	-1.157
	RN	-1.166	-1.227	-1.172	-1.177	-1.165	-1.080	-1.100	-1.060	-0.971	-1.028	-1.081	-1.172	-1.133	-1.225	-1.179
	PB	-1.653	-1.494	-1.416	-1.409	-1.396	-1.383	-1.492	-1.353	-1.303	-1.308	-1.352	-1.300	-1.252	-1.303	-1.292
	PE	-0.937	-1.024	-0.989	-1.038	-1.000	-1.031	-1.091	-1.092	-1.121	-1.065	-1.035	-0.945	-0.932	-0.917	-0.909
	AL	-1.152	-1.181	-1.153	-1.196	-1.150	-1.105	-1.146	-1.113	-1.153	-1.110	-1.166	-1.191	-1.115	-1.119	-1.069
	SE	-1.273	-1.228	-1.074	-1.148	-1.177	-1.125	-1.191	-1.121	-1.137	-1.017	-1.018	-1.044	-0.944	-1.001	-1.022
	BA	-1.040	-1.060	-0.980	-0.969	-0.935	-0.878	-0.923	-0.942	-0.950	-0.917	-0.988	-1.013	-0.978	-1.018	-0.980
	MG	-0.788	-0.742	-0.660	-0.709	-0.687	-0.662	-0.676	-0.687	-0.690	-0.604	-0.619	-0.626	-0.592	-0.606	-0.623
	ES	-1.101	-0.950	-0.859	-0.846	-0.801	-0.768	-0.658	-0.701	-0.656	-0.567	-0.611	-0.663	-0.575	-0.680	-0.624
	RJ	-0.026	0.030	0.176	0.099	0.028	-0.007	-0.154	-0.156	-0.237	-0.238	-0.319	-0.345	-0.268	-0.298	-0.225
	SP	-0.251	-0.357	-0.243	-0.283	-0.414	-0.412	-0.551	-0.563	-0.533	-0.351	-0.364	-0.446	-0.233	-0.313	-0.273
	PR	-0.766	-0.734	-0.622	-0.635	-0.618	-0.541	-0.538	-0.532	-0.523	-0.494	-0.522	-0.496	-0.477	-0.481	-0.446
	RS	-0.543	-0.445	-0.381	-0.379	-0.320	-0.270	-0.345	-0.439	-0.449	-0.389	-0.437	-0.456	-0.441	-0.472	-0.430
	MS	-0.761	-0.719	-0.671	-0.635	-0.543	-0.498	-0.538	-0.520	-0.571	-0.476	-0.479	-0.519	-0.441	-0.513	-0.491
	MT	-0.571	-0.513	-0.457	-0.359	-0.278	-0.260	-0.383	-0.416	-0.446	-0.392	-0.450	-0.503	-0.393	-0.401	-0.408
	GO	-1.002	-0.898	-0.784	-0.765	-0.755	-0.716	-0.684	-0.728	-0.712	-0.618	-0.628	-0.657	-0.588	-0.621	-0.593

This table shows the living cost estimates for each one of the 52 regions analyzed in this study. We chose the Federal District as the omitted State in the regressions, so its index is zero. The higher the value, the higher the living cost in the region. Our Metropolitan Region definition also includes the capital of each State.

Table A3: Probability of physicians choosing certain region after 5 years, given their first practice location

		Generalists' Location 5 years later									
		Metropolitan Regions					Countryside				
		N	NE	SE	S	MW	N	NE	SE	S	MW
First Location											
Metropolitan Regions	N	59.95	3.28	12.65	1.41	2.11	9.84	3.51	4.45	0.70	2.11
	NE	0.30	78.63	6.77	1.35	0.90	0.75	9.63	1.20	0.38	0.08
	SE	0.79	1.59	82.31	1.51	0.99	0.40	1.15	10.10	0.50	0.65
	S	0.30	0.41	5.93	81.15	0.24	0.53	0.36	2.79	7.94	0.36
	MW	0.97	1.16	9.28	3.87	70.02	0.77	1.74	3.09	0.39	8.70
Countryside	N	21.52	4.35	14.13	2.39	2.61	40.87	3.26	5.65	1.30	3.91
	NE	0.64	37.23	6.52	1.19	1.19	1.11	47.81	3.02	0.64	0.64
	SE	0.51	1.05	25.14	1.80	1.21	0.62	1.02	66.14	0.92	1.59
	S	0.45	0.60	4.52	27.71	0.60	0.30	0.45	3.16	60.84	1.36
	MW	0.62	1.86	8.25	3.92	18.56	0.62	2.27	9.28	2.06	52.58

This table describes practice location of physicians who graduated between 2001 and 2009 just after graduation and five years later. Each cell (i, j) in the table has the fraction of physicians who started working in region i (row) just after graduation – metropolitan areas (including capitals) and countryside for the 5 Brazilian geographic regions – that worked in region j (column) five years after graduation. The table shows persistency of practice location, especially for those who initially chose to work in metropolitan areas.

Table A4: Reduced Form Evidence at the Regional Level

	Dep. var.: Physicians Practice Choices per 1,000 people									
	OLS					2SLS				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Avg Hourly Wage	0.001	0.035***	0.014**	0.015**	0.016	-0.043***	0.184***	0.022	0.057	0.123**
	(0.008)	(0.009)	(0.005)	(0.006)	(0.010)	(0.015)	(0.046)	(0.026)	(0.036)	(0.055)
Amenities			0.011**	0.022*	0.010			0.011**	0.016	-0.012
			(0.005)	(0.012)	(0.010)			(0.005)	(0.014)	(0.012)
Health Infrastructure			0.041***	0.093***	0.099***			0.039**	0.077***	0.073**
			(0.015)	(0.018)	(0.025)			(0.019)	(0.019)	(0.031)
Health Insurance			0.015***	0.009	0.017			0.016***	0.003	0.005
			(0.004)	(0.022)	(0.023)			(0.004)	(0.021)	(0.021)
Physicians p.c.			-0.040**	-0.077***	-0.052			-0.037*	-0.063***	-0.033
			(0.017)	(0.023)	(0.032)			(0.020)	(0.024)	(0.030)
Med. School Graduates			0.005*	0.013**	0.013**			0.005*	0.012**	0.009
			(0.003)	(0.006)	(0.006)			(0.003)	(0.005)	(0.007)
Population Weight	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
State-Region FE	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes
1st Stage F-Stat						13.66	16.12	4.29	4.74	3.16

This table presents the results of linear regressions of generalist physicians practice choice per 1,000 people (mean 0.019) on average hourly wage and other covariates. Unit of observations is state-region (i.e., metropolitan areas or countryside) by year (N=676), between 2001 and 2013. All regressions include a constant and state-region trends. Columns 1 to 5 present OLS estimates and columns 6 to 10 present 2SLS estimates instrumenting wages with neighbors' amenities, health infrastructure, health insurance indices and physicians per capita (as described in section 4.3). Regressions estimated using population weights, except columns 5 and 10. Section 3 describes the variables used. Columns indicated with State-Region FE include state-region fixed effects. Standard errors clustered by state-region (52 clusters) in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

B Health Provision and Health Outcomes

This subsection provides descriptive evidence on the correlation between health provision and health outcomes in Brazil. Figure 2 tabulates a few measures of access to healthcare and health outcomes in rural and urban areas in Brazil from *Pesquisa Nacional de Saúde* (PNS) and the Mortality Information System (SIM/Datasus). Figure 2a shows that those living in rural areas are about 38 percent more likely to not have been to a Doctor's appointment in the last 12 months than those living in urban areas. Figure 2b indicates that infants in rural areas are about 20 percent less likely to visit the Doctor in the first 30 days of life, and about 50 percent more likely to not have gone through seven prenatal care visits (the recommended by the Brazilian Ministry of Health) than infants in urban areas. This difference relates to higher infant mortality rates in rural areas. Figures 2c and 2d also point that men over 50 years old in rural areas are more likely to stay more than three years without a DRE exam (rectal examination), and that people in these areas are more likely to have more than three years since the last blood glucose exams than those in urban areas. These exams are important for early detection of prostate cancer and diabetes, respectively, and lower access may be one factor underlying the higher prostate cancer mortality rate and hospitalization because of diabetes in rural areas than in urban ones.

While the descriptive evidence from Figure 2 suggest that rural areas tend to have lower access to healthcare and worse related health outcomes than in urban areas, we cannot infer any causal relationship from them. To get finer evidence on the relationship between the presence of Doctors and local health outcomes, we correlate the number of physicians per capita and infants' health outcomes across Brazilian municipalities between 2005 and 2016 using a linear regression model in Table A5. The table presents three indicators of infants and women health outcomes: share of infants born with less than seven prenatal care visits in Panel A, infant mortality rate in Panel B, and maternal mortality rate in Panel C. All regressions include a constant and year fixed effects.

We derive some stylized facts from this exercise. First, we find a positive correlation between physicians per capita in a given municipality-year and infants/mothers' health outcomes. Column 1 presents the raw correlations which are statistically significant at one percent for our three variables. Column 3 show that these correlations do not disappear or loose statistical significance when add state fixed effects and compare municipalities within states. Second, this correlation does not seem to come from a rural versus urban comparison. Columns 2 and 4 show that adding a countryside fixed effect almost do not affect the point estimates of the correlations between the number of doctors and health outcomes estimated in columns 1 and 3, respectively. The relationship gets weaker, however, when we add municipality fixed effects in Column 5. This may be driven by other local characteristics relevant for infants' health but also by small variation in the number of physicians within underserved municipalities over the decade.

Table A5: Reduced Form Motivation

	Dependent variables indicated in each panel				
	(1)	(2)	(3)	(4)	(5)
Panel A. *Dep.var.:* Infant born with less than 7 prenatal care visits (%)					
Physicians p.c.	-5.44***	-5.33***	-2.57***	-2.53***	-0.64***
	(0.31)	(0.31)	(0.23)	(0.23)	(0.23)
Panel B. *Dep.var.:* Infant mortality rate					
Physicians p.c.	-0.97***	-0.94***	-0.61***	-0.59***	0.16
	(0.11)	(0.11)	(0.11)	(0.11)	(0.30)
Panel C. *Dep.var.:* Maternal mortality rate					
Physicians p.c.	-8.51***	-8.35***	-7.02***	-7.11***	-0.95
	(1.42)	(1.43)	(1.47)	(1.49)	(4.41)
Countryside FE		Yes			
State FE			Yes		
State-Countryside FE				Yes	
Municipality FE					Yes

This table presents the results of OLS regressions using different specifications and dependent variables as indicated in each panel and column. Unit of observations municipality-year (N= 61,211), between 2005 and 2016. All regressions include a constant, year fixed effects, and state-region trends. Dependent variable in *Panel A* is the percentage of infants born that had less than 7 prenatal visits during gestation, in *Panel B* is the infant mortality rate (per 100,000 live births), and in *Panel C* is the mortality rate of mothers during delivery (per 100,000 live births). Mean dependent variables: 39.8 (Panel A), 15.1 (Panel B), and 60.7 (Panel C). Each column present results from a different specification: column 2 includes a fixed effect for municipalities in the countryside (i.e., outside the capital or metropolitan areas), column 3 includes state fixed effects, column 4 includes state-countryside fixed effects, and column 5 includes municipality fixed effects. Standard errors clustered by municipality (5,565 clusters) in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

C Demand Side Model

Alternatively we consider two different formulations for the demand model. We estimate these models and compare the fitting of them to the fitting of our baseline model.

Our model assumes that hospitals in location j choose the wage rate they are going to offer to physicians at period t, y_{jt}, in order to maximize a simple welfare function subject to a budget constraint. Mathematically the problem is written as:

$$\begin{aligned} &\underset{y_{jt}}{max}\ \left(p_{jt} - y_{jt}\right) L_{jt}\left(y_{jt}, \mathbf{y_{-jt}}\right) \\ &s.t.\ \ y_{jt} \cdot L_{jt}\left(y_{jt}, \mathbf{y_{-jt}}\right) \leq \bar{R}_{jt}, \end{aligned}$$

where, $L_{jt}\left(y_{jt}, \mathbf{y_{-jt}}\right)$ is the demand of physicians in location j, period t, which depends on the wage rate offered at that location/period and the wage rate offered in all other locations, $\mathbf{y_{-jt}}$, and $\bar{R}_{jt}$ is the maximum amount of public resources that may be used to hire new physicians in that location/period. We interpret p_{jt} as the marginal value in R$ that public administrators in location j, period t, attribute to a physician working in that location/period.

The first order condition of this problem gives rise to the following equation:

$$\frac{p_{jt}}{1 + \lambda_{jt}} = \left[\frac{L_{jt}\left(y_{jt}, \mathbf{y_{-jt}}\right)}{\tilde{L}_{jt}\left(y_{jt}, \mathbf{y_{-jt}}\right)} + y_{jt}\right], \forall j, t, \tag{9}$$

where $\lambda_{jt} \geq 0$ is the Lagrange multiplier associated to the budget constraint and $\tilde{L}_{jt}\left(y_{jt}, \mathbf{y_{-jt}}\right)$ is the derivative of $L_{jt}\left(y_{jt}, \mathbf{y_{-jt}}\right)$ with respect to y_{jt}. As the elements on the right hand side of this equation can be obtained from our data, this equation can be used to identify $\frac{p_{jt}}{1+\lambda_{jt}}$. These estimates give us the lower bound of the marginal benefit generated by a physician in each location/year. However, p_{jt} and λ_{jt} cannot be independently identified from the data without strong assumptions – see Verboven (1996). Identifying p_{jt} and λ_{jt} independently is necessary to subsequently solve the model and perform counterfactuals.

I. Oligopsony. In light of this identification problem, our first demand model assumes that $\lambda_{jt} = 0$ for all j and t – i.e. the budget constraint is not binding in all regions/years. Then, the vector of equilibrium wages for each region at year t is the solution of the system of the J first order conditions. This is a simple oligopsony model where public administrators of different regions compete against each other for the supply of physicians in order to maximize a given welfare function.

II. Budget constraint. Alternatively, our second demand model assumes that the budget constraint is always binding for all locations. Wages are set to satisfy the budget constraints at each location/year. Equilibrium wages are obtained by solving, for each year, the system of J budget constraints, $y_{jt} \cdot L_{jt}(y_{jt}, \mathbf{y_{-jt}}) = \bar{R}_{jt}$. This model reflects the well known fact that the lack of public resources is an important constraint to investment in public health.

III. Fitting to the data. As Table A6 shows, the fitting to the data and the estimates of the budget constraint model are very similar to the baseline model. The oligopsony model, however, has been running for more than two months and did not converge at the same precision level than the other two models. A plausible explanation is that this model is not good at explaining our data. Conditional on this caveat, column 7 shows that the fitting of the oligopsony model to the data is substantially worse than the other two models.

Table A6: Demand Models

			Baseline model		Budget Constraint		Oligopsony	
		Actual Distrib. (%) (1)	Predicted Distrib. (%) (2)	% Correctly Predicted (3)	Predicted Distrib. (%) (4)	% Correctly Predicted (5)	Predicted Distrib. (%) (6)	% Correctly Predicted (7)
Metropolitan Regions	N	3.3	3.5	48.2	3.5	48.3	1.6	56.3
	NE	12.3	13.7	67.0	13.9	67.1	7.1	74.4
	SE	30.4	32.9	70.3	32.7	70.3	22.5	73.8
	S	10.0	11.4	71.6	11.4	72.4	2.5	85.1
	MW	4.0	3.9	46.5	3.9	46.0	0.2	59.8
Countryside	N	2.9	1.5	20.6	1.5	21.1	0.0	86.7
	NE	8.2	5.7	34.3	5.6	33.8	64.8	1.4
	SE	21.6	21.0	56.2	21.4	57.3	1.2	71.6
	S	3.9	3.5	41.8	3.5	42.3	0.0	0.0
	MW	3.4	2.9	27.3	2.7	27.1	0.0	0.0
Total				58.3		58.4		26.9

This table shows the model performance when we consider three different types of demand for physicians: (i) inelasticity demand – baseline – in coluns 2-3; (ii) all local governments being budget constrained in columns 4-5; and (iii) all governments being part of an oligopsony in columns 6-7. The predictions are obtained by assuming each physician chooses the region that has the highest predicted probability. Although the results are shown using ten aggregated regions, we averaged the choices by region from choices (predicted frequencies) calculated considering the 52 alternatives. We take 10,000 draws from the normal distributions estimated to evaluate the choice of each physician. Sample size: 46,989. Note that the oligopsony model did not converge at the same precision level than the other two models.

D Estimates with Restricted Choice set

Table A7: Restricted Choice Set Estimates – Physicians' Place of Birth and Medical School Region

	Multinomial Logit	Multinomial Logit with Control Function	Random Coefficients	Random Coefficients with Control Function
	(1)	(2)	(3)	(4)
Birth Metrop Region	**1.715*****	**1.713*****	**2.398*****	**2.397*****
	(0.050)	**(0.050)**	**(0.094)**	**(0.094)**
× Male	-0.082**	-0.081**	-0.253	-0.251
	(0.039)	(0.039)	(0.071)***	(0.071)***
× Age	0.294**	0.296**	0.149	0.151
	(0.125)	(0.125)	(0.222)	(0.222)
× Medschool Rank	1.265***	1.266***	1.141***	1.143***
	(0.065)	(0.065)	(0.117)	(0.117)
Birth Countryside Region	**2.947*****	**2.946*****	**2.699*****	**2.700*****
	(0.050)	**(0.050)**	**(0.122)**	**(0.122)**
× Male	0.074*	0.074*	0.172*	0.172*
	(0.038)	(0.038)	(0.092)	(0.092)
× Age	-0.273***	-0.272***	-1.045***	-1.046***
	(0.110)	(0.110)	(0.277)	(0.278)
× Medschool Rank	0.121**	0.121**	0.804***	0.807***
	(0.063)	(0.063)	(0.158)	(0.158)
Medschool Metrop Region	**3.545*****	**3.546*****	**4.714*****	**4.715*****
	(0.046)	**(0.046)**	**(0.078)**	**(0.078)**
× Male	-0.172***	-0.173***	-0.408***	-0.408***
	(0.035)	(0.035)	(0.059)	(0.059)
× Age	0.630***	0.629***	0.213	0.211
	(0.106)	(0.106)	(0.170)	(0.170)
× Medschool Rank	-0.396***	-0.396***	-0.454***	-0.453***
	(0.061)	(0.061)	(0.100)	(0.100)
Medschool Countryside Region	**1.689*****	**1.690*****	**2.770*****	**2.771*****
	(0.073)	**(0.073)**	**(0.088)**	**(0.088)**
× Male	-0.160***	-0.160***	-0.181***	-0.182***
	(0.051)	(0.051)	(0.060)	(0.060)
× Age	1.051***	1.050***	0.631***	0.630***
	(0.158)	(0.158)	(0.187)	(0.187)
× Medschool Rank	-0.096	-0.098	0.012	0.014
	(0.096)	(0.096)	(0.118)	(0.118)

This table displays the preference estimates for a standard and random coefficients logit, both with and without a control function. In this model we consider that choice sets are restricted in the following way: physicians graduated from a certain region have as alternatives all places chosen by physicians that graduated in the same region over the whole sample period plus its own region of birth.Sample size: 46,989. Respective log likelihoods: -80201.54, -80197.88, -77293.52 and -77289.93. Standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. Point estimates using 150 simulation draws. All columns include alternative-specific dummies and region-specific year trends.

Table A8: Preference Estimates – Regions' Characteristics

	Multinomial Logit	Multinomial Logit with Control Function	Random Coefficients	Random Coefficients with Control Function
	(1)	(2)	(3)	(4)
Physicians Ratio	**0.276**	**0.463**	**0.732**	**1.034**
	(0.537)	**(0.541)**	**(0.694)**	**(0.705)**
× Male	-0.314	-0.318	-0.305	-0.309
	(0.189)	(0.189)	(0.244)	(0.244)
× Age	-0.563	-0.551	-1.706**	-1.698**
	(0.581)	(0.581)	(0.748)	(0.748)
× Medschool Rank	-0.115	-0.113	-0.635	-0.625
	(0.321)	(0.321)	(0.409)	(0.409)
Health Infrastructure	**2.388***	**2.574***	**2.912***	**3.100***
	(0.502)	**(0.507)**	**(0.660)**	**(0.665)**
× Male	-0.068	-0.064	-0.076	-0.071
	(0.181)	(0.181)	(0.243)	(0.243)
× Age	-0.580	-0.585	0.398	0.397
	(0.543)	(0.543)	(0.719)	(0.719)
× Medschool Rank	-0.402	-0.399	-0.256	-0.266
	(0.307)	(0.307)	(0.405)	(0.405)
Health Insurance	**0.069**	**-0.087**	**0.098**	**-0.096**
	(0.369)	**(0.374)**	**(0.481)**	**(0.487)**
× Male	-0.258***	-0.256***	-0.323***	-0.322***
	(0.102)	(0.102)	(0.126)	(0.126)
× Age	-1.516***	-1.525***	-1.370***	-1.379***
	(0.322)	(0.322)	(0.399)	(0.399)
× Medschool Rank	1.203***	1.201***	2.157***	2.153***
	(0.174)	(0.174)	(0.213)	(0.212)
Amenity Index	**0.798****	**0.679****	**1.279***	**1.093***
	(0.273)	**(0.277)**	**(0.352)**	**(0.360)**
× Male	0.153**	0.151	0.249**	0.247**
	(0.091)	(0.091)	(0.118)	(0.118)
× Age	0.030	0.030	-0.001	0.001
	(0.275)	(0.275)	(0.351)	(0.351)
× Medschool Rank	-0.391	-0.389	-0.896***	-0.891***
	(0.151)	(0.151)	(0.195)	(0.195)
Avg Hourly Wage	**-0.481**	**2.539****	**-0.840**	**2.532 ****
	(0.231)	**(1.139)**	**(0.285)**	**(1.457)**
× Male	0.210	0.209	0.423	0.421
	(0.138)	(0.138)	(0.173)	(0.173)
× Age	-0.749	-0.751	-0.736	-0.745
	(0.396)	(0.395)	(0.480)	(0.480)
× Medschool Rank	0.531**	0.533***	0.661**	0.653**
	(0.225)	(0.224)	(0.279)	(0.279)
Region Unobs		**-2.290***		**-2.632****
		(0.846)		**(1.077)**

This table displays the preference estimates for a standard and random coefficients logit, both with and without a control function. In this model we consider that choice sets are restricted in the following way: physicians graduated from a certain region have as alternatives all places chosen by physicians that graduated in the same region over the whole sample period plus its own region of birth. Sample size: 46,989. Respective log likelihoods: -80201.54, -80197.88, -77293.52 and -77289.93. Standard deviations are in parenthesis. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Point estimates using 150 simulation draws. All columns include alternative-specific dummies and region-specific year trends.

Table A9: Preference Estimates – Interaction Term β^u

	Random Coefficients	Random Coefficients with Control Function
	(1)	(2)
Birth Metrop Region	3.015***	3.014***
	(0.078)	(0.078)
Birth Countryside Region	4.439***	4.443***
	(0.130)	(0.130)
Medschool Metrop Region	1.942***	1.943***
	(0.106)	(0.106)
Medschool Countryside Region	-0.032	-0.023
	(0.172)	(0.171)
Physicians Ration	-0.006	0.001
	(0.145)	(0.143)
Health Infrastructure	-0.026	-0.026
	(0.123)	(0.123)
Health Insurance	0.045	0.046
	(0.113)	(0.113)
Amenity Index	0.007	0.007
	(0.072)	(0.072)
Avg Hourly Wage	0.054	-0.098
	(0.582)	(0.430)
Region Unobs		0.773*
		(0.455)

This table displays the preference estimates for a random coefficients logit, both with and without a control function. In this model we consider that choice sets are restricted in the following way: physicians graduated from a certain region have as alternatives all places chosen by physicians that graduated in the same region over the whole sample period plus its own region of birth. Sample size: 46,989. Respective log likelihoods: -77293.52 and -77289.93. Standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. Point estimates using 150 simulation draws. All columns include alternative-specific dummies and region-specific year trends.

E Data Appendix

We now describe the data sources and data cleaning in detail.

E.1 Organizing CFM database

Fix manually observations that are wrong in the following variables: med-school graduation year, registry year in CFM, registry cancellation year in CFM, birth year, state of birth, city of birth and medical school name. Ex: 006 is 2006 in med school-graduation-year

Recover missing values for variables: med-school graduation year, medical school name, city of birth, state of birth and birth year. Since some physicians have more than one record (work in more than one state), if an information is blank in one of their records but not in the other we might be able to recover it. This recovery is based mainly on physicians full name. Until now no record was deleted. We have 637,058 records, and 459,740 physicians (uniquely identified by full name [50]).

Keep only physicians that graduated between 2001-2013. Among the records deleted in this step, 2.6% were because medschool_gradyear was missing (16,634/637,058). Considering people with birth year up to 1971 (beginning med-school with 30 years in 2001, a conservative measure) and records that have both birth-year and medschool_gradyear missing we have that we wrongly deleted at most 4,602 observations. Records: 221,240. Physicians (uniquely identified by full name): 153,569. Records that might have been wrongly deleted: 4,602. Physicians (uniquely identified by full name) that might have been wrongly deleted: 4,599. Keep only physicians that were born and did their medical school in Brazil. Records: 219,925. Physicians (uniquely identified by full name): 152,799. Only 770 physicians (uniquely identified by full name) were deleted.

Important variables for us: full name, birth year, gender, city of birth, state of birth, medical school name, medical school state, graduation year. Delete the duplicates in terms of these variables. Records: 171,068. Physicians (uniquely identified by full name): 152,799. Physicians(uniquely identified by full name and birth year): 153,109. We have 310 homonyms physicians that graduated between 2001 and 2013. Since we have very feel we will use only the full name to identify physicians.

Delete observations that have missing birth year, medical school name or city of birth. We lost 3,162 physicians (uniquely identified by full name) because of missing values. It represents 3162/152799=2.1% of our database. Records: 163,588. Physicians (uniquely identified by full name): 149,637. According to SIGRAS/INEP, there are 149,002 physicians graduating between 2001-2013, our final database of physicians is very close to it.

[50]We took all the special characters and spaces from names in order to minimize typos problems

Keep only one record per physician. Criteria used was to first consider records that are ACTIVE and then the PRINCIPAL records. Records: 149,637. Physicians (uniquely identified by full name): 149,637.

Merge birth and medical school cities' names with their codes provided by IBGE (Brazilian Institute of Geography and Statistics).

Merge birth and medical school cities' code with identifiers of metropolitan regions in each state provided by IBGE. Our definition of metropolitan region will also incorporate the capital of each state. Countryside will be defined as the cities that are neither state capitals nor belong to metropolitan regions.

Merge with ranks of university courses published by *Folha de São Paulo* newspaper in 2013. Gama Filho University was closed by the Ministry of Education because of its low academic quality and the serious impairment of its economics and financial situation. Because of that we imputed its ranking as last in our database. Create an excel file with the counts of our final database (CFMcontagemfinal.xls).

E.2 Merge between CFM and CNRM databases

Merge with CNRM was done in three steps: Exact merge using full name and registry number in CFM; (ii) Exact merge using full name; and (iii) Probabilistic merge using full name (2,292 were merged here). We do not consider merges with the following inconsistency: if the beginning of the residency occurs before the med-school graduation date. Homonyms or wrong probabilistic merge gives us 1,332 records or 669 (0.45%) physicians that did not do medical residency but in our database appear as if they did. We cleaned these wrong observations.

We have that 59,450 (40.2% out of 149,637) physicians finished medical residency (until Aug 2014). All specialist physicians were dropped leaving us with 90,187 generalists (uniquely identified by full name).

E.3 Organizing RAIS database

First we select in RAIS (we have data from 2001-2015)[51] the physicians, ie, the observations with the following Brazilian Occupational Codes (CBO): 2002 CBOs beginning with 2231, 2251, 2252, 2253 and 234453, this last one begin medicine professor; and 1994 CBOs beginning with 0.61 and 13770, this last one being professor of occupational medicine.

[51] We had a problem with the CBO of the 2001 identified data provided to us by the Ministry of Labor, so we did not use it to find physicians work city. When constructing the wages, we used the 2001 non-identified data available online at ftp://ftp.mtps.gov.br/pdet/microdados/RAIS/2001/, which did not have this problem.

Remove spaces and special characters from physicians' full name. Then we recover missing name, date of birth, race and gender using the physicians' social security number (known in Brazil as CPF) and PIS (Social Integration Program) number used by the Ministry of Labor to identify workers in the database.

Define as missing value the zeros that appear in working hours, remuneration, birth date, CPF and PIS.

Merge with deflator and convert remuneration variable to 2010 reais and divide it by working hours creating a remuneration per hour variable.

Create our own age variable which is based on RAIS year of birth (available for years 2002-2010) and on RAIS age (available for 2001 and 2011+)

Identify the cities that belong to metropolitan regions or are capitals using IBGE classification.

Merge with population data by municipality-year provided by IBGE.

Keep only the first employment relationship of each physician.

Keep only one choice per physician. Around 17% of physicians in our database chose to work in more than one of our 52 regions in the same year. For those we picked the region that physicians worked more hours.

E.4 Organizing CNES database

Keep only professionals classified as physicians: 2002 CBOs beginning with 2231, 2251, 2252, 2253

Create a variable identifying physicians that are enrolled in a medical residency program: We identify them through CBO 2231F9 and the employment relationship variable in CNES.

Identify the cities that belong to metropolitan regions or are capitals using IBGE classification.

Keep only the first employment relationship of each physician.

Keep only one choice per physician. Around 17% of physicians in our database chose to work in more than one of our 52 regions in the same year. For those we picked the region that physicians worked more hours.

E.5 Merge CFM generalists with RAIS and CNES

Now we want to find the city each generalist physician is working after graduation. We will search for them both in RAIS and CNES.

We start with RAIS performing an exact merge using physicians full name and birth year. We disconsider merged observations in which graduation year > year found working in RAIS. 51,446 physicians (57.0%) were merged with RAIS.

Second, we perform an exact merge with CNES using physicians' full name. We disconsider merged observations in which graduation year > year found working in CNES. CNES informs which physicians are currently a training medical resident in that health facility. We use this information to exclude from our database another 29,624 physicians that are enrolled in a medical residency program, leaving us with 60,563 generalist physicians. In our merge with CNES, we found 56,213 (92.8%) generalist physicians.

When we take both datasets into consideration we were able to find 57,195 (94.4%) physicians at some point in time. We decided to keep only physicians we could see the place of work up to 3 years after graduation, which left us with our final database of 46,989 generalist physicians. We do not consider these physicians because we cannot assure these regions were their first location choice after medical school.

Table A1 shows the comparison between physicians found in RAIS or CNES up to three years after graduation and physicians lost in the merge process. We believe most of our losses were due to: (i) Misspelling and typing errors in the names and dates of birth. Specially for women, since in Brazil it is quite common for them to include their husband's surname after married; (ii) Physicians which *only* work shifts up to 24 hours in hospitals, and A&E departments.

E.6 Construction of the Amenity Index

E.6.1 Transport Index

Data related to the fleet in Brazil. Information came from DENATRAN. We used the fleet information by municipality for December of each year. Appended all the information. All the administrative regions of Brasilia were considered just as Brasilia city (summed them). Data was collapsed (sum) to the UF/(countryside or metropolitan region & capital) level. Data was merged with IBGE database containing population per municipality-year. We constructed four variables: Bus per 1000 people (includes bus and shuttle); Cars per 1000 people (includes cars and pick-ups); Motorcycle per 1000 people. Since we don't have this data for 2000 in the municipality level, we used the information of 2001 in 2000.

Data on Traffic Deaths. Data came from DATASUS/SIM. Death was collected by city of death and year. ICD for traffic accidents: V01-V99. Data was collapsed (sum) to the state/(countryside or metropolitan region & capital) level. Data was merged with IBGE database containing population per municipality-year. We created the variable traffic deaths per 1,000 people

Data on Establishments from RAIS. We obtained the number of establishments by municipality/year (2000-2015) with the following National Classification of Economic Activity (CNAE 2.0 or 1.0, depending on the year): CNAE 2.0: 49-51, 55, 56, 90, 92, 91 and 93; CNAE 1.0: 60-62, 55.1, 55.2, 92.3, 92.5, 92.6. We classified these establishments in the following way: Transport by land, water and air, all together: CNAE 2.0 is 49-51 and CNAE 1.0 is 60-62; Hotels: CNAE 2.0 is 55 and CNAE 1.0 is 55.1; Restaurants: CNAE 2.0 is 56 and CNAE 1.0 is 55.2; Entertainment: CNAE 2.0 is 90-93 and CNAE 1.0 is 92.3, 92.5 and 92.6. Data was merged with IBGE database containing population per municipality-year. Data was collapsed (sum) to the state/(countryside or metropolitan region & capital) level. We created the variable number of establishments per 1,000 people for all the types described above. We have bad data for the state of Pernambuco in 2002 (zero establishments). So I interpolated this year's value for all establishments variable with 2001 and 2003 information.

How we constructed the transportation index. We merged information of fleet, establishments and traffic deaths at the state/(countryside or metropolitan region & capital) level. The index was created using the KKL method: normalize each variable that will be included in the index, than calculate their average. Variables used in the index: Transportation Establishments per 1000 people (DENATRAN); Buses per 1000 people (RAIS); Cars per 1000 people (RAIS); Motorcycles per 1000 people (RAIS); Traffic Deaths per 1000 people (SIM).

E.6.2 Violent Deaths

Data came from DATASUS/SIM. Death was collected by city of death and year. The ICD-10 codes used to classify violent deaths are the ones used in the National Violent Death Reporting System for Deaths ($\leq$ 1year), excluding suicides and terrorism: Assault (homicide): X85-X99, Y00-Y09; Event of undetermined intent: Y10-Y34; Unintentional exposure to inanimate mechanical forces (firearms): W32-W34; Legal Intervention: Y35.

Data was collapsed (sum) to the Sate/(countryside or metropolitan region & capital) level. Data was merged with IBGE database containing population per municipality-year. We created the variable violent deaths per 1,000 people.

E.6.3 Entertainment Index

Data on movie theaters from ANCINE. There is only information about the number of movie theater rooms per city from 2007 on. Between 2000-2006 we did a projection using the evolution of the total number of rooms between 2000-2006 (the national data exists since 1971) and the proportion of rooms each city had in 2007. We kept these proportion the same and adjusted for the national total provided in the period. Data was collapsed (sum) to the state/(countryside or metropolitan region & capital) level. Data was merged with IBGE database containing population per municipality-year. We created the variable number of cinema rooms per 1,000 people.

Data on Establishments. We described how we obtained hotels, restaurant and entertainment establishments per capita in the section we explain the transport index

How we constructed the entertainment index. We merged information of cinemas and establishments at the state/(countryside or metropolitan region & capital) level. Created an entertainment index using KKL method: normalize each variable that will be included in the index, than calculate their average. Variables used in the index: Restaurants per 1000 people (RAIS); Hotels per 1000 people (RAIS); Entertainment Establishments per 1000 people (RAIS); and Cinemas per 1000 people (ANCINE).

E.6.4 Education

We use the Index of Development of Basic Education (IDEB) in Brazil. IDEB is an indicator of the quality of education that combines information from the performance of students in national assessments at the end of school levels of basic education (4/5th, 8/9th and 11/12th grades) with flow rates. This index started in 2005 and is released every two years.

Data was merged with IBGE database containing population per municipality-year. Data was collapsed (mean) to the UF/(countryside or metropolitan region & capital) level using population weights. Then data for the years 2006, 2008, 2010, 2012 and 2014 were interpolated. Data for 2000-2004 were set as being equal to 2005.

E.6.5 Public Investment

Data on public expenditures by state and city comes from. Only one city has missing information in one year (290430). We interpolate. Data was merged with IBGE database containing population per municipality-year. Data was collapsed (sum) to the state/(countryside or metropolitan region & capital) level. We created two variables: state investments per capita and City investments per capita. Values were deflate data to 2010 reais.

The index was created using the KKL method: normalize each variable that will be included in the index, than calculate their average. Variables used in either index: State investments per capita; and Cities investments per capita.

E.6.6 Amenity Index

The amenity index was created using the KKL method. Variables used in the index: Entertainment Index; GDP per capita (data obtained at the year-municipality level at IBGE website); Education Index; Transport Index; Violent Deaths per Capita; and Public Investment Index.

E.7 Construction of other regions' characteristics

E.7.1 Physicians per capita

Data from CNES, 2005-2016 (December). The ratio of total physicians per 1,000 people was calculates for each region. Data from 2000-2004 was imputed as being the same as 2005.

E.7.2 Health Insurance

Data is from the National Regulatory Agency for Private Health Insurance and Plans (ANS). Data from December each year 2000-2016, no imputation. We drop observations in which the beneficiary municipality is not know inside the state (last four digits in the city code are equal to "0000" or city code is equal to zero). This is less than 1% of total beneficiaries between 2000-2016. We divide the number of beneficiaries by the population and obtain the health insurance coverage of each region. One person might be the beneficiary of more than one health insurance plan. He/she will be doubled counted. But this is how ANS calculates health insurance coverage.

E.7.3 Health Infrastructure Index

Data from CNES, 2005-2016 (December). Diagnoses and imaging equipment obtained: mammographs, ultrasound machines, x-ray machines, computed tomography (CT) scanners and magnetic resonance imaging (MRI) scanners. We calculate the number of each one of these equipment per 1,000 people in each region. Data from 2000-2004 was imputed as being the same as 2005. The index was constructed using these five rates and the KKL method.

E.7.4 Physicians Wage per Hour

We keep the records where: 2002 CBO is 225170 and 223129 (generalist physician) or 225125 and 223115 (clinic physician); and 1994 CBO is 0.6105 (General physician). We deflate average remuneration values to 2010 reais. Using working hours we construct the remuneration per hour in 2010 reais. Data is collapsed to a countryside/metropolitan region weighted by city population. We construct two variables: one with the average remuneration of all generalists (data described in steps above) and average remuneration of recently graduated physicians (age $\leq$ 35 years).

E.8 Construction of the Living Cost Index

The living cost index was constructed following Summers (1973); Seabra and Azzoni (2015). Below we describe step by step how to construct it. We use PNAD and Datazoom from PUC-Rio to make the variables compatible over time. Organizing **PNAD** database to run living-cost regression. Keep only the household living arrangements classified as "permanently individual". Delete the collective and temporary individual arrangements (v0201 == 1). Keep only the following household settings (variable v4105 <=3): "Urban - urban area", "Urban - non urban area" and "Urban - isolated area" . We focus only on households in the urban area because we understand that the real estate market of rural areas may not represent the dynamics of local living cost.

Construct two variables: "Number of children up to 24 years in the household" and "Number of children over 24 years old in the household". We use variables v8005 (age in years) and v0403 == 3 (to focus only on son/daughter).

Keep only the heads of household in the sample (variable v0401 == 1).

Create a dummy for rented households (v0207 == 3). Construct variable ln(Monthly Rent) using variable v0208 and deflate it to 2010 reais. Calculate the number of apartments (using variable v0202==4) per State and the number of houses per state (using variable v0202==2), remembering to weight by the household sampling weight (variable v4611). With these two information I calculate the rate "Number of houses/Number of Apartments" by state. We will use variables: v0105: number of people in household; v0206: number of rooms used for sleeping; v2016: number of bathrooms (in 2001 we do not have this information only a dummy if there is a bathroom or not); and v0205: number of rooms.

Create the auxiliary variable "Total number of households" by state summing the variable v4611 (household sampling weight). Create the variable "Proportion of rented households in the state" using variables v0207 and "Total number of households". Create the variable "Proportion of households classified as slums in the state" using variable v0203 (main material used on walls, which need to be 1 "brickwall") and "Total number of households". Create dummy for households connected to sewage (v0217 - sewage treatment, which needs to be either 1 "sewage system" or 2 "septic tank with drain", 0 otherwise). Create variable "Proportion of households connected to sewage in the state" using variable v0217 (sewage treatment, which needs to be either 1 "sewage system" or 2 "septic tank with drain", 0 otherwise) and "Total number of households". Create dummy for households in which garbage is collected directly or indirectly (v0218==1 or v0218 == 2). Create dummy for households in which the energy source used for lightning is eletric - network, generator or solar (v0219 == 1). Create dummy for households in which the source of water supply is through network, well or spring (v0212 == 2 or v0212 == 4). Create a variable with the mean income by state using v4614 (monthly income) and v4611 (household sampling weight). Deflate this variable to 2010 reais.

Create several variables related to the household head: Dummy for gender (v0302); Dummy for white or Asian (v0404 == 2 or v0404 == 6); Variable for years of schooling (v4803, and v4703 for years <= 2006), recoding 17 (not identified) to missing; Dummies if young (between 17 and 29 years old), adult (30 years or more) or old (60 years or more) using variable v8005; Dummy if lives together with another person (v4111); Dummy if married (v4011); Four dummies for the type of family: couple without children, couple with children, mother with children, other (v4723) - in all of these types of family can exist other parents, housekeepers, etc; Dummy if was born or not in the municipality (ie, if migrant or not) (v0501); Dummies for time living in the municipality (considering both people that were and were not born in the municipality: up to four years (v5061 == 2 | v5121 ==2), 5-9 years (v5063 == 4 | v5123 == 4), 10 or more years (v5065 == 6 | v5125 == 6); and Dummies for the mean income: class E, mean income <= 1085 reais; class D, between 1086-1734 reais; class C: between 1735-7475 reais, class B: 7476-9745 reais.

Heckit regression using PNAD. We run a Probit where the dependent variable is a dummy of whether the household is rented or not. The independent variables are: number rooms used for sleeping, number of bathrooms (in 2001 we do not have this information only a dummy if there is a bathroom or not), number of rooms, main material used on walls (1 if brickwall, zero otherwise), dummy for households in which garbage is collected directly or indirectly, dummy for households in which the energy source used for lightning is electric - network, generator or solar; dummy for households in which the source of water supply is through network, well or spring; dummy for households connected to sewage; "Number of houses/Number of Apartments" by state; proportion of rented households in the state; proportion of households classified as slums in the state; proportion of households connected to sewage in the state; mean income by state; dummy for gender; dummy for white or Asian; years of schooling; age; dummies if young (between 17 and 29 years old), adult (30 years or more) or old (60 years or more); four dummies for the type of family: couple without children, couple with children, mother with children, other; number of children up to 24 years in the household; number of children over 24 years old in the household; dummies for time living in the municipality: up to four years, 5-9 years, 10 or more years; dummy if was born or not in the municipality; mean income; dummies for the mean income: class E (mean income <= 1085 reais), class D (between 1086-1734 reais), class C (between 1735-7475 reais); class B(7476-9745 reais). The predicted values from the equation above are retained to calculate inverse mills ratio;

Least Square Regression using only the sample of rented houses and PNAD. Dependent variable: ln(rent value). Independent variables: inverse mills ratio; dummies for each state (DF dummy will be the one dropped); number rooms used for sleeping; number of bathrooms (in 2001 we do not have this information only a dummy if there is a bathroom or not); number of rooms; main material used on walls (1 if brickwall, zero otherwise).

Using **2010 Census** data, also with the help of Datazoom (PUC-Rio), we organize the database to run living-cost regression.

Keep only the household living arrangements classified as "permanently individual". We delete the collective and improvised individual arrangements (v4001 == 1 | v4001 == 2). Keep only households located in urban areas (v1006==1). We focus only on them because we understand that the real estate market of rural areas may not represent the dynamics of local living cost.

Create a variable that identifies the municipalities that belong to the "metropolitan region + capital" and the countryside of each state. Construct two variables: "Number of children up to 24 years in the household" and "Number of children over 24 years old in the household". We use variables v6036 (age in years) and (v0502 == 4 | v0502 == 5, to focus only on son/daughter).

Keep only the heads of household in the sample (variable v0502== 1).

Construct a dummy for rented households (v0201==3). Construct variable ln(Monthly Rent) using variable v0208. Calculate the number of apartments (using variable v4002==13) and the number of houses (using variable v4002==11 | v4002 == 12) per state/(countryside or metropolitan region & capital), remembering to weight by the household sampling weight (variable v0010). With these two information, calculate the rate "Number of houses/Number of Apartments" per state/(countryside or metropolitan region & capital). Create the auxiliary variable "Total number of households" per state/(countryside or metropolitan region & capital) summing the variable v0010 (household sampling weight). Create the variable "Proportion of rented households in the state/(countryside or metropolitan region & capital)" using the dummy for rented households and variable "Total number of households". Create the variable "Proportion of households classified as slums in the state/(countryside or metropolitan region & capital)" using variable v0202 (main material used on walls, which needs to be 1 "brickwall") and "Total number of households". Create the variable "Proportion of households connected to sewage in per state/(countryside or metropolitan region & capital)" using variable v0207 (sewage treatment, which needs to be either 1 "sewage system" or 2 "septic tank with drain", 0 otherwise) and "Total number of households". Dummy for households in which garbage is collected directly or indirectly (v0210 ==1 or v0210 == 2). Dummy for households in which the energy source used for lightning is eletric - network, generator or solar (v0211 == 1 | v0211 == 2). Dummy for households in which the source of water supply is through network, well or spring (v0208 == 1 or v0208 == 2). Create a variable with the mean income per state/(countryside or metropolitan region & capital) using v6529 (monthly income in Jul/2010) and v0010 (household sampling weight).

Create several variables related to the household head: Dummy for gender (v0601); Dummy for white or Asian (v0606 == 1 or v0606 == 3); Variable for years of schooling (v6400), recoding 5 (not identified) to missing ; Dummies if young (between 17 and 29 years old), adult (30 years or more) or old (60 years or more) using variable v6036; Dummy if lives together with another person (v0637 == 1); Dummy if married (v0639 == 1 | v0639 == 2 | v0639 == 3); Four dummies for the type of family: couple without children, couple with children, mother with children, other (v5090) - in all types of family can exist other parents, housekeepers, etc; Dummy if was not born in the municipality (ie, if migrant) (v0618 == 3); Four dummies for time living in the municipality: up to four years, 5-9 years, 10 or more years (variable v0624 and v6036 if not a migrant); Dummies for the mean income: class E, mean income <= 1085 reais; class D, between 1086-1734 reais; class C: between 1735-7475 reais, class B: 7476-9745 reais.

Heckit regression using 2010 Census. We run a Probit where the dependent variable is a dummy of whether the household is rented or not. The independent variables are: number of rooms used for sleeping (v0204); number of bathrooms (v0205); number of rooms (v0203); main material used on walls (1 if brickwall, zero otherwise); dummy for households in which garbage is collected directly or indirectly; dummy for households in which the energy source used for lightning is electric - network, generator or solar; dummy for households in which the source of water supply is through network, well or spring; dummy for households connected to sewage; "Number of houses/Number of Apartments" in the state/(countryside or metropolitan region & capital); Proportion of rented households in state/(countryside or metropolitan region & capital); Proportion of households classified as slums in the state/(countryside or metropolitan region & capital); proportion of households connected to sewage in the state/(countryside or metropolitan region & capital); mean income per state/(countryside or metropolitan region & capital); dummy for gender; dummy for white or Asian; years of schooling; age; dummies if young (between 17 and 29 years old), adult (30 years or more) or old (60 years or more); four dummies for the type of family: couple without children, couple with children, mother with children, other; number of children up to 24 years in the household; number of children over 24 years old in the household; dummies for time living in the municipality: up to four years, 5-9 years, 10 or more years; dummy if was born or not in the municipality; mean income; dummies for the mean income: class E (mean income <= 1085 reais), class D (between 1086-1734 reais), class C (between 1735-7475 reais), class B (7476-9745 reais). The predicted values from the equation above are retained to calculate inverse mills ratio.

OLS using only the sample of rented houses. Dependent variable: ln(rent value). Independent variables: inverse mills ratio; dummies for each state/(countryside or metropolitan region & capital) (DF dummy will be the one dropped); number rooms used for sleeping; number of bathrooms (in 2001 we do not have this information only a dummy if there is a bathroom or not); number of rooms; main material used on walls (1 if brick wall, zero otherwise). Weight the regressions by household weights. Also run this regression with state dummies instead of state/(countryside or metropolitan region & capital).

Constructing the living cost index using the state dummies and the state/(countryside or metropolitan region & capital) dummies obtained from regressions above. The steps above using the 2010 Census data provides living cost indexes for each state as a whole, and its "metropolitan region + capital" (MR) and countryside (CS). The ideal would be for us to have this set of indexes in every year. But unfortunately the 2000 Census don't have rent information and the smallest representative unit of analysis in PNAD is the states and metropolitan regions. So using the living cost indexes obtained through 2010 Census we calculate, for each state, how bigger or smaller are the MR and CS indexes when compared to the one related to the whole state. We will assume that these ratios are the same in every year. This means that if in 2010 the countryside of Rio has a living cost index that is 2/3 of the living cost in the whole state, we will assume that this ratio is maintained in all the other years.

Using the PNAD data and the methodology above, we were able to calculate living costs indexes for the every state as a whole. Then we use the ratios calculated with the 2010 Census to estimate the living costs in 2001-2009 and 2011-2015 for the MR and CS of each state. The 2000 living cost index was assumed to be the same one as 2001. Now a higher index means higher living cost. Graph A1 shows the index calculated for at the state/(countryside or metropolitan region & capital) level using 2010 Census and Table A2 details our final living cost index for each region from 2001-2015. We then invert the logic, so the highest number represents the lowest living cost, and normalize the index to be between 0 and 1. The wages per hour (in 2010 reais) are multiplied by this index to adjust it for the local power purchase.