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

Francisco Costa Letícia Nunes Fabio Miessi Sanches

December 2, 2021

Abstract

This paper exploits location choices of all generalist physicians graduated in Brazil between 2001 and 2013 to study policies aiming at increasing the supply of physicians in underserved areas. We set up and estimate a supply and demand model for physicians. We estimate physicians' locational preferences using a random coefficients discrete choice model. The demand has private establishments competing for physicians with private and public facilities around the country. Policy counterfactuals indicate that quotas in medical schools for students born in underserved areas and the opening of vacancies in medical schools in deprived areas are more cost-effective than financial incentives.

JEL codes: J20, I10, C35, O15, M54.

We thank Kym Ardison, Samuel Bazzi, Luis Braido, Bladimir Carrillo, Claudio Ferraz, Sergio Firpo, Jason Garred, Igal Hendel, Francisco Lima, Claudio Lucinda, Cecilia Machado, Rodrigo Soares, André Trindade, Mar Reguant, Leo Rezende, Rudi Rocha, Marcelo Sant'Anna, Gabriel Ulyssea, Ben Handel, two anonymous referees, and seminar participants at Northwestern, EPGE, USP, Anpec 2015, SBE 2017, and 4th SBE/REAP. We thank the Brazilian Federal Council of Medicine for sharing important data. Costa and Nunes thank for financial support from FAPERJ and from Rede de Pesquisa Aplicada FGV. This study was financed in part by CAPES/Brasil; Grant #001. *Costa*, University of Delaware and FGV EPGE, fcosta@udel.edu; *Nunes*, Insper, leticiafcn@insper.edu.br; *Miessi Sanches*, BI Norwegian Business School and Insper, fmiessi@gmail.com.

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. We focus on generalists because they are directly responsible for the supply of basic healthcare. These professionals are frequently the focus of public policies designed to reduce regional imbalances in the provision of health services.[4] We estimate a model of demand and supply of generalists. We numerically solve the model and use it to simulate different policies.

The supply is based on a random coefficients discrete choice model where physicians choose the region of the country they will work right after graduating from medical school.

[1] Some papers have associated the availability and quality of human resources with better health outcomes, e.g., 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] E.g., Dal Bo et al. (2013); Dunne et al. (2013); Ashraf et al. (2020); Finan et al. (2017).

[4] We do not study medical specialists because these tend to be concentrated in specialized health centers.

The model allows us 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 data 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 local health infrastructure are relevant but less important than these two types of geographic (home and graduation place) preferences. On the demand side, we model the interactions between the public and the private sectors. We characterize the demand for physicians as a Bertrand-Nash game where private health facilities compete for physicians with public and private health facilities operating in the different regions of the country. Our counterfactual experiments indicate that investments in medical schools in underserved areas and the adoption 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 regional imbalances in the distribution of physicians. First, there is a sizable imbalance in the presence of doctors across regions within the country. Figure 1a shows the number of registered physicians per thousand inhabitants in each state's capital and countryside in 2014. While the number of physicians per thousand people ranges from 11.9 to 1.42 across state capitals,[5] 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 imbalances in the distribution of physicians is associated to lower access to preventive care and worse health outcomes.

We assembled a unique dataset with information on all 60,563 generalist physicians that received a medical degree in Brazil between 2001 and 2013. We merge the registries of all

[5]E.g., the United States had 2.6 per thousand people in 2013.

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% 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.

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, (iv) stock of physicians, (v) coverage 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 0.4 and 0.7 in metropolitan areas and the countryside, respectively. Our results also suggest that health infrastructure and amenities impact positively physicians' utility. Importantly, we find that preferences for living in the place of birth and the place of graduation 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 Dal Bo et al. (2013), Agarwal (2015), and 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.

[FIGURE 1 AROUND HERE]

Finally, we find that preferences are heterogeneous according to the rank of the medical school from which the physicians graduated. Those who graduated from better medical schools value more local amenities, are more inelastic to wages, derive lower value for returning to their region of birth, and are the most inclined for staying in their locale of graduation. Because the best 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.

Equipped with physicians' labor supply, we set up a model of regional demand for physicians. We assume that each regional market is composed by an homogeneous set of public and private health facilities. In each market, the private sector chooses private wages to maximize profits considering wages in the public sector in that region as well as wages paid by the public and the private sector in all other regions. We assume that public sector wages in each region are set exogenously according to some bureaucratic process.[6] Equilibrium private wages in each region are then the result of a Bertrand-Nash game between private health facilities operating in all geographic regions.

Lastly, we numerically solve the model and use it to estimate the cost-effectiveness of different public policies on improving the geographic distribution of generalist physicians. Following the World Health Organization, our benchmark geographic distribution of physicians is such that the physicians to population ratio is the same in all regions. We find that policies exploiting physicians' geographic preferences are the most cost-effective. First, quotas in medical schools for students born in underserved areas reduce the geographic imbalance in the distribution of physicians by 63% at minimal costs. Second, the opening

[6]To justify this assumption we present evidence that physicians public sector wages do not respond to private sector wages. This assumption is also in accordance with the literature, e.g., Katz and Krueger (1991), Duggan (2000), and Sanches et al. (2018).

of vacancies in medical schools in areas lacking generalists reduces this imbalance by 65%, but at higher costs compared to quotas. Third, an increase of 50% in wages paid by the public sector to doctors in needy areas is also effective, but at higher costs than the first two alternatives. Last, investing in health infrastructure is less effective and the costliest option.

Our paper directly relates to the literature that studies labor supply of qualified professionals in rural and poorer areas.[7] More broadly, this paper contributes to the development literature on recruitment of workers for service delivery,[8] and the growing literature that applies discrete choice models to study the determinants of migration decisions, demand for neighborhoods and labor sorting.[9] 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 geographic preferences are decisive to explain the locational decisions of qualified health professionals. Importantly, our counterfactuals clarify that the cost-effectiveness of policies leveraging physicians' preferences for living in the place of birth and on preferences for living in the graduation place are substantially different.[10]

[7] Bolduc et al. (1996); Holmes (2005); Dunne et al. (2013); and Agarwal (2015). Studies in a developing country context use stated-preference surveys (de Bekker-Grob et al., 2012).

[8] Dal Bo et al. (2013); Ashraf et al. (2020); Bau and Das (2020); and Deserranno (2019).

[9] Bayer et al. (2007); Kennan and Walker (2011); Bayer et al. (2016); and Diamond (2016).

[10] Kulka and McWeeny (2018) and 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. Different from these papers, we distinguish physicians' preferences for working close to their home region (birthplace) or close to the school from where they graduated. This differentiation is important because policies exploiting the two types of geographic preferences (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. Finally, the demand side of our model takes into account differences in the behavior of the

Methodologically, our model differs in two aspects from models in the literature that study the geographic distribution of high skill professionals. First, in various markets the demand of specialized professionals depends on the public and private sectors. We model the interactions between these two sectors. Second, we explicitly model competition for physicians between firms operating in different markets. Depending on the setting, private sector wage reactions to public policies may have consequences on the results of these policies.

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 and affect retention and performance.[11] Our estimates suggest that policies based on wages may also affect the composition of recruited professionals by attracting those more responsive to financial incentives, in our setting, those who graduated from lower ranked universities.

2 Empirical Context

2.1 The Labor Market for Physicians

The Brazilian market for physicians is dichotomous; both public and private sectors are important providers of health services in the country and physicians can work in both sectors. The Brazilian public health system, known as the Unified Health System (*Sistema Único de Saúde*), 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. The management structure of the public system is decentralized, with the Federal Government transferring responsibilities to states and municipalities to make the health provision more aligned with the local needs (Elias and Cohn, 2003). The private system also plays an relevant role covering close to 25% of the

public and private sector and allows wages in any market to respond to changes in relevant characteristics of other markets.

[11] E.g., Ashraf et al. (2020), and Deserranno (2019).

population through private health insurance plans.

The public and the private sectors offer, roughly, the same types of health services.[12] Services offered by the public health system are free of charge for every Brazilian citizen. Federal, state and municipal governments have autonomy to hire physicians and health workers, typically following budgetary restrictions of each sphere of government. The main admission process to jobs in the public sector is through public exams. The private sector hires health professionals through standard market processes and wages, working journey and fringe benefits are negotiated between workers and employers.

Any physician can choose where to work and is apt to take both 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. A survey conducted by the Federal Council of Medicine (Scheffer et al., 2015) in 2014 with 2,400 doctors shows that around 21.6% of physicians work exclusively in the public sector, 26.9% work exclusively in the private sector and 51.5% have joint appointments in the public and private sectors.

2.2 Geographic Distribution of Physicians

In 1980, Brazil had 1.15 physicians per 1,000 inhabitants. After the inauguration of several medical schools in the past decades, the number of physicians per 1,000 inhabitants increased to about 2.11 by the end of 2015 (Scheffer et al., 2015). Despite the growth in the number of physicians, some regions are still severely underserved. As Figure 1a 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. From a different perspective, Brazilians living in cities with populations smaller than 50,000 people, which

[12]There are, however, differences in the quality of the services offered by both sectors.

represents 32.6% 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).

The Brazilian government implemented different programs to mitigate the undersupply of physicians in disadvantaged areas. The Primary Care Professional Valorization Program (*Provab*), created in 2011, offered competitive and tax-free wages and a 10% increase in the final grade in admission exams to medical specialization programs. The More Physicians Program (*Mais Médicos*), created in 2013, used three strategies: (i) expansion and building of new primary health care units in needy areas; (ii) increasing 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 all new positions. Figure 1b presents the number of vacancies created by the More Physicians Program and the number of physicians who 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 poorer states. Given the excess demand for physicians, the government started to source foreign doctors, especially from Cuba. This suggests that policies based mainly on financial incentives were not sufficient to reduce regional imbalances and that the main hurdle to overcome this imbalance is not the lack of positions with good wages, but some other aspect behind physicians' locational preferences.

In Appendix B, we show evidence that inadequate supply of physicians has immediate implications for local access to healthcare and for the health status of the population.

3 Data and Descriptive Analysis

This section provides preliminary evidence on the importance of different local attributes to physicians choices, that we use to guide the formulation of the structural model in the next sections. We provide detailed information in Appendix G.

3.1 Data

Sample and aggregation level. Our sample consists of all physicians graduated between 2001 and 2013. We aggregate geographical units into pairs metropolitan region/state and countryside/state, amounting to 52 possible choices.[13] All explanatory variables are averaged at this geographical level for the year prior to graduation because we assume physicians graduating and making locational choice at year t observe local characteristics at year $t-1$.

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 physicians that were born and completed medical school in Brazil between 2001 and 2013, our period of study. 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. We have that 40% of recent graduates in our sample 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 registry of all formal job links from the Ministry of Labor (RAIS, Ministry of Labor) from 2001 to 2015 and the National Register of Health Establishments (CNES, Ministry of Health) from 2005 to 2016, using their full name. In the first database we observe detailed information about all employer-employee links, 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 or private sectors. With these two datasets we

[13]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 plus the Federal District that has no countryside. The state of Santa Catarina is composed primarily of metropolitan regions.

know the workplace of all active physicians.[14] 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.[15] Reassuringly, as the descriptive statistics in Table A4 show, physicians that made it into our final sample do not look systematically different from the ones we lose.

Table A1 presents the summary statistics. Column 1 reports where the physicians in our sample chose to work. It depicts a similar scenario to the one in Figure 1. The recent graduates in our sample choose to work more in capitals and metropolitan areas, and more than 50% 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.

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 A1 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 pattern can be seen in physicians' birth location.

Quality of medical schools. We proxy the quality of Brazilian medical schools using the ranks of university courses published by *Folha de São Paulo* newspaper in 2013. Brazil had 214 medical schools in our time frame. The top 25 schools graduated 17% of the doctors in our sample, and these universities are mostly concentrated in metropolitan regions of the South, Southeast, and Midwest.

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 regional markets), and peer effects (more physicians could bring greater learning opportunities and

[14] If a physician works in more than one region, we consider the one with more hours.

[15] We lose 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 were their first location choice after medical school.

networking). The pattern in Table A1 column 4 resemble the one shown in Figure 1.

Health infrastructure. The lack of equipment and supplies may affect physicians' decisions of where to work. To assess how physicians perceive the working environment in each region, we measure the availability of essential medical equipment using CNES data from 2005 to 2015 on the per capita number of ultrasound and x-ray machines, mammographs, computed tomography and magnetic resonance imaging scanners across regions. With these ratios, we developed a normalized index of health infrastructure (Kling et al., 2007). Table A1 column 5 shows that health infrastructure is the worst exactly in regions with fewer physicians.

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.

Amenities. Local amenities play an important role in labor sorting of skilled workers (e.g. Diamond, 2016). We computed a local amenity index which includes: (i) education, comprising the scores in a national exam of local elementary schools (INEP); (ii) entertainment, quantified by the number of cinemas, hotels, restaurants and recreation firms per capita (RAIS and the Cinema Regulatory Agency); (iii) transportation (RAIS and the National Traffic Department); (iv) 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). We combine all these variables into one amenity index (Kling et al., 2007). Table A1 column 7 shows that the imbalance follows the same pattern documented in the health infrastructure index.

Wages. We also construct a measure of physicians' expected compensation. For each of the 52 decision units we calculate the average hourly wage of recently graduated physicians (up to 35 years old) using RAIS. This data allows us to identify the type of contract, that is, whether the physician is a public or private employee. We use this information to obtain the

wages in the private and public sectors, as well as the number of jobs physicians' held in each sector. Average wages, therefore, correspond to the weighted average of wages paid by the public and private sectors in each region. 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 strongly correlated with the earnings opportunities in the informal markets,[16] so this measure captures a meaningful variation in physicians' compensation over time and across regions. We adjust the expected wages for local living costs, calculated using the values of real estate rentals from the National Household Sample Survey (PNAD) and the 2010 Census (see Appendix G for details)

The last column in Table A1 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 imbalance in the number of generalist physicians per capita across the country.

3.2 Descriptive Analysis

We present descriptive evidence to illustrate how the set of local attributes may influence physicians practice location choices. Panel A in Table A2 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. Table A2 reveals two interesting patterns.

First, as the main diagonal of both panels illustrates, most physicians prefer to stay

[16] Figure A3 shows physicians' wages in formal and informal jobs are strongly correlated.

in the same region they were born or completed medical school. Overall, 62.2% (60.6%) of physicians stay in the same region they were born (completed medical school). This number is even larger for physicians that were born (or completed medical school) in the most developed areas of the country (South and Southeast). This evidence suggests that physicians' hold strong preferences over living on their birth place and their graduation place.

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% (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% of physicians born in the North move to the Southeast after graduation. Similar patterns are observed when we look at Panel B.

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.

4 Empirical Framework

This section characterizes our structural model and estimation procedure. The framework is intended to capture key characteristics of physicians' labor supply and of the demand for physicians in Brazil. The labor supply is a discrete choice model that describes physicians' practice location choices as a function of a wide set of information on physicians' characteristics and practice location choices (see Berry et al., 2004). We also develop a model of local demand for physicians in which physicians may work for the public and the private sectors. Our demand model incorporates the different dynamics in these two sectors. In particular,

we consider a profit maximizing private sector that competes for the supply of physicians with the public and the private sector in the different regions of the country. We close the section by describing strategies that we employ to estimate the model.

4.1 Labor Supply

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:

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

$$\tilde{\beta}_{ik} = \beta_k^c + \sum_r z_{ir}\beta_{kr}^o + \beta_k^u v_{ik}. \tag{2}$$

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' average real wages, w_{jt}, are included in the vector of observed attributes of each location.[17] 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},

[17] We discuss the process that determines w_{jt} in the next subsection.

on physician i's indirect utility.[18]

The terms $\tilde{\beta}_{ik}$ are decomposed into a choice specific constant, β_k^c, 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_{kr}^o and β_k^u 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 $\{\beta_k^c\}_k$, $\{\beta_{kr}^o\}_{kr}$ and $\{\beta_k^u\}_k$, respectively. The variables z_{ir} contain physicians attributes that are present in our data, such as age, gender and the rank of the medical school physician i graduated from. The variables v_{ik} contain physicians' unobserved characteristics (e.g. marriage status, number of children, etc).

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

$$u_{ijt} = \delta_{jt} + \sum_{k,r} x_{jtk} z_{ir} \beta_{kr}^o + \sum_k x_{jtk} v_{ik} \beta_k^u + \varepsilon_{ij}, \tag{3}$$

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

This formulation 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. In practice, the inclusion of these interactions produces a model with flexible substitution patterns (see Berry et al., 1995,

[18] 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.

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. From this set we can obtain the probability of any given physician i choosing any given location j, $s_{ijt}\left(\mathbf{x_t}, \mathbf{z_i}; \theta\right)$, 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 for Labor

In many countries, as in Brazil, the public and private sectors are important providers of jobs for newly graduated physicians. As we mentioned in Section 2, most physicians have jobs in public and private health facilities at the same time. Our demand model explicitly incorporates the interactions between these two sectors.

We start by assuming that average wages paid to generalists at region j, period t, w_{jt}, depend on the wages paid by public and private health facilities:

$$w_{jt} = w_{jt}^{pri} \cdot \lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right) + w_{jt}^{pub} \cdot \left[1 - \lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)\right], \tag{5}$$

where, w_{jt}^{pri} (w_{jt}^{pub}) is the hourly wage paid to physicians by private (public) health facilities and $\lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right) \in [0, 1]$ is the fraction of time physicians devote to jobs in the private sector at region j, period t. Consequently, $1 - \lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)$ corresponds to the fraction

of time physicians devote to jobs in the public sector at the same region and period of time.[19] We assume that $\lambda_{jt}(\cdot)$ is a differentiable function of public and private wages. We do not give a structural interpretation to the function $\lambda_{jt}(\cdot)$. We see it as a "reduced form" approximation to physicians' decision process that determines how they split their working journey between the public and the private sectors. This formulation is consistent with the fact that in Brazil most physicians have joint appointment in the public and private sectors. We discuss the estimation of $\lambda_{jt}(\cdot)$ in the next subsection.

Private sector. To characterize the process that determines private wages, we assume that at any region j there exists a set of homogeneous private health facilities that maximizes profits by choosing the wage offered to physicians, w_{jt}^{pri}. Private health facilities solve:

$$\underset{w_{jt}^{pri}}{max} \left(p_{jt} - w_{jt}^{pri}\right) \cdot L_{jt}^{pri}\left(\mathbf{w_t^{pri}}, \mathbf{w_t^{pub}}\right), \tag{6}$$

where $L_{jt}^{pri}\left(\mathbf{w_t^{pri}}, \mathbf{w_t^{pub}}\right)$ is the supply of physicians to *private health facilities* which depends on the wage rate offered by private health facilities in all locations at period t, $\mathbf{w_t^{pri}}$, and the wage offered by public health facilities in all regions, $\mathbf{w_t^{pub}}$. The term p_{jt} is the marginal revenue of a physician for private facilities operating at location j, period t.

We assume that $L_{jt}^{pri}\left(\mathbf{w_t^{pri}}, \mathbf{w_t^{pub}}\right)$ can be factored as $L_{jt}^{pri}\left(\mathbf{w_t^{pri}}, \mathbf{w_t^{pub}}\right) = L_{jt}(\mathbf{w_t}) \cdot \lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)$, where $L_{jt}(\mathbf{w_t})$ is the aggregate supply of physicians for region j, period t, which depends on the vector of average wages, $\mathbf{w_t}$, calculated according to equation (5).[20] Substituting $L_{jt}^{pri}\left(\mathbf{w_t^{pri}}, \mathbf{w_t^{pub}}\right)$ into private facilities' maximization problem we write the

[19] Alternatively, we may interpret $\lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)$ as the share of physicians working in the private sector. We will use these two interpretations interchangeably.

[20] This supply function is obtained from the aggregation of physicians location choice as described in subsection 4.1. We describe how we compute this function in subsection 4.3.

private sector maximization problem as:

$$\underset{w_{jt}^{pri}}{max} \left(p_{jt} - w_{jt}^{pri}\right) \left[L_{jt}\left(\mathbf{w_t}\right) \cdot \lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)\right]. \tag{7}$$

The first order condition of this problem is:

$$p_{jt} = w_{jt}^{pri}\left[1 + \frac{1}{\varepsilon_{L_{jt}}^{pri} + \varepsilon_{\lambda_{jt}}^{pri}}\right], \tag{8}$$

where, $\varepsilon_{L_{jt}}^{pri}$ is the private wage elasticity of $L_{jt}(\cdot)$ and $\varepsilon_{\lambda_{jt}}^{pri}$ is the private wage elasticity of $\lambda_{jt}(\cdot)$. This equation holds for all regions j and all time periods, t. Having estimated $L_{jt}(\cdot)$ and $\lambda_{jt}(\cdot)$, we can recover p_{jt} from the system of first order conditions of this problem, that implicitly defines private wages as a function of public wages in all regions.

Public sector. Modeling the behavior of the public sector is more complex. In particular, differently from the private sector, it is not obvious what is the objective function of the public sector. Indeed, an extensive literature on the behavior of public hospitals finds that they respond differently to financial incentives when compared to private hospitals (e.g., Duggan, 2000). We take an agnostic approach and assume that public wages are exogenously given according to some bureaucratic or political process. Similar assumptions have been used to model the behavior of public enterprises in other settings (e.g., Sanches et al., 2018).

We test the plausibility of this assumption by running regressions of public wages on private wages controlling for location and time fixed effects and using instruments for private wages.[21] Table A5 shows that the coefficients attached to private wages are not significant – the point estimate is equal to -0.03 (p-value 0.84) in our preferred specification. This is consistent with evidence that wages in the public sector are much less responsive to market conditions than wages in the private sector (see Katz and Krueger, 1991).

Under the assumption that public sector wages are set exogenously, we can use the system

[21]These instruments are the same we use to instrument average wages in the aggregate labor supply equation.

of first order conditions of private hospitals in all regions – equation (8) – to solve the model for the (Bertrand-Nash) vector of equilibrium private wages and the fraction of generalists working in the public and private sectors. This completes the description of the demand side of our model. Next we discuss how we estimate the primitives of supply and demand.

4.3 Estimation

The main issue behind the estimation of the supply model is that average 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 health facilities but not by the econometrician). Berry et al. (1995) and Berry et al. (2004) developed methods to estimate discrete choice random coefficients models when observed components of the utility function are endogenous. More recently, Gandhi et al. (2020) showed that standard BLP methods may be biased and inconsistent when there are zeroes in aggregate choice probabilities. In our setting, we observe zero (or close to zero) aggregate choice probabilities for some regions in some years. This pattern is a direct result of the geographic imbalances in the distribution of generalist physicians: for some years we do not observe any generalist physician choosing to work at some regions.

To deal with zeros in aggregate choice probabilities, we apply the control function approach developed in Petrin and Train (2010). The identification assumptions of our baseline model follow closely Agarwal (2015), who also uses a control function to address the endogeneity of wages of medicine residents in an empirical model of the "USA medical match". In Section 5.3, we report a series of robustness checks of the model and estimation procedures.

We define the instruments for average 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[22] as location j, period t – i.e., the average of variables x_{jtk} in equation (1), except

[22]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.

wages, across all $j' \neq j$ that are at the same geographic region as location j (see Berry et al., 1995). Equation (8) provides a justification for this assumption. It shows that private wages at any region respond to observed wages and characteristics of all other regions. This approach will be valid if observed location attributes are determined exogenously (see Nevo, 2000). Notice also that we are already including a full set of region fixed-effects in our model. We believe that this helps to mitigate potential problems with endogeneity of wages.

Mathematically, we assume that average wages are a linear function of the other observed location attributes (except wages) including location and year fixed effects, $\tilde{\mathbf{x}}_{\mathbf{jt}}$, the instruments (as explained in the previous paragraph), $\mathbf{h}_{\mathbf{jt}}$, and an error term, η_{jt}:

$$w_{jt} = \tilde{\mathbf{x}}_{\mathbf{jt}}\gamma_1 + \mathbf{h}_{\mathbf{jt}}\gamma_2 + \eta_{jt}. \tag{9}$$

The instrumental variables, $\mathbf{h}_{\mathbf{jt}}$, do not enter utility directly but affects wages. The vector (γ_1, γ_2) 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 $\mathbf{h}_{\mathbf{jt}}$ 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:

$$\tilde{\xi}_{jt} = \eta_{jt}\psi_1 + \tilde{\eta}_{jt}\psi_2, \tag{10}$$

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

We next define physicians' location choice probabilities based on equations (3), (4), (9) and (10). Given the instruments, $\mathbf{h}_{\mathbf{jt}}$, we can recover the variable η_{jt} via OLS from equation (9), so we proceed as if η_{jt} and (γ_1, γ_2) 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 error 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:

$$s_{ijt}\left(\mathbf{x_t}, \mathbf{z_i}; \theta\right) = \int \frac{exp\left(\delta_{jt} + \sum_{k,r} x_{jtk} z_{ir} \beta^o_{kr} + \sum_k x_{jtk} v_{ik} \beta^u_k\right)}{\sum_q exp\left(\delta_{qt} + \sum_{k,r} x_{qtk} z_{ir} \beta^o_{kr} + \sum_k x_{qtk} v_{ik} \beta^u_k\right)} dF_{\mathbf{v}} dF_{\tilde{\eta}}, \tag{11}$$

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 $\tilde{\eta}_{jt}$.

We first estimate equation (9) by OLS and recover the error term η_{jt}. This term, along with the observed variables, is plugged into the integral in equation (11). The integral is approximated via Monte Carlo 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 $F_{\tilde{\eta}}$.[23] We calculate the integral in (11) as the average across draws of the logit formula. We estimate the vector of parameters, θ, via Simulated Maximum Likelihood. To obtain the aggregate supply of physicians at each location-year, which we call $L_{jt}\left(\mathbf{w_t}\right)$, we integrate equation (11) over the distribution of individuals in our sample.

For the policy analysis we conduct in Section 6, we also need to estimate the following

[23]We also perform robustness checks using 100 and 200 draws. Tables A21 and A22 show that estimates and standard errors change marginally.

elements: *(i)* the fraction of hours physicians work for the private sector at each region and period of time, $\lambda_{jt}(\cdot)$;[24] and *(ii)* marginal revenues of private facilities at each region and period of time, p_{jt}.

We assume that $\lambda_{jt}(\cdot)$ is characterized by the following equation:

$$\lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right) = \frac{exp\left(\alpha_0 + \alpha_1\left(ln\left(w_{jt}^{pri}\right) - ln\left(w_{jt}^{pub}\right)\right) + \alpha_j + \alpha_t + \alpha_{jt}\right)}{1 + exp\left(\alpha_0 + \alpha_1\left(ln\left(w_{jt}^{pri}\right) - ln\left(w_{jt}^{pub}\right)\right) + \alpha_j + \alpha_t + \alpha_{jt}\right)}, \tag{12}$$

where, α_j and α_t are region and year fixed effects and α_{jt} is a region-year effect that is not observed by the econometrician; α_0 is a constant and α_1 is a parameter that captures the effects of (log) wage differentials in the private and public sectors, $ln\left(w_{jt}^{pri}\right) - ln\left(w_{jt}^{pub}\right)$, on the fraction of time physicians work in the private sector. To estimate this model we divide both sides of equation (12) by $1 - \lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)$ and take logs. The resulting equation is linear in the parameters:

$$ln\left[\frac{\lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)}{1 - \lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)}\right] = \alpha_0 + \alpha_1\left(ln\left(w_{jt}^{pri}\right) - ln\left(w_{jt}^{pub}\right)\right) + \alpha_j + \alpha_t + \alpha_{jt}. \tag{13}$$

We allow (log) wage differentials to be correlated with the unobservable α_{jt} and use as instruments for $ln\left(w_{jt}^{pri}\right) - ln\left(w_{jt}^{pub}\right)$ the same instruments we use for average wages, $\mathbf{h_{jt}}$.[25]

Finally, we can compute the marginal revenue of private health facilities, p_{jt}, for each region and period of time. Using the estimates of $\lambda_{jt}(\cdot)$ and $L_{jt}(\cdot)$ we compute the elasticities $\varepsilon_{L_{jt}}^{pri}$ and $\varepsilon_{\lambda_{jt}}^{pri}$. Plugging these elasticities on the right hand side of equation (8), we obtain p_{jt} for each market and period of time.

[24] Because the data on worked hours is noisy, we calculate $\lambda_{jt}(\cdot)$ as the fraction of jobs in the private sector divided by the number of jobs in both sectors in each region and year.

[25] As a robustness check we estimate the model assuming that $\lambda_{jt}(\cdot)$ has a normal distribution. Results in Table A6 are qualitatively and quantitatively close to the results obtained assuming that $\lambda_{jt}(\cdot)$ has a logistic distribution, as in equation (12). An advantage of the logistic model is that it has a closed form, which simplifies the solution of the model.

5 Estimates and Model Fitting

This section presents the estimates of physicians' labor supply and demand parameters and shows the fitting of our model. First, we describe the estimates of the aggregate supply, the function that characterizes the share of physicians working in the private sector, and the wage elasticities of physicians' aggregate supply. We close the section describing the fitting of our supply model to the data, and discussing a series of robustness exercises.

5.1 Estimates

We estimate four different versions of the supply model. Tables 2 and 1 show the estimates of the parameters β^c and β^o in equation (2).[26] The first two columns in Table 2 illustrate 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.[27] All equations include location fixed effects and an interaction between location fixed effects and a time trend. 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.

Overall, the sign of our estimates are as expected in both the Logit and Random Coef-

[26] Table A8 shows the coefficients attached to the interactions between local attributes and unobserved physicians' characteristics, β^u.

[27] The inclusion of the control function in equation (11) 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.

ficient 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 living close to family 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.

[TABLE 1 AROUND HERE]

[TABLE 2 AROUND HERE]

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. Table A9 reports the first stage estimates, parameters (γ_1, γ_2) in equation (2).

The coefficient attached to η_{jt} – from equation (10) 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.[28]

Finally, we show the estimates of $\lambda_{jt}(\cdot)$ in Table A6. Column 3 has the OLS estimates of the effects of wage differentials on the share of physicians working in the private sector. The coefficient is positive and statistically significant at 1%. Column 4 has the estimate when we

[28]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).

instrument wage differentials with the same instruments we used for average wages. With instruments, the coefficient attached to wage differentials increases to 2.45 and remains significant at 1%. As expected, it indicates that increases in private wages relatively to public wages lead to increases in physicians' supply to the private sector.

5.1.1 Heterogeneous preferences

In our model, the interaction terms between local attributes and physicians' observed characteristics capture observable heterogeneous preferences. Our estimates suggest that the fixed cost of migration is lower for men than for women. We find that men derive lower value for staying in the locale of their graduation and those born in metropolitan regions are less inclined to work close their birth place than women.

Importantly, physicians' supply seems to be different according to the rank of the course each physician graduated from. 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 -– note that the top medical school in Brazil has a rank index equal to zero, and those with the worse evaluations have a rank index close to one. Those who graduated from better quality schools in metropolitan regions derive the greatest utility from staying in their local of graduation.

It is important to note, however, that the interpretation of the coefficient of wage and university ranks are subject to an additional caveat. Those who graduated from the very best schools may earn more than the average recent graduate, such that the average wage has smaller influence in their location choice. Likewise, most qualified physicians may have access to higher-end professional networks and expect to have greater earnings trajectories for staying in more competitive markets, such as the metropolitan regions.

5.1.2 Wage elasticities

Finally we report in Table 3 the average wage elasticities of the aggregate supply of physicians, $L_{jt}(\cdot)$. Columns 1 and 2 (3 and 4) show own elasticities and the minimum of cross wage elasticities for metropolitan areas (countryside). Physicians' supply function is inelastic with respect to average wages. The highest own wage elasticity across metropolitan areas is around 0.93 and the lowest is around 0.27. Elasticities in the countryside are relatively higher than in metropolitan areas for most states. These numbers are in line with previous estimates found in the health literature.[29]

[TABLE 3 AROUND HERE]

5.2 Model Fitting

We close the section by discussing the fitting our model to the data. We say the model correctly predicted physician i's choice if we observe physician i choosing location j in the data and, in the model, the probability of going to location j is one of the 3 largest probabilities over the 52 possible locations. Table A10 shows the model predicts correctly the locational choice of 78.9% of the physicians.[30] Furthermore, our model performs well compared to other papers in this literature. For example, in Bayer et al. (2016), 47% of households chose one of the top 10% choices ranked by the model.

Table A7 shows the fitting of the function that describes the proportion of physicians working in the private sector, $\lambda_{jt}(\cdot)$. The estimates of the proportion of physicians working

[29] Baltagi et al. (2005) use a dynamic panel of 1,303 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 jobs packages and find wage elasticities around 0.04.

[30] For 58.3% of physicians, the most likely location to be chosen by the physician is the actual chosen location. A model that randomly allocates physicians to the 52 locations would predict correctly the choice of approximately only 1.92% (1/52) of the physicians.

in the private sector are close to the proportions observed in the data, except in the North region where the model over-predicts the fraction of physicians in the private sector. Table A7 compares average wages computed from equation (5) and average wages observed in the data. Again, our estimates of average wages are close to the observed average wages.

5.3 Robustness Checks

We perform a series of robustness checks of our supply model. Appendix E.1 shows supply estimates using a typical BLP estimator (Berry et al., 1995) instead of the control function. We discuss the key differences between both methods and show that in our setting both models produce similar estimates. Appendix E.2 shows estimates of the baseline model using different sets of instruments. Appendix E.3, shows estimates of the model restricting physicians' choice set according to the historical placement of each medical school. Appendix A.4 shows estimates of the Random Coefficients model with a different number of draws for the unobserved individual characteristics. The results of the models in appendices E.2, E.3, and E.4 are similar to the ones of the baseline model.

6 Counterfactuals and Policy Analysis

What types of policies are more effective to reduce the regional imbalances in the geographic distribution of generalist physicians in Brazil? To answer this question, we compare the counterfactual distribution of physicians produced by four policies: a policy of quotas in medical schools according to students place of birth, redistribution of vacancies in medical schools across the country, improvement of health infrastructure in underdeveloped regions, and financial incentives (increasing public wages in targeted areas).

We also provide a meaningful back-of-the-envelope calculation of the cost-effectiveness of these counterfactual policies. 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. We close the section discussing some important features of our framework.

Technically, we simulate the four counterfactual scenarios by numerically solving the system of first order conditions represented by equation (8) for each policy. We obtain for all periods the vector of equilibrium private wages that is consistent with the counterfactual change using our estimates of the aggregate supply function and of the share of physicians working in the private sector. We then calculate the supply of physicians in each region using the new equilibrium vector of private wages. The results of the counterfactual experiments are robust to changes in the estimation procedure (see Appendix E.1).

[TABLE 4 AROUND HERE]

We evaluate each counterfactual distribution of physicians against the population share in each region. This benchmark follows the WHO recommendation of a minimum of 2.3 health workers per thousand people (WHO, 2006). Table 4 column 1 shows the share of the population in each region. Column 2 shows the predicted distribution of physicians according to our random coefficient model with control function evaluated at average wages computed from equation (5). The quadratic error of 0.57 depicts the baseline distribution imbalance originated from our model relative to the benchmark distribution. Columns 3-6 show the counterfactual distribution of physicians implied by the four policies we discuss next.

6.1 Medical school quotas based on place of birth

The descriptive evidence in Subsection 3.2 and the estimates of the supply function indicate that living close to the place of birth is important to explain physicians' location decisions in Brazil. Based on that, 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% and 70% of the vacancies in all medical schools would be allocated to students born in region "A" and "B", respectively. This policy could be implemented, e.g., through

a centralized admission system resembling the Brazilian Unified Selection System (SISU) already in place.[31]

Table 4 column 3 shows the counterfactual distribution of physicians resulting from this policy. Implementing the quota system would improve the distribution of physicians in the country by approximately 64% – the quadratic error between the observed and benchmark distribution of physicians would fall from 0.57 to 0.20. We would observe, in particular, a substantial increase in the fraction of physicians working in the countryside of the North and Northeast regions, the two most vulnerable areas of the country. Because this policy does not change the location of medical schools, only the composition of their students, it is relatively cheap to be implemented.[32]

6.2 Targeted creation of new vacancies in medical schools

Next, we analyze how the distribution of physicians across the country would be affected by the redistribution of vacancies in medical schools towards regions lacking generalist physicians. Our estimates suggest that migration costs are relatively large, so such policy could help to keep physicians in a specific area.

Table 4 column 4 shows that the targeted creation of vacancies increases the share of physicians in the North and Northeast countryside by more than 60%, attracting mostly

[31] Machado and Szerman (2016) find that SISU enabled universities to attract more students from different states. 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.

[32] 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 (see Arcidiacono and Lovenheim, 2016, for a review). Estevan et al. (2018) show that quotas targeting poorer students in Brazil did not reduce the effort of targeted applicants.

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 66%.

We implement this counterfactual based on the actual expansion of vacancies in medical schools between 2009 and 2016. In this period, the federal government expanded the number of vacancies in public universities and extended government-funded scholarships for private colleges (FIES Program). As a result, from 2009-2016, there were on average additional 10,341 vacancies in medical schools each year relative 1996-2008, the period the physicians in our sample started medical school. Considering that around 40.5% of physicians graduate and remain generalist, we create a scenario where these 4,187 vacancies/year [$0.405 \times 10{,}341$] were created in a way to approximate the distribution of students to the population distribution. 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.[33]

We provide the cost-effectiveness of this policy when implemented in two different ways: through the expansion of vacancies in public universities and vouchers for private universities. Considering the average cost of $13,796 for each undergraduate student in a public university,[34] the total cost of creating these federal vacancies would be $143 million.[35] 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 compute the cost of the program if supplied through vouchers to private universities. We calculate that the cost of implementing this policy using

[33]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 area we randomly duplicate a physician from the same region/CS or region/MR, changing their place of graduation to the one where we are adding a vacancy.

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

[35]All costs deflated to BRL in 2010, and convert to USD.

a voucher policy would be $334 million per year.[36] Thus, to improve the spatial distribution of generalists by one percentage point would cost between 2.2 and 5.0 million dollars per year using the targeted expansion of vacancies in medical schools.

6.3 Improving health infrastructure in N/NE countryside

Our estimates suggest that physicians weigh local health infrastructures when choosing their work place. We, thus, consider a policy that improves local medical and hospital equipment in the countryside of the North and Northeast regions, the two most disadvantaged areas. To implement this counterfactual, we improve the health infrastructure index by 50% in these two regions. Table 4 column 5 shows that better infrastructure helps to attract 14% more physicians to the targeted areas, improving the regional imbalance by 6%.

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 quantify the improvement in the health infrastructure index in the Federal District (DF) over 2005-2012 and calculate the total amount invested in health, excluding the wage bill.[37] 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. We calculate that improving the health infrastructure by 0.1 in the period costed $70.3 million per year.[38]

[36]The average yearly tuition of private medical schools in Brazil is around $32,295 per year. https://www.escolasmedicas.com.br/mensalidades.php accessed on February 21, 2019.

[37]We focus on the Federal District (DF) to avoid double counting in the way municipalities and states inform their health spending. Health expenditure from the System on Public Budgets in Health (SIOPS).

[38]Between 2005 and 2012, the Federal District's health infrastructure index increased

This give use that a 50% increase in the infrastructure index of the North and Northeast regions would cost $600 million per year. Thus, investing in health infrastructure to improve one percentage point in the distribution of physicians would cost $94.2 million per year.[39]

6.4 Increasing wages in N/NE countryside

Last, we consider a 50% bonus on wages paid by the public sector for physicians that take their first job after graduation in the countryside of the North or Northeast. This is likely the most studied policy used to recruit qualified personnel to underserved areas.[40] The last column of Table 4 shows that a 50% increase in wages paid by the public sector attracts about 25% more physicians to these regions, and therefore improves the allocation of generalists by around 12%.

The status quo public wage bill of physicians in our sample in these areas is around $1,760 million per year (considering full time work contracts). The bonus would increase the wage bill on $195 million per year. In sum, using such wage incentive scheme, each percentage point improvement in the distribution of physicians would cost about $15.7 million per year.

6.5 Discussion

We close the section quantifying to which extent taste heterogeneity and demand side responses may influence physicians location choices in our counterfactuals. To quantify the

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. We split the fixed cost over this 7-year period. This overestimates the annual cost of the policy by effectively assuming that the new infrastructure fully depreciates in this 7-year period.

[39]Note that attracting physicians is not the only consequence from improving health infrastructure, as improved health equipment have direct effect on the quality of health provision.

[40]E.g., Dal Bo et al. (2013); Andreassen et al. (2013); Kennan and Walker (2011); Ashraf et al. (2020); Finan et al. (2017).

importance of (observable) taste heterogeneity on the geographic distribution of physicians, we simulate the geographic distribution of physicians assuming that they have the same preferences as physicians graduated from the worst ranked medical school. The quadratic error considering the tastes of graduates from the worst school is 0.58, very close to the benchmark quadratic error of 0.57. This indicates that changing the quality of physicians' training would not produce a major change in the distribution of physicians across the country.

[TABLE 5 AROUND HERE]

In our model, private wages may respond to public wages and exogenous changes in observed characteristics of any region. The possibility of "repricing" may have implications on average wages and, therefore, on the counterfactual equilibrium distribution of physicians. We illustrate the effects of "re-pricing" on the results of the four counterfactual policies we examined. To do this we calculate the distribution of physicians holding private wages fixed at the initial level – i.e., we switch off the possibility of "repricing". Table 5 presents the difference between the counterfactual geographic distribution of physicians in Table 4 with the counterfactual distribution without "re-pricing". We see that in all policies considered, "re-pricing" reduce the number of physicians going to countryside and increase the number of those choosing to work in the metropolitan regions. Column 1, for example, shows that private sector responses to the quota policy increases 0.71% the share of physicians choosing to work in the metropolitan areas of the North region and reduces 0.19% the share of those choosing the Northern countryside.

The last row in the table shows the variation in quadratic errors of the distribution of physicians due to "re-pricing". We find that, in our setting, "re-pricing" has little effect on the equilibrium distribution of physicians. Private wages respond to a relatively small fraction of average wages and, therefore, the effects of changes in private wages on average wages in that region will be small. Second, because cross and own average wage elasticities are low and because the number of competing regions is relatively large, changes in average

wages in a given region will not cause major changes neither in the aggregate supply of physicians to that region nor in private wages of health facilities in other regions.

7 Conclusion

We exploit revealed preferences of all generalist physicians graduated in Brazil between 2001 and 2013 to estimate a regional demand and supply model of generalist physicians. 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 area as where they were born or completed medical school. These two types of geographic preferences are much more important to explain physicians' location choice than wages or quality of health infrastructure.

We estimate the structural model and use it to simulate the effects of different policies on the geographic distribution of generalist physicians in Brazil. Our back-of-the-envelope cost calculations suggest that policies exploiting physicians' willingness to live close to the birth place and place of graduation 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. Increases in public wages for physicians in needy areas also appear to be effective, but at a much higher annual cost than the two previous policies. We highlight that the policies discussed here are related to general practitioners and should not be directly applied more broadly to other more specialized physicians.

References

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

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., Davenport, E., and Lee, S. S. (2020). Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the delivery of Public Services. *American Economic Review*, 110(5):1355–94.

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. (2020). Teacher Value Added in a Low-Income Country. *American Economic Journal: Economic Policy*, 12(1):62–96.

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.

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*, 11(3):156–96.

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. *American Economic Review*, 106(3):479–524.

Duggan, M. G. (2000). Hospital Ownership and Public Medical Spending. *Quarterly Journal of Economics*, 115(4):1343–1373.

Dunne, T., Klimek, S. D., Roberts, M. J., and Xu, D. Y. (2013). Entry, Exit, and the Determinants of Market Structure. *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. *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.

Gandhi, A., Lu, Z., and Shi, X. (2020). Estimating Demand for Differentiated Products with Zeroes in Market Share Data. *Available at SSRN 3503565*.

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

Katz, L. F. and Krueger, A. B. (1991). Changes in the Structure of Wages in the Public and Private Sectors. Technical report, NBER working paper.

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.

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.

Sanches, F., Silva Junior, D., and Srisuma, S. (2018). Banking Privatization and Market Structure in Brazil: a Dynamic Structural Analysis. *RAND Journal of Economics*, 49(4):936–963.

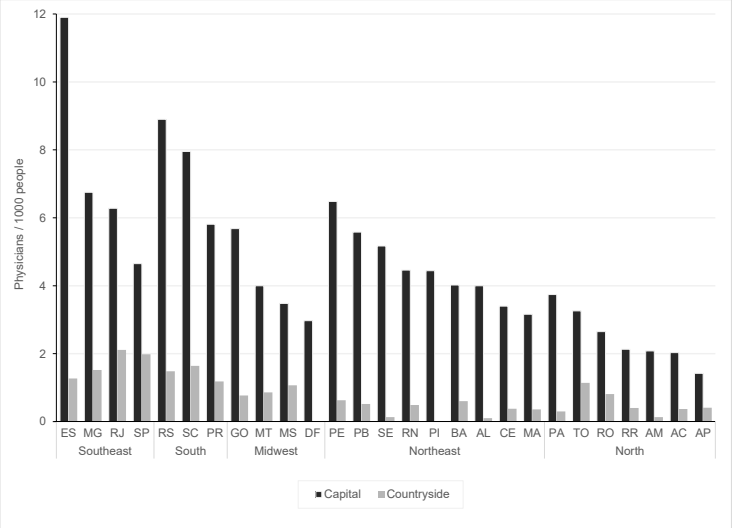
Scheffer, M., Cassenote, A., and Aureliano, B. (2015). Demografia Médica No Brasil 2015. *Departamento de Medicina Preventiva, Faculdade de Medicina da USP. Conselho Regional de Medicina do Estado de São Paulo. Conselho Federal de Medicina.*

WHO (2006). World Health Report 2006: Working Together for Health. *World Health Organization.*

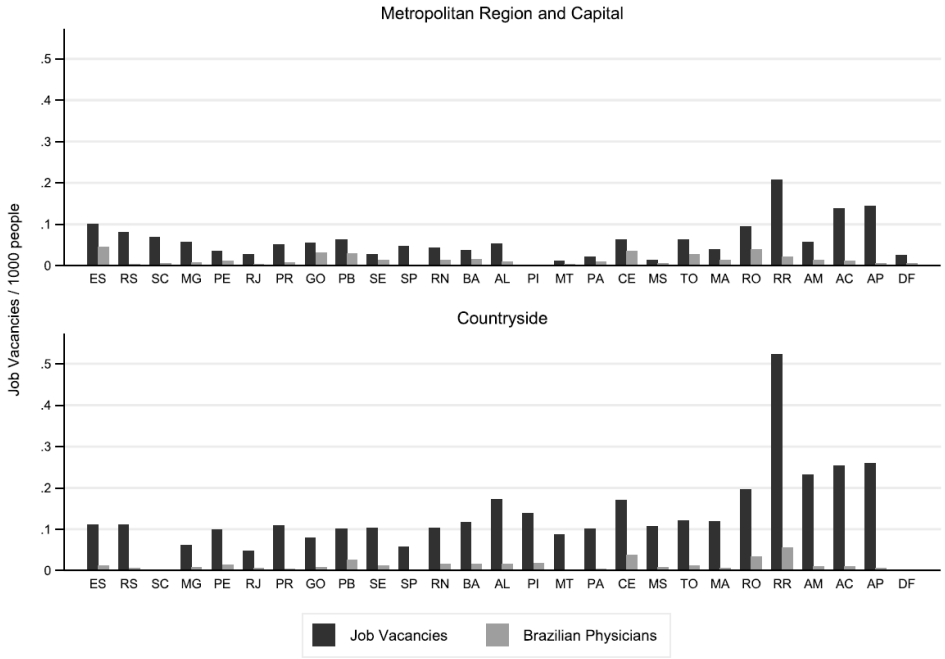
WHO (2010). Increasing access to health workers in remote and rural areas through improved retention. *World Health Organization*, 23(February):3–69.

Figure 1: Physicians and Job Postings in Capitals and Countryside by State

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



(b) Job Vacancies from *Mais Médicos* Program and Positions Filled by Brazilians



Graph (a) shows the number of physicians per capita in capitals (black) and countryside (gray) by state. Graph (b) the number of job vacancies created by the *Mais Médicos* Program (black) between July 2013 to July 2014 and the number of Brazilian physicians that filled the open positions (in gray) per 1,000 people; the upper (bottom) panel shows the metropolitan (countryside) regions across states.

Table 1: 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 shows 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 in parenthesis. Point estimates using 150 simulation draws. All columns include alternative-specific dummies and region-specific year trends. *** p<0.01, ** p<0.05, * p<0.1.

Table 2: 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 shows 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 in parenthesis. Point estimates using 150 simulation draws. All columns include alternative-specific dummies and region-specific year trends. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Wage Elasticities

	Metropolitan Regions		**Countryside**	
	Own	Cross (Min)	Own	Cross (Min)
	(1)	(2)	(3)	(4)
N	0.309	-0.068	0.552	-0.127
NE	0.574	-0.110	0.835	-0.156
SE	0.270	-0.058	0.571	-0.097
S	0.307	-0.051	0.602	-0.089
MW	0.441	-0.103	0.937	-0.162

This table shows the average own and the minimum cross wage elasticity of the supply of physicians, $L_{jt}(\cdot)$, based on the random coefficients logit with control function model. Each row corresponds to one region.

Table 4: Cost-Effectiveness Analysis

				Counterfactuals (%)			
		Population Distrib. (%) (1)	Predicted Distrib. (%) (2)	Birth Region (3)	New Medschool Vacancies (4)	Infra x1.5 (N/NE CS) (5)	Wage x1.5 (N/NE CS) (6)
Metropolitan Regions	N	3.55	3.17	2.41	2.05	3.09	3.08
	NE	12.09	11.92	10.25	9.96	11.66	11.41
	SE	24.98	28.62	29.29	26.96	28.16	28.34
	S	8.54	9.18	9.23	8.01	9.01	9.02
	MW	3.32	3.97	3.39	3.78	3.87	3.83
Countryside	N	4.51	3.30	4.40	4.83	3.71	4.00
	NE	15.77	9.75	13.24	15.97	11.15	12.36
	SE	17.39	21.44	19.01	17.35	21.00	20.31
	S	5.99	4.16	4.10	5.46	4.05	3.90
	MW	3.87	4.48	4.67	5.63	4.30	3.74
Quadratic Error			0.566	0.205	0.193	0.530	0.495
% Reduction in Imbalance				63.76	65.93	6.37	12.40
Cost (1,000 USD) per % Reduction in Imbalance					[2,169 ; 5,066]	94,192	15,727

This table shows the counterfactual distribution and the cost-benefit of different policy simulations. We use the random coefficients model with a control function and 150 draws. Column (1) shows the population distribution (the benchmark to be achieved) and Column (2) where physicians chose to work according to our original model. The following columns show the counterfactual distribution of generalists if: (3) medical schools have quotas based on place of birth; (4) targeted creation of new vacancies in medical schools in places with the lowest student-populaiton ratio; (5) wages and (6) health infrastructure in the North and Northeast countrysides increased by 50%. The quadratic error indicates how far the counterfactual distributions are from the population one. Bellow, there is the percentage reduction in quadratic error each counterfactual would produce relative to the predicted distribution quadratic error. The last row shows the cost (2010 USD) incurred in each policy for a 1% reduction in imbalance. The lower (upper) bound cost of opening new medical school vacancies are calculated with the average cost of vacancies in a public (private) university.

Table 5: Variation in Distribution With and Without Re-pricing

		Birth Region (1)	New Medschool Vacancies (2)	Infra x1.5 (N/NE CS) (3)	Wage x1.5 (N/NE CS) (4)
Metropolitan Regions	N	0.71	1.11	0.05	0.29
	NE	0.27	0.15	-0.01	0.11
	SE	0.01	0.02	0.02	0.51
	S	0.00	0.03	0.01	0.36
	MW	0.23	-0.02	0.00	0.21
Countryside	N	-0.19	0.01	0.03	0.32
	NE	-0.22	-0.25	-0.03	-1.84
	SE	-0.05	-0.01	-0.01	0.07
	S	-0.07	-0.02	-0.03	-0.06
	MW	-0.13	-0.05	-0.04	0.06
% Variation in Quadratic Error		-0.004	-0.704	-0.036	-1.045

This table shows the counterfactual distribution of different policy simulations with and without endogenous changes in private sector wages – "re-pricing". Each cell shows the difference in the counterfactual allocation with re-pricing as in Table 4 and the distribution of the respective policy without re-pricing – i.e., holding private wages constant. We use the random coefficients model with a control function and 150 draws.

Appendix (for online publication only)

Appendix for "How to Attract Physicians to Underserved Areas? Policy Recommendations from a Structural Model" (Costa, Nunes and Sanches, July 2021)

- Section A present the descriptive statistics of the data and of physicians location choice mentioned in the paper.
- Section B discusses correlation between health provision and health outcomes mentioned in Section 2.
- Section C presents supplementary material on the construction of our sample as mentioned in the paper.
- Section D presents additional results of estimates of the supply model.
- Section E contains results of several robustness checks of our baseline results.
- Section F contains supplementary summary statistics.
- Section G describes the data sources and data cleaning process in detail.

A Summary Statistics

Table A1: 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 our main variables. Column 1 shows the decision of practice location after graduation. Columns 2 and 3 display where physicians finished the medical school and where they were born. Columns 4–9 show the regions' attributes, respectively: (4) the ratio of physicians per 1,000 people; (5) the health infrastructure index; (6) the percentage of the population that has health insurance; (7) the amenity index has mean zero and their values vary between: [-1.04,1.98]; (8) the average hourly wage generalists up to 35 years old receive in each region, multiplied by the living cost index.

Table A2: 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 shows the 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) that decided to work in region j (column). Analogously, in *Panel B* indicate the region physicians did medical school (rows) and their practice location choices (column). Numbers in bold mark the diagonal.

B Health Provision and Health Outcomes

This subsection provides descriptive evidence on the correlation between health provision and health outcomes in Brazil. Figure A1 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 A1a shows that those living in rural areas are about 38% more likely to not have been to a Doctor's appointment in the last 12 months than those living in urban areas. Figure A1b indicates that infants in rural areas are about 20% less likely to visit the Doctor in the first 30 days of life, and about 50% 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 A1c and A1d 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 A1 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 A3. 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.

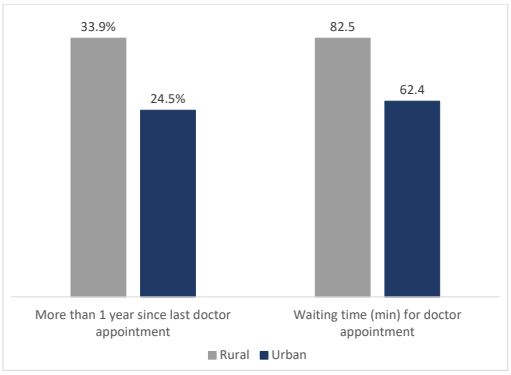
We derive some stylized facts from this exercise. First, we find a positive correlation between physicians per capita in a given municipality-year and infants/mothers' health outcomes. Column 1 presents the raw correlations which are statistically significant at one percent for our three variables. Column 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.

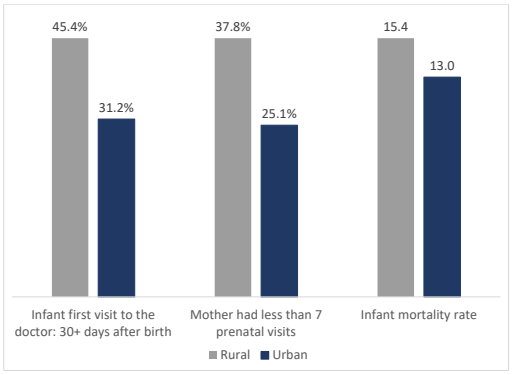
Table A3: Reduced Form Motivation

	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.	-2.26***	-2.29***	-2.34***	-2.36***	-0.64***
	(0.23)	(0.23)	(0.22)	(0.22)	(0.22)
Panel B. *Dep.var.:* Infant mortality rate					
Physicians p.c.	-0.57***	-0.57***	-0.58***	-0.58***	0.16
	(0.10)	(0.10)	(0.10)	(0.10)	(0.29)
Panel C. *Dep.var.:* Maternal mortality rate					
Physicians p.c.	-6.62***	-6.60***	-6.34***	-6.33***	1.13
	(1.47)	(1.46)	(1.46)	(1.46)	(4.63)
Countryside FE		Yes			
State FE			Yes		
State-Countryside FE				Yes	
Municipality FE					Yes

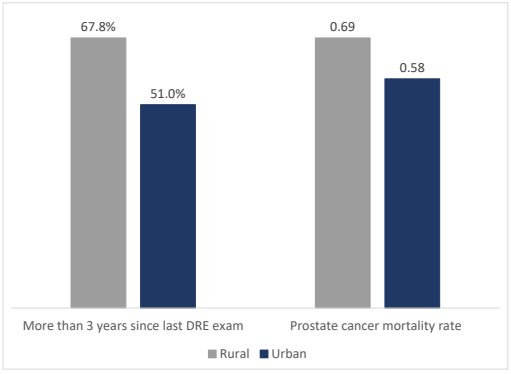
This table presents the results of OLS regressions using different specifications and dependent variables as indicated in each panel and column. Unit of observations municipality-year (N= 66,780), between 2005 and 2016. All regressions include a constant, year fixed effects, and state-countryside trends. 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 1,000 live births), and in *Panel C* is the mortality rate of mothers during delivery (per 100,000 live births). Mean dependent variables: 39.0 (Panel A), 14.9 (Panel B), and 61.6 (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.



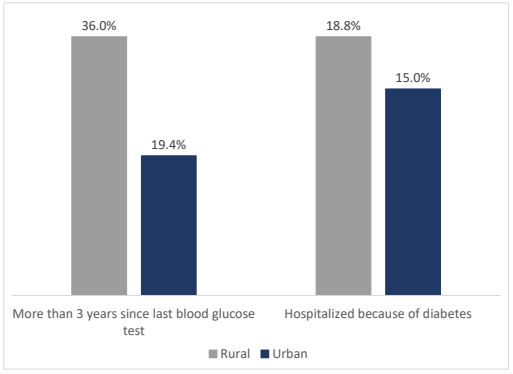
(a) Access to Doctors



(b) Infants' Health



(c) Men's Health



(d) Diabetes

Figure A1: 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 (the recommended by the Brazilian Health Ministry), 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 the Mortality Information System (SIM-Datasus). The others are from the 2013 National Health Survey (PNS).

C Sample

Table A4: Physicians Lost During Merge process - Mean Difference

	Matched (1)	Not Matched (2)	Difference p-value (3)	Normalized Difference (4)
Birth Region (%)				
MR N	3.7	3.8	0.83	0.00
MR NE	13.2	11.6	0.00	0.05
MR SE	27.1	36.2	0.00	-0.19
MR S	8.0	9.8	0.00	-0.07
MR MW	4.6	5.6	0.00	-0.04
CS N	1.7	1.1	0.00	0.05
CS NE	7.4	3.7	0.00	0.16
CS SE	24.8	19.6	0.00	0.13
CS S	5.9	6.1	0.31	-0.01
CS MW	3.6	2.7	0.00	0.05
Medschool Region (%)				
MR N	8.1	5.3	0.00	0.11
MR NE	17.9	15.9	0.00	0.05
MR SE	27.2	32.2	0.00	-0.11
MR S	10.5	12.2	0.00	-0.05
MR MW	4.4	4.9	0.01	-0.02
CS N	0.1	0.0	0.00	0.05
CS NE	0.7	0.2	0.00	0.08
CS SE	27.9	26.3	0.00	0.04
CS S	2.9	2.9	0.84	0.00
CS MW	0.3	0.1	0.00	0.04
Graduation Year (%)				
2001	2.6	14.2	0.00	-0.43
2002	4.4	10.0	0.00	-0.22
2003	5.1	10.8	0.00	-0.21
2004	5.7	10.9	0.00	-0.19
2005	6.1	8.4	0.00	-0.09
2006	6.2	7.8	0.00	-0.06
2007	6.2	7.2	0.00	-0.04
2008	6.4	6.1	0.18	0.01
2009	7.4	4.9	0.00	0.11
2010	8.7	4.8	0.00	0.16
2011	11.5	4.8	0.00	0.25
2012	13.5	4.7	0.00	0.31
2013	16.3	5.4	0.00	0.36
Age	25.7	24.9	0.00	0.29
% Male	52.3	34.0	0.00	0.37
Medschool Rank	83.5	70.4	0.00	0.24
Number of Obs	46,989	13,574		

This table shows the difference in means between physicians we could match to a working region up to three years after graduation and the ones we could not. Most of our losses were due to: (i) Misspelling and errors in the names and dates of birth; (ii) RAIS and CNES having less accurate registries in their early years; (iii) Physicians which only work shifts up to 24 hours in hospitals, and A&E departments. The two major differences are in gender and graduation year. More women were not matched because in Brazil it is quite common for them to include their husbands' surname after getting married. As for graduation year, our match is better in later years because RAIS and CNES databases improved over time. The other variables do not show high differences.

D Baseline Model

Table A5: Correlations between Wages in the Private and Public Sectors

	Log public wage	
	(1)	(2)
Log private wage	0.239***	-0.038
	(0.084)	(0.185)
R-squared	0.316	0.240

This table displays the estimates of regressing the log average public wages on log average private wages controlling for location and time fixed effects. Column 2 shows estimates using instruments for private wages. 674 observations. Robust standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

Table A6: Estimates of $\lambda_{jt}(\cdot)$

	Normal Distribution Instrumented		Logistic Distribution Instrumented	
	(1)	(2)	(3)	(4)
Panel A. Estimates of $\lambda_{jt}(\cdot)$				
Log wage difference (α_1)	0.285**	1.243***	0.561***	2.449***
	(0.125)	(0.384)	(0.155)	(0.714)
Panel B. First stage				
Health Infra Instrument		0.992		0.992
		(0.846)		(0.846)
Health Insurance Instrument		1.401**		1.401**
		(0.691)		(0.691)
Physicians Ratio Instrument		0.415		0.415
		(0.597)		(0.597)
Amenity Index Instrument		-1.168*		-1.168*
		(0.606)		(0.606)

This table displays the estimates of $\lambda_{jt}(\cdot)$ in Panel A, (α_1) from equation (12). All regressions with location and time fixed effects. The first two columns show results assuming that $\lambda_{jt}(\cdot)$ has a normal distribution, and the last two assume it has a logistic distribution. Columns 2 and 4 presentes results using as instruments for $ln\left(w_{jt}^{pri}\right) - ln\left(w_{jt}^{pub}\right)$ the same instruments we use for average wages, $\mathbf{h_{jt}}$. Column 4 Panel A shows the estimate used in couterfactuals. Panel B shows the first stage estimates, KP F-statatistic 4.11. 532 observations. Robust standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. Table A26 presents the summary statatistics of wages in the public and private sectors.

Table A7: Actual and Predicted Wages and Shares of Hours in the Private Sector

		Wages		$\lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)$	
		Actual (1)	Predicted (2)	Actual (3)	Predicted (4)
Metropolitan Regions	N	9.90	9.60	28.29%	28.07%
	NE	17.74	17.59	36.04%	34.22%
	SE	9.86	9.88	42.02%	38.81%
	S	9.84	9.84	52.55%	47.91%
	MW	14.03	13.94	31.64%	30.79%
Countryside	N	18.30	20.33	10.02%	31.51%
	NE	26.20	26.33	15.01%	19.24%
	SE	18.04	18.08	33.22%	29.59%
	S	19.43	19.38	31.40%	25.75%
	MW	27.40	25.83	21.53%	22.82%

This table shows the fitting of our estimated wages (columns 1 and 2) and of the $\lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)$ function (columns 3 and 4). Column 1 shows the average wage as observed in the data. Column 2 shows the average wage computed using the $\lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)$ function. Column 3 has $\lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)$ as observed in the data. Column 4 reports $\lambda_{jt}\left(w_{jt}^{pri}, w_{jt}^{pub}\right)$ as predicted by the model.

Table A8: Preference Estimates – Interaction Term β^u

	Random Coefficients (1)	Random Coefficients with Control Function (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 A9: Control Function First Stage

	Main Estimates
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 of our main estimates in the paper. Regression includes region and year dummies. Robust standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

Table A10: Model Fit – Actual and Predicted Frequencies

		Actual Frequency (1)	Correctly Predicted (2)	% Correctly Predicted (2/1) (3)
Metropolitan Regions	N	1,552	1,259	81.1
	NE	5,759	5,081	88.2
	SE	14,303	13,288	92.9
	S	4,683	3,656	78.1
	MW	1,874	1,376	73.4
Countryside	N	1,374	321	23.4
	NE	3,854	1,732	44.9
	SE	10,143	8,557	84.4
	S	1,845	1,265	68.6
	MW	1,602	555	34.6
Total		46,989	37,090	78.9

This table shows the fitting of the aggregate supply of physicians using the estimated parameters from the random coefficients logit with control function. Column 1 shows the distribution of physicians as observed in the data. Column 2 has the frequency of physician choices that were correctly predicted by the model. We say the model correctly predicted physician i's choice if we observe physician i choosing location j in the data and, in the model, the probability of going to location j is one of the 3 largest probabilities over the 52 possible locations. Columns 3 is the ratio between column 2 and column 1. Sample size: 46,989. We take 10,000 draws from the estimated normal distributions to evaluate the choice of each physician.

E Robustness Checks

This appendix shows results of several robustness checks of our baseline model. Appendix E.1 shows the estimates of our supply model and counterfactuals when we use the BLP estimator instead of the control function approach and a detailed discussion about the differences between the two estimation procedures; Appendix E.2 shows the estimates of the supply model using different sets of instruments; Appendix E.3 shows the estimates of the supply model using different definitions for physicians choice sets; and, Appendix E.4 shows the results of the supply model when we change the number of draws of unobserved individual characteristics used in the estimation of our baseline model.

E.1 Control Function and BLP

To understand the sensitivity of our results to the choice of the estimators, we re-estimate our model using a standard BLP estimation procedure. We estimate the model below:

$$s_{ijt}\left(\mathbf{x_t}, \mathbf{z_i}; \theta\right) = \frac{exp\left(\delta_{jt} + \sum_{k,r} x_{jtk} z_{ir} \beta^o_{kr}\right)}{\sum_q exp\left(\delta_{qt} + \sum_{k,r} x_{qtk} z_{ir} \beta^o_{kr}\right)}, \tag{14}$$

where, $\delta_{jt} = \sum_k x_{jtk}\beta^c_k + \xi_j + \tilde{\xi}_j \cdot t + \tilde{\xi}_{jt}$. This model does not include unobserved individual level heterogeneity – i.e. it restricts β^u in equation (6) in our paper to zero. This restriction reduces considerably the time we need to estimate the model and still produces reasonable estimates. We first estimate (δ, β^o) simultaneously combining the Maximum Likelihood estimator with the BLP contraction mapping. The contraction mapping is useful to reduce the dimensionality of the estimation problem – see also Falcettoni (2020). From the estimates of δ_{jt} obtained in the first step we estimate β^c using the equation for δ_{jt} and BLP instruments (same as used in our baseline specification) for wages. The main advantage of the BLP over the control function approach is that, as δ_{jt} is estimated along with the coefficients of the interactions between observed region characteristics and demographic variables, consistency of the coefficients of these interactions (β^o) does not depend on assumptions about the joint distribution of observed region characteristics and unobserved region characteristics ($\tilde{\xi}_{jt}$), while the estimates of the model with the control function correction for endogeneity does.

Tables A12 and A11 displays the coefficients of these interactions for three models: (i) Multinomial Logit without the control function (i.e. without any correction for endogeneity of wages as shown in column 1 of Table 3); (ii) Multinomial Logit with the control function (shown in column 2 of Table 3); and (iii) the same model estimated using the BLP estimator. They reveal that the estimates of all interactions are very close. For example, the coefficients

of interactions of wages with gender, age and rank in the model with the control function (column 2) are, respectively, 0.215, -0.899 and 0.384; the same estimates in column 3 (BLP estimator) are 0.229, -0.908 and 0.367, respectively. It is also interesting to note that the coefficients of wage interactions in columns 2 and 3 are also very close to the coefficients of the interactions in column 1, which does not control for endogeneity.

Together these results seem to suggest that the coefficients of the interactions are robust to assumptions on the distribution of unobservables. Similar patterns were found in Petrin and Train (2010) and Agarwal (2015), which also use control functions to estimate discrete choice models. Overall, the proximity between the BLP and control function estimates of these interactions indicates that the potential advantage of the BLP over the control function approach does not appear to be relevant in our setting.

To understand the sensitivity of our counterfactual exercises to changes in the estimation procedure we also run all the counterfactuals using the BLP model in column 3 of Tables A12 and A11. As shown in Table A13, the results of this exercise are qualitatively and quantitatively very close to the results obtained from our baseline model.

Finally, we illustrate how the presence of zeroes in aggregate choice probabilities may affect our results. We compute the fitting of model (14) using both, control function and BLP estimates. Table A14 shows that the fitting of the BLP model is worse than the fitting of the model with the control function. In principle, these differences are not expected because the estimates of (δ, β^o) produced by the two models are very close.[41] The key difference is that, as a consequence of zeros in aggregate choice probabilities, δ_{jt} for some regions/years cannot be estimated by the BLP estimator. Therefore to compute the fitting of the BLP model in Table A15 we excluded these alternatives from physicians choice set.

We then investigated how this issue affects the fitting of the BLP model. First, we compute the fitting of the model using (i) the BLP estimates of β^o and (ii) the control function estimates of δ excluding the years/regions with δ_{jt}'s that could not be estimated by the BLP estimator. The fitting measure generated by this exercise is quite close to the fitting of the BLP model in Table A14 – i.e. with BLP estimates of (δ, β^o). Next we recomputed the fitting of the model using the full set of δ_{jt}'s obtained when we estimate the model with the control function (i.e. we do not exclude regions/years with zero aggregate choice probabilities). Table A15 shows that the fitting of this model gets close to the fitting of the model with control function in Table A15. This result indicates that the presence of zero

[41] The fitting of these models were obtained directly from the estimates of (δ, β^o) and, therefore, the differences between the estimates of β^c have not direct effects on our fitting measures.

aggregate choice probabilities for some regions/years may have consequences for our analysis and serves to justify the utilization of the control function.

E.2 Other Instruments

We perform robustness exercises using as additional instruments the wages of university professors (excluding physicians) in the public sector *(i)* in the same region, and *(ii)* in neighboring regions. We use university professors wages because they are also high skill professionals and in, Brazil, wages in the public sector are set based on years of formal education and experience. Implicitly, to use the first (second) instrument we are assuming that unobserved characteristics of a given region – that are relevant to explain physicians choices – are not correlated with wages of public universities professors in that (neighboring) region(s). There are reasons to believe that these assumptions are plausible. Most public universities are controlled by the federal government and wages paid by federal universities, by law, must be the same across regions. This implies that wages paid by federal universities do not depend on characteristics (observed or unobserved) of any particular region. Nonetheless, average wages in federal universities vary across regions because they depend on faculty composition, such as degrees and seniority. The estimates of the supply model when these instruments are used are pretty close to our baseline estimates.

Table A11: Preference Estimates - Regions' Characteristics - Control Function vs BLP

	Multinomial Logit (1)	Multinomial Logit Control Function (2)	Multinomial Logit BLP (3)
Physicians Ratio	**0.341**	**0.522**	**-3.350*****
	(0.536)	**(0.540)**	**(0.520)**
× Male	-0.303	-0.306	-0.313
	(0.188)	(0.188)	(0.189)
× Age	-0.368	-0.356	-0.258
	(0.580)	(0.580)	(0.581)
× Medschool Rank	-0.107	-0.105	-0.032
	(0.322)	(0.322)	(0.324)
Health Infrastructure	**2.291*****	**2.483*****	**1.257****
	(0.500)	**(0.504)**	**(0.532)**
× Male	-0.093	-0.090	-0.089
	(0.182)	(0.182)	(0.184)
× Age	-0.845	-0.850	-0.952
	(0.547)	(0.547)	(0.551)
× Medschool Rank	-0.314	-0.312	-0.339
	(0.310)	(0.310)	(0.314)
Health Insurance	**0.174**	**0.022**	**3.817*****
	(0.371)	**(0.375)**	**(0.364)**
× Male	-0.247**	-0.245**	-0.241**
	(0.102)	(0.102)	(0.102)
× Age	-1.621***	-1.630***	-1.667***
	(0.323)	(0.323)	(0.323)
× Medschool Rank	1.126***	1.124***	1.093***
	(0.175)	(0.175)	(0.176)
Amenity Index	**0.783*****	**0.654****	**1.142*****
	(0.271)	**(0.275)**	**(0.247)**
× Male	0.139	0.138	0.143
	(0.091)	(0.091)	(0.091)
× Age	0.067	0.068	0.099
	(0.273)	(0.273)	(0.274)
× Medschool Rank	-0.466***	-0.463***	-0.478***
	(0.151)	(0.151)	(0.153)
Avg Hourly Wage	**-0.394***	**2.673****	**1.918***
	(0.231)	**(1.147)**	**(0.984)**
× Male	0.216	0.215	0.229*
	(0.137)	(0.137)	(0.137)
× Age	-0.898**	-0.899**	-0.908**
	(0.393)	(0.393)	(0.394)
× Medschool Rank	0.381*	0.384*	0.367
	(0.225)	(0.225)	(0.228)
Region Unobs		**-2.327*****	
		(0.853)	

This table shows the preference estimates for a plain multinomial logit (1) and also for multinomial logits that use the control function method (2) and the BLP method (3) to deal with the wage endogeneity. Sample size: 46,989. Respective log likelihoods: -82395.80 and -81902.94. Standard deviations in parenthesis. Control function approach include alternative-specific dummies and region-specific year trends. *** p<0.01, ** p<0.05, * p<0.1.

Table A12: Preference Estimates - Physicians' Place of Birth and Medical School Region - Control Function vs. BLP

	Multinomial Logit (1)	Multinomial Logit Control Function (2)	Multinomial Logit BLP (3)
Birth Metrop Region	**1.752*****	**1.751*****	**1.772*****
	(0.051)	**(0.051)**	**0.051**
× Male	-0.084**	-0.083**	-0.084**
	(0.040)	(0.040)	(0.040)
× Age	0.290**	0.292**	0.258**
	(0.127)	(0.127)	(0.128)
× Medschool Rank	1.306***	1.307***	1.298***
	(0.066)	(0.066)	(0.067)
Birth Countryside Region	**3.100*****	**3.100*****	**3.109*****
	(0.050)	**(0.050)**	**0.05**
× Male	0.083**	0.083**	0.083**
	(0.038)	(0.038)	(0.038)
× Age	-0.263**	-0.262**	-0.267**
	(0.110)	(0.110)	(0.111)
× Medschool Rank	0.036	0.036	0.025
	(0.063)	(0.063)	(0.064)
Medschool Metrop Region	**3.723*****	**3.723*****	**3.752*****
	(0.046)	**(0.046)**	**0.046**
× Male	-0.169***	-0.169***	-0.174***
	(0.036)	(0.036)	(0.036)
× Age	0.625***	0.624***	0.619***
	(0.107)	(0.107)	(0.107)
× Medschool Rank	-0.425***	-0.425***	-0.422***
	(0.061)	(0.061)	(0.062)
Medschool Countryside Region	**1.726*****	**1.727*****	**1.734*****
	(0.074)	**(0.074)**	**0.074**
× Male	-0.143***	-0.143***	-0.144***
	(0.052)	(0.052)	(0.052)
× Age	1.152***	1.151***	1.136***
	(0.161)	(0.161)	(0.161)
× Medschool Rank	-0.008	-0.009	-0.001
	(0.098)	(0.098)	(0.099)

This table shows the preference estimates for a plain multinomial logit (1) and also for multinomial logits that use the control function method (2) and the BLP method (3) to deal with the wage endogeneity. Sample size: 46,989. Respective log likelihoods: -82395.80 and -81902.94. Standard deviations in parenthesis. Control function approach include alternative-specific dummies and region-specific year trends. *** p<0.01, ** p<0.05, * p<0.1.

Table A13: Counterfactual Physicians' Distribution - Multinomial Logit with the BLP Approach to Endogeneity

				Counterfactuals (%)			
		Population Distrib. (%) (1)	Predicted Distrib. (%) (2)	Birth Region (3)	Medschool Region (4)	Infra x1.5 (N/NE CS) (5)	Wage x1.5 (N/NE CS) (6)
Metropolitan Regions	N	3.55	3.21	2.44	2.36	3.16	3.11
	NE	12.09	12.19	9.38	9.91	12.04	11.59
	SE	24.98	28.53	28.68	27.72	28.26	28.23
	S	8.54	9.42	8.68	8.15	9.33	9.26
	MW	3.32	3.95	3.79	4.01	3.90	3.81
Countryside	N	4.51	3.12	4.73	4.30	3.23	3.55
	NE	15.77	9.51	14.80	15.45	9.90	11.38
	SE	17.39	22.54	19.08	18.34	22.73	22.08
	S	5.99	3.91	3.89	5.06	3.87	3.69
	MW	3.87	3.63	4.54	4.70	3.58	3.30
Quadratic Error			0.566	0.256	0.280	0.567	0.527
% Reduction in Misallocation				54.73	50.52	0.24	6.95

This table shows the counterfactual distribution of different policy simulations. We use the multinomial logit model with the BLP approach to endogeneity. Column (1) shows the population distribution (the benchmark to be achieved) and Column (2) where physicians chose to work according to the model. The following columns show the counterfactual distribution of generalists if: (3) medical schools have quotas based on place of birth; (4) targeted creation of new vacancies in medical schools in places with the lowest student-population ratio; (4) wages and (5) health infrastructure in the North and Northeast countryside increased by 50%. The quadratic error indicates how far the counterfactual distributions are from the population one. Below, there is the percentage reduction in quadratic error each counterfactual would produce relative to the predicted distribution quadratic error.

Table A14: Model Fit – Actual and Predicted Frequencies - Control Function vs. BLP

			Logit with Control Function		Logit with BLP	
		Actual Frequency (1)	Correctly Predicted (2)	% Correctly Predicted (2/1) (3)	Correctly Predicted (4)	% Correctly Predicted (4/1) (5)
Metropolitan Regions	N	1,552	1,196	77.1	1,091	70.3
	NE	5,759	4,986	86.6	4,721	82.0
	SE	14,303	13,202	92.3	11,762	82.2
	S	4,683	3,657	78.1	3,483	74.4
	MW	1,874	1,368	73.0	1,232	65.7
Countryside	N	1,374	319	23.2	297	21.6
	NE	3,854	1,802	46.8	1,658	43.0
	SE	10,143	8,412	82.9	7,754	76.4
	S	1,845	1,265	68.6	1,133	61.4
	MW	1,602	553	34.5	524	32.7
Total		46,989	36,760	78.2	33,655	71.6

This table shows the fitting of the aggregate supply of physicians using the estimated parameters from the logit model with control function and the logit model with the BLP approach to endogeneity. Column 1 shows the distribution of physicians as observed in the data. Columns 2 and 4 have the frequency of physician choices that were correctly predicted by both models. We say the model correctly predicted physician i's choice if we observe physician i choosing location j in the data and, in the model, the probability of going to location j is one of the 3 largest probabilities over the 52 possible locations. Columns 3 (5) is the ratio between column 2 (4) and column 1. Sample size: 46,989. We take 10,000 draws from the estimated normal distributions to evaluate the choice of each physician.

Table A15: Model Fit – Actual and Predicted Frequencies - Logit BLP

			Delta Original Model, Full Choice Set		Delta Original Model, BLP Choice Set	
		Actual Frequency (1)	Correctly Predicted (2)	% Correctly Predicted (2/1) (3)	Correctly Predicted (4)	% Correctly Predicted (4/1) (5)
Metropolitan Regions	N	1,552	1,026	66.1	897	57.8
	NE	5,759	4,795	83.3	4,486	77.9
	SE	14,303	13,331	93.2	11,888	83.1
	S	4,683	3,663	78.2	3,488	74.5
	MW	1,874	1,331	71.0	1,158	61.8
Countryside	N	1,374	311	22.6	295	21.5
	NE	3,854	1,713	44.4	1,630	42.3
	SE	10,143	8,415	83.0	7,751	76.4
	S	1,845	1,229	66.6	1,116	60.5
	MW	1,602	429	26.8	525	32.8
Total		46,989	36,243	77.1	33,234	70.7

This table shows the fitting of the aggregate supply of physicians using the estimated parameters from the logit BLP model with deltas from the original model (random coefficients with control function). Results using the full choice set and also the restricted choice set imposed by the BLP approach (excluding region-years with zero physicians). Column 1 shows the distribution of physicians as observed in the data. Columns 2 and 4 have the frequency of physician choices that were correctly predicted by both models. We say the model correctly predicted physician i's choice if we observe physician i choosing location j in the data and, in the model, the probability of going to location j is one of the 3 largest probabilities over the 52 possible locations. Columns 3 (5) is the ratio between column 2 (4) and column 1. Sample size: 46,989. We take 10,000 draws from the estimated normal distributions to evaluate the choice of each physician.

Table A16: Control Function First Stage with Alternative Instruments

		Alternative Instrument	
	Main Estimates	Univ. Professor Wages in Own Region	Univ. Professor Wages in Other Regions
	(1)	(2)	(3)
Constant	0.016	0.026	0.024
	(0.030)	(0.030)	(0.033)
Health Infrastructure	-0.060	-0.066	-0.057
	(0.052)	(0.052)	(0.052)
Physicians Ratio	0.018	0.032	0.014
	(0.102)	(0.101)	(0.102)
Health Insurance	0.013	0.014	0.014
	(0.072)	(0.072)	(0.072)
Amenity Index	0.073	0.068	0.069
	(0.047)	(0.049)	(0.050)
Health Infra Instrument	0.212**	0.226**	0.215**
	(0.095)	(0.096)	(0.096)
Physicians Ratio Instrument	-0.479***	-0.497***	-0.480***
	(0.139)	(0.141)	(0.138)
Health Insurance Instrument	0.062	0.065	0.067
	(0.094)	(0.095)	(0.095)
Amenity Index Instrument	0.302*	0.291*	0.302*
	(0.157)	(0.161)	(0.162)
Univ. Prof. Wages		-0.047**	
		(0.020)	
Univ. Prof. Wages Instrument			-0.032
			(0.048)
Observations	676	676	676
F-Statistics	78.97	83.61	76.64

This table displays the control function first stage of our main estimates in the paper (column 1) and for alternative instruments (column 2 and 3) presented in Tables A22 and A21. Regression includes region and year dummies. Robust standard deviations are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

Table A17: Preference Estimates with Alternative Instruments – Place of Birth and Medical School Region

	Multinomial Logit with Control Function Alternative Instruments	
	Univ. Professor Wages in Own Region (1)	Univ. Professor Wages in Other Regions (2)
Birth Metrop Region	**1.750*****	**1.750*****
	(0.051)	**(0.051)**
× Male	-0.085**	-0.085**
	(0.040)	(0.040)
× Age	0.297**	0.295**
	(0.127)	(0.127)
× Medschool Rank	1.308***	1.308***
	(0.066)	(0.066)
Birth Countryside Region	**3.100*****	**3.100*****
	(0.050)	**(0.050)**
× Male	0.084**	0.084**
	(0.038)	(0.038)
× Age	-0.264**	-0.263**
	(0.110)	(0.110)
× Medschool Rank	0.034	0.035
	(0.063)	(0.063)
Medschool Metrop Region	**3.724*****	**3.723*****
	(0.046)	**(0.046)**
× Male	-0.171***	-0.171***
	(0.036)	(0.036)
× Age	0.627***	0.627***
	(0.107)	(0.107)
× Medschool Rank	-0.424***	-0.423***
	(0.061)	(0.061)
Medschool Countryside Region	**1.729*****	**1.729*****
	(0.074)	**(0.074)**
× Male	-0.14***	-0.14***
	(0.052)	(0.052)
× Age	1.149***	1.149***
	(0.161)	(0.161)
× Medschool Rank	-0.014	-0.015
	(0.098)	(0.098)

This table shows robustness estimates for a multinomial logit model with control function considering two alternative instruments. Sample size: 46,989. Standard deviations in parenthesis. All columns include alternative-specific dummies and region-specific year trends.*** p<0.01,** p<0.05,* p<0.1.

Table A18: Preference Estimates with Alternative Instruments – Regions' Characteristics

	Multinomial Logit with Control Function Alternative Instrument	
	Univ. Professor Wages in Own Region (1)	Univ. Professor Wages in Other Regions (2)
Physicians Ratio	**0.609**	**0.565**
	(0.541)	**(0.543)**
× Male	-0.319*	-0.318*
	(0.188)	(0.188)
× Age	-0.345	-0.348
	(0.578)	(0.578)
× Medschool Rank	-0.077	-0.08
	(0.321)	(0.321)
Health Infrastructure	**2.556** ***	**2.492** ***
	(0.504)	**(0.504)**
× Male	-0.073	-0.074
	(0.184)	(0.184)
× Age	-0.891	-0.889
	(0.55)	(0.55)
× Medschool Rank	-0.312	-0.313
	(0.313)	(0.313)
Health Insurance	**0.023**	**0.019**
	(0.373)	**(0.375)**
× Male	-0.231**	-0.231**
	(0.101)	(0.101)
× Age	-1.642***	-1.641***
	(0.32)	(0.32)
× Medschool Rank	1.097***	1.097***
	(0.173)	(0.173)
Amenity Index	**0.624** **	**0.635** **
	(0.28)	**(0.282)**
× Male	0.116	0.116
	(0.092)	(0.092)
× Age	0.113	0.112
	(0.279)	(0.279)
× Medschool Rank	-0.452***	-0.451***
	(0.154)	(0.154)
Avg Hourly Wage	**3.106** ***	**2.433** **
	(0.986)	**(1.05)**
× Male	0.219	0.219
	(0.137)	(0.137)
× Age	-0.909**	-0.908**
	(0.394)	(0.394)
× Medschool Rank	0.395*	0.391*
	(0.225)	(0.225)
Region Unobs	**-2.661** ***	**-2.146** ***
	(0.727)	**(0.775)**

This table shows robustness estimates for a multinomial logit model with control function considering two alternative instruments. Sample size: 46,989. Standard deviations in parenthesis. All columns include alternative-specific dummies and region-specific year trends.*** p<0.01,** p<0.05,* p<0.1.

E.3 Estimates with Restricted Choice Set

We also estimate the model restricting physicians' choice set according to the placement of the alumni of each medical school. Precisely, we identify the placement history of all programs and restricted the choice set of graduates of a given program to places that have, at any point in time, been chosen by at least one graduate of that program. We, then, estimate a version of the model restricting the choice set of all students in each medical school to this set locations. Tables A19 and A20 show that the point estimates are very similar.

Table A19: 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 A20: Restricted Choice Set 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.

E.4 Random Coefficients with Different Number of Draws

We also estimated the supply model using different number of draws of unobserved individual characteristics. The results of the model estimated with 100 and 200 draws are close to our baseline results.

Table A21: Preference Estimates with Different Draws – Place of Birth and Medical School Region

	Random Coefficients with Control Function Number of Draws	
	100	200
	(1)	(2)
Birth Metrop Region	**2.508*****	**2.514*****
	(0.099)	**(0.099)**
× Male	-0.257***	-0.256
	(0.075)	(0.075)
× Age	0.175	0.151***
	(0.232)	(0.233)
× Medschool Rank	1.186***	1.189***
	(0.123)	(0.123)
Birth Countryside Region	**2.847**	**2.839***
	(0.133)	**(0.133)**
× Male	0.199**	0.191***
	(0.100)	(0.099)
× Age	-1.137	-1.120***
	(0.300)	(0.299)
× Medschool Rank	0.800	0.820***
	(0.171)	(0.171)
Medschool Metrop Region	**4.999**	**4.996*****
	(0.084)	**(0.084)**
× Male	-0.457	-0.451
	(0.064)	(0.064)
× Age	0.156	0.150***
	(0.184)	(0.183)
× Medschool Rank	-0.543	-0.535***
	(0.108)	(0.107)
Medschool Countryside Region	**2.827**	**2.816****
	(0.094)	**(0.097)**
× Male	-0.158**	-0.165***
	(0.063)	(0.065)
× Age	0.766	0.777
	(0.195)	(0.199)
× Medschool Rank	0.137	0.125**
	(0.123)	(0.125)

This table shows robustness estimates for logit model with control function using different numbers of simulation draws. Sample size: 46,989. Standard deviations in parenthesis. All columns include alternative-specific dummies and region-specific year trends.*** p<0.01,** p<0.05,* p<0.1.

Table A22: Preference Estimates with Different Draws – Regions' Characteristics

	Random Coefficients with Control Function Number of Draws	
	100	200
	(1)	(2)
Physicians Ratio	**1.053**	**1.026**
	(0.716)	**(0.716)**
× Male	-0.316	-0.316
	(0.247)	(0.247)
× Age	-1.498**	-1.501**
	(0.759)	(0.759)
× Medschool Rank	-0.374	-0.354
	(0.415)	(0.415)
Health Infrastructure	**3.076*****	**3.056 *****
	(0.674)	**(0.675)**
× Male	-0.087	-0.080
	(0.248)	(0.248)
× Age	0.092	0.119
	(0.735)	(0.736)
× Medschool Rank	-0.287	-0.309
	(0.415)	(0.415)
Health Insurance	**0.051**	**0.065**
	(0.496)	**(0.496)**
× Male	-0.297**	-0.300**
	(0.127)	(0.127)
× Age	-1.551***	-1.546***
	(0.404)	(0.404)
× Medschool Rank	1.970***	1.964***
	(0.215)	(0.215)
Amenity Index	**1.121*****	**1.118*****
	(0.366)	**(0.366)**
× Male	0.239**	0.242**
	(0.120)	(0.120)
× Age	0.016	0.009
	(0.356)	(0.356)
× Medschool Rank	-1.036***	-1.028***
	(0.199)	(0.199)
Avg Hourly Wage	**2.742***	**2.783***
	(1.480)	**(1.481)**
× Male	0.432**	0.432**
	(0.173)	(0.173)
× Age	-0.972**	-0.943*
	(0.482)	(0.483)
× Medschool Rank	0.549*	0.559**
	(0.281)	(0.281)
Region Unobs	**-2.708****	**-2.730*****
	(1.097)	**(1.099)**

This table shows robustness estimates for logit model with control function using different numbers of simulation draws. Sample size: 46,989. Standard deviations in parenthesis. All columns include alternative-specific dummies and region-specific year trends.*** p<0.01,** p<0.05,* p<0.1.

F Supplementary Descriptive Evidence

This appendix provides supplementary tables and evidence that complement Section 3. Table A4 shows that physicians in our sample (matched) are not systematically different from the ones we lose when merging the different data sources (not matched). Most of our losses were probably due to: (i) misspelling and errors in the names and dates of birth; (ii) RAIS and CNES having less accurate registries in their early years; (iii) physicians which only work shifts up to 24 hours in hospitals, and A&E departments. In Table A4 we find that the two major differences are in gender and graduation year. We believe more women were not matched because in Brazil it is quite common for them to include their husbands' surname after getting married. As for graduation year, we believe our match is better in later years because RAIS and CNES databases improved over time. The other variables do not show high differences. More details about the merge process can be found in Appendix G.

Figure A2 shows a strong correlation between physicians' wages in formal and informal jobs. Figure A2 and Table A23 display the living costs we use (see Appendix G for details). Table A24 presents the transition matrix of physicians' practice location choice just after graduation and practice location 5 years later.

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

$$y_{jt} = \alpha + \beta X_{jt-1} + \gamma_j + \delta_j \times t + \varepsilon_{jt} \tag{15}$$

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.[42]

The first five columns of Table A25 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 percentage of medical school generalist students graduating in that area contributes

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

to attract more new graduates to that region – both across and within state-region.[43] The results from the specification without population weights (column 5) is qualitatively similar.

Our instrumental variables estimates – shown in Table A25 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 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.

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

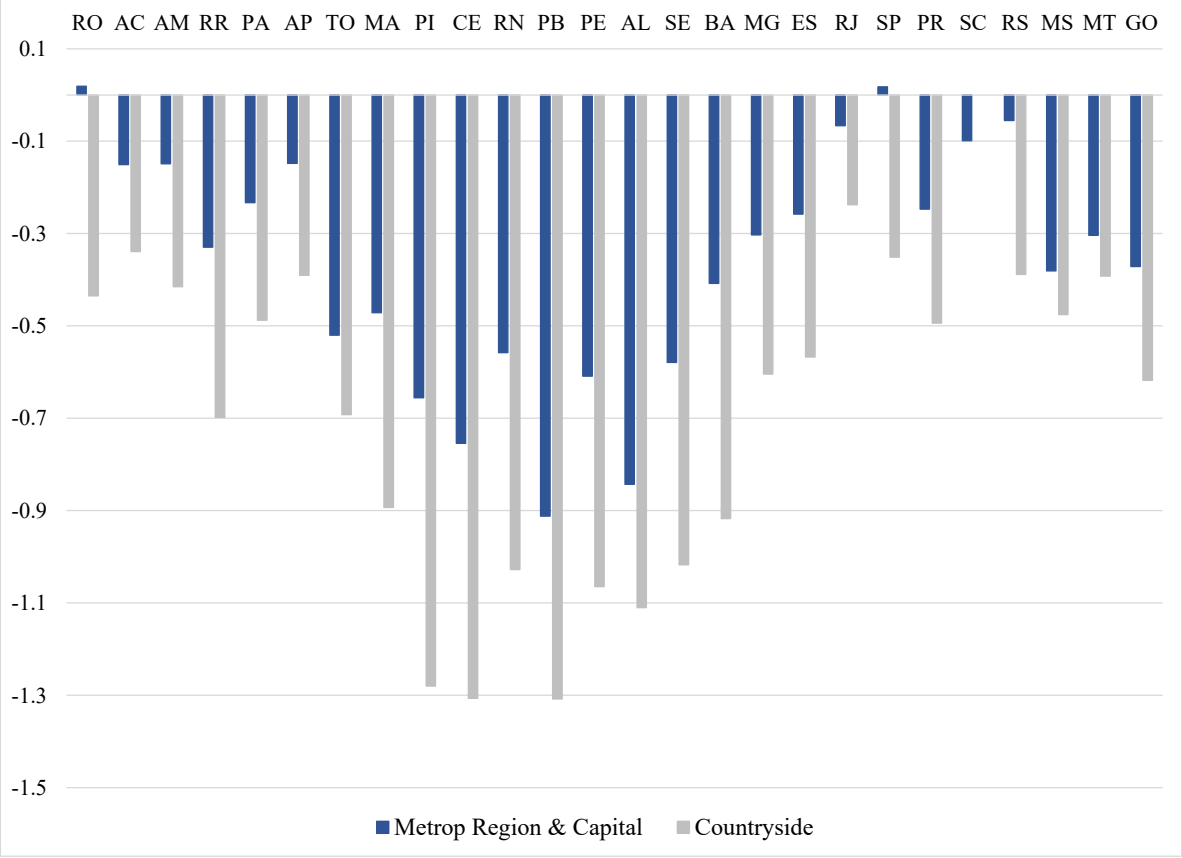


Figure A2: 2010 Living Cost Index

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

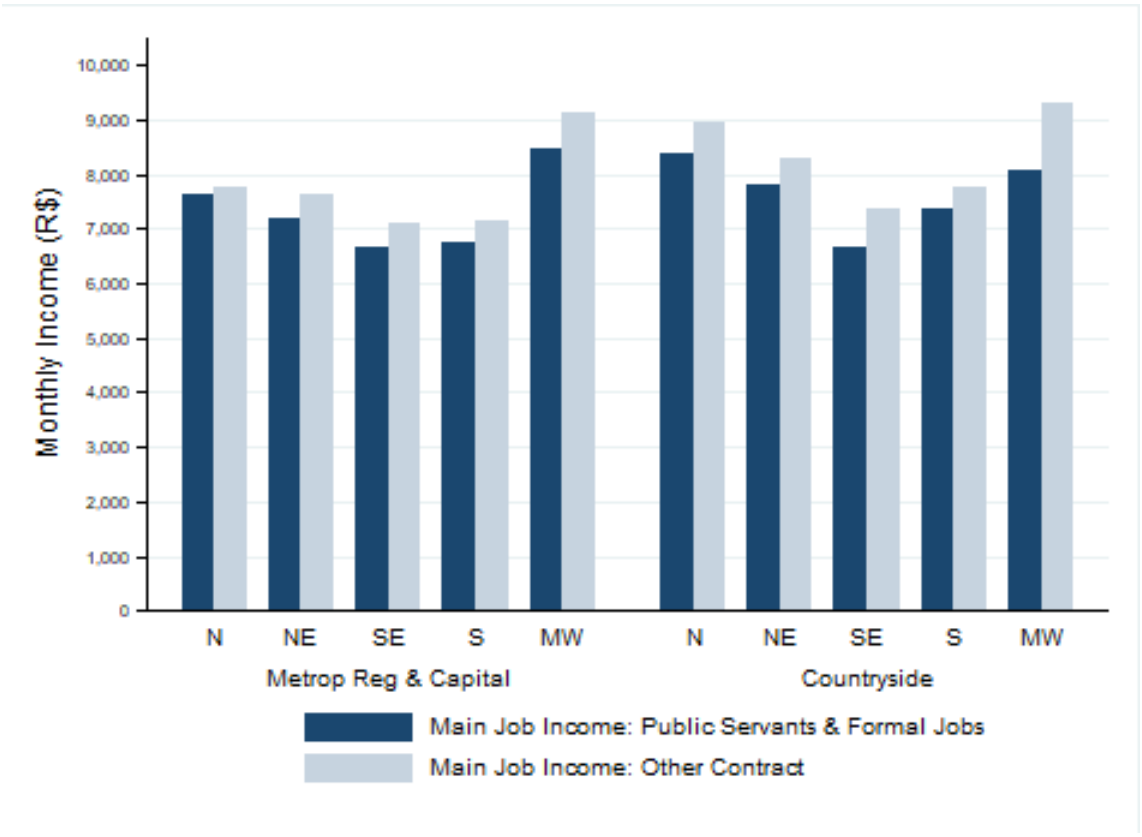


Figure A3: Average Labor Income by contract Type

This graph shows the average labor income of physicians in Census 2010 (under 35 years old) by region (countryside/metropolitan region) and by type of contract: formal or public sector jobs, and other forms of contracts (e.g., self-employed or informal). We see that average wages in other forms of contract is highly correlated with average public and private sectors wages.

Table A23: 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 A24: 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											
	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
Metropolitan Regions	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**
	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
Countryside	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 A25: 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.

Table A26: Summary Statistics of Wages in the Public and Private Sectors

	by Contract Type			
	Avg Public Wage/hr (1)	% Public Sector (2)	Avg Private Wage/hr (3)	% Private Sector (4)
North				
	9.27	64.18	10.27	28.13
	(5.22)	(36.39)	(11.23)	(32.30)
	16.54	73.45	24.79	10.07
	(10.59)	(39.64)	(18.70)	(22.70)
Northeast				
	13.97	59.63	24.30	36.10
	(5.40)	(27.32)	(19.80)	(25.39)
	25.37	79.77	27.04	15.10
	(9.46)	(30.94)	(21.30)	(24.96)
Sountheast				
	8.63	56.03	9.95	42.04
	(5.28)	(27.35)	(4.44)	(26.84)
	18.82	64.85	19.11	33.23
	(9.01)	(21.21)	(8.68)	(19.69)
South				
	10.59	47.29	9.73	52.71
	(4.18)	(25.28)	(3.54)	(25.28)
	20.17	69.09	18.12	30.91
	(5.01)	(24.32)	(3.84)	(24.32)
Midwest				
	14.64	64.32	10.63	31.83
	(11.28)	(29.63)	(4.51)	(27.39)
	29.15	73.46	20.26	21.41
	(20.22)	(28.02)	(5.85)	(22.61)
Brazil				
	16.59	66.52	19.09	27.71
	(10.99)	(31.89)	(16.38)	(28.16)

This table displays the summary statatistics of wages in the public and private sectors.

G Data Appendix

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

G.1 Organizing CFM database

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

Recover missing values for variables: medical 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 some 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 names. Until now no record was deleted. We have 637,058 records, and 459,740 physicians (uniquely identified by full name [44]).

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 it is a small number, we 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

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

identified by full name): 149,637. According to SIGRAS/INEP, 149,002 physicians graduated between 2001-2013, and our final database of physicians is very close to it.

Keep only one record per physician. The criteria used was to consider first 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' codes with identifiers of metropolitan regions in each state provided by IBGE. Our definition of metropolitan regions will also incorporate the capital of each state. Therefore, countrysides are 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.

G.2 Merge between CFM and CNRM databases

Merge with CNRM was performed in three steps: (i) exact merge using the full name and the CFM registry number; (ii) exact merge using the full name; and (iii) probabilistic merge using the full name (2,292 were merged in this last step). We do not consider merges with the following inconsistency: if the beginning of the residency occurred before the medical 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).

G.3 Organizing RAIS database

First, we extract the physicians from RAIS (data from 2001-2015)[45], i.e., the observations with the following Brazilian Occupational Codes (CBO): 2002 CBOs beginning with 2231,

[45]There was a problem with the Brazilian Occupational Classification (CBO) in the identified 2001 RAIS 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

2251, 2252, 2253, and 234453, this last one begin medicine professor; and 1994 CBOs beginning with 0.61 and 13770, this last one representing "professor of occupational medicine".

Remove spaces and special characters from physicians' full names. Then we recover missing names, dates 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. Then, divide it by working hours and create 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.

G.4 Organizing CNES database

Keep only workers 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.

available online at ftp://ftp.mtps.gov.br/pdet/microdados/RAIS/2001/, which did not have this problem.

G.5 Merge CFM generalists with RAIS and CNES

Now, we want to find the cities in which generalist physicians are working after graduation. We searched for them both in RAIS and CNES.

We started with RAIS by performing an exact merge using physicians' full names and birth years. We did not consider merged observations in which graduation year > year found working in RAIS. We manage to merge 51,446 physicians (57.0%) with RAIS.

Second, we performed an exact merge with CNES using physicians' full names. Again, we did not consider merged observations in which graduation year > year found working in CNES. CNES also informs which physicians are currently training medical residents 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 consider both datasets, we were able to find 57,195 (94.4%) physicians at some point in time. We decided to keep only physicians we could see up to 3 years after graduation, which left us with our final database of 46,989 generalist physicians. We did not keep these physicians because we could not be sure that these regions were their first choice of workplace after medical school. Table A4 shows the comparison between physicians found in RAIS or CNES up to three years after graduation and physicians lost in the merge process.

G.6 Construction of the Amenity Index

G.6.1 Transport Index

Vehicle fleet data. We used the December fleet information from DENATRAN at the municipality level for each year. All the administrative regions of Brasilia were considered just as Brasilia city. Data was collapsed (sum) to the UF/(countryside or metropolitan region & capital) level and merged with the IBGE database containing population per municipality-year. We constructed three variables: (i) bus per 1000 people (includes bus and shuttle); (ii) cars per 1,000 people (includes cars and pick-ups); (iii) and motorcycle per 1,000 people. Since we don't have this data for 2000 at the municipality level, we used the 2001 information instead.

Data on Traffic Deaths. Data came from DATASUS/SIM and was obtained by city of death and year. ICD-10 codes used for traffic accidents: V01-V99. Data was collapsed (sum) to the state/(countryside or metropolitan region & capital) level and merged with the 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: (i) transport by land, water, and air, all together: CNAE 2.0 is 49-51 and CNAE 1.0 is 60-62; (ii) hotels: CNAE 2.0 is 55 and CNAE 1.0 is 55.1; (iii) restaurants: CNAE 2.0 is 56 and CNAE 1.0 is 55.2; (iv) entertainment: CNAE 2.0 is 90-93 and CNAE 1.0 is 92.3, 92.5 and 92.6. Data was merged with IBGE population per municipality-year and collapsed (sum) to the state/(countryside or metropolitan region & capital) level. We created the variable "number of establishments per 1,000 people" for all four types described above. There was no information for the state of Pernambuco in 2002 (zero establishments). So we interpolated this year's value for all establishments types with 2001 and 2003 information.

How we constructed the transportation index. We merged information on fleet, transportation establishments and traffic deaths at the state/(countryside or metropolitan region & capital) level. The index was built using the KKL method: normalize each variable and 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).

G.6.2 Violent Deaths

Data came from DATASUS/SIM and was obtained 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, excluding suicides and terrorism: (i) assault (homicide): X85-X99, Y00-Y09; (ii) event of undetermined intent: Y10-Y34; (iii) unintentional exposure to inanimate mechanical forces (firearms): W32-W34; (iv) legal Intervention: Y35.

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

G.6.3 Entertainment Index

Data on movie theaters are from ANCINE. There is only information about the number of movie theater rooms per city from 2007 on. We projected data between 2000-2006 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 proportions 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 and merged with IBGE 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 transport index section.

How we constructed the entertainment index. We merged information on cinemas and entertainment establishments at the state/(countryside or metropolitan region & capital) level. Then, we created an entertainment index using the KKL method: normalize each variable and 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).

G.6.4 Education

We use the Basic Education Development Index (IDEB) in Brazil. IDEB measures the quality of basic education and combines information on (i) the school flow (promotion, grade retention, school evasion) obtained annually through the School Census, and (ii) the performance achieved by the students in the assessments applied nationally. The index started in 2005 and is released every two years.

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

G.6.5 Public Investment

Data on public expenditures by state and city comes from the Brazilian Public Sector Accounting and Tax Information System (SICONFI). Only one city had missing information in one year (290430), which we interpolated. Data was merged with IBGE population per municipality-year and 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 deflated to 2010 reais.

The index was created using the KKL method: normalize each variable and calculate their average. Variables used in both indexes: State investments per capita; and cities investments per capita.

G.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.

G.7 Construction of other regions' characteristics

G.7.1 Physicians per capita

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

G.7.2 Health Insurance

Data is from the National Regulatory Agency for Private Health Insurance and Plans (ANS). We use data from 2000 to 2016 (December), without any imputation. We drop observations in which the beneficiaries' municipality is unknown inside the state (the last four digits in the city code are equal to "0000" or the city code is equal to zero). It represents 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.

G.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 equipments 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.

G.7.4 Physicians Wage per Hour

Data comes from RAIS. We keep the records in which (i) 2002 CBO is 225170 and 223129 (generalist physician) or 225125 and 223115 (clinic physician), and (ii) 1994 CBO is 0.6105 (general physician). We construct the average remuneration of recently graduated physicians (age $\leq$ 35 years) and deflate it to 2010 reais. Using working hours we construct the remuneration per hour in 2010 reais. Then, data is collapsed to a countryside/metropolitan region

weighted by city population. We also distinguish the remuneration between the public and private sectors. The public sector encompasses establishments with legal nature defined as either 1 (public administration) or 201-1 (public enterprise). The private sector includes all the other establishments.

G.8 Construction of the Living Cost Index

The living cost index was constructed following Summers (1973)[46] and Seabra and Azzoni (2015).[47] We regress the rental value against many property characteristics and a set of dummies that identifies in which region the property is located (Summers, 1973). Potential sample selection issues related to the fact that rented households may differ in many ways from those not rented are addressed using Heckman (1979).[48] Below we describe step by step how to construct it. We use data from PNAD and the 2010 Census. The Datazoom package from PUC-Rio is used to make the variables compatible over time.

Organize PNAD database. 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"

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

[47]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.

[48]Heckman, J. (1979). Sample Selection Bias as a Specification Error. Econometrica, 47(1):153– 161.

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 the 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).

Organize 2010 Census Data. 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).

Linving Cost Index. We construct the living cost index using the state dummies and the state/(countryside or metropolitan region & capital) dummies obtained from the regressions above. The steps using the 2010 Census data provide 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 every year. But unfortunately, the 2000 Census does not 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 the 2010 Census we calculate, for each state, how bigger or smaller the MR and CS indexes are when compared to the one related to the whole state. We will assume that these ratios are the same 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 other years.

Using the PNAD data and the methodology described above, we were able to calculate

living costs indexes for every state as a whole. Then we used 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 costs. Graph A2 shows the index calculated at the state/(countryside or metropolitan region & capital) level using 2010 Census and Table A23 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 purchasing power.