The Geography of Medical Care and the Economics of Who Pays

Yumin Hu [*] Guoen (Gordon) Liu [†] Mingrui Liu [‡] Yuhang Pan [§]
Mingzhi (Jimmy) Xu [¶]

June 10, 2026

Abstract

Medical care is produced locally but increasingly traded across space through patient mobility. Its financing, however, remains local. In China, origin cities reimburse their residents wherever they seek treatment, while destination cities receive medical demand and experience the resulting scale and congestion effects. This separation creates a fiscal externality because local insurance authorities do not fully internalize how their financing decisions reshape medical markets elsewhere. We combine comprehensive inpatient records with transportation-network changes to document the geography of patient flows and estimate how cross-city demand affects medical provision. We then develop a multi-region general-equilibrium model in which patients choose treatment destinations, medical providers exhibit scale economies and capacity constraints, and local governments finance origin-based insurance through payroll taxes. The model embeds medical-service trade within broader goods and labor markets and distinguishes the effects of local and nonlocal patients on medical productivity. Quantitative counterfactuals show that decentralized financing leaves substantial welfare gains unrealized. Coordinating local insurance financing improves the allocation of fiscal burdens and medical activity across cities, while national insurance pooling generates additional gains by directing resources toward places where they have greater social value.

Keywords: Healthcare Access; Trade in Medical Services; Policy Coordination
JEL Codes: I11; I13; R13

[*]Yumin Hu: Nankai University, China (yumin.hu@nankai.edu.cn)
[†]Guoen (Gordon) Liu: Peking University, China (gordonliu@nsd.pku.edu.cn)
[‡]Mingrui Liu: Peking University, China (mrliu2023@nsd.pku.edu.cn; Tel: +86 13232068072)
[§]Yuhang Pan: Peking University, China (yhpan@pku.edu.cn)
[¶]Mingzhi (Jimmy) Xu: Peking University, China (mingzhixu@nsd.pku.edu.cn)

1 Introduction

Medical services have traditionally been considered local nontradable services, but improvements in transportation have made them increasingly tradable through patient mobility. In 2024, cross-regional medical visits in China reached approximately 400 million.[1] Patient mobility allows individuals to seek care beyond the medical resources available in their home regions and to choose among destinations that differ in prices, quality, congestion, and access costs. These choices reallocate medical demand across space and may lead to substantial patient concentration in cities offering more attractive medical services. The welfare consequences of such concentration are theoretically ambiguous. On the one hand, larger medical markets may generate economies of scale, foster specialization, and improve medical productivity or quality. On the other hand, when medical capacity is constrained, patient inflows may increase congestion and reduce the medical resources available per patient. Policies that affect patient flows therefore shape not only access to care, but also the balance between scale economies and congestion.

These externalities generated by trade in medical services create a fundamental coordination problem for decentralized policymakers. Policies chosen by one region can redirect patients and medical spending across space, altering caseloads, medical-market outcomes, and welfare in other regions. Local governments, however, typically evaluate such policies based primarily on local welfare and fiscal constraints and may therefore fail to internalize their broader effects. We study this coordination problem through the lens of local health-financing taxes. In the absence of patient mobility, taxes used to finance local health insurance primarily affect local residents and the local medical market. Once patients can seek care across regions, however, these taxes are no longer purely local instruments: by affecting insurance financing and patients' effective prices, they reshape patient flows and transmit their effects across regional medical markets. Trade in medical services can therefore drive a wedge between decentralized tax choices and the socially optimal tax structure, creating potential gains from policy coordination.

In this paper, we study the role of policy coordination in internalizing the externalities generated by trade in medical services. We first provide empirical evidence that motivates the key mechanisms in our theoretical framework. Instrumenting patient volume with transportation-induced changes in market access, we document evidence of congestion and heterogeneous scale effects. In particular, we find that an exogenous increase in patient volume significantly reduces length of stay (LOS). After ruling out several alternative explanations, we interpret the decline in LOS as evidence of congestion: under capacity constraints, hospitals accommodate additional demand by accelerating bed turnover and reducing the amount of inpatient care provided per admission. We also provide suggestive evidence that the scale effects generated by local and nonlocal demand

[1] https://www.nhsa.gov.cn/art/2025/7/14/art_7_17248.html

differ.

Second, we develop and estimate a gravity-style model of patient flows to recover revealed-preference measures of destination-specific medical service attractiveness. We validate using external quality metrics such as mortality and health service index.[2] Using these recovered attractiveness measures, we estimate the scale elasticity of medical production and the service adequacy elasticity governing the welfare cost of congestion.

Finally, using the estimated and calibrated parameters, we conduct a series of counterfactual exercises to examine how local taxation can help internalize the externalities generated by trade in medical services. We first show that local taxation complements, rather than substitutes for, the existing medical insurance system. We then compute the optimal local tax rates and the associated welfare gains under national policy coordination. Finally, we quantify the welfare effects of shifting from city-level health-insurance financing to national pooling.

This paper makes two main contributions to the literature. First, it contributes to the growing literature on trade in medical services and patient mobility. Existing studies identify several key determinants of patients' destination choices, including provider quality (Tay, 2003; Gutacker et al., 2016; Dingel et al., 2023), geographic distance (Acton, 1975; Beckert et al., 2012; Lu and Slusky, 2016), and out-of-pocket costs (Brot-Goldberg et al., 2017; Prager, 2020; Ackley, 2022). By shaping patients' destination choices, differences in quality, prices, and access costs reallocate medical demand across regions and may lead to the spatial concentration of patients, with potentially opposing welfare consequences. On the demand side, greater patient mobility can improve patient welfare by expanding access to preferred providers and treatment options (Ho, 2006; Gaynor et al., 2016). On the supply side, the concentration of patients induced by trade in medical services may generate economies of scale and foster medical specialization (Gaynor et al., 2005; Chandra and Staiger, 2007; Dingel et al., 2023). At the same time, patient inflows may create congestion when local medical capacity is constrained (Sá et al., 2019), potentially worsening health outcomes (Kc and Terwiesch, 2009; Hoe, 2022). While previous studies have typically examined these determinants and consequences separately, we integrate them into a unified multi-region general equilibrium framework. This framework allows us to quantify the opposing welfare effects of trade in medical services and examine whether policy coordination can internalize the cross-regional externalities generated by medical service trade.

Among studies of trade in medical services, our paper is most closely related to Dingel et al. (2023), who use U.S. Medicare data to show that roughly one-fifth of medical care is traded across regions and that larger markets export greater quantities, higher-

[2]The health service index comes from the Tsinghua Urban Health Index, a dataset developed by Tsinghua University. Higher index means better health services. Website: https://tuhi.tsinghua.edu.cn/managementPaper.html

quality services, and a wider variety of medical services. We build on their analysis in two ways. First, we examine not only the heterogeneous scale benefits generated by local and nonlocal demand, but also the congestion costs that arise when medical capacity is constrained. Second, we introduce local health-financing policies and show how coordination of local taxation can internalize the cross-regional spillovers generated by patient mobility.

Second, this paper contributes to the literature on policy coordination in the presence of cross-jurisdictional externalities generated by trade and mobility. A large literature shows that governments choosing policies noncooperatively fail to internalize their effects on other jurisdictions. These inefficiencies can be mitigated through appropriately designed taxes that directly target cross-border externalities (Harstad, 2024), reciprocal trade agreements that constrain terms-of-trade manipulation (Bagwell and Staiger, 1999), coordinated rules governing trade and domestic policies (Ederington, 2001), and multilateral policy coordination (Ossa, 2014). One strand of related work studies how trade-based incentives can sustain international cooperation in addressing global externalities (Nordhaus, 2015). Another examines how imperfect coordination among subnational governments distorts policy choices in the presence of cross-jurisdictional spillovers. Within this strand, Jiang (2026) quantifies distortions in U.S. transportation investment arising from network, terms-of-trade, and fiscal externalities. We extend this literature to trade in medical services by examining how patient mobility alters local policy incentives and creates scope for welfare-improving coordination.

Our paper is closely related to Jiang (2026) in its emphasis on policy coordination across jurisdictions. Both papers study how decentralized local governments may fail to internalize cross-regional externalities. The settings and policy margins, however, differ in two important respects. First, Jiang (2026) examines distortions in local investment in transportation infrastructure that facilitates trade, whereas we examine distortions in local health-financing policies arising from trade in medical services. Second, while Jiang (2026) focuses on externalities associated with the decentralized provision of trade-supporting infrastructure, we study those generated by trade in medical services itself, as patient concentration creates both scale economies and congestion under limited medical capacity.

2 Background

China's basic medical insurance system delivers near-universal, mandatory coverage through multiple schemes designed for distinct population groups. Its two central components are the Urban Employee Basic Medical Insurance (UEBMI), which is financed by payroll contributions from employers and employees, and the Urban and Rural Resident Basic Medical Insurance (URRBMI), which covers other urban residents and rural popu-

lations and is funded by individual premiums alongside substantial government subsidies. Coverage primarily includes inpatient care, with varying levels of benefits for outpatient services and chronic disease treatment. Reimbursement policies are largely standardized through nationally defined drug and medical consumables catalogs.

Medical care in China is delivered through an extensive provider network dominated by public hospitals and complemented by a broad base of primary care facilities. Public hospitals are stratified into primary, secondary, and tertiary levels, with many high-tier institutions concentrated in major metropolitan areas. Alongside hospitals, community health centers and clinics play a central role in providing primary care. Recent reforms have emphasized a tiered system of diagnosis and treatment, directing patients to seek initial care at the primary level and to be referred to higher-tier hospitals for more complex conditions.

To address regional disparities in medical quality and resource availability, policymakers have increasingly promoted policies that facilitate cross-city access to care. Before the mid-2010s, patients seeking treatment outside their city of enrollment typically paid expenses upfront and applied for reimbursement ex post, a process involving substantial paperwork and long delays. In 2016, the National Healthcare Security Administration launched a nationwide standardization reform that established a national information platform, unified data standards, and enabled real-time direct settlement for inpatient care across cities and later across provinces. Subsequent reforms expanded direct settlement to more hospitals and to outpatient services, including chronic, special-disease, and eventually routine outpatient care, while complementary measures—such as electronic insurance credentials and adjustments to deductibles and copayment rules—sought to balance access with fund sustainability. Under the current system, providers at the treatment location initially cover medical expenses, which are then periodically settled between enrollment and treatment locations according to the enrollment location's reimbursement policies.

Over the same period, China has undertaken substantial upgrades to its transportation infrastructure. Expansions of highway and high-speed rail networks, together with increased frequencies of intercity trains and flights, have significantly lowered travel times and costs, enabling patients to more easily obtain medical care outside their home cities.

3 Data

This section introduces the data used in the empirical analysis. We first describe the main data sources and their key characteristics, then explain the data-linkage process, and finally present several stylized facts. For the first-difference analysis, we use data from 2017 and 2019 to construct a city–disease–year-level dataset. The core data consist of detailed inpatient records that identify both patients' places of residence and their

treatment destinations. This origin–destination structure allows us to measure intercity patient flows, distinguish local from nonlocal demand, and examine trade in medical services at the disease level. We further link the inpatient records to measures of local medical capacity, medical productivity, and transportation-based market access used in the empirical analysis.

3.1 Data Description

3.1.1 Electronic Medical Records

Our dataset is constructed from transaction-level inpatient medical records, which we aggregate to the patient level by linking each patient's registered place of residence to the location where care is received. The first difference analysis relies on data from 2017 and 2019, and comprises over 30 million patient-level observations. Each hospitalization record includes admission and discharge dates, ICD-10 codes for primary and secondary diagnoses, medical expenses, patient demographics (age, gender, and occupation), and hospital characteristics (tier, class, and type).[3] To form a consistent city-level sample, we exclude cities with incomplete or missing reporting; the remaining cities account for roughly 95% of China's population. For consistency, we also exclude patients residing in the omitted cities.

The dataset includes approximately 21% of Tier 2 and Tier 3 public hospitals across 30 provinces, forming the most comprehensive national sample currently available for public inpatient care in China, and has been used in existing studies (Lin et al., 2022; Wu et al., 2025).[4] Data quality is consistently high, as records are tied to insurance reimbursement and subject to mandatory digital reporting protocols enforced by China's National Health Commission.[5]

3.1.2 Travel Time

Prefecture-to-prefecture travel-time data are obtained from Ma and Tang (2024). The dataset covers 279 prefectures from 1994 to 2024 and reports travel times for three transportation modes: road, freight rail, and passenger rail. We define intercity travel

[3]In China's healthcare system, hospitals are classified into a three-tier system (tier 1, 2, and 3) and within each tier, there are three classes (A, B, and C), resulting in a total of nine levels (3 tiers × 3 classes). Tier-3 class A hospitals usually have the best medical resources and most skillful physicians among all hospitals. In addition, there are two types of hospitals in China: general or specialized hospitals.

[4]China's hospital system is organized into a three-tier classification reflecting differences in infrastructure, service capacity, and quality of care. Public hospitals dominate inpatient service provision, with Tier 2 and Tier 3 institutions accounting for the vast majority of admissions. Private hospitals account for less than 24% of admissions, while Tier 1 hospitals contribute only 5.7% (National Health Commission of the PRC, 2022, 2023).

[5]Inaccurate diagnoses can result in insurance denial and financial penalties due to insurance verification requirements.

time as the minimum of road and passenger-rail travel times. To account for travel within the origin and destination cities, we add a constant 15 minutes to the travel time for each city pair. This adjustment is motivated by two institutional benchmarks. First, China has promoted the development of a *15-minute medical-service circle*, under which residents should be able to reach a nearby medical facility within 15 minutes.[6] Second, national policy guidelines require the average response time for prehospital emergency medical services to be less than 15 minutes.[7] City-pair travel times are used to construct the city-year level market-access measure, whose construction is described in detail in the following subsection.

3.1.3 Population

Our city-level population data for 2015 are primarily drawn from the *2016 China City Statistical Yearbook*. To better capture the potential demand for medical services, we use the resident population rather than the registered population.

3.1.4 Data Linkage

We construct a city-disease-year level dataset following a five-step procedure. First, we remove inpatient records with implausibly high or low medical expenditures attributable to decimal-entry errors, which eliminates approximately 0.5% of the original observations. Second, we sort the disease groups mainly following the first digit of the ICD-code. Third, we match the geographic information of patients and hospitals to distinguish local and nonlocal patients. Fourth, we aggregate the data to the city-disease-year level. Fifth, we combine the city-year level market access variable with the hospitalization dataset.

3.2 Key Variables and Measurements

Patient Flows We use the geographic information of both hospitals and patients in our transaction-level dataset to construct a city-to-city matrix that records patient flows between cities with disease information. We then aggregate it into city-disease level for each year.

Productivity of Medical Services Productivity is measured by the ratio of the number of patients to the number of doctors: $\frac{\text{No. of Patients}}{\text{No. of Doctors}}$.

[6] According to a report by People's Daily Online, by 2018, more than 80% of residents were able to reach the nearest medical facility within 15 minutes. See http://health.people.com.cn/n1/2018/1130/c14739-30433728.html.

[7] This information is reported on the website of the Chinese People's Political Consultative Conference. See http://cppcc.china.com.cn/2019-06/24/content_74914333.htm.

Market Access (MA) we construct a travel-time based market access variable following Donaldson and Hornbeck (2016):

$$MA_{i,t} = \sum_{j \neq i} \tau_{i,j,t-1}^{-\theta} \times N_j. \tag{1}$$

where $MA_{i,t}$ is the market access for city i in year t . $\tau_{i,j,t-1}$ denotes travel time between city i and j in year $t-1$. θ is the trade elasticity, calibrated at $\theta = 4$ (Simonovska and Waugh, 2014), and N_j is the population of city j in 2015, which is treated as a city-specific fixed value.

Length of Stay (LOS) We calculate the average LOS by diseases. For each disease, we calculate the average length of stay (LOS) separately for local and nonlocal patients.

Diagnosis Related Groups (DRG) Weights We calculate the average DRG weights by diseases. DRG weights are commonly used as a measure of case complexity and expected medical resource use.

3.3 Stylized Facts

We now present some data patterns. From producer perspective, we define each city's export share to measure the share of production that is supplied to nonlocal patients. We then define import share, from consumer perspective, to measure the share of consumption that is supplied by nonlocal hospitals. Figure 1 shows the distributions of export share and import share. We find that most cities exhibit low export shares and high import shares, indicating that medical demand in the majority of cities is met by importing medical services from other regions.

Using GuangZhou and ChangSha as illustrative cases, Figure 2 depicts the destinations of medical services produced in these cities. In the left panel, representing GuangZhou, darker shading indicates greater consumption of services originating from GuangZhou in each region. We observe that consumption shares are higher in regions closer to GuangZhou and decline gradually with increasing distance. A similar pattern is evident for ChangSha, although consumption is more geographically concentrated. Taken together, these figures suggest that medical service flows are broadly consistent with the predictions of the gravity model, with distance emerging as an important determinant of trade.

4 Empirical Analysis

In this section, we examine the externalities generated by patient concentration and provide motivating evidence for the theoretical framework. We first present the empirical framework and then show some evidence of congestion effects and heterogeneous scale effects associated with local and nonlocal patient demand.

4.1 Empirical Specification

We estimate the effects of patient volume on medical outcomes using an instrumental-variables (IV) strategy. Because patient volume may be endogenous, we instrument for the number of patients using a transportation-based market-access measure. The identifying variation in market access is driven by changes in travel times induced by transportation infrastructure improvements. These changes affect a city's accessibility to potential patients and therefore provide a strong predictor of patient volume. The exclusion restriction requires that transportation-induced changes in market access affect medical outcomes only through their effects on patient volume. This assumption is plausible because the instrument is constructed from changes in bilateral travel times and predetermined population levels, rather than from contemporaneous changes in local medical supply or demand.

The first stage of IV estimation is:

$$ln(Num_patient_{idt}) = \beta_0 + \beta_1 ln(MA_{it}) + \beta_2 C_{it} + \mu_{id} + \kappa_{pt} + \epsilon_{it} \tag{2}$$

where $Num_patient_{idt}$ is city i's patients of disease d in year t. MA_{it} denotes market access for city i in year t. C_{it} represents control variables, including the number of city i's secondary and tertiary hospitals in year t. μ_{id} and κ_{pt} are city-disease and province-year fixed effects, respectively.

The second stage of IV estimation is:

$$ln(S_{idt}) = \beta_0 + \beta_1 ln(\widehat{Num_patient}_{idt}) + \beta_2 C_{it} + \mu_{id} + \kappa_{pt} + \epsilon_{it} \tag{3}$$

where S_{idt} is city i's outcomes of disease d in year t. When examining negative externalities, the main outcome is average LOS; when examining positive externalities, the main outcome is productivity of medical service. We conduct a first-difference analysis using data from 2017 and 2019.

4.2 Empirical Results

4.2.1 Congestion Effect

Table 1 reports the instrumental-variable estimates of the congestion effects of patient volume. Columns (1) and (2) present the first-stage results. Market access strongly predicts patient volume: a 10% increase in market access raises the number of patients by approximately 13.8%. The corresponding first-stage and Kleibergen–Paap F-statistics are generally well above conventional thresholds, alleviating concerns about weak instrument variable.

Columns (3)–(6) report the second-stage estimates. Column (3) shows that a 10% increase in patient volume reduces the average length of stay (LOS) by approximately 0.85%. Relative to the sample mean of 9.47 days, this effect corresponds to a reduction of about 0.08 days, or nearly two hours, per admission.

We interpret the decline in LOS as evidence of congestion-induced adjustment. When medical capacity is constrained, hospitals may accommodate additional demand by accelerating bed turnover and reducing the amount of inpatient care provided per admission. An alternative explanation is that the reduction can be the increase of the efficiency. If this is the case, the improving efficiency should not have heterogeneity between local and nonlocal patients. Columns (4)–(5) show that the coefficients only statistically significant in local patients. These results provide some suggestive evidence ruling out this explanation. Another possibility is that the decline in LOS reflects a shift toward admitting less severe or less resource-intensive patients. Column (6), however, shows no significant change in average DRG weights, providing little evidence of such a compositional shift. Taken together, the findings are consistent with hospitals responding to rising patient volume by compressing inpatient treatment under capacity pressure.

4.2.2 Heterogeneous Scale Effects of Local and Nonlocal Demand

We next examine whether patient concentration generates regional scale economies and whether these effects vary with the importance of cross-regional demand. Figure 3 first shows a positive and statistically significant scale effect in the full sample. The estimated elasticity is approximately 0.49, implying that a 10% increase in patient volume raises medical productivity by about 4.9%. This result provides evidence that larger medical markets improve the productivity of medical-service provision. The corresponding regression estimates underlying Figure 3 are reported in Table C1.

We then explore heterogeneity across diseases with different propensities for cross-regional treatment. We select neoplasms as a representative high-mobility disease category and skin diseases as a representative low-mobility category. The estimated scale elasticity is approximately 0.58 for neoplasms, compared with about 0.35 for skin diseases.

Both estimates are positive, but the larger point estimate for neoplasms suggests that scale economies may be stronger for medical services that are more frequently traded across regions. Since the confidence intervals overlap and the comparison is based on selected representative disease categories, we interpret this pattern as suggestive rather than conclusive evidence of heterogeneity by disease mobility.

The first-stage results in columns (2) and (3) of Table C1 provide complementary support for this interpretation. The first-stage F-statistic is smaller in the low-mobility disease sample, indicating that market access has weaker predictive power for patient volume in diseases for which patients are less likely to seek treatment across regions. This pattern is consistent with our classification of diseases by mobility. Moreover, because the two specifications are estimated using the same empirical framework, the larger second-stage point estimate for the high-mobility disease provides suggestive evidence that the productivity gains from patient concentration vary with the tradability of medical services.

The patient-flow responses reported in Figure 4 provide complementary evidence. Greater market access significantly increases total, local, and nonlocal patient volumes. The estimated elasticity is approximately 0.71 for local patients and 0.94 for nonlocal patients. Although the latter estimate is less precise and the confidence intervals overlap, its larger magnitude suggests that improvements in market access attract nonlocal demand particularly strongly. Combined with the larger scale estimate for the high-mobility disease category, these patterns are consistent with nonlocal demand playing an important role in generating regional scale economies. They should nevertheless be interpreted as suggestive evidence, rather than as a formal test that the scale effect of nonlocal patients exceeds that of local patients. The detailed regression results underlying Figure 4 are reported in Table C2.

5 Theoretical Framework

The paper develops a multi-region model of patient flows for a single medical procedure. Patients choose destinations based on out-of-pocket prices, quality, congestion (capacity per caseload), and bilateral access costs. Medical services are produced competitively using local labor and a traded manufacturing composite, with region-level scale forces related to caseload composition.

The model combines three ingredients: patient mobility across regions, endogenous medical cost and quality, and a trade-linked regional general equilibrium. Patient flows determine medical scale and congestion; these in turn affect medical costs, wages, and sectoral labor allocation. Local governments finance insurance through taxes, creating strategic interactions across regions.

5.1 Model Setup

Medical Choice probabilities Patients reside in origin regions n and seek treatment in destination regions i. Let $N_n = L_n s_n$ denotes the number of potential patients in origin n, with s_n the probability of becoming patients. $M_i > 0$ denotes effective capacity (e.g., staffed procedure rooms × hours, or a composite of beds, surgeons), and Q_i the caseload in region i. A patient ω receiving services in region i from n obtains utility

$$U_{ni}(\omega) \;=\; -\alpha \ln p_{ni} \;+\; \ln \delta_i \;+\; \beta \ln \frac{M_i}{Q_i} \;-\; \ln \rho_{ni} \;+\; \varepsilon_{ni}(\omega), \tag{4}$$

Given the origin-specific reimbursement rate R_n, a patient from region n who receives care in destination i faces an out-of-pocket price

$$p_{ni} = (1 - R_n) c_i. \tag{5}$$

Here, $\delta_i > 0$ indexes the quality of care available in region i, and $\frac{M_i}{Q_i}$ is the capacity per caseload, a measure of service adequacy (higher values indicate more resources per patient, shorter waits, and greater comfort). The parameter $\beta > 0$ captures the utility gain from higher service adequacy. $\rho_{ni} > 0$ denotes access costs (smaller values mean easier access). Idiosyncratic taste shocks ε_{ni} are i.i.d. across (i, ω) and follow a Type-I extreme-value distribution.

Define the origin-specific inclusive value of medical service $\Psi_n = \sum_k p_{nk}^{-\alpha} \delta_k M_k^{\beta} Q_k^{-\beta} \rho_{nk}^{-1}$. Then the multinomial logit share for region i is

$$\lambda_{ni} \;=\; \frac{p_{ni}^{-\alpha} \delta_i M_i^{\beta} Q_i^{-\beta} \rho_{ni}^{-1}}{\displaystyle\sum_k p_{nk}^{-\alpha} \delta_k M_k^{\beta} Q_k^{-\beta} \rho_{nk}^{-1}} \;=\; \frac{p_{ni}^{-\alpha} \delta_i M_i^{\beta} Q_i^{-\beta} \rho_{ni}^{-1}}{\Psi_n}. \tag{6}$$

Expected medical service flows from n to i is given by $Q_{ni} = N_n \lambda_{ni}$. Providers in region i treat a total caseload

$$Q_i = \sum_n Q_{ni} = \sum_n N_n \lambda_{ni}. \tag{7}$$

Let Q_{ii} be local cases and define inbound nonlocal cases as $Q_{-i} \equiv \sum_{n \neq i} Q_{ni} = Q_i - Q_{ii}$.

Worker Sector Choice Workers in each region choose between two sectors: medical services (S) and manufacturing (M). For each worker ω in region i, the utility from working in sector $k \in \{S, M\}$ is

$$U_{ki}(\omega) = \varphi_{ki}(\omega)\, b_{ki}\, w_{ki}, \tag{8}$$

where $w_{S,i}$ and $w_{M,i}$ are sector-specific wages, $b_{ki} > 0$ is a sector-city shifter capturing nonpecuniary attractiveness or persistent mobility frictions, and $\varphi_{ki}(\omega)$ is an idiosyncratic taste draw. Assume φ_{ki} is i.i.d. Fréchet with shape parameter $\epsilon > 0$:

$$\Pr(\varphi_{ki} \leq x) = \exp(-x^{-\epsilon}), \qquad x > 0. \tag{9}$$

Each worker chooses the sector that gives higher utility. The resulting fraction of region-i workers employed in the medical sector is

$$\ell_i \equiv \frac{L_{S,i}}{L_i} = \frac{(b_{S,i}\, w_{S,i})^{\epsilon}}{(b_{S,i}\, w_{S,i})^{\epsilon} + (b_{M,i}\, w_{M,i})^{\epsilon}}, \tag{10}$$

and the manufacturing employment share is $1 - \ell_i = L_{M,i}/L_i$. The parameter ϵ governs the wage-responsiveness of sectoral labor supply: higher ϵ implies that workers are more sensitive to relative wage differences across sectors.

Medical Services Production Medical services are produced under perfect competition using local labor and a composite manufacturing input. We assume free entry and region-level increasing returns that are external to individual providers. In particular, each price-taking provider chooses its service quality and production scale, taking as given aggregate regional medical output, input prices, and the exogenous regional productivity shifter. Production of medical services in region i is given by

$$q_{S,i} = A_i L_{S,i}^{\upsilon} m_{M,i}^{1-\upsilon} \upsilon^{-\upsilon} (1-\upsilon)^{\upsilon-1}, \tag{11}$$

where $L_{S,i}$ and $m_{M,i}$ denote labor and manufacturing inputs used in medical production. The term A_i is the productivity of medical services in region i, defined as

$$A_i = \frac{\bar{A}_i H(Q_{ii}, Q_{-i})}{K(\delta_i)}, \tag{12}$$

where $\bar{A}_i$ is an exogenous region-specific productivity shifter, $H(Q_{ii}, Q_{-i})$ captures region-level external increasing returns. We specify the externality term as

$$H(Q_{ii}, Q_{-i}) = Q_{ii}^{\alpha_1} Q_{-i}^{\alpha_2}, \tag{13}$$

where $K(\delta_i)$ is an increasing quality-cost term reflecting the idea that higher service quality is more costly to provide. In the baseline specification, set

$$K(\delta_i) = \delta_i. \tag{14}$$

Under perfect competition, the corresponding unit cost of medical services in region i is therefore

$$c_i = \frac{w_{S,i}^{\upsilon} P_{M,i}^{1-\upsilon} \delta_i}{\bar{A}_i \, Q_{ii}^{\alpha_1} Q_{-i}^{\alpha_2}}. \tag{15}$$

In equilibrium, service volume Q_i and service quality δ_i are jointly determined with patient demand, factor prices, and regional market-clearing conditions.[8]

Manufacturing Production Manufacturing follows a standard Eaton–Kortum structure. In each region i, a continuum of tradable varieties $x \in [0,1]$ is produced competitively using local labor and a composite manufacturing intermediate. The production function for variety x is

$$q_{M,i}(x) = \xi_i(x) \, B_i \, l_{M,i}(x)^{\eta} \, m_{M,i}(x)^{1-\eta} \, \eta^{-\eta} (1-\eta)^{\eta-1}, \tag{16}$$

where $l_{M,i}(x)$ and $m_{M,i}(x)$ denote, respectively, the labor and intermediate input used in producing variety x in region i, and B_i is a region-specific productivity shifter. Aggregate manufacturing labor employed in region i is

$$L_{M,i} = \int_0^1 l_{M,i}(x) \, dx. \tag{17}$$

Productivity draws $\xi_i(x)$ are i.i.d. across varieties and regions and follow a unit Fréchet distribution with shape parameter θ:

$$\Pr(\xi_i(x) \le z) = \exp\left(-z^{-\theta}\right), \qquad z > 0. \tag{18}$$

Shipping one unit of a variety from region i to region n requires $d_{ni} \ge 1$ units of the good, where d_{ni} is an iceberg trade cost.

Under perfect competition, each variety is supplied by the lowest-cost region. The share of manufacturing expenditure in region n allocated to goods produced in region i is therefore

$$\pi_{ni} = \frac{\left(w_{M,i}^{\eta} P_{M,i}^{1-\eta} d_{ni}\right)^{-\theta} B_i^{\theta}}{\sum_k \left(w_{M,k}^{\eta} P_{M,k}^{1-\eta} d_{nk}\right)^{-\theta} B_k^{\theta}}. \tag{19}$$

The corresponding price index for the manufacturing composite in region n is

$$P_{M,n} = \bar{\gamma} \left[\sum_i \left(w_{M,i}^{\eta} P_{M,i}^{1-\eta} d_{ni}\right)^{-\theta} B_i^{\theta} \right]^{-1/\theta}, \tag{20}$$

[8] In equilibrium, a higher quality level δ_i increases demand for treatment in region i, but it also raises the unit cost of supplying care. The equilibrium values of δ_i and Q_i are therefore jointly pinned down by the interaction of patient demand, production costs, and regional market-clearing conditions.

where

$$\bar{\gamma} = \left[\Gamma\left(\frac{\theta + 1 - \sigma}{\theta}\right)\right]^{1/(1-\sigma)}. \tag{21}$$

Local Government Define the average labor income per worker in region n as

$$\bar{w}_i \equiv \frac{w_{S,i} L_{S,i} + w_{M,i} L_{M,i}}{L_i}. \tag{22}$$

Under origin-based insurance pays, that is, assigns the cost of treated cases to the origin's local tax base, the local government budget is

$$R_n\Big(\sum_i Q_{ni} c_i\Big) = \tau_n \bar{w}_n L_n + Z_n \tag{23}$$

Define the average unit cost of medical services incurred by origin-n patients as

$$\bar{c}_n \equiv \sum_i \lambda_{ni} c_i \tag{24}$$

Using $Q_{ni} = N_n \lambda_{ni}$ with $N_n = L_n s_n$, per capita deficit $z_n \equiv Z_n / L_n$, the balanced-budget condition simplifies to

$$s_n R_n \bar{c}_n = z_n + \tau_n \bar{w}_n \tag{25}$$

We assume the deficit is always fixed at baseline level. And in the baseline, we define the deficit-to-tax ratio $\zeta_n \equiv z_n / (\tau_n \bar{w}_n)$.

Equilibrium The market-clearing conditions in the manufacturing goods market imply the following system of N equations

$$w_{M,i} L_{M,i} = \eta \sum_n X_{M,n} \pi_{ni}. \tag{26}$$

Here, total spending on the manufacturing composite in destination n is given by

$$X_{M,n} = (1 - \tau_n)\bar{w}_n L_n - \sum_k Q_{nk} p_{nk} + \frac{1 - \upsilon}{\upsilon} w_{S,n} L_{S,n} + \frac{1 - \eta}{\eta} w_{M,n} L_{M,n}. \tag{27}$$

$X_{M,n}$ has three components: i) disposable income net of out-of-pocket medical spending, ii) medical sector intermediate demand, and iii) manufacturing's own intermediate demand. Using $Q_{ni} = N_n \lambda_{ni}$ with $N_n = L_n s_n$, the spending can be rewritten as

$$X_{M,n} = \bar{w}_n L_n - s_n \bar{c}_n L_n + z_n L_n + \frac{1 - \upsilon}{\upsilon} w_{S,n} L_{S,n} + \frac{1 - \eta}{\eta} w_{M,n} L_{M,n}. \tag{28}$$

Similarly, for the medical services market, market-clearing conditions imply:

$$w_{S,i} L_{S,i} = \upsilon c_i Q_i. \tag{29}$$

The sectoral labor allocation is determined by the worker sector choice condition (10).

Given the exogenous primitives $\{\bar{A}_i, B_i, M_i, s_i, L_i, R_i, z_i, b_{S,i}, b_{M,i}\}$, and $\{\rho_{ni}, d_{ni}\}$, an equilibrium is a collection of prices, sectoral wages, employment levels, and medical-service quantities and qualities, $\{w_{S,i}, w_{M,i}, c_i, \bar{c}_i, P_{M,i}, L_{M,i}, L_{S,i}, \delta_i, Q_{ni}, \lambda_{ni}\}$ that jointly satisfy (10), (7), (15), and (20)–(25).

The full equilibrium system, the counterfactual hat system, the auxiliary baseline shares, and the numerical solution procedure are reported in Appendix A.

5.2 Welfare Impact

The model lends itself naturally to welfare analysis. We measure welfare in region n by the expected utility of a representative resident worker. The worker cares about both the consumption of manufactured goods and the access to medical services, with the relative preference weight on the medical inclusive value denoted as ξ. Expected welfare is therefore

$$\ln W_n \;=\; \ln \frac{y_{M,n}}{P_{M,n}} \;+\; \xi \ln \Psi_n, \tag{30}$$

where

$$y_{M,n} \;\equiv\; (1-\tau_n)\,\bar{w}_n \;-\; s_n\,(1-R_n)\,\bar{c}_n \tag{31}$$

is the per-capita household budget for the manufacturing composite, equal to disposable labor income net of out-of-pocket medical spending. $P_{M,n}$ is the manufacturing price index, and Ψ_n is the inclusive value of medical destination choice. The first term in Equation 30 is therefore real consumption of the manufacturing composite, and the second term captures the welfare value of medical-service access. By construction, $y_{M,n}$ is the per-capita counterpart of $X_{M,n}/L_n$ in Equation A.13 stripped of firm intermediate purchases, so the welfare aggregator is consistent with the household budget constraint underlying manufacturing demand.

We compute welfare changes in counterfactual equilibrium in log differences. The welfare change for region n is

$$\Delta \ln W_n \;\equiv\; \ln \hat{W}_n \;=\; \ln \hat{y}_{M,n} \;-\; \ln \hat{P}_{M,n} \;+\; \xi\, \ln \hat{\Psi}_n. \tag{32}$$

Two equivalent expressions for $\hat{y}_{M,n}$ are useful below. Define the baseline shares

$$y_{1,n} \equiv \frac{(1-\tau_n^0)\,\bar{w}_n^0}{y_{M,n}^0}, \qquad y_{2,n} \equiv \frac{s_n\,(1-R_n^0)\,\bar{c}_n^0}{y_{M,n}^0}, \qquad y_{1,n} - y_{2,n} = 1. \tag{33}$$

Then the change in $y_{M,n}$ relative to baseline can be written as

$$\hat{y}_{M,n} \;=\; y_{1,n}\,\frac{1-\tau_n'}{1-\tau_n^0}\,\hat{\bar{w}}_n \;-\; y_{2,n}\,\frac{1-R_n'}{1-R_n^0}\,\hat{\bar{c}}_n. \tag{34}$$

Welfare Response to Taxes Define the dot operator as the log-differential of a variable:

$$\dot{x} \equiv d\ln x = dx/x.$$

Log-differentiating Equation 31 and dividing by $y_{M,n}$ yields

$$\dot{y}_{M,n} \;=\; y_{1,n}(\tau)\,\dot{\bar{w}}_n \;-\; y_{2,n}(\tau)\,\dot{\bar{c}}_n \;-\; y_{1,n}(\tau)\,\frac{\tau_n}{1-\tau_n}\,\dot{\tau}_n \;+\; y_{2,n}(\tau)\,\frac{R_n}{1-R_n}\,\dot{R}_n, \tag{35}$$

where the shares $y_{1,n}(\tau) \equiv (1-\tau_n)\bar{w}_n(\tau)/y_{M,n}(\tau)$ and $y_{2,n}(\tau) \equiv s_n(1-R_n)\bar{c}_n(\tau)/y_{M,n}(\tau)$ are now evaluated at the counterfactual equilibrium $E(\tau)$ rather than at baseline, and continue to satisfy $y_{1,n}(\tau) - y_{2,n}(\tau) = 1$. The baseline shares in Equation 33 are the special case $y_{1,n}(\tau^0)$ and $y_{2,n}(\tau^0)$. In the tax counterfactual ($\dot{R}_n = 0$) the last term drops, and the marginal welfare change in region n is

$$\dot{W}_n \;=\; y_{1,n}(\tau)\,\dot{\bar{w}}_n \;-\; y_{2,n}(\tau)\,\dot{\bar{c}}_n \;-\; y_{1,n}(\tau)\,\frac{\tau_n}{1-\tau_n}\,\dot{\tau}_n \;-\; \dot{P}_{M,n} \;+\; \xi\,\dot{\Psi}_n. \tag{36}$$

The closed log-differentiated equilibrium system used to compute the marginal welfare response is reported in Appendix B.

Next, we solve for the implied vector of endogenous dot variables,

$$\dot{y} \equiv \big(vec(\dot{\pi}),\; vec(\dot{\lambda}),\; \dot{w}_S,\; \dot{w}_M,\; \dot{P}_M,\; \dot{Q},\; \dot{Q}_{ii},\; \dot{Q}_{-i},\; \dot{c},\; \dot{\bar{c}},\; \dot{\delta},\; \dot{\Psi},\; \dot{L}_S,\; \dot{L}_M,\; \dot{X}_M\big)^\top \in \mathbb{R}^{2N^2+13N},$$

given the exogenous shock vector $(\dot{\tau},\; \dot{M},\; \dot{\rho},\; \dot{d})$. Here, $vec(\dot{\pi})$ and $vec(\dot{\lambda})$ stack the $N \times N$ matrices $(\dot{\pi}_{ni})$ and $(\dot{\lambda}_{ni})$ into $N^2 \times 1$ vectors, and the remaining terms are $N \times 1$ vectors. The system is linear in $\dot{y}$, so the solution can be expressed as a linear mapping from the exogenous shocks to the full set of equilibrium log-differentials, conditional on the equilibrium $E(\tau)$. For the tax counterfactual where only $\dot{\tau}$ is nonzero, the welfare response is $\dot{W} = G(\tau)\,\dot{\tau}$, where $G(\tau)$ is the linear operator implied by the equilibrium system evaluated at $E(\tau)$. The effect of a marginal change in region i's tax rate on the welfare vector is therefore

$$\frac{d\ln W}{d\ln \tau_i} = G(\tau)\,e_i, \tag{37}$$

where e_i denotes the i-th standard basis vector. Accordingly, the effect of region i's own tax change on its own welfare is given by the i-th element of this vector, namely,

$$\frac{d\ln W_i}{d\ln \tau_i} = e_i^\top\, G(\tau)\,e_i. \tag{38}$$

Unilateral Tax Setting We characterize how each region would set its tax rate when acting unilaterally. Region i chooses its tax rate τ_i to maximize its own welfare, taking the tax rates of all other regions, τ_{-i}, as given. For an interior solution, the necessary first-order condition is

$$\left.\frac{d \ln W_i}{d \ln \tau_i}\right|_{\tau=(\tau_i^{\mathrm{u}}, \tau_{-i})} = 0. \tag{39}$$

Cooperative Tax Setting Consider next the case in which each region unilaterally chooses its own tax rate τ_i to maximize aggregate welfare across all regions, $\sum_{n=1}^{N} \ln W_n(\tau)$, taking the tax rates of all other regions as given. For each region i, the corresponding first-order condition is

$$\left.\frac{d \sum_{n=1}^{N} \ln W_n}{d \ln \tau_i}\right|_{\tau=(\tau_i^{\mathrm{c}}, \tau_{-i})} = 0, \qquad \forall i. \tag{40}$$

Cooperative Preference To characterize the full spectrum between unilateral and cooperative tax setting, we allow each region's objective to place weight on both local and aggregate welfare. Specifically, when choosing its tax rate τ_i, region i maximizes

$$\mathcal{W}_i(\tau) \;=\; (1-\gamma)\, \ln W_i(\tau) \;+\; \gamma \sum_{n=1}^{N} \ln W_n(\tau), \tag{41}$$

where $\gamma \in [0,1]$ governs the degree of cooperative preference. When $\gamma = 0$, each region values only its own welfare, recovering the unilateral tax setting. When $\gamma = 1$, each region fully internalizes the national welfare externality, recovering the cooperative tax setting. Intermediate values of γ interpolate between these two polar cases, capturing varying degrees of interregional altruism or policy coordination.

6 Quantitative Results

6.1 Estimation

This subsection outlines the calibration of key model parameters and reports the quantitative results. We first describe how the model is disciplined by initial conditions, and then present the counterfactual results. Before using the model for counterfactual policy experiments, we calibrate it to a pre-pandemic equilibrium in China. The calibration combines bilateral trade flows from 2017 with bilateral patient flows from 2019, the closest available years for which both networks can be consistently measured. We interpret these data as capturing the baseline spatial structure of market access and medical access in China. This section describes the data sources and the procedures used to calibrate the main model parameters. Our final calibration sample consists of $N = 92$ cities,

which serve as the basic geographic units and have valid information on both baseline patient flows and trade flows.

6.1.1 Patient Flows and Trade Flows

Patient Flows We measure intercity patient mobility using bilateral patient-flow records in 2019 from the hospitalization data. These records identify the residence city of patients and the city in which they receive medical treatment, enabling us to observe cross-city medical visits at the origin–destination level. We aggregate the raw records to the city-pair level and use them to characterize the spatial structure of patient mobility across China.

Trade flows We measure bilateral trade flows using transaction-level records from Taobao, one of China's largest e-commerce platforms. The data cover the top 100 exporting cities in 2017. We aggregate all transactions to the city-pair level based on the origin and destination cities of delivery.[9] Since Taobao captures only online transactions within this platform, we scale these flows using the ratio of aggregate domestic trade to total transactions on Taobao.[10] Among the 100 exporting cities covered by the trade-flow data, 92 can be matched to cities with valid patient-flow information. These 92 cities constitute the final city sample used in the calibration.

6.1.2 Other city statistics

We collect additional city-level variables required for calibration, including the initial values of L_i, s_i, c_i, $w_{S,i}$, $w_{M,i}$, and R_i. We obtain city employment from the 2019 *China City Statistical Yearbook* to measure L_i. Using the patient-flow records, we calculate the total number of patients originating from each city, which allows us to infer s_i, the city-specific probability of becoming a patient. The patient-flow records also report medical expenses at the individual level. We use the pre-discount and pre-reimbursement medical charges to measure the cost of medical services. We aggregate these records by treatment destination and compute the average medical charge in each destination city, which serves as our measure of the unit cost of medical services, c_i. We obtain city-level sectoral wage

[9] Taobao consists of a Business-to-Consumer (B2C) platform, Tmall.com, and a Consumer-to-Consumer (C2C) platform, Taobao.com. Our analysis combines transactions from both platforms.

[10] Although Taobao e-commerce data account for only a fraction of total trade volume, they provide a useful lens through which to observe intercity trade patterns in China. The key insight is that while the platform does not capture the absolute level of trade flows, the relative shares across city pairs closely reflect the underlying structure of trade costs and market accessibility. These bilateral trade shares therefore serve as a meaningful proxy for spatial trade frictions, especially in settings where consistent micro-level transaction data are otherwise unavailable. Similar reasoning underlies the use of e-commerce or digital platform data in prior studies such as Lin (2019); Couture et al. (2021); Hu and Xu (2024), which show that online transaction networks can effectively approximate traditional trade linkages in the offline economy.

data for manufacturing and medical services primarily from the *China City Statistical Yearbook*. Missing city-level values are imputed using the corresponding provincial-level data.

Finally, we manually collect city-level reimbursement rates from local health-insurance policy documents. These policy schedules specify the reimbursement rules applicable to medical treatment in each city. For cities where city-specific policy documents are unavailable, we impute the reimbursement rate using the closest available policy schedule from the same province. For example, in Shandong Province, we use the reimbursement policy observed for Qingdao as the benchmark schedule for other cities in the province when their own city-specific rules cannot be consistently recovered.

6.1.3 Elasticities

The model delivers estimating equations for scale elasticities α_1, α_2, the service adequacy elasticity β (elasticity of utility with respect to capacity per caseload M_i/Q_i), and α, elasticity with respect to medical costs.

Equation (6) implies a revealed-preference identification strategy. We parameterize bilateral access costs as $\ln \rho_{ni} = \gamma_0 + \gamma_1 \ln times_{ni} + \nu_{ni}$, where $times_{ni}$ denotes the travel time and ν_{ni} is an orthogonal unobserved component. Taking expectations and logs of $Q_{ni} = \lambda_{ni} N_n$ yields a gravity equation for gross bilateral patient flows:

$$\ln \mathbb{E}[Q_{ni}] = \underbrace{\ln\left(\delta_i c_i^{-\alpha} M_i^{\beta} Q_i^{-\beta}\right)}_{\text{Destination Fixed Effects}} + \underbrace{\ln\left(\frac{N_n (1-R_n)^{-\alpha}}{\Phi_n}\right)}_{\text{Origin Fixed Effects}} - \gamma_0 - \gamma_1 \ln times_{ni}. \tag{42}$$

where $\mathbb{E}[Q_{ni}]$ denotes the number of patients. In estimation, we absorb the first two terms with destination and origin fixed effects.

Combining (15) with (42) and replacing $\ln\left(\delta_i c_i^{-\alpha} M_i^{\beta} Q_i^{-\beta}\right)$ by its fixed-effect estimate yields the second-stage estimating equation:

$$\ln\left(\widehat{\delta_i c_i^{-\alpha} M_i^{\beta} Q_i^{-\beta}}\right) = \alpha_1 \ln Q_{ii} + \alpha_2 \ln Q_{-i} + \beta \ln \frac{M_i}{Q_i} + (1-\alpha) \ln c_i - \upsilon \ln w_{S,i} - (1-\upsilon) \ln P_{M,i} + \xi_i, \tag{43}$$

from which we estimate α, α_1, α_2, β, and υ jointly. Here Q_{ii} is the local caseload in region i, Q_{-i} is the inbound caseload, and $Q_i = Q_{ii} + Q_{-i}$. The term ξ_i collects first-stage sampling error and any unobserved regional supply shifters orthogonal to the included controls.

Quality Estimates Our revealed-preference measures of medical service attractiveness are the destination fixed effects from estimating equation (42). Table 2 presents the estimation results from two different specifications. Our parameters of interest are the

exporter fixed effects. The fixed effects are normalized to have production-weighted mean of zero.

We then compare the fixed-effect estimates from the two specifications by examining the relationship between their relative ranks. Figure 5 shows that the ranks of the fixed-effect estimates are similar across the two specifications. Since Column (2) includes additional controls for same-city links and within-province travel-time effects, it better captures regional access frictions. We therefore use it as our preferred specification.

Figure 6 illustrates the geographic distribution of the estimated attractiveness in baseline specification, with higher values indicating better quality. Cities such as GuangZhou, ChangSha, and Chongqing are in the highest estimated quality group.

We finally relate these attractiveness measures to two external benchmarks: health service index and the mortality rate. Panels (b) and (c) of Figure 7 show a positive association between attractiveness and health service index and a negative association between attractiveness and mortality, respectively. Together, these patterns suggest that our attractiveness estimates capture clinically relevant characteristics.

Scale Elasticities and Service Adequacy Elasticity We use (43) to assess the strength of regional increasing returns. Table 3 reports the estimation results. Columns (1) and (2) use two alternative measures of medical capacity. The coefficients on both local and nonlocal patient volumes are positive and statistically significant. The coefficient on nonlocal patient volume implies that a 10% increase in nonlocal demand raises destination attractiveness by 0.043 points, equivalent to approximately 3.7% of the absolute value of its sample mean. Moreover, the coefficient on nonlocal patient volume is significantly larger than that on local patient volume, suggesting that regional scale economies are driven disproportionately by nonlocal demand.

Column (1), which measures medical capacity using hospital beds per capita, reports a positive and statistically significant coefficient of approximately 0.5. This estimate implies that a 10% increase in hospital beds per capita raises destination attractiveness by approximately 0.05 points, equivalent to about 4.3% of the absolute value of its sample mean. In contrast, column (2) shows that the coefficient on medical-institution beds per capita is positive but statistically insignificant. This contrast suggests that hospital capacity is more strongly associated with destination attractiveness than capacity provided by the broader set of medical institutions, which includes lower-tier, non-hospital providers.

The estimated effect of medical costs reflects two opposing forces. On the one hand, higher costs may proxy for higher-quality or more intensive medical services. On the other hand, they increase the effective cost of treatment and may discourage patients from seeking care. The coefficients on unit medical costs are positive and statistically significant in both columns, suggesting that the former channel dominates in the data.

This finding is consistent with the prediction of our model. As a robustness check, we re-estimate the specifications using the destination fixed-effect estimates recovered from column (1) of Table 2. The results are similar to the baseline estimates.

6.1.4 Other Model Parameters

We set the labor share in medical production to $\upsilon = 0.41$ and the labor share in manufacturing production to $\eta = 0.36$. The value of 0.41 is calculated as the ratio of hospital personnel expenses to medical revenue, using data from the *2021 China Health Statistics Yearbook*. The value of 0.36 is calculated using data from the *2020 China Input–Output Table*. Both values are measured for the year 2020. The trade elasticity is set to $\theta = 4$, following Simonovska and Waugh (2014). The weight on medical value in the welfare function, ξ, is set to 0.2. This value is chosen as a parsimonious benchmark and is broadly consistent with the substantial share of aggregate resources devoted to health care—around 18 percent of GDP in the United States—as well as the health-valuation framework of Hall and Jones (2007). The shape parameter governing workers' sectoral choice, ϵ, determines the elasticity of labor allocation across sectors with respect to relative wages. We set $\epsilon = 1.5$, in line with Galle et al. (2023).

6.1.5 Baseline tax rates

City-level effective tax rates, τ_i^0, are not directly observed in China. Although statutory contribution rules exist, they do not map cleanly into the effective tax burden relevant for our model. This concern is consistent with a large public-finance literature showing that actual tax liabilities depend on enforcement capacity, information reporting, evasion opportunities, and administrative institutions, rather than statutory rates alone (Kleven et al., 2011; Pomeranz, 2015; Best et al., 2015; Slemrod, 2019). In the case of social insurance contributions, effective burdens may further differ from statutory schedules because of incomplete participation, wage underreporting, informal employment, exemptions, and weak or uneven enforcement (Kumler et al., 2020). These concerns are particularly relevant in China, where social insurance and health insurance systems have historically been decentralized and fragmented across localities and population groups (Wagstaff et al., 2009; Meng et al., 2015).

Given the data $\{L_i, s_i, R_i, c_i, \Lambda, \Pi\}$ and a candidate vector of tax rates τ, the manufacturing market-clearing condition Equation A.11, together with the expenditure decomposition Equation A.13, reduces to a linear system in the vector of manufacturing-sector income, $Y^M \equiv (w_{M,i}L_{M,i})_{i=1}^N$:

$$\left[I \;-\; \eta\,\Pi^\top \operatorname{diag}\!\left(\tfrac{1}{\eta} - \tau\right)\right] Y^M \;=\; \eta\,\Pi^\top \left[\left(\tfrac{1}{\upsilon} - \tau\right) \odot Y^S \;-\; s \odot L \odot \bar{c} \odot (1 - R)\right], \qquad (44)$$

where $Y_i^S \equiv \upsilon c_i Q_i$ denotes medical-sector income, and $\bar{c}_i = \sum_k \lambda_{ik} c_k$ is the patient-flow-weighted average medical cost faced by residents of city i, both can be recovered from the data. For any candidate τ, solving Equation 44 yields the implied average wage,

$$\bar{w}_i(\tau) = \frac{Y_i^S + Y_i^M}{L_i}.$$

The local government budget identity Equation A.10 then pins down the implied per-capita deficit,

$$z_i(\tau) = s_i R_i \bar{c}_i - \tau_i \bar{w}_i(\tau),$$

and the deficit-to-revenue ratio,

$$\zeta_i(\tau) \equiv \frac{z_i(\tau)}{\tau_i \bar{w}_i(\tau)}.$$

We calibrate the baseline tax vector to minimize city-level deficit-to-revenue ratios, subject to mild regularization that keeps effective tax rates close to a common benchmark and avoids implausibly large cross-city dispersion.

6.1.6 City-level cooperative preferences

We estimate the cooperative-preference parameter γ by exploiting the analytical first-order conditions of each city's tax-setting problem. The key advantage is that γ enters the first-order condition linearly and separately from the equilibrium gradients. We can therefore recover γ from local comparative statics evaluated at the calibrated baseline, without re-solving the full equilibrium under alternative values of γ.

The identifying restriction is that the observed tax vector τ^0 is an interior solution to the local-planner game, in which city i maximizes the weighted objective $\mathcal{W}_i$ defined in Equation (41). Evaluated at $\tau = \tau^0$, the first-order condition for city i is

$$(1-\gamma) \underbrace{\left.\frac{d \ln W_i}{d \ln \tau_i}\right|_{\tau^0}}_{d_{1,i}} + \gamma \underbrace{\left.\frac{d \sum_{n=1}^{N} \ln W_n}{d \ln \tau_i}\right|_{\tau^0}}_{d_{2,i}} = 0, \tag{45}$$

where $d_{1,i}$ is the own-welfare gradient with respect to city i's tax rate, and $d_{2,i}$ is the aggregate-welfare gradient with respect to the same tax rate. Both gradients are recovered from the model's comparative-static system evaluated at the calibrated baseline τ^0, so each pair $(d_{1,i}, d_{2,i})$ is a known function of the data and calibrated parameters.

Equation 45 can be rewritten as

$$d_{1,i} + \gamma\big(d_{2,i} - d_{1,i}\big) = 0, \qquad i = 1, \ldots, N.$$

This gives an N-dimensional set of restrictions for the scalar parameter γ. We estimate γ by minimizing the sum of squared first-order-condition residuals:

$$\widehat{\gamma} = \arg\min_{\gamma \in \mathbb{R}} \sum_{i=1}^{N} \Big[d_{1,i} + \gamma \big(d_{2,i} - d_{1,i} \big) \Big]^2 = - \frac{\sum_{i=1}^{N} d_{1,i} \big(d_{2,i} - d_{1,i} \big)}{\sum_{i=1}^{N} \big(d_{2,i} - d_{1,i} \big)^2}. \tag{46}$$

The estimator in Equation 46 is equivalent to the slope coefficient from an ordinary-least-squares regression through the origin of $d_{1,i}$ on $(d_{1,i} - d_{2,i})$. The point estimate is $\widehat{\gamma} = 0.541$ with $\mathrm{SE}(\widehat{\gamma}) = 0.053$ ($t = 10.21$, $N = 92$). The estimate lies strictly between the non-cooperative ($\gamma = 0$) and fully cooperative ($\gamma = 1$) benchmarks, and is significantly different from both, indicating that the observed pattern of local tax rates is consistent with cities internalizing a substantial but incomplete share of the cross-city welfare spillovers generated by their own tax decisions. This is consistent with an institutional environment in which local fiscal decisions are shaped by both local incentives and broader central-government coordination (Jiang, 2026).

6.2 Counterfactual

In this subsection we use the calibrated model to evaluate a sequence of counterfactual exercises. We begin by validating the model's ability to reproduce the observed adjustment in patient flows between 2016 and 2019. We then quantify the welfare value delivered by the existing decentralized insurance system relative to a no-insurance benchmark. With these reference points in hand, we study two policy reforms that take the structure of the insurance system as given but change how it is coordinated across cities: national tax coordination, in which a central planner jointly chooses all city-level tax rates, and national insurance pooling, in which both the reimbursement rate and the payroll contribution rate are equalized and the financing budget is consolidated nationally.

6.2.1 Model Performance

Armed with the calibrated parameters, we first benchmark the model's ability to replicate key data moments. In particular, we exploit changes in road connectivity between 2016 and 2019 and ask whether the model can reproduce the observed spatial adjustment in patient flows over this period. Starting from the calibrated 2019 equilibrium, we "turn back the clock" by downgrading the road network to approximate connectivity in early 2016. This approach mirrors Heblich et al. (2020) and Hu et al. (2026), who reconstruct historical transport networks to simulate past equilibria.

Based on the reconstructed network, we compute the implied changes in bilateral

patient access costs and iceberg trade costs,

$$\hat{\rho}_{ni} \equiv \rho_{ni,2016}/\rho_{ni,2019}, \qquad \hat{d}_{ni} \equiv d_{ni,2016}/d_{ni,2019}.$$

both can be estimated with the change in travel times together with the elasticity of cost with respect to time following:

$$\ln \hat{\rho}_{ni} = \beta^{\rho} \times \ln \hat{times}_{ni}, \qquad \ln \hat{d}_{ni} = \beta^{d} \times \ln \hat{times}_{ni}.$$

The elasticity β^{ρ}, according to the baseline estimates, is 2.0, and the elasticity β^{d}, according to a similar gravity estimates on taobao trade data 2017, is 0.3.

We then solve for the 2016 equilibrium using exact-hat algebra, keeping all other calibrated parameters at their 2019 values and the local insurance tax at τ^0. Figure 8 validates the model against three city-level moments of the 2016 equilibrium: the total destination-side caseload Q_i, the inbound caseload from other cities Q_{-i}, and the average treatment cost per visit c_i. Each panel plots the model-implied 2016 value against the observed counterpart in logs, with a fitted line reporting the slope and its t-statistic. The slopes are 0.24 ($t = 2.27$), 0.35 ($t = 3.13$), and 0.17 ($t = 2.29$), respectively for total caseload, inbound caseload, and treatment cost—all positive and statistically significant. Taken together, the three panels show that the model reproduces the cross-sectional ordering of destination-level medical activity and treatment cost under the historical road network, capturing the interplay of travel-cost frictions and general-equilibrium feedbacks. This retrospective fit underpins the credibility of the forward-looking counterfactuals that follow.

6.2.2 Welfare Gains from the Current Insurance System

We first quantify the welfare value delivered by China's existing decentralized insurance system relative to a counterfactual in which no public insurance is provided. In the no-insurance counterfactual, we simultaneously set the local payroll tax and the public reimbursement rate to zero,

$$\tau_n = R_n = 0, \qquad \forall n,$$

while holding the exogenous lump-sum component of household resources, z_n^0, fixed at its baseline level. Households therefore face the full treatment cost out of pocket at every destination, and local governments operate neither a payroll levy nor a reimbursement subsidy. This counterfactual should be interpreted as removing the current financing and reimbursement system while holding fixed baseline fiscal resources. Two features of this exercise are worth noting. First, when R_n is set to zero, patients face the full medical cost c_i, rather than the subsidized out-of-pocket price $(1 - R_n^0)c_i$. This directly lowers the medical inclusive value Ψ_n through Equation A.4. Second, holding z_n^0 fixed ensures

that the comparison does not mechanically remove baseline fiscal resources.

We evaluate welfare changes city by city using Equation 32. Averaged across cities, removing the current insurance system reduces resident welfare by 11.5% in consumption-equivalent terms. Equivalently, the existing decentralized insurance system delivers an average welfare gain of approximately 11.5% relative to the no-insurance benchmark, despite the financing distortions emphasized by the model. This result indicates that the current insurance system is welfare-enhancing relative to no public insurance. The optimal-tax counterfactuals below should therefore be interpreted as improving the spatial and fiscal design of an already valuable system, rather than as substitutes for insurance itself.

6.2.3 Policy Coordination and National Insurance Pooling

We next use the calibrated model to quantify the welfare gains from two progressively more integrated policy regimes: nationally coordinating local insurance taxes and consolidating city-level insurance budgets into a single national pool. In the baseline equilibrium, insurance taxes are set by local governments whose objectives partially internalize cross-city welfare spillovers. Specifically, the estimated cooperative-preference parameter is $\widehat{\gamma} = 0.54$, indicating that local governments place substantial, but incomplete, weight on the welfare consequences of their tax choices for other cities. As a result, the baseline tax system generally fails to fully internalize the cross-city effects generated by patient mobility, medical scale economies, and general-equilibrium wage responses. A city's tax choice affects not only its own residents, but also patients and workers elsewhere through changes in patient flows, medical costs, service quality, and trade-linked factor prices.

Coordination benchmark. We first consider a coordination benchmark in which each city fully internalizes the national welfare consequences of its own tax choice. City i continues to choose only its own tax rate τ_i, taking the tax rates of all other cities as given, but now evaluates that choice according to national welfare. In the notation of the cooperative-preference objective in Equation 41, this corresponds to setting $\gamma = 1$, so that city i's objective becomes

$$\mathcal{W}_i(\tau) = \sum_{n=1}^{N} \ln W_n(\tau).$$

For an interior solution, the resulting coordinated tax vector τ^{Coord} satisfies the system of first-order conditions

$$\left.\frac{\partial \sum_{n=1}^{N} \ln W_n(\tau)}{\partial \ln \tau_i}\right|_{\tau=\tau^{\text{Coord}}} = 0, \qquad i = 1, \ldots, N. \tag{47}$$

At this allocation, no city can increase national welfare through a unilateral change in its own tax rate.

Although implemented through city-level policy choices, this benchmark implements the same first-order conditions as a centralized government choosing the full vector of city-specific insurance taxes. It therefore isolates the gains from coordinating local tax policy while leaving the underlying policy instrument—city-level insurance taxes—unchanged.[11]

Welfare gains and decomposition. Moving from the observed baseline tax vector τ^0 to the coordinated tax vector τ^{Coord} generates an average welfare gain of 0.150 log points across cities, equivalent to a 16.2% increase in welfare-equivalent consumption for the representative resident of the average city. Of the $N = 92$ cities, 83 experience welfare gains under national tax coordination. Weighting cities by employment L_n reduces the average gain to 0.102 log points, or approximately 10.8%.[12]

The welfare-change formula in Equation 32 provides a natural decomposition into two channels: real consumption of the manufacturing composite and the inclusive value of medical destinations,

$$\Delta \ln W_n = \underbrace{\ln \frac{\hat{y}_{M,n}}{\hat{P}_{M,n}}}_{\text{real consumption}} + \xi \underbrace{\ln \hat{\Psi}_n}_{\text{medical value}}. \tag{48}$$

Panel A of Table 4 reports the average decomposition across cities. The real-consumption channel accounts for roughly 90% of the welfare gain from national tax coordination, while the medical inclusive-value channel accounts for the remaining 10%.

Tax adjustments and the spatial pattern of medical activity. National tax coordination does not uniformly raise or lower local insurance taxes. Among the $N = 92$ cities in the sample, 24 increase their tax rates relative to the baseline, while 68 reduce them. Figure 9 illustrates the corresponding spatial pattern of medical-sector responses. We report both changes in medical supply, measured by caseloads, and changes in medical value, measured by the inclusive value of medical access. Warm colors indicate increases and cool colors indicate declines. Caseload expansions are concentrated among smaller cities, particularly those located near existing coastal hubs and provincial capi-

[11] The conditions in Equation 47 stack, city by city, the first-order conditions of a central government choosing the full vector of city-level tax rates to maximize aggregate welfare. Under the regularity conditions ensuring an interior and unique optimum, the fixed point of the coordination game therefore coincides with the planner's solution over this policy instrument. Importantly, this is not an unrestricted first best: only the tax vector is chosen, while patient flows, medical quality, treatment costs, labor allocation, wages, and trade flows continue to be determined by the model's equilibrium conditions.

[12] The gap between the unweighted and employment-weighted averages suggests that smaller cities gain disproportionately from national tax coordination. One interpretation is that large cities are already relatively well served as medical destinations in the baseline, whereas smaller cities benefit more from improved access to regional medical hubs under the nationally coordinated allocation.

tals. National tax coordination therefore expands medical activity in second-tier cities surrounding major incumbents, broadening the spatial supply of regional medical services rather than simply reinforcing concentration in the largest medical hubs.

The pattern for medical inclusive value is similar, but not identical. Inclusive value captures residents' overall access to lower-cost, higher-quality, and more accessible treatment destinations rather than the supply response of any single destination city. Consequently, a city may experience only a modest increase in its own caseload while still benefiting substantially from improved access to nearby or better-connected medical providers. This distinction highlights an important feature of the coordinated allocation: welfare gains need not arise from expanding medical supply everywhere, but can instead come from reorganizing the spatial network of medical services so that patients gain better access to more efficient treatment locations.

Table 5 quantifies this reallocation by reporting the average log changes in four production-side outcomes of the local medical sector, grouped by quartile of baseline caseload Q_i^0: total caseload Q_i, nonlocal inbound flows Q_{-i}, medical quality δ_i, and medical-sector labor $L_{S,i}$. Cities in the smallest-caseload quartile (Q1) experience the largest increases along all four margins. By contrast, cities in the larger-caseload quartiles (Q3 and Q4) contract along the quality and labor margins even though their caseloads change little. Because the aggregate number of treatments is anchored by the population requiring care, much of the production-side reallocation therefore occurs through medical quality and labor rather than through the total number of treatments.

Table 6 provides complementary evidence by reporting the cross-city Pearson correlations between the tax adjustment under national coordination,

$$\Delta\tau_i \equiv \tau_i^{\text{Coord}} - \tau_i^0,$$

and four equilibrium outcomes: growth in total caseloads, growth in nonlocal inbound patient flows, growth in medical quality, and the reallocation of workers into the medical sector. All four correlations are positive and sizable. Thus, cities experiencing larger increases in local tax rates also tend to experience larger expansions in medical caseloads, inbound patient flows, medical quality, and the medical-sector labor share. This pattern is consistent with coordinated policy directing greater financing effort toward locations whose medical sectors expand and serve a larger number of nonlocal patients.

National insurance pooling. We next ask whether consolidating insurance budgets, rather than merely coordinating tax rates, generates additional welfare gains. Under both the baseline and the coordinated allocation τ^{Coord}, insurance remains financed at the origin-city level: each city's reimbursement expenditures must be covered by its own tax base. Consequently, the resources available to a local insurance fund remain tied to

local fiscal capacity rather than to where those resources generate the greatest welfare gains. Building on the coordinated allocation, we therefore consolidate the N independent city budgets into a single national pool while holding each city's tax rate fixed at τ_n^{Coord} and its reimbursement rate fixed at R_n^0. Operationally, the national pool implements a vector of per-capita transfers t_n across city insurance funds, so that city n's budget constraint becomes

$$s_n R_n \bar{c}_n \;=\; z_n + t_n + \tau_n^{\text{Coord}} \bar{w}_n, \qquad \sum_{n=1}^{N} L_n t_n = 0, \tag{49}$$

where $t_n > 0$ denotes a net recipient from the pool and $t_n < 0$ denotes a net contributor. The adding-up condition ensures that the national pool only reallocates existing insurance revenue across cities and neither creates nor destroys aggregate resources. It is therefore the resource constraint implied by replacing separate city-level insurance budgets with a single national budget.

We choose the transfer vector to maximize aggregate welfare, $\sum_n \ln W_n$. Because $t = 0$ remains feasible and reproduces the coordinated allocation without pooling, the incremental welfare gain from national pooling is weakly nonnegative by construction. Its magnitude therefore measures the welfare cost of tying each city's insurance resources to its own fiscal capacity even after local tax rates have been optimally coordinated.

Allowing national pooling on top of coordinated taxes raises average welfare by a further 0.028 log points per city, equivalent to a 2.9% increase in consumption-equivalent terms. This gain is of the same order of magnitude as those estimated for spatial fiscal-policy reforms in the United States (Fajgelbaum and Gaubert, 2020).[13] Combining this additional gain with that from national tax coordination brings the total welfare gain relative to the baseline to 19.5% in consumption-equivalent terms.

Among the $N = 92$ cities, 72 experience welfare gains relative to tax coordination alone. Panel B of Table 4 decomposes this incremental welfare gain into the two channels in Equation 48. The composition differs markedly from that under tax coordination. Whereas roughly 90% of the welfare gains from coordinating tax rates operate through real consumption, 58% of the additional gains from national pooling arise through the medical inclusive-value channel. This contrast highlights the distinct roles of the two policy interventions: coordinating tax rates primarily improves welfare through the consumption channel, whereas pooling insurance budgets primarily relaxes local financing constraints, allowing insurance resources to be reallocated toward locations where they generate greater medical value.

The redistribution required to achieve these gains is modest. The gross size of the

[13] Fajgelbaum and Gaubert (2020) estimate that moving from observed to optimal spatial transfers across U.S. cities raises welfare by approximately 4% in their benchmark specification with cross-skill spillovers, with gains ranging from 2% to 6% across alternative specifications.

pool—defined as the sum of all positive transfers,

$$\sum_n L_n \max(t_n, 0),$$

—equals 0.31% of aggregate labor income. By comparison, the additional welfare gain is equivalent to 0.33% of aggregate labor income. Thus, each yuan reallocated through the pool generates approximately 1.1 yuan of welfare-equivalent income.

Financing is highly concentrated on the contributor side but widely dispersed on the recipient side. Of the 92 cities, 15 are net contributors to the national pool, while the remaining 77 are net recipients. The five largest contributors—Wuhan, Guangzhou, Chongqing, Foshan, and Nanjing—provide 81% of the gross pool, whereas the five largest recipients account for only 21% of total inflows. The national pool therefore operates as a broad but relatively thin redistribution from a small number of large-city insurance funds to a wide set of smaller cities. Figure 10 displays the geography of this redistribution. Each city is shaded according to its transfer as a share of the gross pool—blue for net contributors and red for net recipients—so that, by construction, the blue shares and the red shares each sum to one. Outflows are concentrated in a small number of large hub cities, whereas inflows are dispersed across a broad set of smaller cities. The transfer map therefore complements Figure 9: the national pool directs financing toward many of the same second-tier cities in which caseloads and medical value expand under coordinated taxes, thereby supporting the spatial reallocation of medical activity that decentralized city-level budgets cannot fully finance.

7 Conclusion

This paper studies the cross-regional externalities generated by trade in inpatient medical services and the role of policy coordination in addressing them. Under origin-based health insurance, patients' home cities finance reimbursement, while destination cities provide medical services and bear the consequences of patient inflows for local production and capacity. Patient mobility therefore separates the location of financing from the location of service provision, turning local health-financing policies into instruments with cross-regional effects.

Using inpatient records from China, we provide evidence that patient concentration generates opposing forces. Instrumental-variable estimates show that greater patient volume reduces length of stay, a pattern consistent with hospitals accelerating bed turnover under capacity pressure. At the same time, larger patient volumes raise medical productivity. The scale effects are positive for both local and nonlocal demand, with the estimates suggesting that nonlocal demand plays a particularly important role. We incorporate these mechanisms into a multi-region general equilibrium model disciplined by

bilateral patient flows, medical capacity, costs, and regional economic conditions.

Our counterfactual results show that decentralized local tax choices do not fully internalize the spillovers generated by patient mobility. Coordinating local health-financing taxes and pooling insurance financing at the national level can both improve welfare, although they operate through different margins. The findings suggest that facilitating cross-regional access to care should be accompanied by policies that better align financing with the destination-side responsibilities created by patient inflows. More broadly, when locally provided services become tradable through consumer mobility, effective policy design must account for both the gains from market expansion and the costs imposed on destination capacity.

References

Ackley, C. A. (2022). Tiered cost sharing and health care demand. Journal of Health Economics 85, 102663.

Acton, J. P. (1975). Nonmonetary factors in the demand for medical services: Some empirical evidence. Journal of Political Economy 83(3), 595–614.

Bagwell, K. and R. W. Staiger (1999). An economic theory of GATT. American Economic Review 89(1), 215–248.

Beckert, W., M. Christensen, and K. Collyer (2012). Choice of NHS-funded hospital services in England. The Economic Journal 122(560), 400–417.

Best, M. C., A. Brockmeyer, H. J. Kleven, J. Spinnewijn, and M. Waseem (2015). Production versus revenue efficiency with limited tax capacity: Theory and evidence from pakistan. Journal of Political Economy 123(6), 1311–1355.

Brot-Goldberg, Z. C., A. Chandra, B. R. Handel, and J. T. Kolstad (2017). What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics. The Quarterly Journal of Economics 132(3), 1261–1318.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2024). On binscatter. American Economic Review 114(5), 1488–1514.

Chandra, A. and D. O. Staiger (2007). Productivity spillovers in healthcare: Evidence from the treatment of heart attacks. Journal of Political Economy 115(1), 103–140.

Couture, V., B. Faber, Y. Gu, and L. Liu (2021). Connecting the countryside via e-commerce: evidence from china. American Economic Review: Insights 3(1), 35–50.

Dekle, R., J. Eaton, and S. Kortum (2007). Unbalanced trade. American Economic Review 97(2), 351–355.

Dingel, J. I., J. D. Gottlieb, M. Lozinski, and P. Mourot (2023). Market size and trade in medical services. Technical report, National Bureau of Economic Research.

Donaldson, D. and R. Hornbeck (2016). Railroads and American economic growth: A "market access" approach. The Quarterly Journal of Economics 131(2), 799–858.

Ederington, J. (2001). International coordination of trade and domestic policies. American Economic Review 91(5), 1580–1593.

Fajgelbaum, P. D. and C. Gaubert (2020). Optimal spatial policies, geography, and sorting. The Quarterly Journal of Economics 135(2), 959–1036.

Galle, S., A. Rodríguez-Clare, and M. Yi (2023). Slicing the pie: Quantifying the aggregate and distributional effects of trade. Review of Economic Studies 90(1), 331–375.

Gaynor, M., C. Propper, and S. Seiler (2016). Free to choose? reform, choice, and consideration sets in the English National Health Service. American Economic Review 106(11), 3521–3557.

Gaynor, M., H. Seider, and W. B. Vogt (2005). The volume–outcome effect, scale economies, and learning-by-doing. American Economic Review 95(2), 243–247.

Gutacker, N., L. Siciliani, G. Moscelli, and H. Gravelle (2016). Choice of hospital: Which type of quality matters? Journal of Health Economics 50, 230–246.

Hall, R. E. and C. I. Jones (2007). The value of life and the rise in health spending. The Quarterly Journal of Economics 122(1), 39–72.

Harstad, B. (2024). Trade and trees. American Economic Review: Insights 6(2), 155–175.

Heblich, S., S. J. Redding, and D. M. Sturm (2020). The making of the modern metropolis: evidence from london. The Quarterly Journal of Economics 135(4), 2059–2133.

Ho, K. (2006). The welfare effects of restricted hospital choice in the U.S. medical care market. Journal of Applied Econometrics 21(7), 1039–1079.

Hoe, T. P. (2022). Does hospital crowding matter? evidence from trauma and orthopedics in England. American Economic Journal: Economic Policy 14(2), 231–262.

Hu, Y., Y. Li, and M. Xu (2026). The trans-sumatra toll road and economic geography in indonesia. Journal of Economic Geography. Advance online publication.

Hu, Y. and M. Xu (2024). High-speed rail and the demand for skill: Evidence from chinese exports. Available at SSRN 4976973.

Jiang, Y. (2026). Strategic investment in transportation networks: Evidence from the U.S. highway system. Working Paper.

Kc, D. S. and C. Terwiesch (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. Management Science 55(9), 1486–1498.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. Econometrica 79(3), 651–692.

Kumler, T., E. Verhoogen, and J. Frías (2020). Enlisting employees in improving payroll-tax compliance: Evidence from mexico. Review of Economics and Statistics 102(5), 881–896.

Lin, Y. (2019). E-urbanism: E-commerce, migration, and the transformation of taobao villages in urban china. Cities 91, 202–212.

Lin, Z., F. Ba, H. Allore, G. G. Liu, and X. Chen (2022). Geographic variation in inpatient care utilization, outcomes and costs for dementia patients aged 65 years or older—China, 2017–2019. China CDC Weekly 4(45), 997.

Lu, Y. and D. J. G. Slusky (2016). The impact of women's health clinic closures on preventive care. American Economic Journal: Applied Economics 8(3), 100–124.

Ma, L. and Y. Tang (2024). The distributional impacts of transportation networks in China. Journal of International Economics 148, 103873.

Meng, Q., H. Fang, X. Liu, B. Yuan, and J. Xu (2015). Consolidating the social health insurance schemes in china: Towards an equitable and efficient health system. The Lancet 386(10002), 1484–1492.

National Health Commission of the PRC (2022). 2022 statistical bulletin on the development of China's health care industry.

National Health Commission of the PRC (2023). China health statistics yearbook 2023.

Nordhaus, W. (2015). Climate clubs: Overcoming free-riding in international climate policy. American Economic Review 105(4), 1339–1370.

Ossa, R. (2014). Trade wars and trade talks with data. American Economic Review 104(12), 4104–4146.

Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value added tax. American Economic Review 105(8), 2539–2569.

Prager, E. (2020). Healthcare demand under simple prices: Evidence from tiered hospital networks. American Economic Journal: Applied Economics 12(4), 196–223.

Sá, L., L. Siciliani, and O. R. Straume (2019). Dynamic hospital competition under rationing by waiting times. Journal of Health Economics 66, 260–282.

Simonovska, I. and M. E. Waugh (2014). The elasticity of trade: Estimates and evidence. Journal of international Economics 92(1), 34–50.

Slemrod, J. (2019). Tax compliance and enforcement. Journal of Economic Literature 57(4), 904–954.

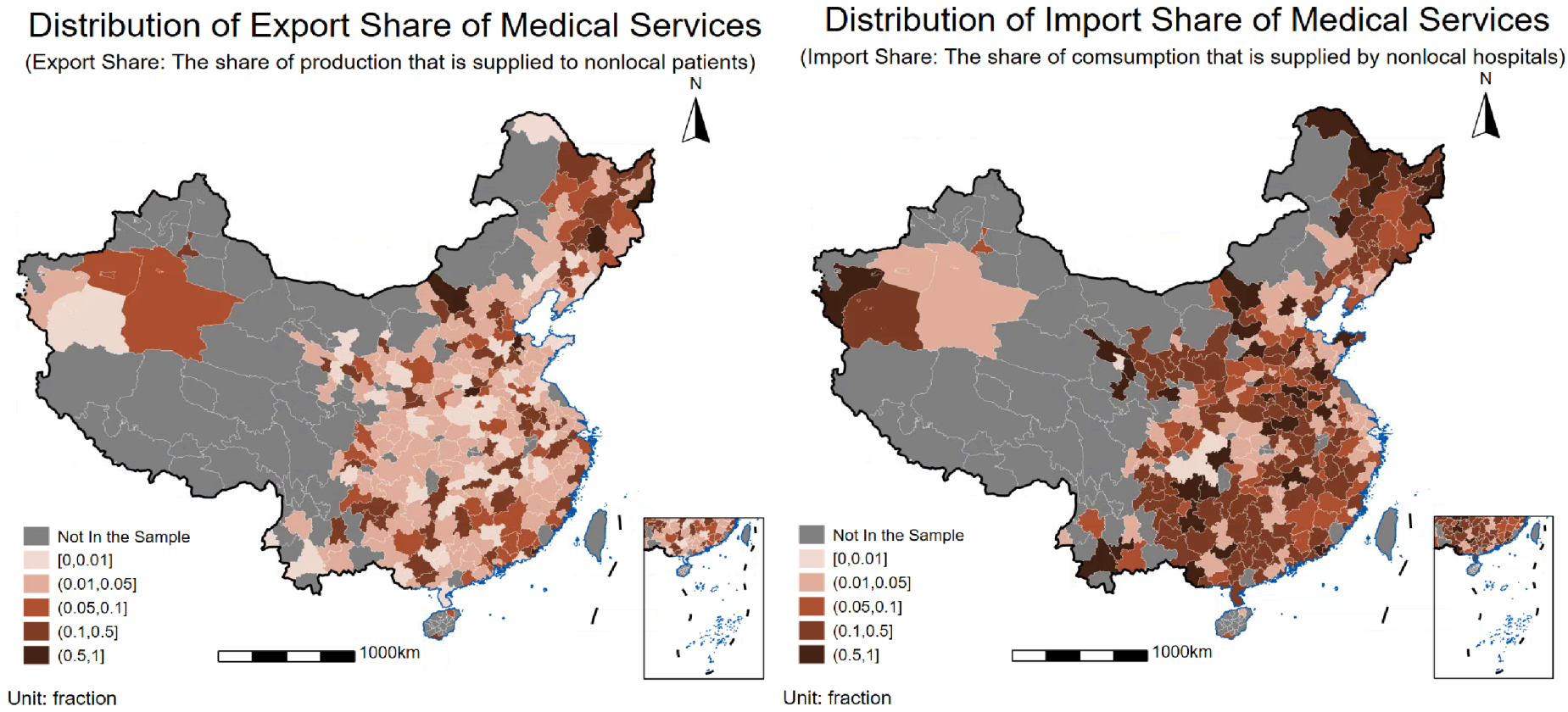
Tay, A. (2003). Assessing competition in hospital care markets: The importance of accounting for quality differentiation. The RAND Journal of Economics 34(4), 786–814.

Wagstaff, A., M. Lindelow, J. Gao, L. Xu, and J. Qian (2009). Extending health insurance to the rural population: An impact evaluation of china's new cooperative medical scheme. Journal of Health Economics 28(1), 1–19.

Wu, W., G. Liu, and Y. Pan (2025). Hospital workload and adaptation under climate change: Evidence from the People's Republic of China. Adb economics working paper series, Asian Development Bank.

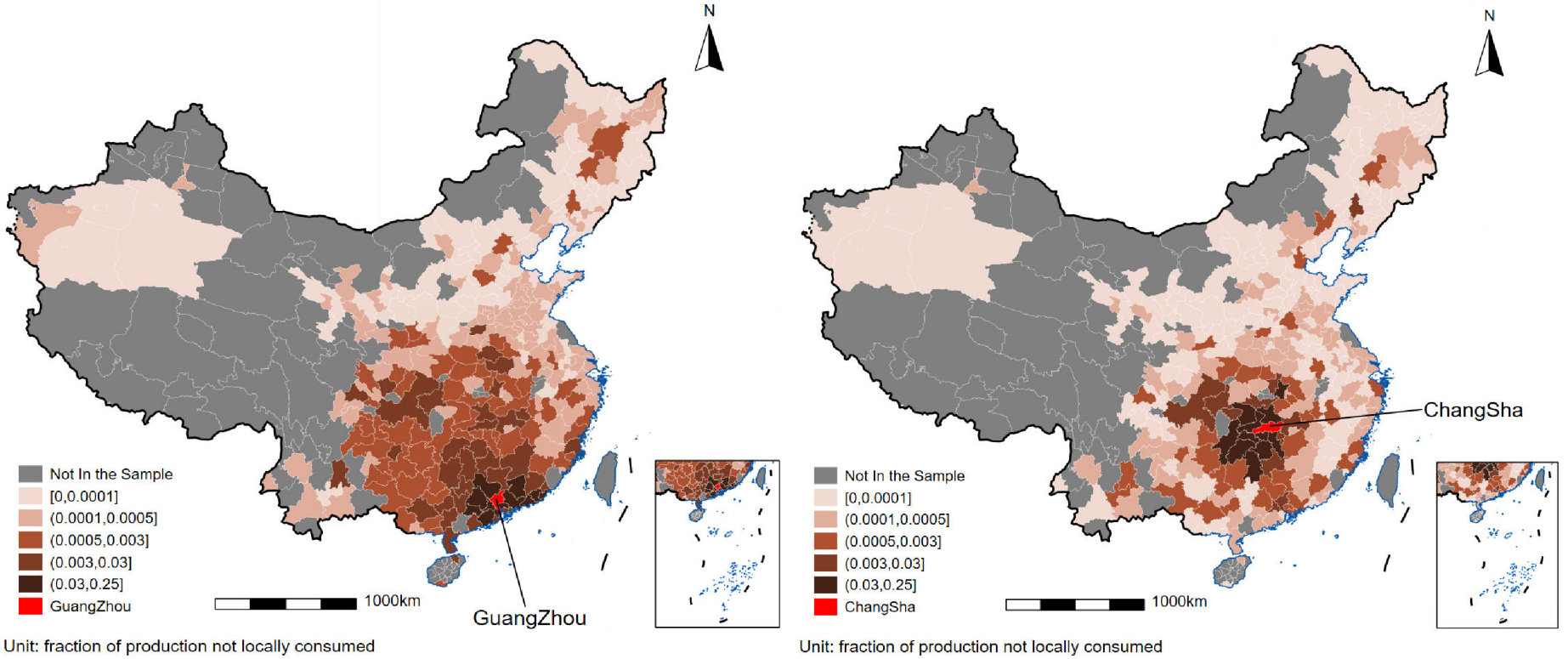
Figures & Tables

Figure 1: Distributions of export share and import share of medical services



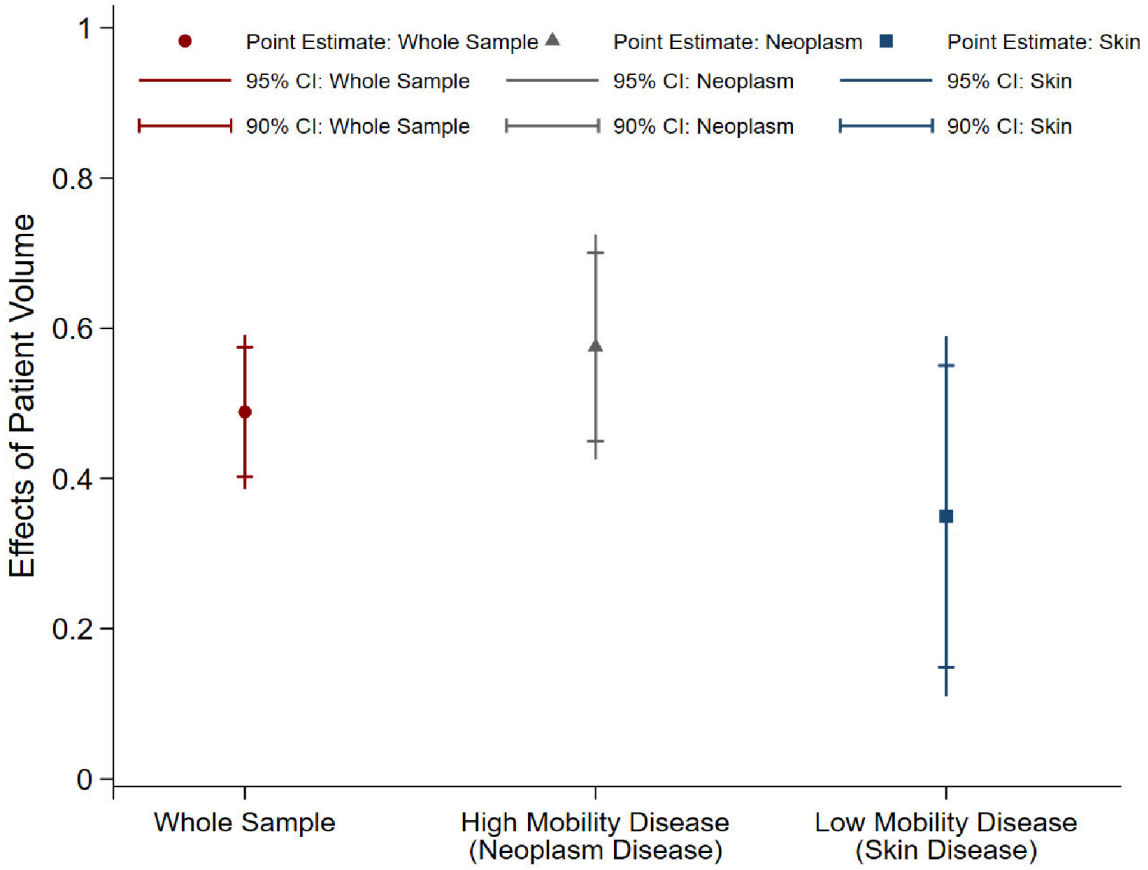
Notes: These figures present the export share and import share of medical services. Darker shading means higher share.

Figure 2: Destinations of medical services



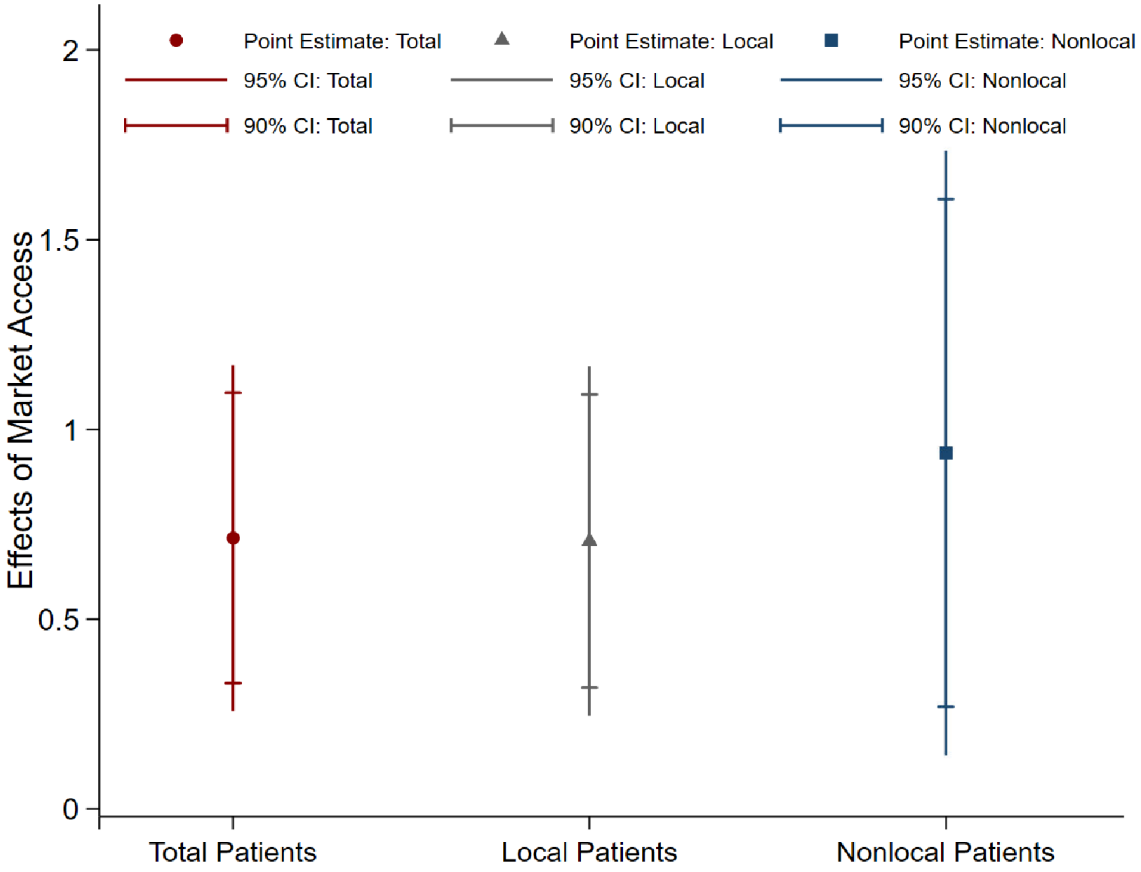
Notes: These figures show the destinations of medical services produced in two representative cities: GuangZhou and ChangSha. Darker shading indicates higher consumption of services originating from these two cities in each region.

Figure 3: Heterogeneous scale effects by disease mobility



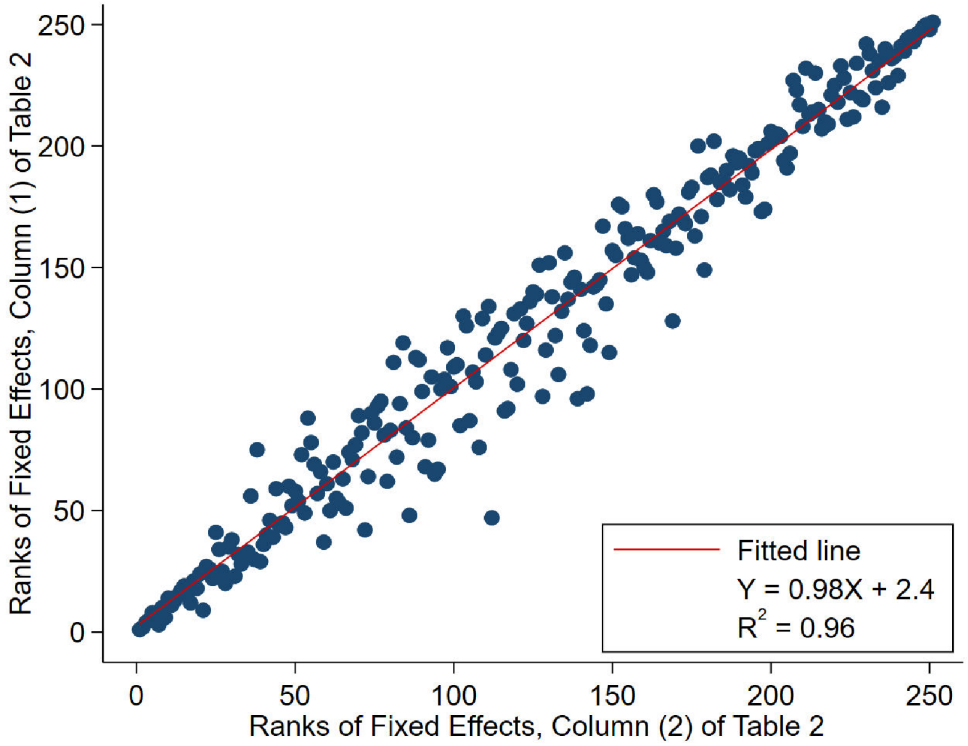
Notes: This figure reports the estimated effects of patient volume on productivity for the full sample and for two representative disease categories with different levels of cross-regional patient mobility. Neoplasms are selected as a high-mobility disease category, whereas skin diseases are selected as a low-mobility disease category. The markers denote point estimates, and the vertical lines report 90% and 95% confidence intervals. The detailed results are reported in Table C1.

Figure 4: Effects of market access on patient volume by patient origin



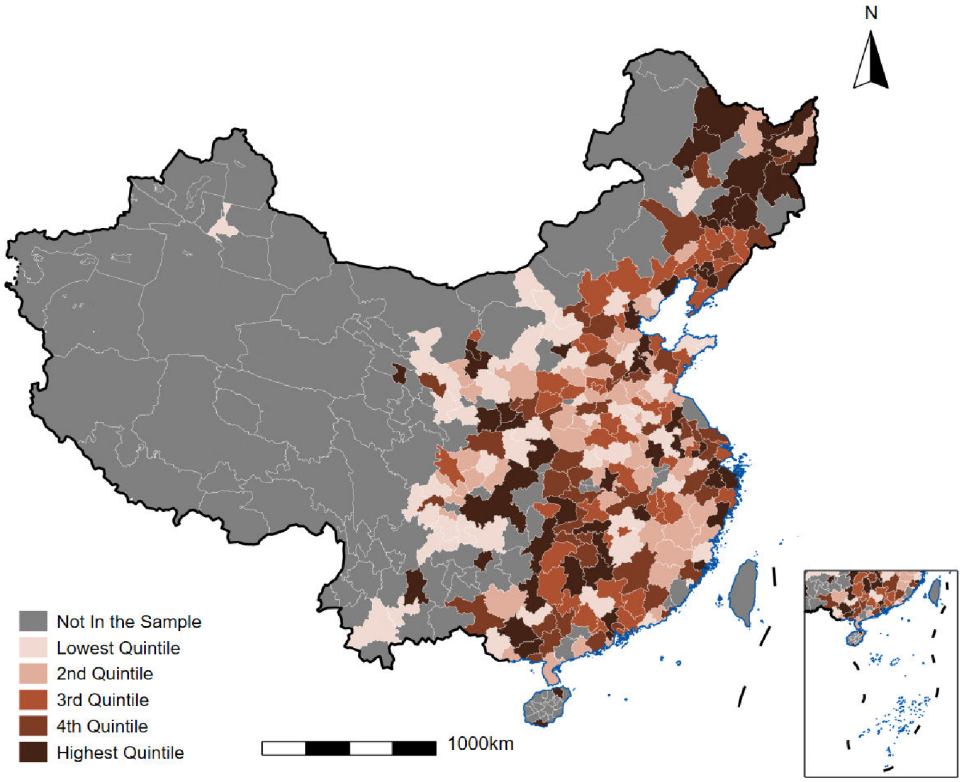
Notes: This figure reports the estimated effects of market access on total, local, and nonlocal patient volumes. Local patients are those residing in the destination city, whereas nonlocal patients are those traveling from other cities for treatment. The markers denote point estimates. The longer vertical lines indicate 95% confidence intervals, while the shorter capped lines indicate 90% confidence intervals. The detailed results are reported in Table C2.

Figure 5: Correlation of destination fixed effects ranks across specifications



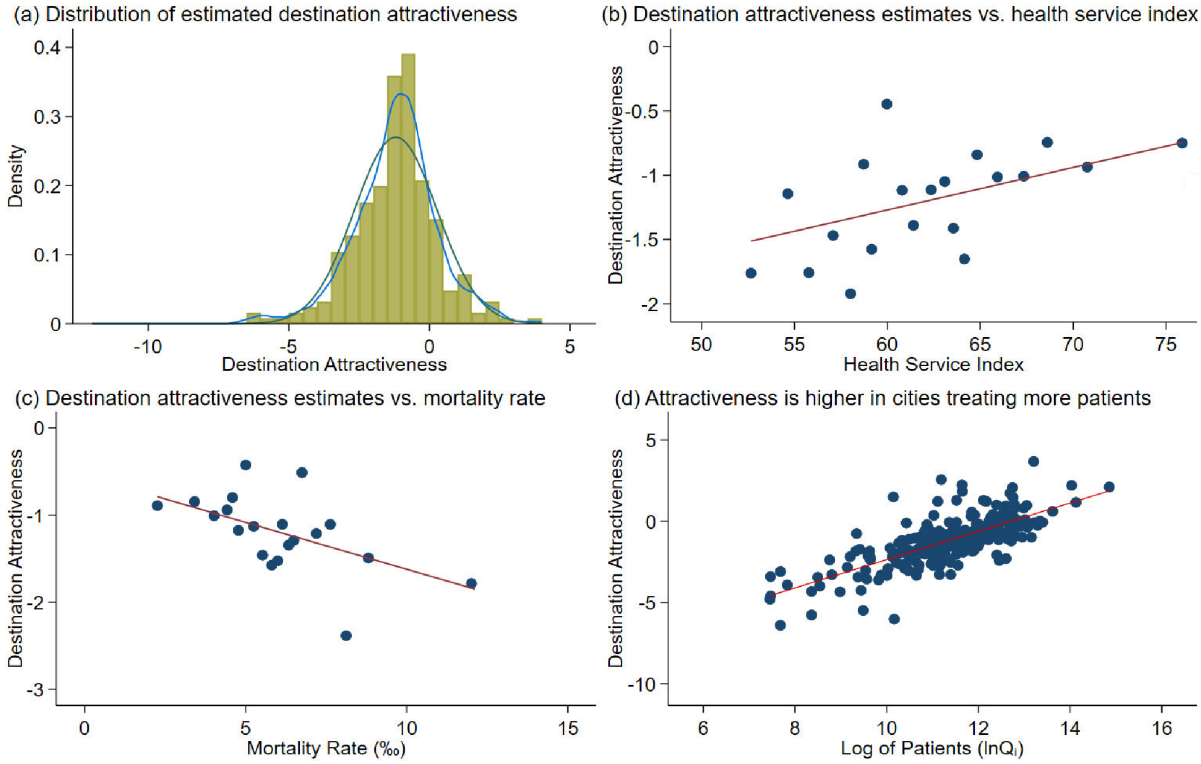
Notes: This figure compares the rankings of the estimated destination fixed effects in columns (1) and (2) of Table 2. Each point represents a destination city. The horizontal axis reports the rank based on column (2), while the vertical axis reports the rank based on column (1). The red line is the fitted linear relationship. The estimated slope of 0.98 and the R^2 of 0.96 indicate that the rankings of destination fixed effects are highly stable across the two specifications.

Figure 6: Distribution of city-level attractiveness estimates



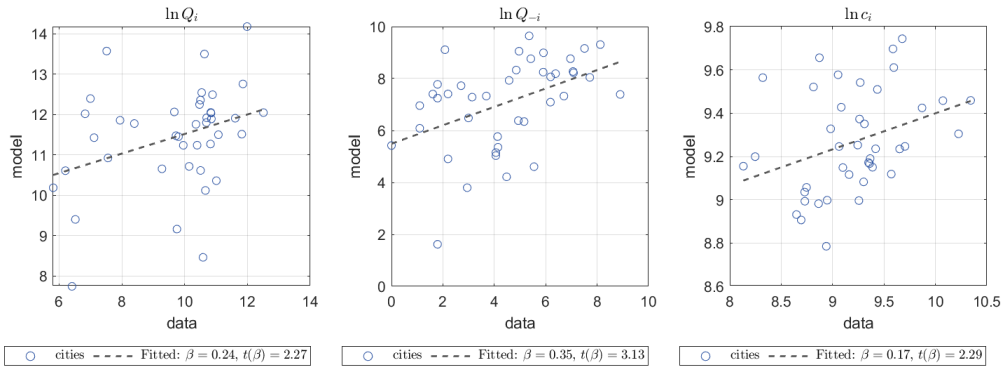
Notes: This figure presents the geographic distribution of attractiveness estimates in column (2) of Table 2. The estimates are ranked and divided into five approximately equal-sized groups, with each group containing about one-fifth of the sample cities. Darker shading indicates higher estimated attractiveness. Cities shaded in gray are not included in the estimation sample.

Figure 7: Validation of attractiveness estimates



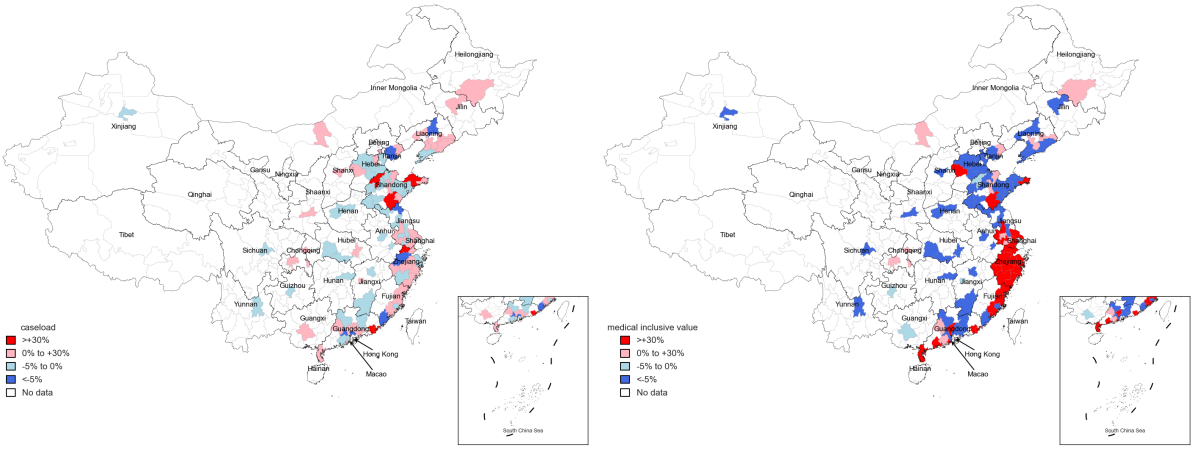
Notes: Panel A displays the distribution of the attractiveness estimates. Panels B and C, using the binscatter approach (Cattaneo et al., 2024), illustrate the relationships between attractiveness and health service index and between attractiveness and the mortality rate, respectively. Panel D shows the relationship between attractiveness and patients.

Figure 8: Model Fit: City-Level 2016 Moments, Model vs Data



Note: Each panel plots the model-implied 2016 value (vertical axis) of a city-level moment against its observed counterpart (horizontal axis), in natural logs. Panel (a) shows the total destination-side caseload $\ln Q_i$, panel (b) the inbound caseload from other cities $\ln Q_{-i}$, and panel (c) the average treatment cost per visit $\ln c_i$. The model-implied 2016 equilibrium is obtained by exact-hat algebra, applying the estimated travel-cost shifters $\widehat{\rho}_{ni}$ and $\widehat{d}_{ni}$ derived from the 2016–2019 change in bilateral road travel times, while holding all other calibrated parameters and the local insurance tax fixed at their 2019 levels. The dashed line in each panel is the OLS regression of the log model value on the log data value; the legend reports the slope coefficient and its t-statistic.

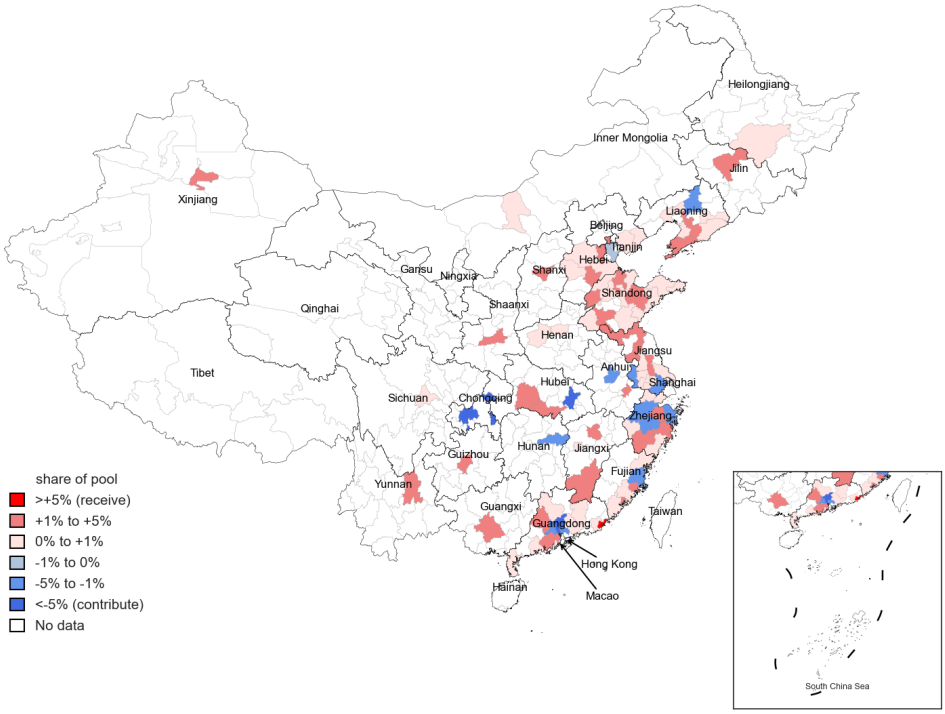
Figure 9: National Tax Coordination: Spatial Pattern of Medical-Side Responses



(a) Change in caseloads (b) Change in inclusive medical value

Note: Each map plots city-level log changes in counterfactual equilibrium objects as local insurance taxes move from the calibrated baseline vector τ^0 to the coordinated tax vector τ^{Coord}. Panel (a) reports the change in total medical caseload treated in destination i, $\ln \hat{Q}_i$. Panel (b) reports the change in the inclusive value of medical destinations available to residents of origin city n, $\ln \hat{\Psi}_n$. Color intensity represents the magnitude of the change: warm colors indicate increases and cool colors indicate declines. Cities with missing or incomplete data are shaded in white.

Figure 10: National Insurance Pooling: Geography of Cross-City Transfers



Note: The map plots each city's optimal transfer under national insurance pooling, expressed as a share of the gross pool, $L_n t_n / \sum_m L_m \max(t_m, 0)$. Red shades denote net recipients ($t_n > 0$) and blue shades denote net contributors ($t_n < 0$), with darker colors indicating larger shares of the pool. By construction, the red shares sum to one and the blue shares sum to minus one. Tax rates are held at the coordinated vector τ^{Coord} and reimbursement rates at their baseline values. Cities with missing data are shaded in white.

Table 1: Congestion effects of patient volume

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable:	log(No. of patients)		log(Average LOS)	log(Average LOS of local patients)	log(Average LOS of nonlocal patients)	log(Average DRG weights)
log(Market access)	1.372***	1.384***				
	(0.313)	(0.300)				
log(No. of patients)			-0.085***	-0.080***	-0.090	-0.020
			(0.030)	(0.030)	(0.056)	(0.015)
Province-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
City-disease fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
No. of secondary hospitals		Yes	Yes	Yes	Yes	Yes
No. of tertiary hospitals		Yes	Yes	Yes	Yes	Yes
Mean of outcome in levels	2801	2801	9.47	9.46	9.10	0.93
F statistics	19.28	21.29				
First-stage F statistics			21.29	23.12	16.41	21.29

Notes: This table reports instrumental-variable estimates of the effects of patient volume on length of stay and average DRG weights. The full balanced panel would contain 7,518 observations (179 cities × 21 diseases × 2 years). Because information on some diseases is unavailable for certain cities, the regression of the full sample contains 7,426 observations. Columns (1) and (2) report the first-stage estimates, in which the dependent variable is the log number of patients and the excluded instrument is log market access. Columns (3)–(6) report the corresponding second-stage estimates. The F-statistics in columns (1) and (2) test the significance of the excluded instrument in the first-stage regressions. The first-stage F-statistics reported in columns (3)–(6) are the Kleibergen–Paap rk Wald F-statistics for weak identification. The row labeled "Mean of outcome in levels" reports the sample mean of the corresponding untransformed dependent variable. Standard errors, clustered at the city level, are reported in parentheses. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Table 2: Estimating destination fixed effects from revealed preferences

	(1)	(2)
Dependent Variable:	Bilateral patient flows from city i (origin) to city j (destination)	
log(Travel time)	-2.716***	-2.049***
	(0.061)	(0.160)
A dummy indicating whether same province	2.236***	2.233***
	(0.220)	(0.310)
A dummy indicating whether same province * log(Travel time)		0.752***
		(0.252)
A dummy indicating whether same region		2.770***
		(0.426)
	Exporter FE and importer FE are included	
No. of observations	63001	63001
Pseudo R-squared	0.979	0.981

Notes: This table reports estimates of destination fixed effects using two different specifications. Column (2) serves as the baseline. Estimates are obtained using a pseudo–Poisson maximum likelihood (PPML) model. Standard errors, are reported in parentheses. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Table 3: Scale Elasticities, Service Adequacy Elasticity, and Cost Sensitivity

	(1)	(2)
log(No. of local patients): α_1	0.275***	0.263***
	(0.060)	(0.060)
log(No. of nonlocal patients): α_2	0.426***	0.439***
	(0.039)	(0.039)
Medical capacity per caseload:	0.500**	
log(No. of hospital beds per 1000 people)	(0.219)	
Medical capacity per caseload:		0.408
log(No. of medical institution beds per 1000 people)		(0.265)
Unit cost of medical services: log(cost per inpatient visit)	0.691**	0.759***
	(0.269)	(0.266)
	Wage in the healthcare sector and producer price index are controlled	
p-value for H_0: $\alpha_2 \leq \alpha_1$	0.099	0.053
No. of observations	249	251

Notes: This table reports estimates of scale elasticities, service adequacy elasticity, and cost sensitivity. The dependent variable is the destination fixed effect estimated in column (2) of Table 2. Two different variables are used to measure medical capacity. Standard errors, are reported in parentheses. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Table 4: Welfare Decomposition: Policy Coordination and National Insurance Pooling (log points)

	(1) Total $\frac{1}{N}\sum_n \Delta \ln W_n$	(2) Real consumption $\frac{1}{N}\sum_n \ln(\hat{y}_{M,n}/\hat{P}_{M,n})$	(3) Medical value $\frac{\xi}{N}\sum_n \ln \hat{\Psi}_n$
Panel A: National tax coordination ($\tau^0 \to \tau^{\text{Coord}}$)			
Gains	0.150	0.132	0.017
Contribution	100.0%	88.4%	11.6%
Panel B: Additional gain from national insurance pooling			
Gains	0.028	0.012	0.017
Contribution	100.0%	41.9%	58.1%
Panel C: Total (coordination and pooling)			
Gains	0.178	0.144	0.034
Contribution	100.0%	81.0%	19.0%

Notes: Decomposition of the welfare change $\Delta \ln W_n = \ln(\hat{y}_{M,n}/\hat{P}_{M,n}) + \xi \ln \hat{\Psi}_n$, averaged across the $N = 92$ cities. Panel A reports the change from the baseline tax vector τ^0 to the coordinated vector τ^{Coord} with city-level budgets. Panel B reports the additional change from consolidating the city insurance budgets into a single national pool through budget-neutral transfers, holding tax rates at τ^{Coord} and reimbursement rates at R_n^0. The bottom panel reports the cumulative change relative to the baseline; Panels A and B sum to the total up to rounding.

Table 5: Medical-Sector Response by Baseline-Q_i Quartile

	(1) Caseload $\ln \hat{Q}_i$	(2) Inbound patients $\ln \hat{Q}_{-i}$	(3) Medical quality $\ln \hat{\delta}_i$	(4) Medical labor share $\ln \hat{\ell}_i$
Q1 (smallest)	+0.307	+0.838	+0.989	+0.197
Q2	+0.005	+0.199	+0.118	−0.176
Q3	−0.008	−0.002	−0.172	−0.355
Q4 (largest)	+0.001	−0.024	−0.228	−0.247

Notes: Cities are sorted by baseline destination caseload and split into four equal-size quartiles (Q1 = smallest). Entries are within-quartile means of the log change in each medical-sector outcome between the baseline and the national tax coordination counterfactual.

Table 6: Cross-City Correlations Between $\Delta\tau_i$ and Medical-Sector Responses

	(1) Caseload $\ln \hat{Q}_i$	(2) Inbound patients $\ln \hat{Q}_{-i}$	(3) Medical quality $\ln \hat{\delta}_i$	(4) Medical labor share $\ln \hat{\ell}_i$
Correlation r	+0.363	+0.578	+0.730	+0.294
t-statistic	+3.65	+6.64	+10.02	+2.88

Notes: Pearson cross-city correlations between the change in the local insurance tax under national tax coordination, $\Delta\tau_i = \tau_i^{\text{Coord}} - \tau_i^0$, and each of four medical-sector outcomes. The t-statistic is reported in the second row.

A Model Equations and Counterfactual Systems

A.1 Equilibrium System

$$\pi_{ni} = \frac{\left(w_{M,i}^{\eta} P_{M,i}^{1-\eta} d_{ni}\right)^{-\theta} (B_i)^{\theta}}{\sum_k \left(w_{M,k}^{\eta} P_{M,k}^{1-\eta} d_{nk}\right)^{-\theta} (B_k)^{\theta}} \tag{A.1}$$

$$P_{M,n} = \bar{\gamma} \left[\sum_i \left(w_{M,i}^{\eta} P_{M,i}^{1-\eta} d_{ni}\right)^{-\theta} (B_i)^{\theta} \right]^{-\frac{1}{\theta}} \tag{A.2}$$

$$\lambda_{ni} = \frac{c_i^{-\alpha} \delta_i M_i^{\beta} Q_i^{-\beta} \rho_{ni}^{-1}}{\sum_k c_k^{-\alpha} \delta_k M_k^{\beta} Q_k^{-\beta} \rho_{nk}^{-1}} \tag{A.3}$$

$$\Psi_n = (1 - R_n)^{-\alpha} \sum_k c_k^{-\alpha} \delta_k M_k^{\beta} Q_k^{-\beta} \rho_{nk}^{-1} \tag{A.4}$$

$$Q_{ni} = L_n s_n \lambda_{ni} \tag{A.5}$$

$$Q_{-i} = Q_i - Q_{ii} \tag{A.6}$$

$$Q_i = \sum_n Q_{ni} \tag{A.7}$$

$$c_i = \frac{w_{S,i}^{\upsilon} P_{M,i}^{1-\upsilon} \delta_i}{\bar{A}_i \, Q_{ii}^{\alpha_1} Q_{-i}^{\alpha_2}} \tag{A.8}$$

$$\bar{c}_n = \sum_i \lambda_{ni} c_i \tag{A.9}$$

$$s_n R_n \bar{c}_n \;=\; z_n + \tau_n \bar{w}_n \tag{A.10}$$

$$w_{M,i} L_{M,i} = \eta \sum_n X_{M,n} \pi_{ni} \tag{A.11}$$

$$w_{S,i} L_{S,i} = \upsilon c_i Q_i \tag{A.12}$$

$$X_{M,n} = \bar{w}_n L_n - s_n L_n \bar{c}_n + z_n L_n + \frac{1-\upsilon}{\upsilon} w_{S,n} L_{S,n} + \frac{1-\eta}{\eta} w_{M,n} L_{M,n} \tag{A.13}$$

$$\ell_i = \frac{(b_{S,i}\, w_{S,i})^{\epsilon}}{(b_{S,i}\, w_{S,i})^{\epsilon} + (b_{M,i}\, w_{M,i})^{\epsilon}} \tag{A.14}$$

$$\bar{w}_i = \frac{w_{S,i}\, L_{S,i} + w_{M,i}\, L_{M,i}}{L_i} \tag{A.15}$$

A.2 Counterfactual Hat System

To create the counterfactual equilibrium, we model a world with the exact same exogenous determinants as in the current-day baseline equilibrium, except that we assume that hat change in mobility costs and tax rate. To solve for these counterfactual equilibria, we follow the methodology introduced by Dekle et al. (2007) and express the equilibrium

conditions of the model in changes relative to their baseline values. The equilibrium conditions in terms of relative (hat) variables are then given by:

$$\hat{\pi}_{ni} = \frac{\left(\hat{w}_{M,i}^{\eta}\hat{P}_{M,i}^{1-\eta}\hat{d}_{ni}\right)^{-\theta}}{\sum_k \left(\hat{w}_{M,k}^{\eta}\hat{P}_{M,k}^{1-\eta}\hat{d}_{nk}\right)^{-\theta}\pi_{nk}} \tag{A.16}$$

$$\hat{P}_{M,n} = \Big[\sum_k \left(\hat{w}_{M,k}^{\eta}\hat{P}_{M,k}^{1-\eta}\hat{d}_{nk}\right)^{-\theta}\pi_{nk}\Big]^{-1/\theta} \tag{A.17}$$

$$\hat{\lambda}_{ni} = \frac{\hat{c}_i^{-\alpha}\hat{\delta}_i\hat{M}_i^{\beta}\hat{Q}_i^{-\beta}\hat{\rho}_{ni}^{-1}}{\sum_k \left(\hat{c}_k^{-\alpha}\hat{\delta}_k\hat{M}_k^{\beta}\hat{Q}_k^{-\beta}\hat{\rho}_{nk}^{-1}\right)\lambda_{nk}} \tag{A.18}$$

$$\hat{\Psi}_n = \sum_k \left(\hat{c}_k^{-\alpha}\hat{\delta}_k\hat{M}_k^{\beta}\hat{Q}_k^{-\beta}\hat{\rho}_{nk}^{-1}\right)\lambda_{nk} \tag{A.19}$$

$$\hat{Q}_{ii} = \hat{\lambda}_{ii} \tag{A.20}$$

$$Q_{ii}\hat{Q}_{ii} + Q_{-i}\hat{Q}_{-i} = Q_i\hat{Q}_i \tag{A.21}$$

$$\hat{Q}_i = \sum_n \mu_{ni}\hat{\lambda}_{ni} \tag{A.22}$$

$$\hat{c}_i = \frac{\hat{w}_{S,i}^{\upsilon}\hat{P}_{M,i}^{1-\upsilon}\hat{\delta}_i}{\hat{Q}_{ii}^{\alpha_1}\hat{Q}_{-i}^{\alpha_2}} \tag{A.23}$$

$$\hat{\bar{c}}_n = \sum_i \omega_{ni}\hat{\lambda}_{ni}\hat{c}_i \tag{A.24}$$

$$\hat{\bar{c}}_n = \frac{\zeta_n}{1+\zeta_n} + \frac{1}{1+\zeta_n}\hat{\tau}_n\hat{\bar{w}}_n \tag{A.25}$$

$$\hat{w}_{M,i}\hat{L}_{M,i} = \sum_n \phi_{ni}\hat{X}_{M,n}\hat{\pi}_{ni} \tag{A.26}$$

$$\hat{w}_{S,i}\hat{L}_{S,i} = \hat{c}_i\hat{Q}_i \tag{A.27}$$

$$\hat{X}_{M,n} = a_{1n}\hat{\bar{w}}_n - a_{2n}\hat{\bar{c}}_n + a_{3n} + a_{4n}\hat{w}_{S,n}\hat{L}_{S,n} + a_{5n}\hat{w}_{M,n}\hat{L}_{M,n} \tag{A.28}$$

$$\hat{L}_{S,i} = \frac{\hat{w}_{S,i}^{\epsilon}}{\ell_i\,\hat{w}_{S,i}^{\epsilon} + (1-\ell_i)\,\hat{w}_{M,i}^{\epsilon}} \tag{A.29}$$

$$\hat{L}_{M,i} = \frac{\hat{w}_{M,i}^{\epsilon}}{\ell_i\,\hat{w}_{S,i}^{\epsilon} + (1-\ell_i)\,\hat{w}_{M,i}^{\epsilon}} \tag{A.30}$$

where the average wage hat is

$$\hat{\bar{w}}_n = f_{S,n}\,\hat{w}_{S,n}\,\hat{L}_{S,n} + f_{M,n}\,\hat{w}_{M,n}\,\hat{L}_{M,n}, \tag{A.31}$$

and $f_{S,n} \equiv w_{S,n}L_{S,n}/(\bar{w}_n L_n)$, $f_{M,n} \equiv w_{M,n}L_{M,n}/(\bar{w}_n L_n) = 1 - f_{S,n}$ are the baseline income shares from each sector.

Several auxiliary shares, all evaluated at the baseline equilibrium, are useful for the

hat system:

$$\mu_{ni} = \frac{Q_{ni}}{Q_i}, \quad \omega_{ni} = \frac{\lambda_{ni} c_i}{\bar{c}_n}, \quad \phi_{ni} = \frac{X_{M,n}\pi_{ni}}{\sum_k X_{M,k}\pi_{ki}}, \quad \zeta_n = \frac{z_n}{\tau_n \bar{w}_n}.$$

In addition, define the expenditure-component shares in $X_{M,n}$ at baseline:

$$a_{1n} = \frac{\bar{w}_n L_n}{X_{M,n}}, \quad a_{2n} = \frac{s_n L_n \bar{c}_n}{X_{M,n}}, \quad a_{3n} = \frac{z_n L_n}{X_{M,n}}, \quad a_{4n} = \frac{\frac{1-\upsilon}{\upsilon} w_{S,n} L_{S,n}}{X_{M,n}}, \quad a_{5n} = \frac{\frac{1-\eta}{\eta} w_{M,n} L_{M,n}}{X_{M,n}}.$$

Solving the Model

- Step 1: Guess a vector of sectoral wages $(\hat{w}_S, \hat{w}_M)$.
- Step 2 (Inner loop): Use the sector choice equations (A.29)–(A.30) to compute $\hat{L}_S, \hat{L}_M$ from the guessed wages. Then use equations (A.16)–(A.25) to solve for prices index $\hat{P}_M$, quality $\hat{\delta}$, quantity $\hat{Q}$, unit cost $\hat{c}$ and the shares hat ($\hat{\pi}$, $\hat{\lambda}$) consistent with $(\hat{w}_S, \hat{w}_M)$. Then, solve for expenditures $\hat{X}_{M,n}$ using (A.28).
- Step 3 (Outer loop): Verify if equations (A.26) and (A.27) hold. If not, adjust the guess of $(\hat{w}_S, \hat{w}_M)$ and proceed to step 1 again until both equilibrium conditions are obtained.

B Marginal Welfare System

To evaluate this condition at a given policy vector τ, we first solve the hat system (A.16)–(A.30) to obtain the counterfactual equilibrium $E(\tau)$, with equilibrium objects $\pi_{ni}(\tau)$, $\lambda_{ni}(\tau)$, $w_{S,i}(\tau)$, $w_{M,i}(\tau)$, $P_{M,n}(\tau)$, $Q_i(\tau)$, $c_i(\tau)$, $\bar{c}_n(\tau)$, $\delta_i(\tau)$, $L_{S,i}(\tau)$, $L_{M,i}(\tau)$, $X_{M,n}(\tau)$, $\bar{w}_n(\tau)$, and $\Psi_n(\tau)$. We then log-differentiate the equilibrium conditions around $E(\tau)$ to obtain the following closed linear system. All shares and level variables appearing as coefficients are evaluated at $E(\tau)$:

$$\dot{\pi}_{ni} = -\theta \left[\eta \dot{w}_{M,i} + (1-\eta)\dot{P}_{M,i} - \sum_k \pi_{nk}(\tau) \left(\eta \dot{w}_{M,k} + (1-\eta)\dot{P}_{M,k} \right) \right] \tag{A.32}$$

$$\dot{P}_{M,n} = \sum_k \pi_{nk}(\tau) \left(\eta \dot{w}_{M,k} + (1-\eta)\dot{P}_{M,k} \right) \tag{A.33}$$

$$\dot{\lambda}_{ni} = \left(-\alpha \dot{c}_i + \dot{\delta}_i + \beta \dot{M}_i - \beta \dot{Q}_i \right) - \sum_k \lambda_{nk}(\tau) \left(-\alpha \dot{c}_k + \dot{\delta}_k + \beta \dot{M}_k - \beta \dot{Q}_k \right) \tag{A.34}$$

$$\dot{\Psi}_n = \sum_k \lambda_{nk}(\tau) \left(-\alpha \dot{c}_k + \dot{\delta}_k + \beta \dot{M}_k - \beta \dot{Q}_k \right) \tag{A.35}$$

$$\dot{Q}_{ii} = \dot{\lambda}_{ii} \tag{A.36}$$

$$Q_{ii}(\tau)\, \dot{Q}_{ii} + Q_{-i}(\tau)\, \dot{Q}_{-i} = Q_i(\tau)\, \dot{Q}_i \tag{A.37}$$

$$\dot{Q}_i = \sum_n \mu_{ni}(\tau)\, \dot{\lambda}_{ni} \tag{A.38}$$

$$\dot{c}_i = \upsilon \dot{w}_{S,i} + (1-\upsilon)\dot{P}_{M,i} + \dot{\delta}_i - \alpha_1 \dot{Q}_{ii} - \alpha_2 \dot{Q}_{-i} \tag{A.39}$$

$$\dot{\bar{c}}_n = \sum_i \omega_{ni}(\tau)\, (\dot{c}_i + \dot{\lambda}_{ni}) \tag{A.40}$$

$$\dot{\bar{c}}_n = \frac{\tau_n\, \bar{w}_n(\tau)}{z_n + \tau_n\, \bar{w}_n(\tau)}\, (\dot{\tau}_n + \dot{\bar{w}}_n) \tag{A.41}$$

$$\dot{w}_{M,i} + \dot{L}_{M,i} = \sum_n \phi_{ni}(\tau)\, (\dot{X}_{M,n} + \dot{\pi}_{ni}) \tag{A.42}$$

$$\dot{w}_{S,i} + \dot{L}_{S,i} = \dot{c}_i + \dot{Q}_i \tag{A.43}$$

$$\dot{X}_{M,n} = a_{1n}(\tau)\, \dot{\bar{w}}_n - a_{2n}(\tau)\, \dot{\bar{c}}_n + a_{4n}(\tau)\, (\dot{w}_{S,n} + \dot{L}_{S,n}) + a_{5n}(\tau)\, (\dot{w}_{M,n} + \dot{L}_{M,n}) \tag{A.44}$$

$$\dot{L}_{S,i} = \epsilon\, (1 - \ell_i(\tau))\, (\dot{w}_{S,i} - \dot{w}_{M,i}) \tag{A.45}$$

$$L_{M,i}(\tau)\, \dot{L}_{M,i} + L_{S,i}(\tau)\, \dot{L}_{S,i} = 0 \tag{A.46}$$

where the average wage log-differential is

$$\dot{\bar{w}}_n = f_{S,n}(\tau)\, (\dot{w}_{S,n} + \dot{L}_{S,n}) + f_{M,n}(\tau)\, (\dot{w}_{M,n} + \dot{L}_{M,n}). \tag{A.47}$$

Here, $\dot{M}_i$, $\dot{\rho}_{ni}$, and $\dot{d}_{ni}$ are exogenous shocks to medical capacity, patient access costs, and trade costs, respectively. In the tax counterfactual, $\dot{M} = \dot{\rho} = \dot{d} = 0$, so only $\dot{\tau}_i$ enters the right-hand side.

The auxiliary shares $\mu_{ni}(\tau), \omega_{ni}(\tau), \phi_{ni}(\tau), \ell_i(\tau), f_{S,n}(\tau), f_{M,n}(\tau)$ and the expenditure-component coefficients $a_{1n}(\tau), a_{2n}(\tau), a_{4n}(\tau), a_{5n}(\tau)$ share the same definitions as their baseline counterparts, but with all equilibrium objects evaluated at $E(\tau)$. At the calibrated baseline $\tau = \tau^0$, the hat system yields identity (all hats equal one), so $E(\tau^0)$ coincides with the observed baseline equilibrium and all shares reduce to their baseline values.

C Appendix Figures & Tables

Table C1: Scale effects by disease mobility

Dependent Variable:	(1)	(2)	(3)	(4)
	log(Number of patients)	log(Per-physician number of patients)		
	Full sample	Full sample	Neoplasm diseases	Skin diseases
log(Market access)	1.002***			
	(0.278)			
log(Number of patients)		0.489***	0.575***	0.350***
		(0.052)	(0.076)	(0.121)
Mean of outcome in levels	2,639	22.4	13.8	5.3
Province-year fixed effects	Yes	Yes	Yes	Yes
City-disease fixed effects	Yes	Yes	Yes	Yes
Number of secondary hospitals	Yes	Yes	Yes	Yes
Number of tertiary hospitals	Yes	Yes	Yes	Yes
F statistic	13.02			
First-stage F statistic		13.02	9.804	5.981

Notes: This table shows the scale effects by disease mobility. The full balanced panel would contain 7,518 observations (179 cities × 21 diseases × 2 years). Because information on some diseases is unavailable for certain cities, the regression of the full sample contains 7,246 observations. Column (1) reports the first-stage relationship between market access and the number of patients. Columns (2)–(4) report IV estimates of the effect of patient volumes on the per-physician number of patients for the whole sample, neoplasm diseases, and skin diseases, respectively. Province-year fixed effects, city-disease fixed effects, and controls for the numbers of secondary and tertiary hospitals are included in all specifications. The F-statistics in column (1) test the significance of the excluded instrument in the first-stage regressions. The first-stage F-statistics reported in columns (2)–(4) are the Kleibergen–Paap rk Wald F-statistics for weak identification. The row labeled "Mean of outcome in levels" reports the sample mean of the corresponding untransformed dependent variable. Standard errors, clustered at the city level, are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table C2: Effects of market access on patient flows

	(1)	(2)	(3)
	PPML		
Dependent Variable:	No. of patients	No. of local patients	No. of nonlocal patients
log(Market access)	0.706***	0.938**	0.714***
	(0.235)	(0.407)	(0.232)
Province-year fixed effects	Yes	Yes	Yes
City-disease fixed effects	Yes	Yes	Yes
Number of secondary hospitals	Yes	Yes	Yes
Number of tertiary hospitals	Yes	Yes	Yes
Mean of outcome in levels	2,617	2,370	257

Notes: This table reports pseudo-Poisson maximum likelihood (PPML) estimates of the effects of market access on the numbers of total, local, and nonlocal patients. The full balanced panel would contain 7,518 observations (179 cities $\times$ 21 diseases $\times$ 2 years). Because information on some diseases is unavailable for certain cities, and the PPML estimator eliminate singleton or separated groups, the regression of the full sample contains 7,412 observations. All specifications include province-year fixed effects, city-disease fixed effects, and controls for the numbers of secondary and tertiary hospitals. The row labeled "Mean of outcome in levels" reports the sample mean of the corresponding untransformed dependent variable. Standard errors, clustered at the city level, are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table C3: Robustness of scale elasticities, service adequacy elasticity, and cost sensitivity estimates

	(1)	(2)
log(No. of nonlocal patients): α_1	0.503***	0.518***
	(0.047)	(0.047)
log(No. of local patients): α_2	0.197***	0.188**
	(0.073)	(0.073)
Medical capacity per caseload:	0.456*	
log(No. of hospital beds per 1000 people)	(0.264)	
Medical capacity per caseload:		0.249
log(No. of medical institution beds per 1000 people)		(0.319)
Unit cost of medical services: log(cost per inpatient visit)	0.765**	0.828**
	(0.324)	(0.320)
	Wage in the healthcare sector and producer price index are controlled	
p-value for H_0: $\alpha_1 \leq \alpha_2$	0.006	0.003
No. of observations	249	251

Notes: This table reports estimates of scale elasticities, service adequacy elasticity, and cost sensitivity. The dependent variable is the destination fixed effects estimate recovered from column (1) of Table 2. The results reported in this table provide robustness checks for the baseline estimates in Table 3. Two different variables are used to measure medical capacity. Standard errors, are reported in parentheses. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.