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
[§]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; **?**), 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; **?**). 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 **?**, 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-quality services, and a wider variety of medical services. We build on their analysis in two ways. First,

[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

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 populations 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 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

[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 $\times$ 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.

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: $\frac{\text{No. of Patients}}{\text{No. of Doctors}}$.

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

[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.

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

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 C3.

5 Theoretical Framework

We develop a multi-region general-equilibrium model of care for a single medical procedure. Patients choose where to receive treatment based on out-of-pocket prices, provider quality, service adequacy, and travel costs. Hospitals use local labor and traded goods to produce care. Their costs and quality depend on the scale of medical activity, with local and nonlocal patients contributing differently, while a larger caseload can strain medical capacity. Insurance is administered in patients' regions of residence. Each local authority sets reimbursement for its residents, including those treated elsewhere, and finances the resulting expenditure through a payroll tax. Its policy therefore changes both the distribution of patients across destinations and the medical-market conditions they face. These changes feed back into wages, goods prices, and the allocation of labor across sectors. The framework allows us to compare decentralized insurance decisions with coordinated financing and to measure their welfare effects in general equilibrium.

5.1 Model Setup

We consider a set of regions in which residents may travel for medical care, while manufacturing goods are traded across regions. Each region has a fixed population L_i. Residents work in either medical services or manufacturing and do not move across regions

for work.

Patients: destination choice and medical trade. A fraction s_n of residents in origin n require treatment, so the number of patients is $N_n = L_n s_n$. Each patient chooses a treatment destination i. Let Q_i denote the caseload in destination i and $M_i > 0$ its effective medical capacity. A patient ω from n obtains utility

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

where p_{ni} is the out-of-pocket price, $\delta_i > 0$ is quality, and M_i/Q_i measures capacity per patient. The parameter $\beta > 0$ governs the value of service adequacy, while $\rho_{ni} > 0$ captures bilateral access costs. The shocks $\varepsilon_{ni}(\omega)$ are independent Type-I extreme-value draws across patients and destinations.

Insurance is based on residence. Given the reimbursement rate $R_n \in [0, 1)$ in origin n, a patient treated in i pays

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

where c_i is the unit cost of care in destination i. Define the origin-specific medical inclusive value

$$\Psi_n \equiv \sum_k p_{nk}^{-\alpha} \delta_k \left(\frac{M_k}{Q_k} \right)^{\beta} \rho_{nk}^{-1}. \tag{6}$$

The share of patients from n choosing i is

$$\lambda_{ni} = \frac{p_{ni}^{-\alpha} \delta_i (M_i/Q_i)^{\beta} \rho_{ni}^{-1}}{\Psi_n}. \tag{7}$$

Patient flows and destination caseloads then satisfy

$$Q_{ni} = N_n \lambda_{ni}, \qquad Q_i = \sum_n Q_{ni} = \sum_n N_n \lambda_{ni}. \tag{8}$$

We distinguish patients treated in their region of residence, Q_{ii}, from inbound patients, $Q_{-i} \equiv \sum_{n \neq i} Q_{ni}$, so that $Q_i = Q_{ii} + Q_{-i}$.

Medical providers: scale and capacity externalities. Medical services are produced competitively using local labor and a traded manufacturing composite. At the regional level, production takes the form

$$Q_i = A_i L_{S,i}^{\upsilon} m_{S,i}^{1-\upsilon} \upsilon^{-\upsilon} (1 - \upsilon)^{\upsilon - 1}, \tag{9}$$

where $L_{S,i}$ is medical-sector employment and $m_{S,i}$ is the manufacturing composite used in medical production. Providers take the region-level productivity term

$$A_i = \frac{\bar{A}_i H(Q_{ii}, Q_{-i})}{K(\delta_i)} \tag{10}$$

as given when choosing inputs. Here, $\bar{A}_i$ is an exogenous productivity shifter, H captures scale effects external to individual providers, and K is the cost of supplying quality.

We depart from Dingel et al. (2026) by allowing local and nonlocal caseloads to have distinct effects on medical productivity:

$$H(Q_{ii}, Q_{-i}) = Q_{ii}^{\alpha_1} Q_{-i}^{\alpha_2}. \tag{11}$$

Both types of patients contribute to regional scale, potentially at different rates. Both also add to Q_i and thereby reduce capacity per patient, M_i/Q_i, holding capacity fixed. The scale effect operates through production costs, while capacity pressure enters patients' destination choices. Following Dingel et al. (2026), we assume a linear quality-cost function, $K(\delta_i) = \delta_i$. Under perfect competition, the unit cost of treatment is

$$c_i = \frac{w_{S,i}^{\upsilon} P_{M,i}^{1-\upsilon} \delta_i}{\bar{A}_i Q_{ii}^{\alpha_1} Q_{-i}^{\alpha_2}}, \tag{12}$$

where $w_{S,i}$ is the medical-sector wage and $P_{M,i}$ is the price of the manufacturing composite.

Manufacturing production and goods trade. Manufacturing follows an Eaton–Kortum structure. In each region i, a continuum of varieties $x \in [0,1]$ is produced competitively using local labor and the manufacturing composite:

$$q_{M,i}(x) = \xi_i(x) B_i l_{M,i}(x)^{\eta} m_{M,i}(x)^{1-\eta} \eta^{-\eta} (1-\eta)^{\eta-1}. \tag{13}$$

The term B_i is a region-specific productivity shifter. Productivity draws are independent across varieties and regions and follow a unit Fréchet distribution with shape parameter θ:

$$\Pr(\xi_i(x) \leq z) = \exp(-z^{-\theta}), \qquad z > 0. \tag{14}$$

Aggregate manufacturing employment is

$$L_{M,i} = \int_0^1 l_{M,i}(x)\, dx. \tag{15}$$

Shipping one unit from i to n requires $d_{ni} \geq 1$ units at the origin.

Buyers source each variety from the lowest-cost supplier. Consequently, the share of

manufacturing expenditure in n spent on goods produced in i is

$$\pi_{ni} = \frac{B_i^\theta \left(w_{M,i}^\eta P_{M,i}^{1-\eta} d_{ni}\right)^{-\theta}}{\sum_k B_k^\theta \left(w_{M,k}^\eta P_{M,k}^{1-\eta} d_{nk}\right)^{-\theta}}. \tag{16}$$

With a CES elasticity of substitution σ, the corresponding manufacturing price index is

$$P_{M,n} = \bar{\gamma} \left[\sum_i B_i^\theta \left(w_{M,i}^\eta P_{M,i}^{1-\eta} d_{ni}\right)^{-\theta}\right]^{-1/\theta}, \tag{17}$$

where

$$\bar{\gamma} = \left[\Gamma\left(\frac{\theta + 1 - \sigma}{\theta}\right)\right]^{1/(1-\sigma)}. \tag{18}$$

Occupational choice and labor allocation. Workers choose between medical services (S) and manufacturing (M) within their region of residence. Worker ω in region i obtains sector-specific utility

$$V_{ki}(\omega) = \varphi_{ki}(\omega) b_{ki} w_{ki}, \qquad k \in \{S, M\}, \tag{19}$$

where $b_{ki} > 0$ captures nonpecuniary sector attractiveness and $\varphi_{ki}(\omega)$ is an idiosyncratic taste draw. These draws are independent and Fréchet distributed with shape parameter $\epsilon > 0$:

$$\Pr(\varphi_{ki} \leq x) = \exp(-x^{-\epsilon}), \qquad x > 0. \tag{20}$$

The resulting medical employment share is

$$\ell_i \equiv \frac{L_{S,i}}{L_i} = \frac{(b_{S,i} w_{S,i})^\epsilon}{(b_{S,i} w_{S,i})^\epsilon + (b_{M,i} w_{M,i})^\epsilon}, \tag{21}$$

with $L_{M,i}/L_i = 1 - \ell_i$. Thus, changes in patient demand and goods trade can alter medical employment through their effects on relative sectoral wages.

Local insurance financing with mobile patients. Each origin finances reimbursement for its residents regardless of where they receive care. Define average labor income and average treatment cost per patient from origin n as

$$\bar{w}_n \equiv \frac{w_{S,n} L_{S,n} + w_{M,n} L_{M,n}}{L_n}, \qquad \bar{c}_n \equiv \sum_i \lambda_{ni} c_i. \tag{22}$$

Given its reimbursement rate R_n, the origin's insurance budget is

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

where τ_n is the payroll tax and Z_n is deficit financing. Writing $z_n \equiv Z_n / L_n$ and using $Q_{ni} = L_n s_n \lambda_{ni}$ yields

$$s_n R_n \bar{c}_n = \tau_n \bar{w}_n + z_n. \tag{24}$$

We hold z_n fixed at its baseline level in the counterfactuals.

General equilibrium. The preceding choices and prices are linked by goods and medical-service market clearing. Let $X_{M,n}$ denote total expenditure on the manufacturing composite in region n, including household purchases and intermediate demand from both sectors. Household expenditure equals labor income net of payroll taxes and out-of-pocket medical payments. Cobb–Douglas input demand therefore gives

$$X_{M,n} = (1 - \tau_n)\bar{w}_n L_n - \sum_i Q_{ni} p_{ni} + \frac{1 - \upsilon}{\upsilon} w_{S,n} L_{S,n} + \frac{1 - \eta}{\eta} w_{M,n} L_{M,n}. \tag{25}$$

Using the insurance budget Equation 24, this can equivalently be written as

$$X_{M,n} = \bar{w}_n L_n - s_n \bar{c}_n L_n + z_n L_n + \frac{1 - \upsilon}{\upsilon} w_{S,n} L_{S,n} + \frac{1 - \eta}{\eta} w_{M,n} L_{M,n}. \tag{26}$$

Manufacturing labor payments in each region equal the labor share of revenue from all destinations:

$$w_{M,i} L_{M,i} = \eta \sum_n X_{M,n} \pi_{ni}. \tag{27}$$

Likewise, medical labor payments equal the labor share of treatment revenue:

$$w_{S,i} L_{S,i} = \upsilon c_i Q_i. \tag{28}$$

Given the policy vector $\{R_i, \tau_i\}_i$ and the remaining primitives, an equilibrium consists of patient destination shares and flows, medical caseloads and quality levels,[8] sectoral wages and employment, treatment costs, manufacturing trade shares and prices, and medical and manufacturing expenditures. These variables satisfy patient choice Equation 7, caseload accounting Equation 8, medical unit costs Equation 12, manufacturing sourcing and pricing Equation 17, occupational choice Equation 21, insurance budgets Equation 24, and goods and medical-service market clearing Equation 27–28. Section A.1 presents the full equilibrium system.

[8] With reimbursement rates and payroll taxes specified, the insurance budgets constrain average treatment costs. Patient destination choices, competitive medical unit costs, occupational choices, and market clearing then jointly determine quality, costs, and patient flows.

5.2 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

$$\ln W_n \;=\; \ln \frac{y_{M,n}}{P_{M,n}} \;+\; \xi \ln \Psi_n, \tag{29}$$

where

$$y_{M,n} \;\equiv\; (1-\tau_n)\,\bar{w}_n \;-\; s_n\,(1-R_n)\,\bar{c}_n \tag{30}$$

is household expenditure on the manufacturing composite per resident. It equals labor income net of payroll taxes and out-of-pocket medical payments. The inclusive value Ψ_n captures the expected value of the medical destinations available to residents of n.

Welfare responses to tax policy. For any positive variable x, let $\hat{x} \equiv x'/x$ denote its counterfactual change. The exact change in regional welfare is

$$\ln \hat{W}_n = \ln \hat{y}_{M,n} - \ln \hat{P}_{M,n} + \xi \ln \hat{\Psi}_n. \tag{31}$$

We obtain these changes by solving the exact-hat equilibrium system in Section A.2 for a specified vector of tax changes.

To characterize marginal tax incentives, let $\dot{x} \equiv d \ln x$ and evaluate the log-differentiated equilibrium system at a given tax vector τ. Appendix A.3 derives the resulting linear mapping

$$\dot{\boldsymbol{W}} = G(\tau)\dot{\boldsymbol{\tau}}, \tag{32}$$

where the nth element of $\dot{\boldsymbol{W}}$ is $d \ln W_n$. Here,

$$G_{ni}(\tau) \equiv \frac{\partial \ln W_n}{\partial \ln \tau_i} \tag{33}$$

includes the direct effect of region i's tax change and all induced changes in patient flows, medical costs and quality, wages, and goods prices.

Tax coordination. We characterize unilateral and coordinated tax setting through a common policy objective. Region i chooses τ_i, taking τ_{-i} as given, to maximize

$$\mathcal{W}_i(\tau) = \ln W_i(\tau) + \gamma \sum_{n \neq i} \ln W_n(\tau), \qquad \gamma \in [0,1]. \tag{34}$$

The parameter γ governs how much region i internalizes the effects of its tax policy on residents elsewhere. At an interior equilibrium τ^{γ}, its first-order condition is

$$G_{ii}(\tau^{\gamma}) + \gamma \sum_{n \neq i} G_{ni}(\tau^{\gamma}) = 0, \qquad i = 1, \ldots, N. \tag{35}$$

At $\gamma = 0$, each region considers only its own welfare, yielding $G_{ii}(\tau^{\mathrm{u}}) = 0$ for every i. At $\gamma = 1$, all regions share the objective $\sum_n \ln W_n(\tau)$, yielding $\sum_n G_{ni}(\tau^{\mathrm{c}}) = 0$ for every i.

6 Quantitative Results

We calibrate the model to a pre-pandemic equilibrium in China and use the calibrated economy to evaluate counterfactual insurance policies. This section describes the data, calibration, and quantitative results in turn.

6.1 Data and Calibration

We calibrate the model using bilateral patient flows, bilateral goods transactions, and city-level economic and insurance data. The common sample contains $N = 92$ cities with valid information on both patient mobility and goods trade.

Patient flows. We measure patient mobility using origin–destination records from the 2019 hospitalization data. Each record identifies a patient's city of residence and city of treatment. Aggregating the records by city pair gives baseline flows Q_{ni}. We use these flows to construct the patient destination-share matrix $\Lambda = (\lambda_{ni})$, local caseloads Q_{ii}, and inbound caseloads Q_{-i}.

Goods trade. We construct bilateral goods flows from 2017 Taobao transaction records covering the top 100 shipping-origin cities. For each transaction, we identify the shipping-origin city i and delivery-destination city n, then aggregate transactions by city pair. Normalizing shipments from each origin by total platform purchases delivered to destination n gives the baseline trade-share matrix $\Pi = (\pi_{ni})$.[9]

City-level inputs and tax calibration. We obtain city employment L_i and sectoral wages $w_{M,i}$ and $w_{S,i}$ from the 2019 *China City Statistical Yearbook*. Dividing the number of patients residing in city i by L_i gives the patient-incidence rate s_i used in the model.

[9] We combine transactions from Tmall.com, Taobao's business-to-consumer platform, and Taobao.com, its consumer-to-consumer platform. Platform transactions cover only part of domestic trade; using their bilateral shares to calibrate Π assumes that their geographic sourcing pattern is informative about broader intercity trade. For related uses of digital transaction data to study spatial trade patterns, see Lin (2019); Couture et al. (2021); Hu and Xu (2024).

The hospitalization records also report medical charges before discounts and insurance reimbursement. We average these gross charges by treatment city to proxy for destination-specific unit cost c_i. Combining destination costs with patient shares gives the average cost of treating residents of origin i, $\bar{c}_i = \sum_k \lambda_{ik} c_k$.

We construct the origin-specific reimbursement rate R_i from local health-insurance policy documents.[10] Effective city-level payroll tax rates, by contrast, are not directly observed. Statutory contribution rates may differ from the effective burden represented by τ_i because participation, reported wages, and enforcement vary across localities.[11] We therefore infer baseline tax rates from manufacturing market clearing. Define manufacturing-sector income as $Y_i^M \equiv w_{M,i} L_{M,i}$ and medical-sector income as $Y_i^S \equiv \upsilon c_i Q_i$. For a candidate tax vector τ, the market-clearing conditions form a linear system in Y^M:

$$\left[I - \eta \Pi^\top \operatorname{diag}\left(\eta^{-1}\mathbf{1} - \tau\right)\right] Y^M = \eta \Pi^\top \left[\left(\upsilon^{-1}\mathbf{1} - \tau\right) \odot Y^S - s \odot L \odot \bar{c} \odot (\mathbf{1} - R) \right], \quad (36)$$

where $\mathbf{1}$ is a vector of ones and $\odot$ denotes elementwise multiplication. The solution $Y^M(\tau)$ implies average labor income $\bar{w}_i(\tau) = (Y_i^S + Y_i^M(\tau))/L_i$, and an insurance deficit relative to payroll-tax revenue of $\zeta_i(\tau) = s_i R_i \bar{c}_i / (\tau_i \bar{w}_i(\tau)) - 1$. We choose the baseline tax vector to minimize the sum of squared deficit-to-revenue ratios:

$$\tau^0 \in \arg\min \sum_{i=1}^{N} \zeta_i(\tau)^2. \quad (37)$$

Table C2 reports the calibrated baseline tax rates and reimbursement rates for all 92 cities.

Other parameters. We set the labor share in medical production to $\upsilon = 0.41$, calculated as hospital personnel expenses divided by medical revenue using 2020 data reported in the *2021 China Health Statistics Yearbook*. We set the manufacturing labor share to $\eta = 0.36$ using the *2020 China Input–Output Table*. The trade elasticity is $\theta = 4$, following Simonovska and Waugh (2014). We set the Fréchet shape parameter governing occupational choice to $\epsilon = 1.5$, following Galle et al. (2023). Finally, we set the welfare weight on medical access to $\xi = 0.2$ as a benchmark. Source, Mingrui.

[10] When a city's own schedule cannot be recovered, we use an available schedule from the same province. For example, for cities in Shandong without a recoverable local schedule, we use the observed Qingdao schedule as the provincial benchmark.

[11] Effective tax liabilities depend on enforcement and reporting as well as statutory rules (Kleven et al., 2011; Pomeranz, 2015; Best et al., 2015; Slemrod, 2019). For social insurance, incomplete participation and wage underreporting can further separate effective contributions from statutory schedules (Kumler et al., 2020). Insurance arrangements in China have also varied across localities and population groups (Wagstaff et al., 2009; Meng et al., 2015).

6.2 Medical Attractiveness and Model Elasticities

Patient flows discipline the price sensitivity α, the service-adequacy elasticity β, and the scale elasticities α_1 and α_2. We first recover each destination's attractiveness from bilateral patient choices. We then use the medical unit-cost equation to relate that attractiveness to local and nonlocal caseloads, capacity, treatment costs, and input prices.

Let D_{ni}^{prov} indicate whether origin n and destination i are in the same province. We parameterize bilateral access costs as

$$\ln \rho_{ni} = \varphi_0 + \varphi_t \ln t_{ni} + \varphi_p D_{ni}^{\text{prov}} + \nu_{ni}, \tag{38}$$

where t_{ni} is travel time. Because origin-based reimbursement applies equally to all destinations, it cancels from the conditional destination shares. An origin fixed effect absorbs the number of potential patients and the destination-choice denominator. Under the conditional-mean restriction for the bilateral residual, the model motivates

$$\ln \mathbb{E}[Q_{ni} \mid n, i, t_{ni}] = a_n + \chi_i - \varphi_t \ln t_{ni} - \varphi_p D_{ni}^{\text{prov}}, \qquad \chi_i \equiv \ln \delta_i - \alpha \ln c_i + \beta \ln \left(\frac{M_i}{Q_i} \right). \tag{39}$$

We estimate this conditional mean by PPML with origin and destination fixed effects. The estimated destination fixed effects $\hat{\chi}_i$ recover relative revealed medical attractiveness, combining quality, treatment cost, and capacity per patient. Substituting the medical unit-cost Equation 12 into χ_i gives the second-stage relationship

$$\hat{\chi}_i = a + \alpha_1 \ln Q_{ii} + \alpha_2 \ln Q_{-i} + \beta \ln \left(\frac{M_i}{Q_i} \right) + (1 - \alpha) \ln c_i - \upsilon \ln w_{S,i} - (1 - \upsilon) \ln P_{M,i} + u_i. \tag{40}$$

The constant a accommodates the normalization of the first-stage destination effects. The residual u_i contains unobserved medical productivity, $\ln \bar{A}_i$, and estimation error in $\widehat{\chi}_i$. Using the calibrated value of υ, we estimate α_1, α_2, β, and $1 - \alpha$ from Equation 40. Interpreting these estimates as structural elasticities requires the regressors in Equation 40 to be uncorrelated with unobserved destination-specific medical productivity.

Revealed destination attractiveness. Table 2 reports the first-stage PPML estimates of Equation 39. Because the destination fixed effects are identified only relative to one another, we normalize their treatment-volume-weighted mean to zero. Column (1) controls for travel time and whether the two cities are in the same province. Column (2) further allows the effect of travel time to differ within provinces and includes a same-region indicator. As Figure C1 shows, the resulting rankings of destination attractiveness are similar across columns. We use the destination fixed effects from column (2) in the second-stage estimation. Figure 5 maps the spatial distribution of the attractiveness estimates across cities. Guangzhou, Changsha, and Chongqing are among the cities with the highest estimated values. As an external comparison, Figure 6 relate these estimates

to a health-service index and mortality. More attractive destinations tend to have higher health-service indices and lower mortality rates. These associations are consistent with the fixed effects capturing meaningful differences in medical provision, although the fixed effects also incorporate prices and service adequacy, and mortality may reflect differences in patient composition.

Scale and capacity elasticities of destination attractiveness. Table 3 reports the second-stage estimates of Equation 40. Column (1) uses hospital beds per capita as a proxy for service capacity. Local and inbound caseloads both have positive, statistically significant coefficients, but the coefficient on inbound caseloads is significantly larger. Conditional on the other regressors, a 10% increase in inbound patients is associated with approximately 0.043 log points higher destination attractiveness, consistent with a stronger scale contribution from nonlocal patients, where a 10% increase in local patients is associated with approximately 0.028 log points higher destination attractiveness. The capacity coefficient is about 0.5: at a given caseload, a 10% higher bed adequacy is associated with roughly 0.05 log points higher attractiveness.

The coefficient on $\ln c_i$ is also positive and significant. Under the maintained linear quality-cost specification, it estimates $1-\alpha$, implying $\alpha < 1$. This net coefficient combines a direct price component of $-\alpha$, as higher out-of-pocket costs deter patients, with a quality component of $+1$: holding input prices, medical productivity, and caseloads fixed, a higher unit cost corresponds to proportionately higher quality. Replacing hospital beds with beds across all medical institutions leaves the caseload estimates similar, although the capacity coefficient is no longer statistically significant.[12]

6.3 Cooperation Implied by Baseline Tax Rates

We use the calibrated baseline tax vector to assess the degree of cross-city welfare internalization that would rationalize local tax choices. This exercise assumes that each city's baseline tax rate is an interior best response under the policy objective in Equation 34. The cooperative-preference parameter γ enters that objective linearly, so the model's marginal welfare responses can be evaluated at the calibrated baseline without resolving the equilibrium for each candidate value of γ.

At the baseline tax vector τ^0, define $d_{1,i} \equiv G_{ii}(\tau^0)$ as the gradient of city i's log welfare with respect to $\ln \tau_i$, and $d_{2,i} \equiv \sum_{n=1}^{N} G_{ni}(\tau^0)$ as the corresponding gradient of the sum of regional log welfare. Their difference, $d_{2,i} - d_{1,i} = \sum_{n \neq i} G_{ni}(\tau^0)$, measures the welfare spillover to other cities. An interior best response satisfies

$$d_{1,i} + \gamma\big(d_{2,i} - d_{1,i}\big) = 0. \tag{41}$$

[12] Re-estimating Equation 40 with the destination fixed effects from column (1) of Table 2 produces similar results.

We choose the common value of γ that minimizes the sum of squared first-order-condition residuals, yield the solution of γ

$$\hat{\gamma} = -\frac{\sum_i d_{1,i}(d_{2,i} - d_{1,i})}{\sum_i (d_{2,i} - d_{1,i})^2}; \tag{42}$$

Using all 92 calibrated city gradients gives $\hat{\gamma} = 0.541$, with a conventional least-squares standard error of 0.053. The value lies between the unilateral and fully cooperative benchmarks. It describes the degree of spillover internalization that best fits the calibrated baseline under the interior first-order-condition restriction.[13]

6.4 Counterfactual Analysis

We use the calibrated model for three exercises. First, we restore the 2016 road network and compare the model-implied patient flows and treatment costs with observed 2016 outcomes. Second, we assess the welfare contribution of existing public insurance by removing reimbursement and the associated payroll taxes. Finally, we quantify the gains from coordinating city-level tax policies and then pooling insurance budgets nationally, holding reimbursement rules fixed in both reform exercises.

6.4.1 External Validation with the 2016 Road Network

We begin with a retrospective exercise. Starting from the calibrated 2019 equilibrium, we replace the 2019 road network with its 2016 counterpart and use the resulting travel times to construct changes in bilateral patient access costs and goods trade costs. Let $\hat{\rho}_{ni} \equiv \rho_{ni,2016}/\rho_{ni,2019}$ and $\hat{d}_{ni} \equiv d_{ni,2016}/d_{ni,2019}$. The preferred patient-flow specification implies

$$\ln \hat{\rho}_{ni} = \varphi_d \ln\left(\frac{t_{ni,2016}}{t_{ni,2019}}\right), \qquad \ln \hat{d}_{ni} = \kappa_d \ln\left(\frac{t_{ni,2016}}{t_{ni,2019}}\right), \tag{43}$$

where $t_{ni,y}$ denotes travel time in year y. We set the travel-time elasticity of medical access costs to $\varphi_d = 2.049$, according to the estimate in column (2) of Table 2. A comparable gravity regression using 2017 Taobao transactions yields a trade-cost elasticity of $\kappa_d = 0.074$.[14] Starting from the calibrated 2019 equilibrium, we solve for the counterfactual equilibrium under the 2016 road network, holding all other primitives and insurance policies at their 2019 values.

[13] An equality first-order condition applies only to a tax rate strictly inside its feasible policy range. A rate at a bound instead satisfies a one-sided optimality condition.

[14] We estimate the gravity relationship by PPML using the 2017 Taobao data. For each destination n and origin i, the dependent variable is the share of n's purchases sourced from i, divided by n's own-city share. The specification includes origin and destination fixed effects. Standard errors are two-way clustered by origin and destination. The coefficient on log travel time is -0.296 (standard error 0.015; 9,506 bilateral observations; pseudo-$R^2 = 0.78$). Since the model implies an elasticity of trade shares with respect to iceberg trade costs of $-\theta$, this estimate gives a trade-cost elasticity of $\kappa_d = 0.296/\theta = 0.074$ when $\theta = 4$.

Figure 7 compares the model-implied 2016 levels of total destination caseloads, inbound caseloads, and treatment costs with their observed counterparts. The fitted slopes are 0.24 ($t = 2.27$), 0.35 ($t = 3.13$), and 0.17 ($t = 2.29$), respectively. The positive relationships show that the model captures some cross-city variation under the earlier network.

6.4.2 The Welfare Value of Existing Public Insurance

We next compare the calibrated insurance system with an economy without public reimbursement. In the no-insurance counterfactual, we set $R_n = \tau_n = z_n = 0$ for every city. Patients pay the full treatment cost at their chosen destination, while households make no payroll-tax contributions. This comparison captures two immediate changes in household welfare. Removing reimbursement raises out-of-pocket treatment prices, lowering the medical inclusive value at given destination characteristics. Removing the payroll tax raises disposable labor income, while greater out-of-pocket spending reduces the amount available for manufactured consumption. The net effect also incorporates the resulting changes in medical and goods markets. Because R_n applies equally to every destination available to residents of n, its removal does not directly change their relative destination choices; any change in those choices operates through the equilibrium responses of costs, quality-adjusted service adequacy, and other destination characteristics.

Under this benchmark, the no-insurance equilibrium has average consumption-equivalent welfare 11.5% below the insured baseline. The comparison measures the value of the existing reimbursement and financing arrangement conditional on baseline medical quality. It provides a reference point for the reforms below, which retain public insurance and change how its financing is organized.

6.4.3 Welfare Gains from Tax Coordination and National Insurance Pooling

The estimate $\hat{\gamma} = 0.541$ indicates that partial internalization of cross-city welfare effects best fits the calibrated baseline under the interior first-order-condition restriction. We compare this baseline with two increasingly integrated financing arrangements. The first coordinates city-specific taxes while retaining separate local insurance budgets. The second permits transfers across those budgets through a national pool.

Coordinated taxes. Under tax coordination, the vector of city payroll taxes maximizes the sum of regional log welfare, $\sum_n \ln W_n(\tau)$. Reimbursement rates R_n^0 and deficit financing z_n^0 remain fixed. At an interior solution τ^{Coord}, the tax chosen for city i satisfies

$$\sum_{n=1}^{N} G_{ni}\big(\tau^{\text{Coord}}\big) = 0, \qquad i = 1, \ldots, N. \tag{44}$$

A tax rate at the boundary of its feasible range instead satisfies the corresponding one-sided condition. This is a restricted policy optimum over city-specific taxes; patient flows, medical quality and costs, wages, labor allocation, and trade respond in equilibrium.

Relative to the calibrated baseline, coordination raises average city log welfare by 0.150 points, equivalent to a 16.2% consumption-equivalent gain. Eighty-three of the 92 cities gain. Weighting cities by employment yields an average gain of 0.102 log points, or approximately 10.8%. Panel A of Table 4 applies the decomposition in (31): higher real consumption of manufactured goods accounts for roughly 90% of the average gain, while the medical inclusive value accounts for the remaining 10%.

Coordination raises tax rates in 24 cities and lowers them in 68. Figure 8 shows the associated changes in caseloads and residents' medical inclusive values. Caseload growth is concentrated in smaller destinations near established hubs and provincial capitals. Improvements in the medical inclusive value extend beyond cities whose own medical sectors expand, since residents can benefit from changes in the costs, quality, and service adequacy of providers elsewhere.

Table 5 makes the production-side reallocation more explicit. Cities in the smallest quartile of baseline caseloads experience the largest increases in total caseloads, inbound flows, quality, and medical employment. Quality and employment decline in the two largest-caseload quartiles, even where caseloads change little. With population and treatment incidence fixed, the national number of patients does not change; coordination reallocates where they receive care and how many medical resources are devoted to treating them.

Table 6 relates each city's tax adjustment, $\Delta\tau_i \equiv \tau_i^{\text{Coord}} - \tau_i^0$, to its equilibrium outcomes. Tax increases are positively correlated with growth in total caseloads, inbound patients, medical quality, and the medical-sector employment share. These correlations describe the joint spatial adjustment of financing and medical activity under coordination.

National insurance pooling. Coordinated taxes leave each origin responsible for financing its residents' reimbursement. To assess the additional value of sharing insurance resources, we introduce transfers across city funds while holding tax rates at τ^{Coord}, reimbursement rates at R^0, and deficit financing at z^0. Let t_n be the per-resident net transfer received by city n. The local accounting conditions become

$$s_n R_n^0 \bar{c}_n = \tau_n^{\text{Coord}} \bar{w}_n + z_n^0 + t_n, \qquad \sum_{n=1}^{N} L_n t_n = 0. \tag{45}$$

Thus, $t_n > 0$ identifies a net recipient and $t_n < 0$ a net contributor. The adding-up condition ensures that the pool reallocates resources across funds without changing their aggregate financing. We choose transfers to maximize $\sum_n \ln W_n$ subject to these

conditions and the equilibrium system. The allocation with $t = 0$ remains feasible, so the resulting welfare difference isolates the additional gain available from pooling at the coordinated tax vector.

Pooling raises average city log welfare by another 0.028 points, or approximately 2.9%, relative to coordination alone. The combined gain relative to the calibrated baseline is 0.178 log points, equivalent to 19.5%. Seventy-two cities gain from pooling beyond tax coordination. Panel B of Table 4 shows that the medical inclusive value accounts for 58% of the incremental gain, a larger share than under tax coordination. Fifteen cities are net contributors, while 77 are net recipients. Wuhan, Guangzhou, Chongqing, Foshan, and Nanjing together provide 81% of gross transfers; the five largest recipients receive 21%. Figure 9 shows this concentration of contributions among a few large cities and the wider distribution of receipts.

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 (2026, July). Market size and trade in medical services. NBER Working Paper 31030, National Bureau of Economic Research. Revised July 2026.

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.

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.

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

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. 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.

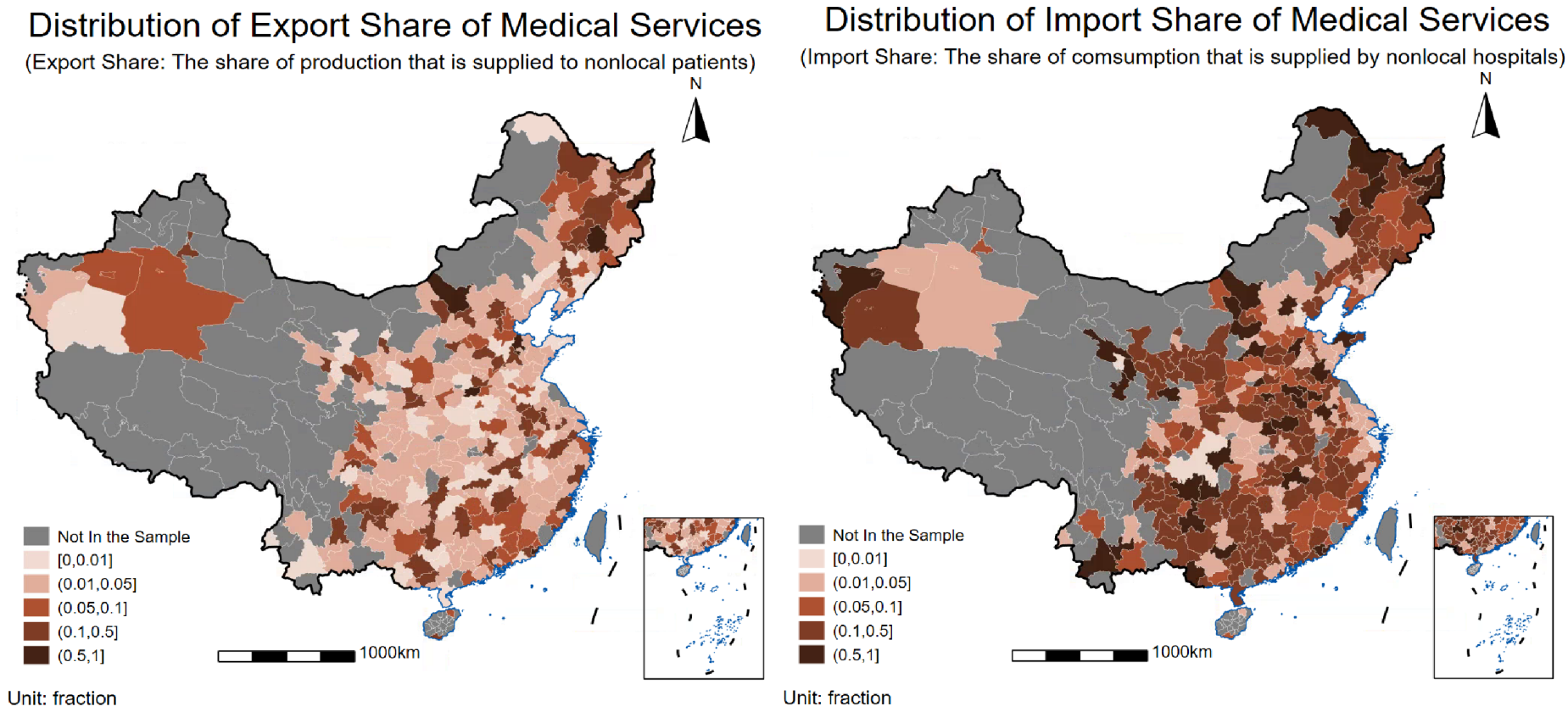
Tay, A. (2003). Assessing competition in hospital care markets: The importance of accounting for quality differentiation. The RAND Journal of Economics 34(4), 786–814.

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.

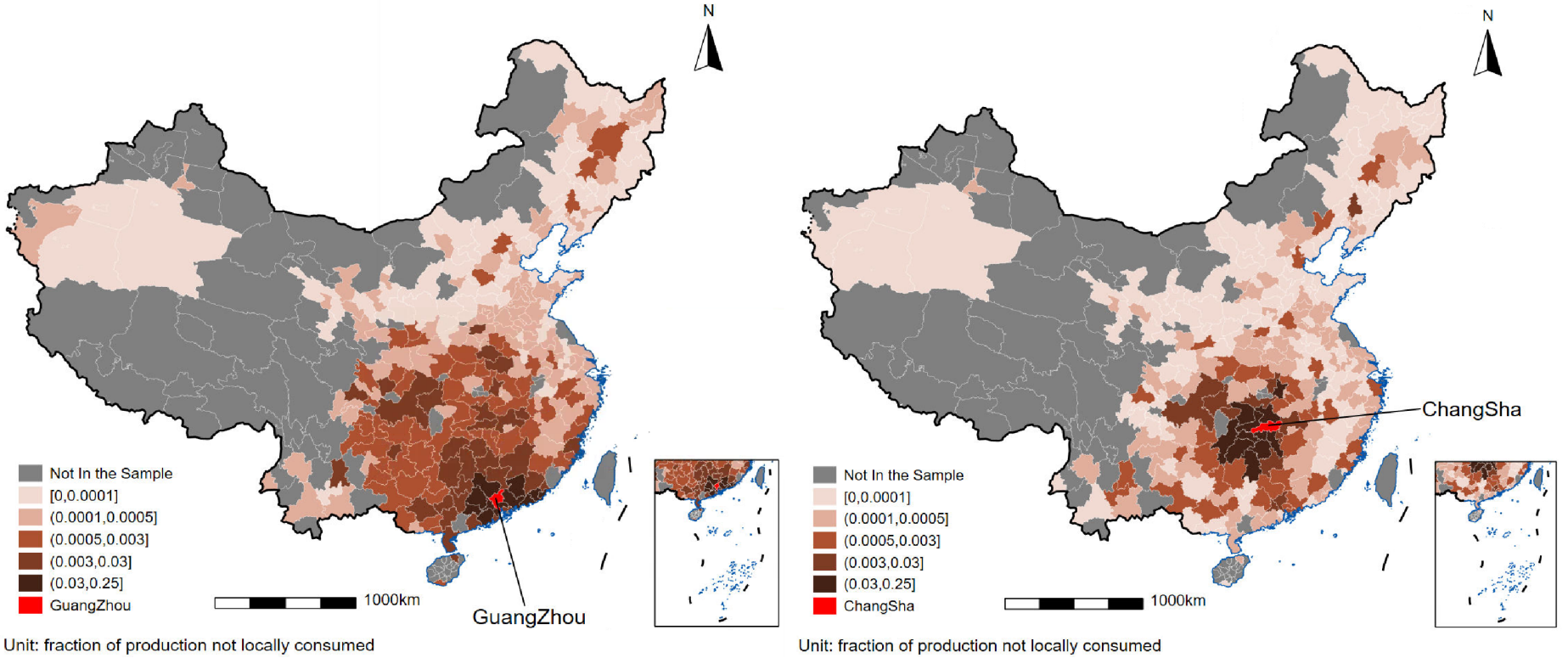
Figures & Tables

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



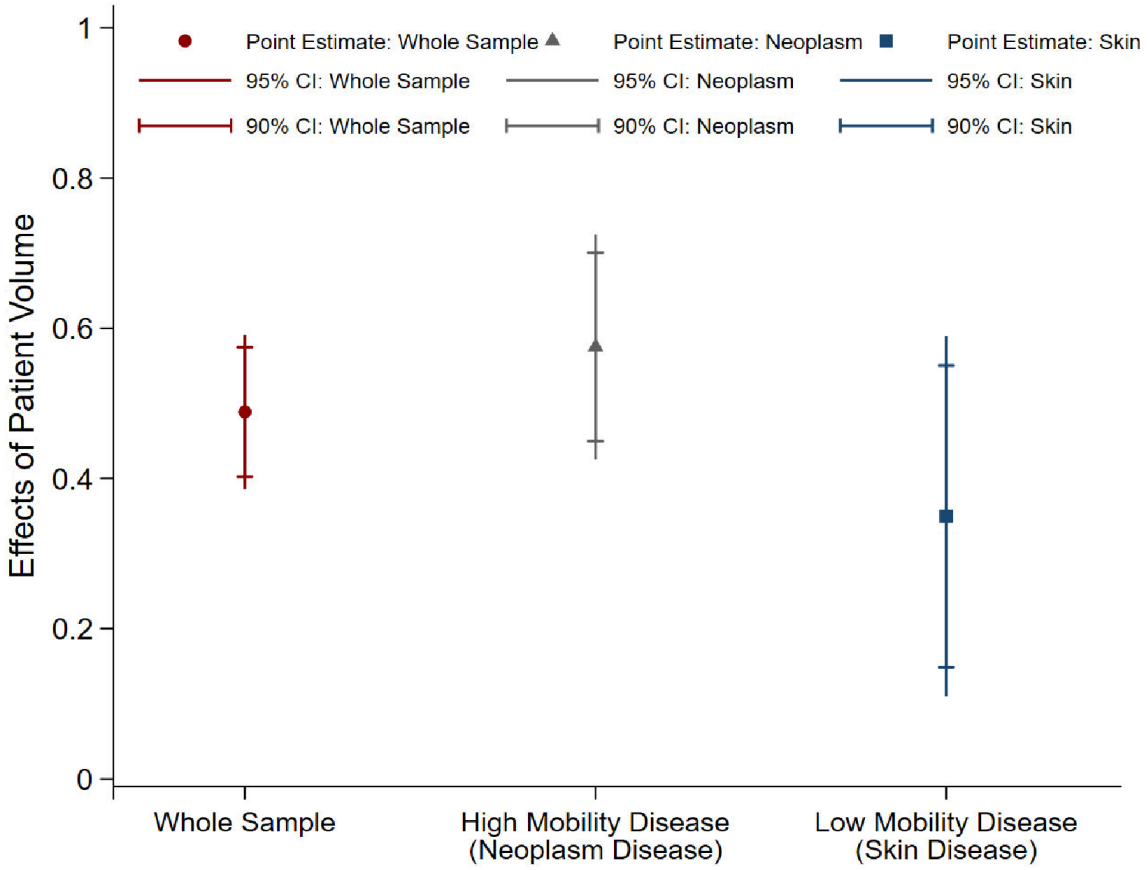
Notes: These figures present the export share and import share of medical services. Darker shading means higher share.

Figure 2: Destinations of medical services



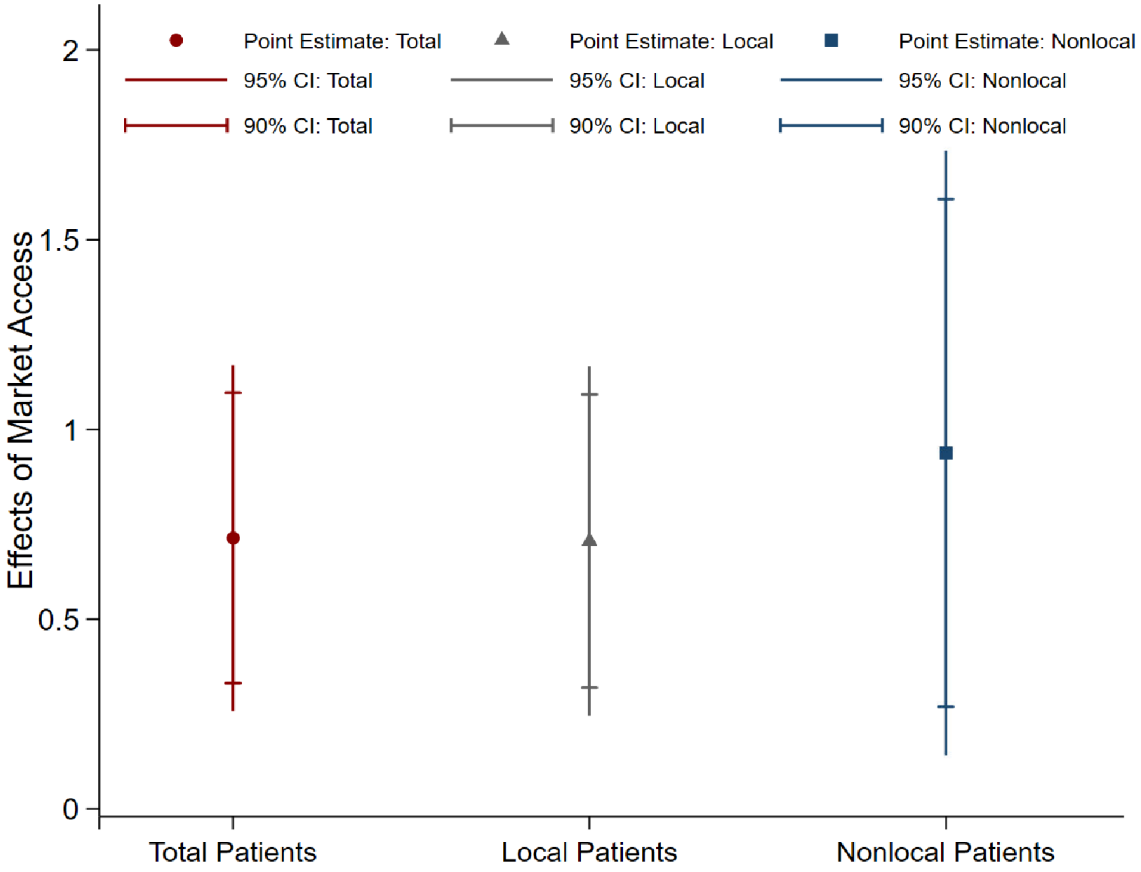
Notes: These figures show the destinations of medical services produced in two representative cities: GuangZhou and ChangSha. Darker shading indicates higher consumption of services originating from these two cities in each region.

Figure 3: Heterogeneous scale effects by disease mobility



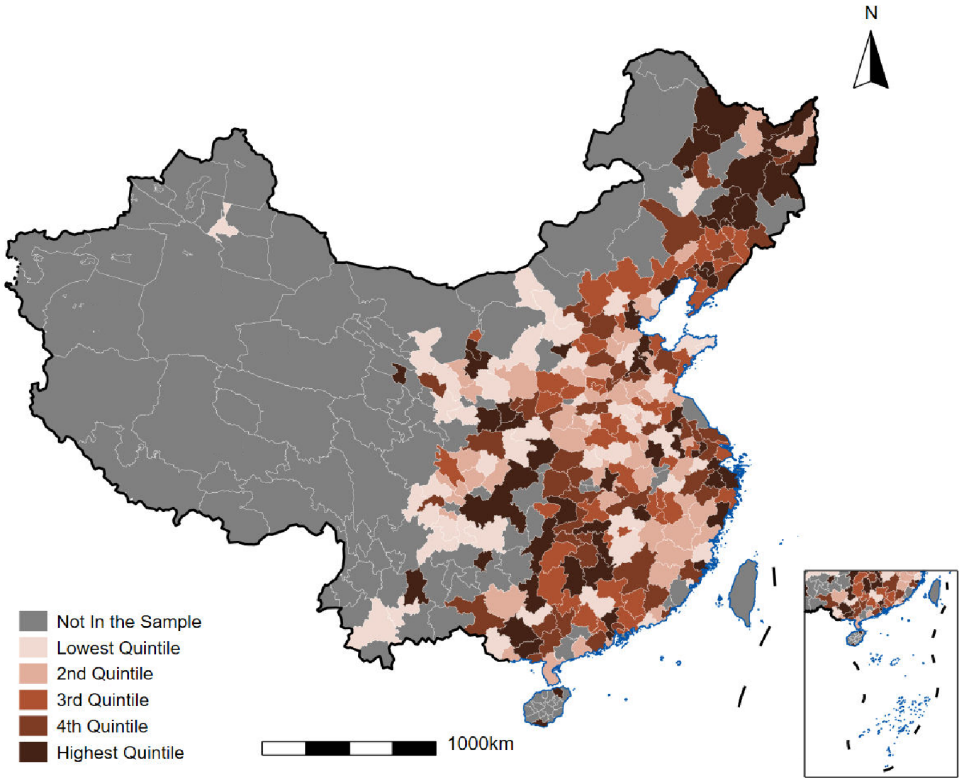
Notes: This figure reports the estimated effects of patient volume on productivity for the full sample and for two representative disease categories with different levels of cross-regional patient mobility. Neoplasms are selected as a high-mobility disease category, whereas skin diseases are selected 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



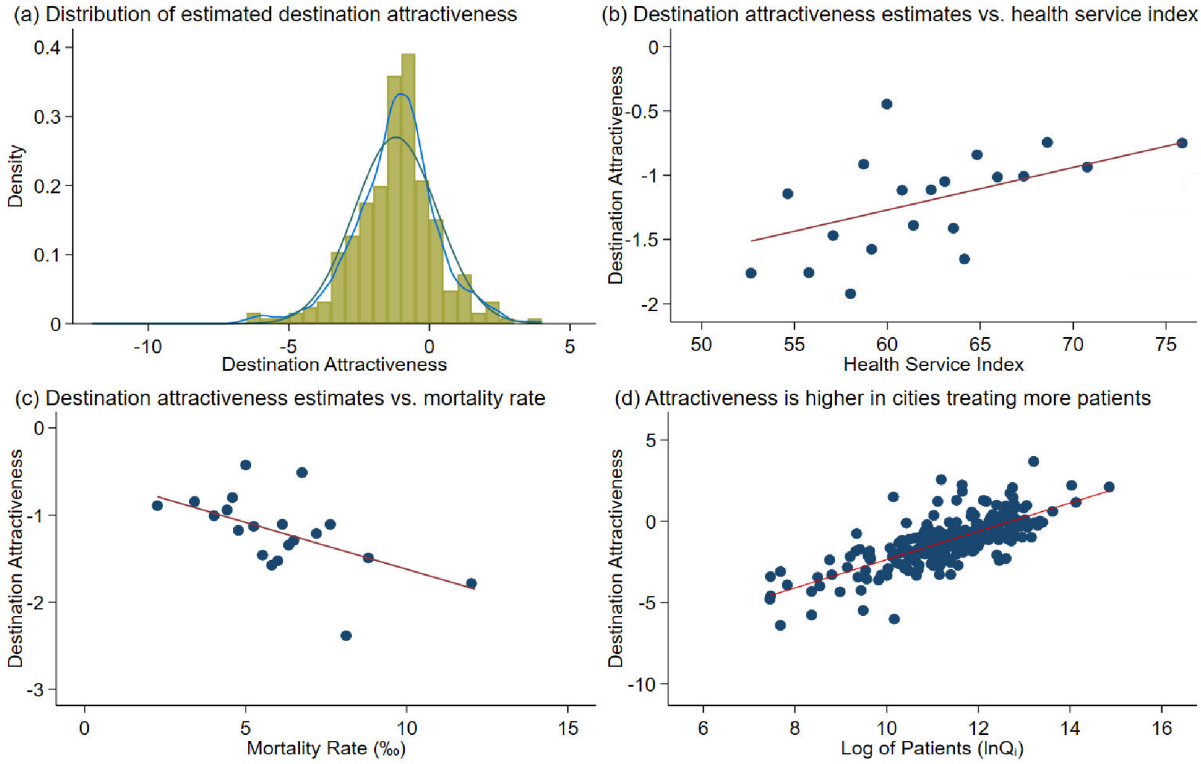
Notes: This figure reports the estimated effects of market access on total, local, and nonlocal patient volumes. Local patients are those residing in the destination city, whereas nonlocal patients are those traveling from other cities for treatment. The markers denote point estimates. The longer vertical lines indicate 95% confidence intervals, while the shorter capped lines indicate 90% confidence intervals. The detailed results are reported in Table C3.

Figure 5: Distribution of city-level attractiveness estimates



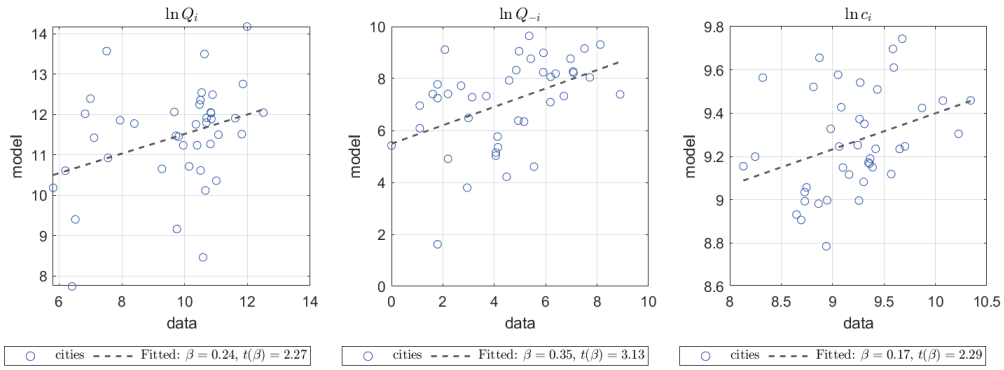
Notes: This figure presents the geographic distribution of attractiveness estimates in column (2) of Table 2. The estimates are ranked and divided into five approximately equal-sized groups, with each group containing about one-fifth of the sample cities. Darker shading indicates higher estimated attractiveness.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

Figure 6: Validation of attractiveness estimates



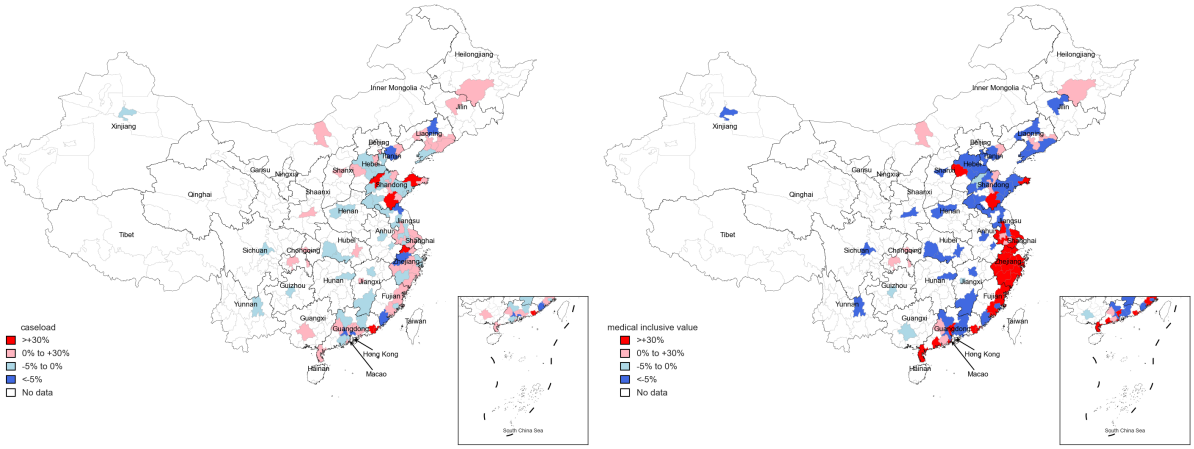
Notes: Panel A displays the distribution of the attractiveness estimates. Panels B and C, using the binscatter approach (Cattaneo et al., 2024), illustrate the relationships between attractiveness and health service index and between attractiveness and the mortality rate, respectively. Panel D shows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

Figure 7: 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.

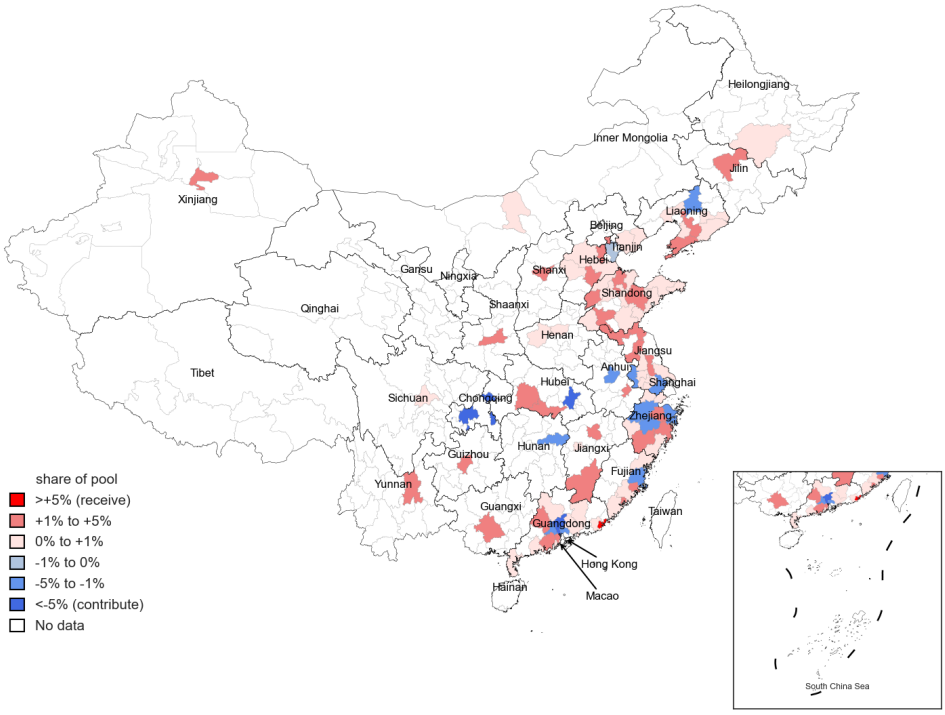
Figure 8: National Tax Coordination: Spatial Pattern of Medical-Side Responses



(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 9: 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 Theoretical Appendix

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 Equation A.29–A.30 to compute $\hat{L}_S, \hat{L}_M$ from the guessed wages. Then use Equation 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 Equation A.28.
- Step 3 (Outer loop): Verify if equations Equation A.26 and Equation 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.

A.3 Welfare Responses to Tax Policy

This appendix derives the welfare response to marginal changes in local payroll taxes. Section A.2 uses the exact-hat system to compute discrete counterfactuals. Here, we log-differentiate the equilibrium conditions at a policy vector τ, holding reimbursement rates, per-capita deficit financing, and all other primitives fixed. We denote the resulting equilibrium by $E(\tau)$. Unless indicated otherwise, all shares and coefficients below are evaluated at this equilibrium.

Exact welfare changes. Let superscript 0 denote the baseline and primes denote a counterfactual. Define the baseline household-expenditure shares

$$y^0_{1,n} \equiv \frac{(1-\tau^0_n)\bar{w}^0_n}{y^0_{M,n}}, \qquad y^0_{2,n} \equiv \frac{s_n(1-R^0_n)\bar{c}^0_n}{y^0_{M,n}}, \qquad y^0_{1,n} - y^0_{2,n} = 1. \tag{A.32}$$

The exact change in household manufacturing expenditure is

$$\hat{y}_{M,n} = y^0_{1,n}\frac{1-\tau'_n}{1-\tau^0_n}\hat{\bar{w}}_n - y^0_{2,n}\frac{1-R'_n}{1-R^0_n}\hat{\bar{c}}_n. \tag{A.33}$$

For completeness, the corresponding exact change in the medical inclusive value is

$$\hat{\Psi}_n = \left(\frac{1-R_n'}{1-R_n^0}\right)^{-\alpha} \sum_i \lambda_{ni}^0 \hat{c}_i^{-\alpha} \hat{\delta}_i \hat{M}_i^{\beta} \hat{Q}_i^{-\beta} \hat{\rho}_{ni}^{-1}. \tag{A.34}$$

In a tax-only counterfactual, R_n, M_i, and ρ_{ni} remain fixed, so their hat terms equal one. Substituting Equation A.33 and Equation A.34 into Equation 31 gives the exact welfare change.

Marginal welfare. For a positive variable x, write $\dot{x} \equiv d \ln x$. At $E(\tau)$, define

$$y_{1,n} \equiv \frac{(1-\tau_n)\bar{w}_n}{y_{M,n}}, \qquad y_{2,n} \equiv \frac{s_n(1-R_n)\bar{c}_n}{y_{M,n}}, \qquad y_{1,n} - y_{2,n} = 1. \tag{A.35}$$

Log-differentiating the household budget Equation 30 yields

$$\dot{y}_{M,n} = y_{1,n}\dot{\bar{w}}_n - y_{2,n}\dot{\bar{c}}_n - y_{1,n}\frac{\tau_n}{1-\tau_n}\dot{\tau}_n + y_{2,n}\frac{R_n}{1-R_n}\dot{R}_n. \tag{A.36}$$

With reimbursement held fixed, $\dot{R}_n = 0$, so the welfare differential is

$$\dot{W}_n = y_{1,n}\dot{\bar{w}}_n - y_{2,n}\dot{\bar{c}}_n - y_{1,n}\frac{\tau_n}{1-\tau_n}\dot{\tau}_n - \dot{P}_{M,n} + \xi\dot{\Psi}_n, \tag{A.37}$$

where $\dot{W}_n \equiv d \ln W_n$. To evaluate this expression at a candidate policy vector τ, we first solve the exact-hat system for $E(\tau)$ and then solve the following log-differentiated equilibrium system around that equilibrium.

Manufacturing prices and trade shares. Let

$$g_{M,ni} \equiv \eta\dot{w}_{M,i} + (1-\eta)\dot{P}_{M,i} + \dot{d}_{ni}. \tag{A.38}$$

Log-differentiating the Eaton–Kortum trade shares and price indices gives

$$\dot{\pi}_{ni} = -\theta\left[g_{M,ni} - \sum_k \pi_{nk} g_{M,nk}\right], \tag{A.39}$$

$$\dot{P}_{M,n} = \sum_k \pi_{nk} g_{M,nk}. \tag{A.40}$$

Patient choices and caseloads. Define the change in the destination-specific component of patient utility as

$$g_{S,ni} \equiv -\alpha\dot{c}_i + \dot{\delta}_i + \beta\dot{M}_i - \beta\dot{Q}_i - \dot{\rho}_{ni}. \tag{A.41}$$

Because reimbursement is common across destinations for a given origin and is held fixed here, patient choice and the medical inclusive value imply

$$\dot{\lambda}_{ni} = g_{S,ni} - \sum_k \lambda_{nk} g_{S,nk}, \tag{A.42}$$

$$\dot{\Psi}_n = \sum_k \lambda_{nk} g_{S,nk}. \tag{A.43}$$

With population and sickness rates fixed, $\dot{Q}_{ni} = \dot{\lambda}_{ni}$. Let $\mu_{ni} \equiv Q_{ni}/Q_i$ denote the share of destination i's patients originating in n. Caseload accounting then gives

$$\dot{Q}_{ii} = \dot{\lambda}_{ii}, \tag{A.44}$$

$$\dot{Q}_i = \sum_n \mu_{ni} \dot{\lambda}_{ni}, \tag{A.45}$$

$$Q_{ii}\dot{Q}_{ii} + Q_{-i}\dot{Q}_{-i} = Q_i \dot{Q}_i. \tag{A.46}$$

Medical costs and insurance budgets. The medical unit-cost equation implies

$$\dot{c}_i = \upsilon \dot{w}_{S,i} + (1-\upsilon)\dot{P}_{M,i} + \dot{\delta}_i - \alpha_1 \dot{Q}_{ii} - \alpha_2 \dot{Q}_{-i}. \tag{A.47}$$

Define $\omega_{ni} \equiv \lambda_{ni} c_i / \bar{c}_n$, the share of origin n's treatment expenditure incurred in destination i. Average treatment cost changes according to

$$\dot{\bar{c}}_n = \sum_i \omega_{ni} \big(\dot{c}_i + \dot{\lambda}_{ni}\big). \tag{A.48}$$

Let $f_{S,n} \equiv w_{S,n} L_{S,n}/(\bar{w}_n L_n)$ and $f_{M,n} \equiv w_{M,n} L_{M,n}/(\bar{w}_n L_n) = 1 - f_{S,n}$. The change in average labor income is

$$\dot{\bar{w}}_n = f_{S,n}\big(\dot{w}_{S,n} + \dot{L}_{S,n}\big) + f_{M,n}\big(\dot{w}_{M,n} + \dot{L}_{M,n}\big). \tag{A.49}$$

Since R_n, z_n, and s_n are fixed, log-differentiating the origin insurance budget yields

$$\dot{\bar{c}}_n = \frac{\tau_n \bar{w}_n}{z_n + \tau_n \bar{w}_n}\big(\dot{\tau}_n + \dot{\bar{w}}_n\big). \tag{A.50}$$

The tax change is specified by the policy experiment; Equation A.50 constrains the resulting equilibrium response of average treatment costs.

Labor allocation and market clearing. Occupational choice and fixed regional population imply

$$\dot{L}_{S,i} = \epsilon(1-\ell_i)\big(\dot{w}_{S,i} - \dot{w}_{M,i}\big), \tag{A.51}$$

$$L_{S,i}\dot{L}_{S,i} + L_{M,i}\dot{L}_{M,i} = 0. \tag{A.52}$$

Medical-service market clearing gives

$$\dot{w}_{S,i} + \dot{L}_{S,i} = \dot{c}_i + \dot{Q}_i. \tag{A.53}$$

For manufacturing expenditure, define the component shares

$$\begin{aligned} a_{1,n} &\equiv \frac{\bar{w}_n L_n}{X_{M,n}}, & a_{2,n} &\equiv \frac{s_n L_n \bar{c}_n}{X_{M,n}}, & a_{3,n} &\equiv \frac{z_n L_n}{X_{M,n}}, \\ a_{4,n} &\equiv \frac{(1-\upsilon) w_{S,n} L_{S,n}}{\upsilon X_{M,n}}, & a_{5,n} &\equiv \frac{(1-\eta) w_{M,n} L_{M,n}}{\eta X_{M,n}}. \end{aligned} \tag{A.54}$$

They satisfy $a_{1,n} - a_{2,n} + a_{3,n} + a_{4,n} + a_{5,n} = 1$. With z_n fixed,

$$\begin{aligned} \dot{X}_{M,n} = a_{1,n}\dot{\bar{w}}_n - a_{2,n}\dot{\bar{c}}_n \\ + a_{4,n}\big(\dot{w}_{S,n} + \dot{L}_{S,n}\big) + a_{5,n}\big(\dot{w}_{M,n} + \dot{L}_{M,n}\big). \end{aligned} \tag{A.55}$$

Finally, define the share of region i's manufacturing revenue earned from destination n as

$$\phi_{ni} \equiv \frac{X_{M,n}\pi_{ni}}{\sum_k X_{M,k}\pi_{ki}}. \tag{A.56}$$

Manufacturing market clearing requires

$$\dot{w}_{M,i} + \dot{L}_{M,i} = \sum_n \phi_{ni}\big(\dot{X}_{M,n} + \dot{\pi}_{ni}\big). \tag{A.57}$$

The welfare-response matrix. In the tax experiment, $\dot{M} = \dot{\rho} = \dot{d} = 0$; only the vector $\dot{\boldsymbol{\tau}}$ is exogenous. Stack the endogenous differentials in $\dot{\boldsymbol{x}}$. Equation A.39–A.57, together with Equation A.49, can be written as

$$A(\tau)\dot{\boldsymbol{x}} = B_\tau(\tau)\dot{\boldsymbol{\tau}}. \tag{A.58}$$

Here $A(\tau)$ and $B_\tau(\tau)$ collect the coefficients of the log-differentiated equilibrium conditions at $E(\tau)$. Using an independent set of equilibrium equations and the model's price normalization, the system determines the local response of endogenous variables whenever the resulting Jacobian is nonsingular.

Write the welfare differential Equation A.37 as

$$\dot{\boldsymbol{W}} = C_x(\tau)\dot{\boldsymbol{x}} + D_\tau(\tau)\dot{\boldsymbol{\tau}}, \qquad D_\tau(\tau) = -\operatorname{diag}\left(y_{1,n}\frac{\tau_n}{1-\tau_n}\right)_{n=1}^{N}. \tag{A.59}$$

Combining this expression with Equation A.58 gives

$$\dot{\boldsymbol{W}} = \underbrace{\left[C_x(\tau)A(\tau)^{-1}B_\tau(\tau) + D_\tau(\tau)\right]}_{G(\tau)}\dot{\boldsymbol{\tau}}. \tag{A.60}$$

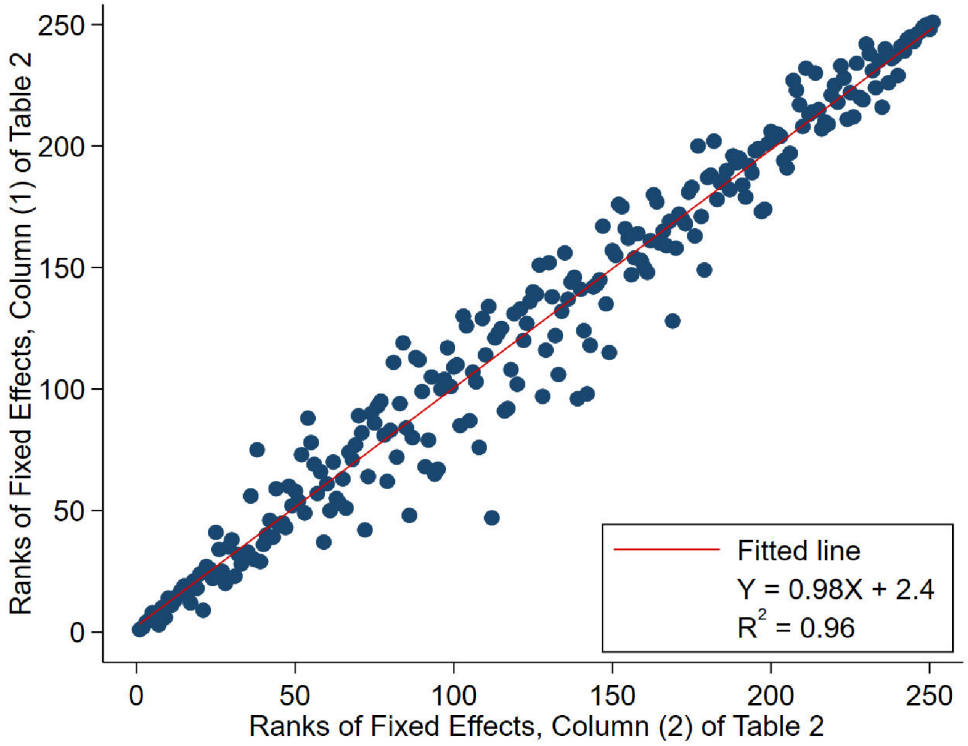
For a change in region i's tax alone, set $\dot{\boldsymbol{\tau}} = e_i\, d\ln\tau_i$. Hence,

$$\frac{\partial \ln \boldsymbol{W}}{\partial \ln \tau_i} = G(\tau)e_i. \tag{A.61}$$

The own-region effect is $e_i^\top G(\tau)e_i$; the effect on the equally weighted cooperative objective is $\mathbf{1}^\top G(\tau)e_i$.

B Appendix Figures & Tables

Figure C1: 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.

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: Baseline Payroll Tax and Reimbursement Rates by City

City	τ_n^0	R_n^0	City	τ_n^0	R_n^0	City	τ_n^0	R_n^0
Tianjin	0.230	0.775	Yingkou	0.300	0.868	Wenzhou	0.042	0.822
Shijiazhuang	0.066	0.816	Changchun	0.300	0.820	Jiaxing	0.070	0.822
Tangshan	0.142	0.816	Harbin	0.300	0.819	Huzhou	0.030	0.822
Qinhuangdao	0.192	0.816	Nanjing	0.082	0.897	Shaoxing	0.128	0.822
Baoding	0.230	0.816	Wuxi	0.065	0.897	Jinhua	0.086	0.822
Cangzhou	0.215	0.816	Xuzhou	0.085	0.897	Quzhou	0.118	0.822
Langfang	0.300	0.816	Changzhou	0.062	0.897	Taizhou (ZJ)	0.072	0.822
Hengshui	0.256	0.816	Suzhou	0.083	0.897	Lishui	0.113	0.822
Taiyuan	0.300	0.811	Nantong	0.066	0.897	Hefei	0.300	0.899
Baotou	0.300	0.893	Lianyungang	0.227	0.897	Wuhu	0.191	0.899
Shenyang	0.216	0.868	Huai'an	0.300	0.897	Fuzhou	0.060	0.814
Dalian	0.126	0.868	Yangzhou	0.071	0.897	Xiamen	0.052	0.814
Anshan	0.300	0.868	Zhenjiang	0.081	0.897	Putian	0.300	0.814
Benxi	0.300	0.868	Taizhou (JS)	0.300	0.897	Quanzhou	0.077	0.814
Dandong	0.300	0.868	Hangzhou	0.044	0.822	Zhangzhou	0.111	0.814
Jinzhou	0.300	0.868	Ningbo	0.073	0.822	Ningde	0.101	0.814
Nanchang	0.300	0.855	Binzhou	0.185	0.825	Huizhou	0.063	0.849
Xinyu	0.300	0.855	Heze	0.300	0.825	Heyuan	0.300	0.849
Ganzhou	0.300	0.855	Zhengzhou	0.069	0.825	Yangjiang	0.098	0.849
Jinan	0.147	0.825	Luoyang	0.188	0.825	Qingyuan	0.300	0.849
Qingdao	0.183	0.825	Wuhan	0.285	0.848	Dongguan	0.071	0.849
Zibo	0.064	0.825	Yichang	0.300	0.848	Zhongshan	0.166	0.849
Dongying	0.300	0.825	Jingzhou	0.300	0.848	Jieyang	0.045	0.849
Yantai	0.175	0.825	Changsha	0.300	0.822	Nanning	0.300	0.763
Weifang	0.248	0.825	Guangzhou	0.300	0.849	Chongqing	0.300	0.792
Jining	0.300	0.825	Zhuhai	0.060	0.849	Chengdu	0.098	0.877
Weihai	0.031	0.825	Shantou	0.042	0.849	Guiyang	0.300	0.770
Rizhao	0.300	0.825	Foshan	0.079	0.849	Kunming	0.300	0.733
Linyi	0.034	0.825	Jiangmen	0.300	0.849	Xi'an	0.300	0.822
Dezhou	0.142	0.825	Zhanjiang	0.147	0.849	Urumqi	0.300	0.833
Liaocheng	0.300	0.825	Zhaoqing	0.300	0.849			

Notes: τ_n^0 is the calibrated baseline payroll tax rate, and R_n^0 is the baseline reimbursement rate. Both are reported as fractions. JS and ZJ distinguish Taizhou in Jiangsu and Zhejiang, respectively.

Table C3: 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 C4: 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.