Vital Links, Divided Gains

Jing Li[a*], Lin Ma[a†], Xiao Lin Ong[b‡], Wei Yan[c§], Junjian Yi[d¶]

[a] *Singapore Management University, Singapore*

[b] *University of Rochester, U.S.*

[c] *Renmin University, China*

[d] *Peking University, China*

January 15, 2026

Abstract

We study the impact of transportation infrastructure on healthcare access and health outcomes. Using administrative data on over 600,000 hospitalizations for cerebral-cardiovascular diseases (CCVD) in Sichuan, China, we show that patients travel from low-medical-resource to high-medical-resource cities, but travel time imposes substantial barriers, especially for low-income patients. We develop and structurally estimate a dynamic spatial model in which individuals choose treatment locations by weighing the expected effectiveness of care against travel and financial costs. Counterfactual simulations indicate that, holding medical resources fixed at the 2010 level, improvements in the national transport network between 2010 and 2018 would reduce CCVD mortality by approximately 10,000 cases per year. While geographic disparities in health outcomes narrowed, gains accrued disproportionately to high-income patients.

Keywords: Transport Infrastructure, Travel Time, Healthcare Access, Health Outcomes, Mortality, CCVD.

JEL classifications: R4, I12, I14.

*Address: 90 Stamford Road, Singapore 178903. Phone: +65-6808-5454. E-mail: lijing@smu.edu.sg.

†Address: 90 Stamford Road, Singapore 178903. Phone: +65-6828-0876. E-mail: linma@smu.edu.sg.

‡Address: 280 Hutchison Road, Rochester, NY 14627. Phone: +1-5856152736. E-mail: xong@ur.rochester.edu.

§Address: 59 Zhongguancun Street, Beijing 100872, China. Phone: +86-15652990056. E-mail: yanwivy@gmail.com.

¶Address: 5 Yiheyuan Road, Haidian District, Beijing, 100871, China. Phone: +86-13802763753. E-mail: junjian.yi@gmail.com.

1 Introduction

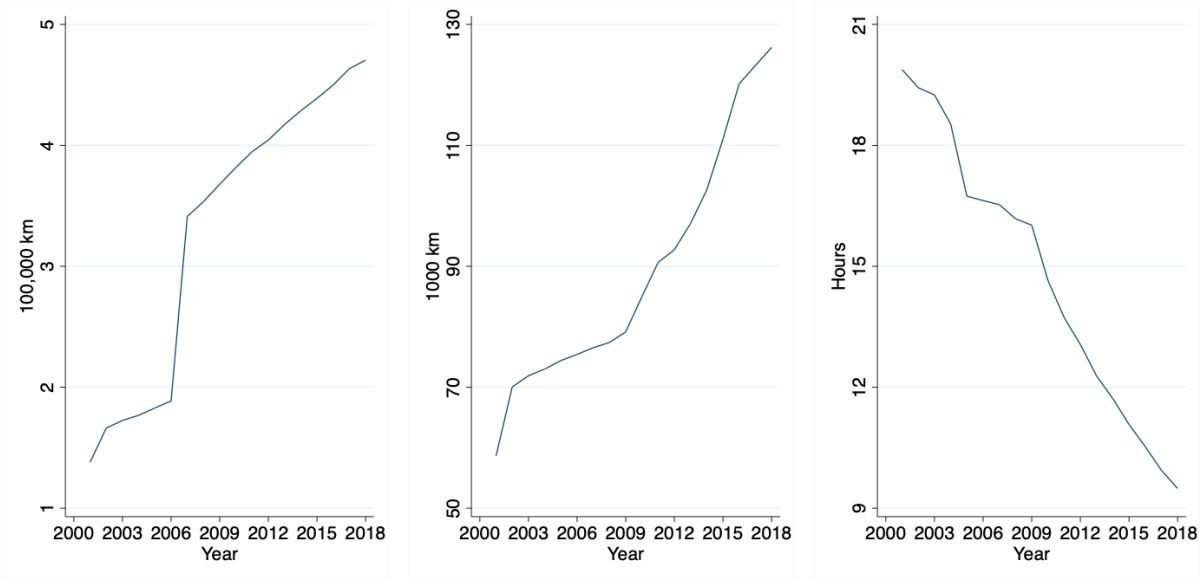
In recent decades, many countries around the world have dramatically expanded their transportation infrastructure (Fay et al., 2019). New highways, upgraded arterial roads, and modern rail systems have extended the reach of national networks, shortened travel times, and tightened economic integration across regions. These investments are typically motivated by their promise to unlock growth: by lowering trade and migration costs, transport improvements can expand market access, facilitate labor mobility, and accelerate the diffusion of ideas (Faber, 2014; Donaldson and Hornbeck, 2016; Donaldson, 2018; Banerjee et al., 2020; Andersson et al., 2023). Yet, transport networks do more than connect product and factor markets. They also serve as critical conduits for accessing essential non-tradable services. One particularly important but underexplored dimension is healthcare, where transport systems operate as vital links — timely access to high-quality medical resources can be the difference between life and death (Dingel et al., 2023).

In this paper, we study the impact of transportation infrastructure on health outcomes through its role in improving access to healthcare services. We study this question in the context of China for the following two main reasons. First, in recent decades, China has witnessed the most rapid expansion of transportation infrastructure globally (Egger et al., 2023). As illustrated in Panels (a) and (b) of Figure 1, between 2001 and 2018, the country's road network more than tripled in length, growing from around 1.4 million kilometers to approximately 4.7 million kilometers, while its total railway network expanded by about 2.2 times, from 59 thousand kilometers to 126 thousand kilometers. This substantial growth in infrastructure has significantly enhanced regional connectivity across the country. Specifically, Panel (c) shows that during the same period, the average pairwise travel time across all prefecture-level city pairs declined by approximately 52%, from 19.9 to 9.5 hours.[1]

Second, China has substantial spatial disparities in the distribution of healthcare resources and, consequently, uneven access to care. Figure 2 illustrates the distribution of

[1]Throughout the paper, we refer to a prefecture-level city simply as a "city."

tertiary hospital beds across cities in 2018.[2] As shown, more developed eastern provinces had significantly greater numbers of tertiary hospital beds compared to their less developed western counterparts. When adjusted for population, the availability of tertiary hospital beds per 10,000 population still varies from fewer than ten to more than two hundred.[3] Given the substantial spatial disparity in high-quality healthcare services, it is important to understand the extent to which the development of the transport network may facilitate access to care across locations, which further shapes both the overall level and distribution of health outcomes in China.



(a) Road Network (b) Railway Network (c) City-Pair Travel Time

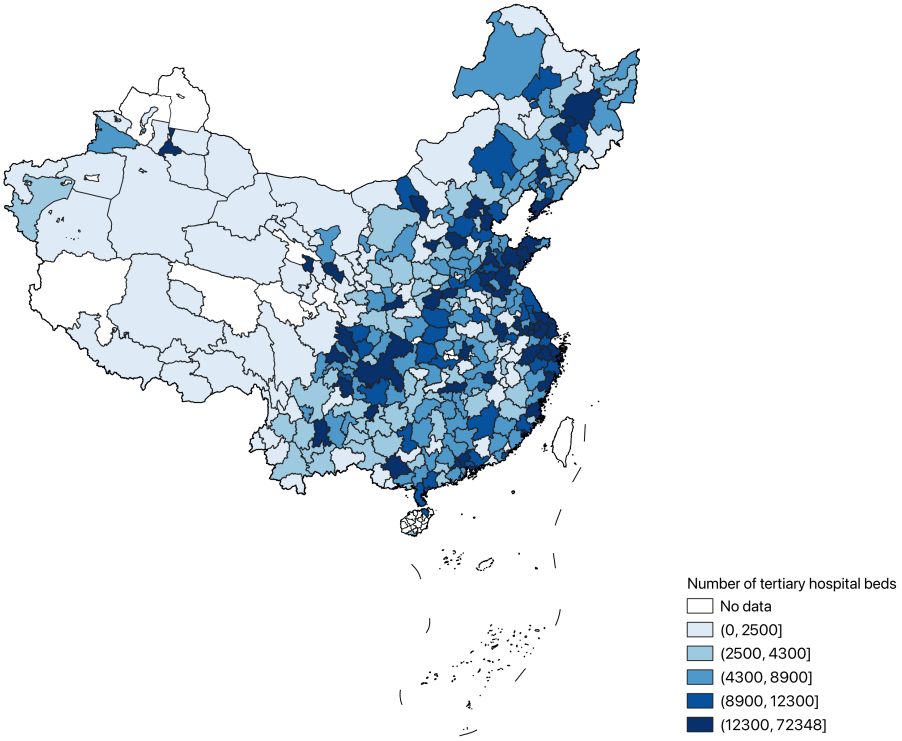
Figure 1: Transport Infrastructure and Travel Time in China (2001-2018)

Notes: Panels (a) and (b) plot the total lengths of road and railway networks, respectively, from 2001 to 2018 in China. Data Source: China Statistical Yearbook. Panel (c) plots the average travel time across all city pairs over the same period. Data source: Ma and Tang (2024).

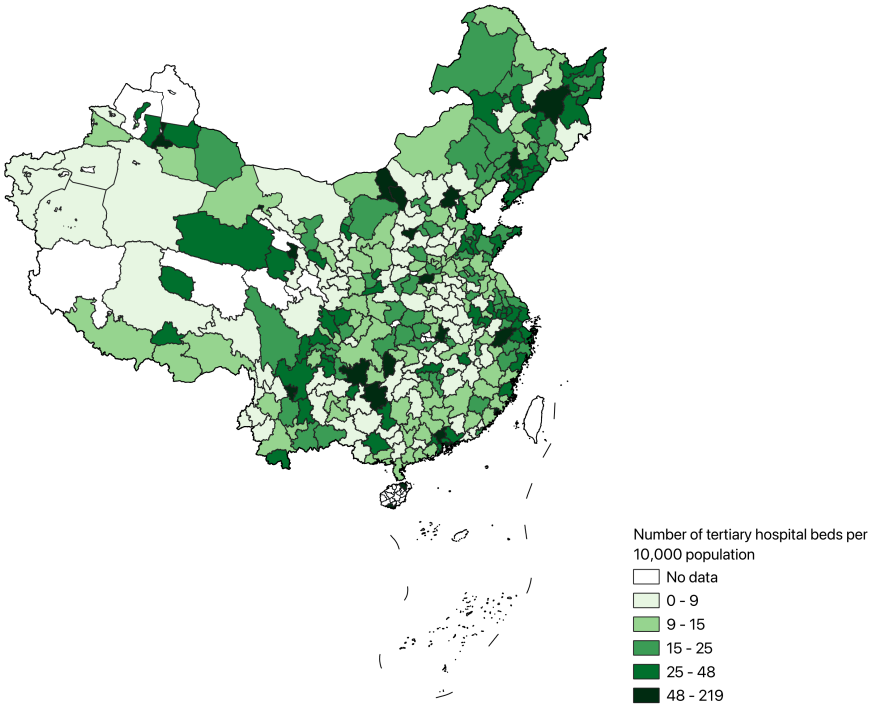
We draw on a unique administrative dataset that covers the universe of inpatient admissions in 2017 and 2018 for one critically important disease category, cerebral-cardiovascular diseases (CCVD), in Sichuan province, China. CCVDs are a leading cause

[2]Tertiary hospitals are the highest-level hospitals of the healthcare system. They are typically large, comprehensive institutions with advanced medical equipment, specialized departments, and highly trained staff, capable of providing the most sophisticated acute care and high-quality medical services. Tertiary hospital bed counts are thus used as a proxy for high-quality healthcare capacity, distinguishing them from secondary or primary hospitals that focus on more basic or community-level care.

[3]Appendix Figure A1 presents the distribution of tertiary hospital beds across cities in 2010. While the number of tertiary hospital beds in each city—both in total and per 10,000 population—is substantially lower in 2010 than in 2018, the spatial unevenness in their distribution is evident in both years.



(a) Number of Tertiary Hospital Beds by Cities



(b) Number of Tertiary Hospital Beds per 10,000 Population by Cities

Figure 2: Spatial Distribution of Medical Resources in 2018

Notes: The figure displays the number of tertiary hospital beds and that adjusted for population across cities in China in 2018. Data sources: 2019 Hospital Annual Report, China Health Commission. "No data" indicate cities that are excluded because at least one required variable is missing or cannot be reliably matched across sources. Administrative boundaries are shown at the prefecture level.

of death in China, accounting for the highest disease-related mortality rate nationally, and thus represent a salient setting for studying the mortality implications of healthcare access. For CCVD patients, timely access to acute care is essential. Delays in treatment can lead to rapid deterioration and irreversible health damage. As a result, patient outcomes are highly sensitive to travel time.[4] Our dataset records more than 600,000 admissions to hospitals in 21 cities in Sichuan. We further enrich these data by linking them to annual hospital reports that document the availability of medical resources at the city level, as well as to cross-city travel time measures compiled by Ma and Tang (2024).

We uncover three key stylized facts that guide our understanding of CCVD patients' travel behaviors. First, the availability of medical resources strongly shapes patient flows: patients travel from locations with more limited medical resources to locations with more abundant medical resources. Second, travel time is a critical barrier. Estimates from a gravity equation specification show that a 1% increase in pairwise travel time is associated with a roughly 2% reduction in the share of traveling patients. Finally, patient characteristics interact with these frictions: higher-income individuals are more likely to travel for care, particularly when facing severe health shocks, whereas lower-income patients appear constrained even in serious cases. Together, these patterns highlight the roles of medical resource availability and travel costs in shaping access to high-quality care, as well as the importance of heterogeneity in income and severity when modeling patient travel decisions.

Guided by the stylized facts, we develop a dynamic spatial model that captures how individuals access healthcare services across a transportation network. In the model, individuals vary in their home location, income, and health status, while locations differ in three key dimensions: the availability of medical resources (exogenous), the level of patient volume (endogenous), and their position within the transportation network. Health

[4]The data contain hospital admissions due to 1) cerebrovascular diseases (ICD-10: I63) which are conditions caused by blood clots or blockages in the brain's blood vessels that can lead to strokes and related brain damage, and 2) coronary heart disease (ICD-10: I20) which refers to conditions caused by reduced blood flow to the heart muscle, typically due to narrowed or blocked coronary arteries, which can result in chest pain or heart attacks.

outcomes, proxied by mortality and recovery probabilities, are jointly determined by a location's medical resources and its patient volume. In equilibrium, individuals experiencing a health shock with varying degrees of severity decide where to seek treatment by weighing the expected effectiveness of care at alternative locations against the costs of traveling. This framework allows us to evaluate how improvements in transportation infrastructure influence the spatial allocation of patients and, ultimately, health outcomes across regions.

We structurally estimate the key parameters of the model using an indirect inference approach, matching model-generated moments to their empirical counterparts. Specifically, we target estimates of the impact of various determinants on the tendency to travel, the distance elasticity in patients' treatment location choices, and the impact of medical resources and patient volumes on treatment outcomes. The model performs well in replicating observed patterns in the data and produces robust estimates that capture the key trade-offs individuals face when navigating healthcare access across space.

We extend our counterfactual analysis to the entire country. We simulate the impact of transportation network improvements from 2010 to 2018. Using the structurally estimated model, we conduct counterfactual simulations to isolate the effect of enhanced transport infrastructure while holding the distribution of medical resources and other factors fixed at 2010 levels. Our findings suggest that the expansion of transportation networks from 2010 to 2018 would have saved about 10,000 lives from CCVDs per year. Improved connectivity enables faster access to high-quality care from remote areas, resulting in a measurable reduction in the spatial disparity of health outcomes. However, this convergence in regional health outcomes is largely driven by benefits accruing to high-income individuals, who are more capable of overcoming the financial barriers to seek out-of-city care. As a result, while geographic inequality in health has declined, income-related disparities in health outcomes have widened, highlighting the regressive distributional effects of transportation-led healthcare improvements in the absence of targeted financial support.

This study contributes to both the urban economics and health economics litera-

ture. While much of the urban economics literature focuses on the economic impacts of transportation infrastructure through trade, labor mobility, and market competition (Faber, 2014; Allen and Arkolakis, 2014; Donaldson, 2018; Banerjee et al., 2020; Asher and Novosad, 2020; Allen and Arkolakis, 2022; Fang et al., 2025), our study adds to a smaller but growing body of work on accessing critical non-tradable services such as healthcare (Li, 2014; Dingel et al., 2023). Complementing this strand, a growing literature in health economics emphasizes the importance of travel costs in shaping healthcare demand and outcomes (Ho, 2006; Ho and Pakes, 2014; Hackmann, 2019; Prager, 2020; Fang et al., 2020). Our study contributes to this literature by studying not only the aggregate health gains from improved transportation connectivity, but also their distributional consequences. Based on a dynamic spatial model of hospital choice, we show that shorter travel times reduce spatial disparities in health by encouraging cross-regional healthcare utilization, yet simultaneously widen income-related health disparities by disproportionately benefiting higher-income patients. Our findings underscore the importance of pairing transportation investments with means-tested subsidies to ensure more equitable health benefits.

2 Institutional Background

2.1 Development of Transport Infrastructure in China

China's transportation infrastructure underwent an unprecedented transformation starting from the 1990s, characterized by massive investments in both highway and rail networks. Over the following three decades, the central government prioritized the construction of inter-provincial expressways and high-speed rail lines as part of its broader agenda to integrate inland and coastal regions and to promote spatial equity. Notably, the expansion of the National Trunk Highway System (NTHS) brought many remote inland regions into the national transportation network (Faber, 2014). As documented in Egger et al. (2023), in 2000, the total length of China's highway system was less than two-thirds of the length of the US system; by 2014, however, China's highway network had already

grown to 183 percent of the total length of the US Interstate Highway System.

An equally remarkable, and perhaps more extensively studied, development is the introduction of the high-speed rail (HSR) system. While the construction was at the trial stages during the first decade of the 21st century, with only a few pilot lines in operation (see the discussions in Ma and Tang, 2024), the pace of HSR expansion accelerated dramatically in the 2010s. By the end of that decade, the Chinese HSR system had become the largest in the world, exceeding the second-largest system by a margin of 40 percent (Egger et al., 2023). The network enabled efficient passenger mobility at unprecedented speeds, especially between major urban centers, leading to a profound impact on the local economy (Lin, 2017). While existing studies have primarily emphasized the economic implications of transportation infrastructure, relatively little attention has been devoted to its effects on medical access, thereby underscoring the novel contribution of our study.

2.2 Healthcare System in China

In China, public hospitals play a dominant role in the healthcare system, providing 85.3% of outpatient services and 81.7% of inpatient services in 2018 (National Health Commission, 2019). Hospitals are classified in three tiers. Primary hospitals primarily provide generalist clinical care and basic public health services, typically with fewer than 100 beds. Secondary hospitals, often equipped with 100 to 500 beds, provide comprehensive healthcare services. Tertiary hospitals are usually general hospitals with more than 500 beds that provide the most sophisticated specialist services. They also play an important role in medical education and research, and serve as prominent medical centers for surrounding regions.

While hospitals at different tiers have designated functions, there is no gatekeeping referral system to triage patients by medical severity, and patients typically seek hospital care on a walk-in basis (Milcent, 2018). Tertiary hospitals do indeed charge higher prices for their services compared to lower-tier hospitals, but the price differences are not substantial. Moreover, patients are reimbursed by public health insurance for care sought

without referrals.[5] As a result, despite accounting for only 7.7% of all hospitals in 2018, tertiary hospitals provided 46% of inpatient care, amounting to 92.9 million inpatient admissions (National Health Commission, 2019).

2.3 Inequality of Healthcare Resources

China's healthcare system has long exhibited spatial disparity in healthcare resources—including infrastructure, technology, and medical personnel—and, consequently, uneven access to high-quality care. This disparity is not only due to regional differences in population structure and healthcare demand, but more importantly, due to institutional arrangements and economic reforms.

During China's central planning period (1949–1978), healthcare infrastructure expanded substantially but remained skewed towards urban areas (Burns and Huang, 2017). Urban public health insurance schemes—the Government Insurance Scheme (GIS) for public sector employees and the Labor Insurance Scheme (LIS) for state-owned enterprise workers—secured high-quality, heavily subsidized healthcare resources for urban populations. Compared with urban areas, rural areas received far less government spending on healthcare. The healthcare system in rural areas was based on the rural commune system, which focused on promoting basic and preventive healthcare through various public health campaigns—such as expanding nationwide immunization, and training indigenous rural health workers (so-called bare-foot doctors). By the late 1970s, healthcare in rural areas was mainly provided by bare-foot doctors and local clinics (Zhang and Kanbur, 2009).

Economic reforms since 1978 exacerbated healthcare inequality across regions. With fiscal decentralization, local governments became increasingly responsible for funding healthcare, resulting in significant regional variation in resource allocation (Milcent, 2018). Wealthier eastern provinces and urban areas could allocate more resources to

[5] Almost all residents in China are covered by public health insurance, which consists of two schemes: urban employee basic medical insurance (UEBMI) and urban and rural residents basic medical insurance (URRBMI). The first is a mandatory scheme for all employees and retirees (and their dependents) with urban *hukou*; the second covers over 95% the remaining population in 2018. Data sources: https://www.gov.cn/xinwen/2019-03/02/content_5369865.htm

healthcare infrastructure and professional training. In contrast, poorer western provinces and rural areas lacked sufficient fiscal capacity to sustain healthcare investments. In 1980, urban areas had 4.57 hospital beds and 7.82 healthcare personnel per 1,000 people, compared to just 1.48 beds and 1.81 personnel in rural areas; these gaps had increased further by 2000.[6]

Since the 2009 healthcare reform, China has expanded public insurance coverage and increased investment in healthcare infrastructure. However, regional disparities persist due to uneven implementation. For example, public hospitals, which operate under strict regulations, require lengthy approvals from multiple regulatory levels for infrastructure investment and medical equipment procurement (Zhang, 2011). These processes heavily rely on the local government's fiscal capacity. Developed regions with more effective local governance can expedite these processes, allowing hospitals to expand more quickly. Conversely, hospitals in less developed regions often face significant challenges in obtaining timely regulatory approval and adequate financial support. Consequently, hospitals equipped with advanced medical equipment, more beds, and better-trained medical professionals are predominantly concentrated in developed urban and eastern regions. For instance, in 2018, Beijing had 47 tertiary hospitals per 10 million residents, whereas the provincial-level average across the rest of China was 19 (National Health Commission, 2019).

3 Data

We use data from the fifth most populous province in China, Sichuan. Located in the Southwest hinterlands, Sichuan covers a total area of 486,000 square kilometers with approximately 83 million residents as of 2024. Home to 21 cities, Sichuan is a large province: its population is equivalent to that of Germany—exceeding that of the two largest states combined in the U.S.. The province's healthcare system and public health insurance arrangements are broadly representative of those in China as a whole. The geography

[6]Data source: Comprehensive Statistical Data and Materials on 50 Years of New China (China State Statistical Bureau, 2000).

of Sichuan province can be divided into two distinct regions: the east and west. Economically advanced cities are typically located in the east, owing to the Sichuan basin characterized by its flat terrain and fertile soil (and thus able to support a large population). On the contrary, less prosperous cities are located in the western mountainous regions which are much less accessible. This province on a whole is landlocked and enclosed by surrounding hills and mountains. As a result, transportation developments have long been bottlenecks in the province before advancements in engineering technology.

Admission-Level Information. The patient admission-level information is drawn from the hospital admission and discharge summary dataset, compiled by the Sichuan Health Commission. This dataset includes comprehensive administrative hospitalization records from Sichuan. It contains detailed inpatient records for all patients diagnosed with CCVDs across all hospitals in the province for the years 2017 and 2018, including demographic characteristics (such as age, gender, place of residence, and occupation) and medical details (such as diagnosis, disease severity, procedure, and mode of discharge in the form of recovery or death), as well as hospital identifier. We impute each patient's monthly income based on their occupation, age, and gender.[7]

Panel A of Table A1 presents summary statistics from the inpatient records. In our sample, 45.7% of patients are female. Approximately 38.7% of cases are classified as severe.[8] We construct three commonly used measures of health outcomes: mortality, recovery, and non-recovery within a 30-day window. Mortality is defined as an indicator equal to 1 if a patient either (a) dies during hospitalization or (b) is readmitted within 30 days and subsequently dies during the second stay. This definition may underestimate true mortality, as it does not capture deaths occurring outside the hospital. Recovery is defined as an indicator equal to 1 if a patient is discharged and not readmitted within 30 days. The remaining cases—patients who are discharged, readmitted within 30 days, and do not die during the second stay—are classified as non-recovered. Across the full sample, 6% of patients died and 92.6% recovered during the sample period.

[7]Specifically, we impute income using the mean earnings of individuals in the same occupation–age-gender group within the same province, where the mean earnings are estimated from the China Family Panel Studies 2016 survey.

[8]Severity is defined based on admission labels indicating "critical" or "urgent" conditions.

Hospital-Level Information. The hospital-specific information is from the 2018 hospital annual report, also obtained from the Sichuan Health Commission. This dataset identifies hospital name, classification tier, ownership type, and the amount of medical resources, such as the number of doctors, nurses, and inpatient beds. As summarized in Panel B of Table A1, 8.9% of hospitals in Sichuan are classified as tertiary, 25.7% as secondary, and the remainder as primary or ungraded institutions. When combined with the patient-level admission and discharge records described above, we find that despite representing a small share of total hospitals, tertiary institutions account for 60% of CCVD admissions in our sample. This concentration reflects their greater capacity to manage complicated health conditions, with more specialized staff, advanced technology, and infrastructure. For example, tertiary hospitals have an average of 853 beds—substantially more than lower-tier institutions. Given these advantages, we use the number of tertiary hospital beds as our primary measure of local medical resources.

City-Level Information. The annual hospital report dataset also provides the geographic location of hospitals, which allows us to aggregate the availability of medical sources at the city level and assess its spatial variation. Panel B of Table A1 shows that the average number of tertiary hospital beds per city is 9,230, with substantial variation across locations (standard deviation is 10,600). Given the matched information on patient hospitalization location, we find that on average, cities receive 15,860 CCVD patients per year. By comparing hospital locations with patients' places of residence, we find that 4.4% of CCVD patients in our sample received treatment in a city different from their city of residence.

City-Pair-Level Information. Conditional on the distribution of medical resources, travel costs also play a critical role in determining patient flows. To assess the importance of travel costs, we link patient-level travel patterns with travel times between cities in Sichuan for relevant years. The travel time data are obtained from Ma and Tang (2024), who carefully account for infrastructure quality when estimating road and rail travel times. Panel D of Table A1 reports that the average road travel time between city pairs is 4.4 hours, while the average rail travel time is 7.9 hours. When using the

minimum of the two modes as the effective travel time, the mean is 4.3 hours.

4 Stylized Facts

In this section, we present three stylized facts that motivate our model of patients' hospital location choices.

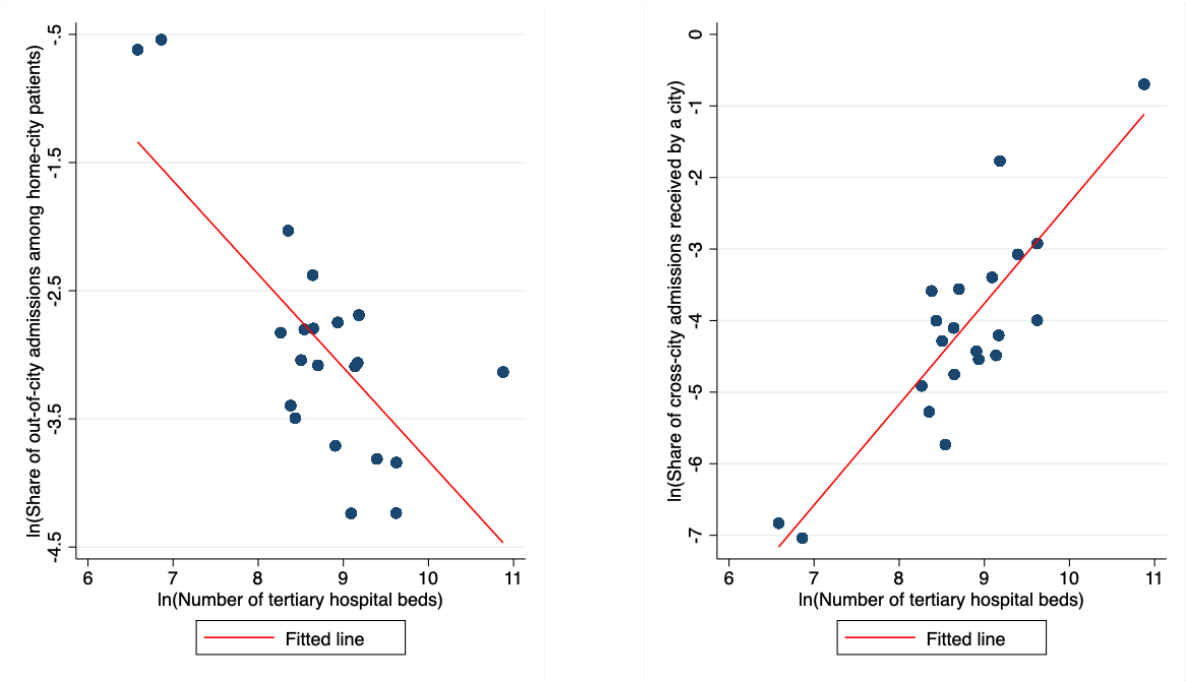
4.1 Medical Resources

The majority of patients travel from areas with fewer medical resources to those with more abundant ones, highlighting the role of healthcare resources in driving out-of-city travel behaviors.[9] Panel A of Figure 3 illustrates the relationship between the number of tertiary hospital beds and the share of patients who seek hospital care outside their city of residence. The figure indicates that patients residing in cities with fewer tertiary hospital beds are more likely to travel for medical treatment. Panel B further shows that these out-of-city patients are more likely to seek care in cities with more abundant medical resources.

4.2 Travel Time

To isolate the effect of travel time, we estimate a standard gravity model in which the share of patients traveling between city pairs is regressed on pairwise travel time, controlling for origin and destination fixed effects. As reported in Columns (1)-(3) of Table 1, a 1 percent increase in travel time is associated with a 1.92-2.05 percent decrease in the share of patients traveling for hospital care, where the exact magnitude depends on whether travel time is measured by rail, road, or the minimum of the two. When travel time is instrumented using pairwise geographic distance, as reported in Columns (4)-(6), the magnitudes of the estimated coefficients become slightly larger but remain qualitatively consistent.

[9] We find that less than 6% of out-of-city patients in our sample travel from high-resource cities to low-resource ones for hospital care.



(a) Home city: share of out-of-city admissions and number of tertiary hospital beds

(b) Destination city: share of cross-city admissions and number of tertiary hospital beds

Figure 3: Medical Travel and Number of Tertiary Hospitals Beds

Notes: Panel (a) plots the share of out-of-city admissions among all patients from the home city against the number of tertiary hospital beds in that city. Panel (b) plots the share of cross-city admissions received by the destination city, among all cross-city admissions in the province, against the number of tertiary hospital beds in the destination city. Each dot represents one city. We plot the figure using the hospital annual report and hospital admission and discharge summary datasets in 2018. Both panels focus on admissions for CCVD.

4.3 Severity and Income

In this section, we explore the impact of income and severity on a patient's tendency to travel out-of-city to seek hospital care. Figure 4 presents the relationship between the monthly income (in logarithms) and the share of out-of-city hospital admissions (in logarithms), stratified by disease severity. Conditional on severity, patients are grouped into bins of 0.1 units in log income, and within each bin, we compute the share of admissions that are out-of-city. The plotted dots represent these bin-level averages, with red solid dots indicating severe cases and blue hollow dots indicating less severe cases. The figure reveals a positive correlation between income and the likelihood of seeking care outside the patient's city of residence for both severity groups. Moreover, the slope of the fitted line is notably steeper for severe patients, suggesting that higher-income individuals

Table 1: Estimation results for the gravity equation

	ln(Share of admissions outside the home city)					
	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
ln(Minimum travel time)	-2.031***			-2.362***		
	(0.278)			(0.334)		
ln(Travel time by rail)		-1.919***			-3.282***	
		(0.290)			(0.727)	
ln(Travel time by road)			-2.050***			-2.529***
			(0.273)			(0.303)
Origin-by-year FE	Yes	Yes	Yes	Yes	Yes	Yes
Destination-by-year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	653	653	653	653	653	653

Notes: This table reports the regression results of a gravity equation in which we regress the share of patients traveling between city pairs on pairwise travel time, controlling for origin and destination city fixed effects. The number of observations used in the regressions (653) is smaller than the number of city-pair observations in reported Table A1 (882), because 99 (88) city pairs have zero cross-city admissions in 2017 (2018), and because Table A1 includes same-city pairs (i.e., a city paired with itself). Standard errors clustered at the origin-by-destination-city level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

with more severe conditions are particularly more likely to travel for medical care. This pattern implies that income constraints may limit travel options for lower-income patients even when faced with serious health conditions, while higher-income patients are more responsive to medical need in their travel decisions.

5 A Model of Hospital Care Location Choice

Motivated by the empirical facts, we develop a dynamic discrete choice model to describe patients' hospital location choice decisions. The model incorporates spatial variation in treatment quality and accounts for heterogeneity in patients' income and disease severity. We highlight the trade-off patients face between travel costs and expected treatment outcomes when choosing where to seek care under dynamic health risks.

5.1 Model Setup

We study an infinite-horizon economy in discrete time, indexed by $t = 0, 1, 2, \ldots$, with a fixed set of cities indexed by $k = 1, 2, \ldots, K$. Each city represents a distinct local

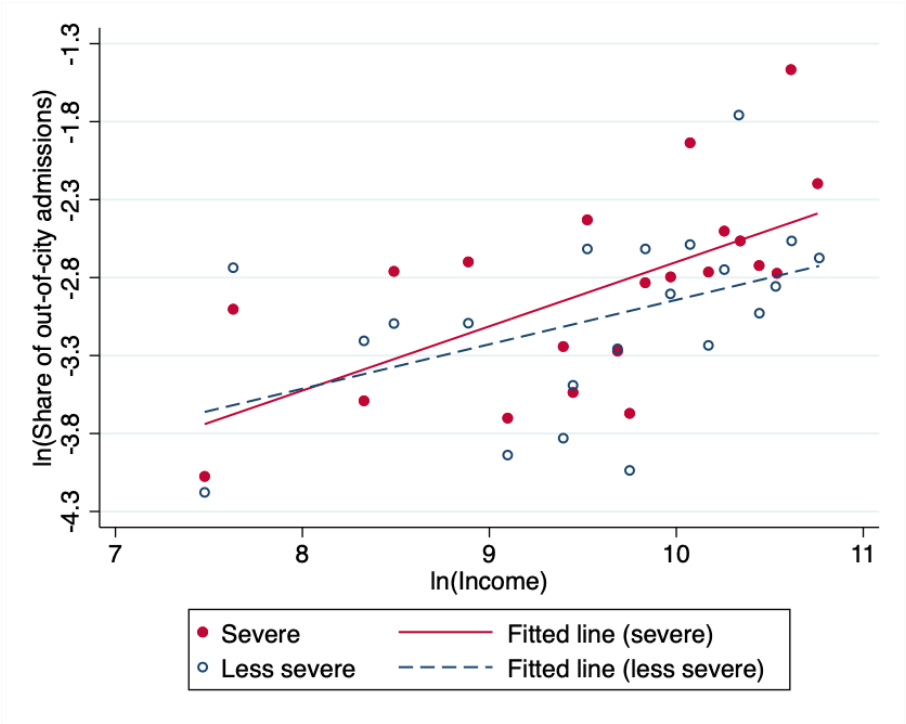


Figure 4: Tendency to Medical Travel and Patient Characteristics

Notes: This figure plots patients' tendency to seek out-of-city medical care by their monthly income and severity in 2018. Conditional on severity, we divide admissions into bins of width 0.1 in ln(income). Within each bin, we calculate the share of out-of-city admissions, and plot the values for severe (less severe) admissions in red (blue hollow) dots. The lines show fitted values separately for each severity group.

healthcare market and is inhabited by a constant population of size P_k. We assume that all individuals reside permanently in their home city k, but upon falling ill, they seek hospital care either in their home city or in another city.[10] Each individual i at time t is described by two state variables: health status s_{it} and productivity level z_{it}.[11]

Health Status. Health status s_{it} is a discrete variable taking values in $-1, 0, 1, \ldots, S$, where $s = -1$ denotes death; $s = 0$ indicates good health; and $s > 0$ represents illness, with larger values of s corresponding to greater severity.

Health status evolves stochastically over time. We use $\pi_{kt}^{s,\bar{s}}$ to denote the probability that an individual's health status transitions from s in time t to $\bar{s}$ in time $t+1$. These transition probabilities are specific to city k in time t. For example, a healthy individual

[10] We abstract from labor migration and focus on individuals' location choice for hospital care.

[11] We broadly interpret "individuals" in our model as either individuals or households, acknowledging that decisions about seeking hospital care are typically made at the family level, subject to the family's budget constraint.

residing in city k faces a probability $\pi_{kt}^{0,s}$ of becoming sick with severity level s ($s > 0$). For any patient with severity s ($s > 0$) who seeks hospital care in city k, regardless of whether the patient resides in k or elsewhere, the mortality rate is $\pi_{kt}^{s,-1}$, and the recovery rate is $\pi_{kt}^{s,0}$. We assume that greater severity leads to higher mortality rate and lower recovery rate: $\pi_{kt}^{s',-1} \geq \pi_{kt}^{s,-1}$ and $\pi_{kt}^{s',0} \leq \pi_{kt}^{s,0}$, $\forall s' > s > 0$.

Productivity. A healthy individual's productivity z_{it} evolves over time according to an AR(1) process in logarithmic form:

$$\log z_{it} = \rho_z \log z_{i,t-1} + \nu_t, \quad \nu_t \sim \mathcal{N}(0, \sigma_\nu) \tag{1}$$

where ρ_z captures the autocorrelation of the productivity, ν_t is the productivity shock, and σ_ν its standard deviation.

Sickness with severity s ($s > 0$) reduces labor productivity by a factor $\delta^s \in (0, 1]$, so that a sick individual's productivity is only $\delta^s z_{it}$. Higher severity leads to lower productivity, with $\delta^{s'} \leq \delta^s$, $\forall s' > s > 0$.

A healthy individual's income is $\omega_{ikt} = w_{kt} z_{it}$, while a sick individual's income is $\omega_{ikt}^s = \delta^s w_{kt} z_{it}$, where w_{kt} is the wage rate for one efficiency unit of labor productivity in city k at time t.

5.2 Individual Decisions

5.2.1 Healthy Individuals

Healthy individuals do not seek hospital care. Their behavior is passive: they earn income, consume it fully in period t, and then transition either to another healthy state or to an illness of severity $s (s > 0)$ in $t + 1$.[12]

The expected utility of a healthy individual i in city k at time t is

$$v_{kt}^0(z_{it}) = u(w_{kt} z_{it}) + \beta \mathbf{E} \left[\left(1 - \sum_{s=1}^{S} \pi_{kt}^{0,s} \right) v_{k,t+1}^0(z_{t+1}) + \sum_{s=1}^{S} \pi_{kt}^{0,s} v_{k,t+1}^s(z_{t+1}) \right], \tag{2}$$

[12] We assume that healthy individuals do not transition directly to death.

where the expectation $\mathbf{E}$ is taken over future productivity shocks. $v^s_{k,t+1}(z_{i,t+1})$ is the value function of the individual with health status s in time $t+1$ ($s \geq 0$).

5.2.2 Sick Individuals

When an individual residing in city k becomes sick (i.e., $s > 0$), she chooses a destination city l for hospital care. This decision depends on treatment quality, associated travel costs, and location preferences across cities.

Treatment Quality. The treatment quality in city l is reflected by the recovery rate ($\pi^{s,0}_{lt}$) and mortality rate ($\pi^{s,-1}_{lt}$) for inpatient cases in this city. We assume that these rates are functions of local medical resources (m_{lt}) and patient volume (p_{lt}). Greater medical resources are expected to improve quality, whereas the effect of patient volume is theoretically ambiguous: congestion may reduce effectiveness, while experience and learning effects from treating more patients may enhance treatment quality. We empirically examine how treatment quality relates to m_{lt} and p_{lt} using micro-level data, as discussed in detail later.

Travel Costs. We consider two types of costs for medical travel from city k to l. The first is a fixed monetary cost, denoted by $\lambda > 0$. We assume that this fixed cost is uniform across cities. It primarily stems from non-geographical administrative barriers that discourage patients from seeking hospital care out of their home city—such as higher co-payment rates or reduced insurance coverage for certain procedures received outside the local provider network. This cost directly reduces consumption. For a sick individual i with severity $s(s > 0)$, disposable income when receiving care in city l is $\delta^s w_{kt} z_{it} - \mathbb{1}(l \neq k) \cdot \lambda$. The second type of costs increases with travel distance. It captures transportation expenses, the opportunity cost of travel, emotional discomfort of being far from home, and the loss of informal support from family and friends. Following the urban economics literature (Stillwell et al., 2014; Bryan and Morten, 2019; Tombe and Zhu, 2019), we assume that these distance-related costs reduce utility directly, and

the associated disutility is denoted as τ_{klt}.[13]

Idiosyncratic Preference. We follow the dynamic spatial literature, such as in Caliendo et al. (2019) and Kleinman et al. (2023), and model the location choice as a dynamic discrete choice problem. We allow for unobserved heterogeneity in individual preferences for medical treatment across cities. In each period, individuals draw an independent and identically distributed ($i.i.d.$) vector of idiosyncratic utility shocks across cities, $\{\varepsilon_l\}_{l=1}^{K}$, and the shocks follow a Type-I Generalized Extreme Value (GEV-I) distribution with cumulative distribution function:

$$F(\varepsilon) = \exp\big\{ - \exp[-(\varepsilon + \bar{\gamma})]\big\}, \tag{3}$$

where $\bar{\gamma}$ is the Euler's constant. The expression is equivalent to a Gumbel distribution with location parameter $-\bar{\gamma}$ and a shape parameter of 1.

Choice Set. Unlike standard dynamic discrete choice models, individuals' choice sets in our model vary by home city, income, and severity. This variation arises from the feasibility constraints on disposable income: medical travel is feasible for an individual only if disposable income after incurring the travel costs is strictly positive (i.e., $\delta^s w_{kt} z_{it} - \mathbb{1}(l \neq k) \cdot \lambda > 0$). In addition, we impose a second constraint grounded in empirical patterns observed in the data: individuals do not travel to cities with strictly worse medical resources than those available in their home city k. That is, the medical resources in the destination city l must satisfy $m_l \geq m_k$. Taken together, we define the feasible choice set for individual i in city k with severity $s(s > 0)$ and productivity z_{it} as:

$$\mathbb{F}_{kt}^{s}(z_{it}) \equiv \{l \in \mathbb{K} \,|\, m_l \geq m_k \text{ and } \delta^s w_{kt} z_{it} - \mathbb{1}(l \neq k) \cdot \lambda > 0\}, \tag{4}$$

where $\mathbb{K}$ denotes the set of all possible destination cities indexed up to K.

Importantly, this choice set is guaranteed to be non-empty, as the home city k is

[13]The urban economics literature typically models migration costs that increase with travel distance as disutilities rather than as direct monetary costs. We follow this convention to facilitate comparison between our estimate of distance elasticity and those in the literature.

always feasible for all individuals. By allowing individual-specific choice sets, the model captures income-induced disparities in healthcare access and accounts for the sparsity of patient flows between certain city pairs observed in the data.[14] However, this flexible setup also introduces subtleties in the recursive formulation of value functions, which we discuss below.

Recursive Problem. We formulate the sick individual's decision as a recursive problem. In each period, a sick individual i with severity $s(s > 0)$, residing in city k, chooses a destination city l for hospital care to maximize expected lifetime utility. Let $v^s_{kt}(z_{it})$ denote the value function of the individual. The recursive problem is given by:

$$v^s_{kt}(z_{it}) = \max_{l \in \mathbb{F}^s_{kt}(z_{it})} \Bigg\{ u(\delta^s w_{kt} z_{it} - \mathbb{1}(l \neq k) \cdot \lambda) - \tau_{klt} + \kappa \varepsilon_l \\ + \mathbf{E}\Big[\beta \left[\pi^{s,0}_{lt} v^0_{k,t+1}(z_{i,t+1}) + \left(1 - \pi^{s,0}_{lt} - \pi^{s,-1}_{lt}\right) v^s_{k,t+1}(z_{i,t+1}) \right] \Big] - \bar{\varepsilon}^s_{kt}(z_{it}) \Bigg\}. \tag{5}$$

The individual's flow utility at time t includes consumption utility, travel disutility, and an idiosyncratic preference shock for city l. The parameter κ captures the importance of preference shock relative to travel disutility in determining hospital choices; as will be shown in Section 5.5, κ is the inverse of the distance elasticity of seeking hospital care. A higher κ implies that hospital choices are more influenced by the preference shock and less sensitive to travel disutility. The individual's future value depends on the probabilities of recovery, death, and remaining sick. Specifically, the individual recovers with probability $\pi^{s,0}_{lt}$ and returns to the healthy value function; dies with probability $\pi^{s,-1}_{lt}$, yielding zero future utility ($v^{-1}_{k,t+1} = 0$); or remains ill with probability $1 - \pi^{s,0}_{lt} - \pi^{s,-1}_{lt}$.[15] This value function specification yields a multinomial logit form for location choice, providing closed-form expressions for expected choice values.

The last term in Eq. (5) corrects the utility drift due to the variability of the choice

[14] Approximately 20% of city pairs have zero observed patient flows in the data.

[15] For tractability, we assume that individuals who remain ill stay at the same severity level $s(s > 0)$, but the model can be flexibly extended to allow transitions across different severity levels without altering the baseline framework.

set $\mathbb{F}^s_{kt}(z_{it})$ across individuals. While this variability helps capture the role of income in location choices, it also causes the value function $v^s_{kt}(z_{it})$ to vary mechanically with the size of the choice set. Intuitively, in discrete choice models, individuals value the number of choices—the more choices, the higher the ex-ante utility. In our context, this feature implies that with a large enough K, individuals might even *prefer* being sick, due to the additional utilities derived from the preference shocks. Recall that healthy individuals are passive and do not draw any ε shocks. To address this issue, we adjust the sick individual's value function by subtracting $\bar{\varepsilon}^s_{kt}(z_{it}) = \kappa \cdot \log \overline{\mathcal{F}}^s_{kt}(z_{it})$, where $\overline{\mathcal{F}}^s_{kt}(z_{it})$ is the cardinality of the choice set $\mathbb{F}^s_{kt}(z_{it})$. $\bar{\varepsilon}^s_{kt}(z_{it})$ represents the expectation of ε_l conditional on city l being the optimal choice within the feasibility set. Appendix B.1 provides more mathematical details.

Figure 5 presents the timeline of the sick individual. At the beginning of the time period t, the individual observes the realization of idiosyncratic preference shocks $\{\varepsilon_l\}_{l=1}^K$ and the distribution of medical resources (m_{lt}) across all potential destination cities, while forming expectations about patient volumes (p_{lt}) in these cities. Based on this information and expectation, the individual selects a city l for hospital care. She then incurs the fixed monetary cost λ, consumes the residual income $\delta^s w_{kt} z_{it} - \mathbb{1}(l \neq k) \cdot \lambda$, and experiences travel disutility τ_{klt}. At the end of the period, the treatment outcome is realized, and the individual transitions to one of the three states in the next period: recovery (returning to health), continued illness, or death (exiting the model). The entire model can be represented as a Markov process over $S + 2$ health states, as detailed in Online Appendix Section B.3.

5.3 Aggregation

Law of Motion for Population. We now describe the evolution of the population across health states and productivity levels in each city. Let $L^s_{kt}(z_t)$ denote the number of individuals whose home city is k at time t, with health status s and productivity level z_t. For notational simplicity, we suppress the individual index i in the subscript whenever doing so does not create confusion.

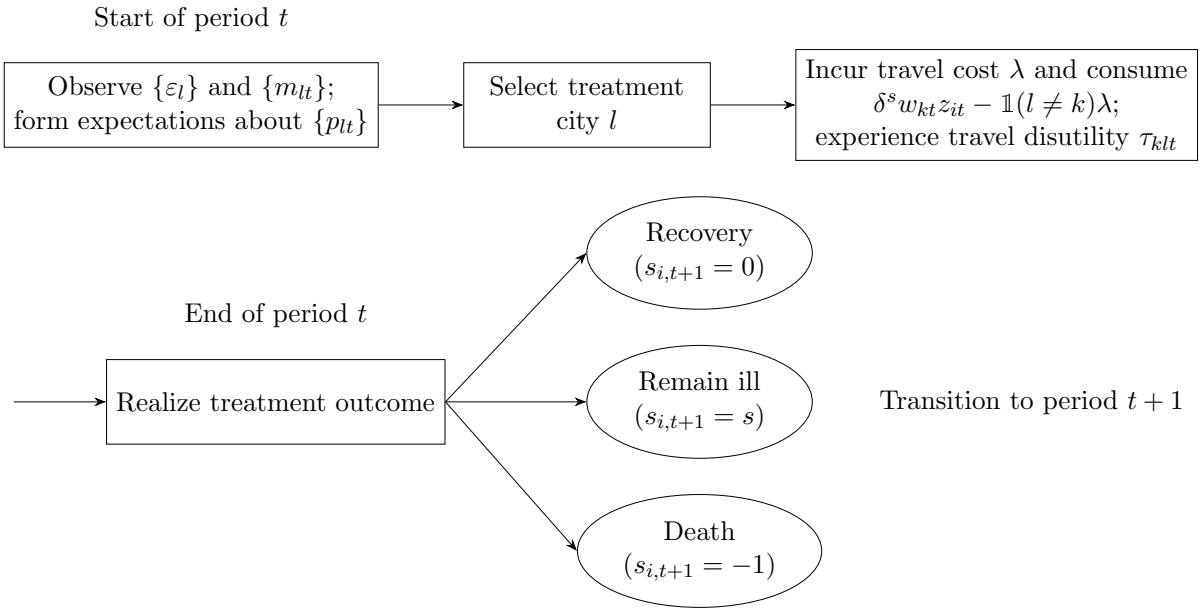


Figure 5: Timeline of Sick Individual's Decision and Health Transition at Time t

Healthy population in home city k at time t, $L^0_{kt}(z_t)$, consists of three groups: (i) healthy individuals at $t-1$ who remain healthy; (ii) sick individuals who sought care either inside or outside home city at $t-1$ and recover by time t; and (iii) healthy newborns who replace deceased individuals. Here, we assume that each deceased individual at time $t-1$ is immediately replaced by a healthy newborn in the same city at the beginning of time t.[16] This ensures a constant total population and facilitates a stationary analysis. Accordingly, the law of motion for the healthy population is given by:

$$
\begin{aligned}
L^0_{kt}(z_t) = & \underbrace{\left(1 - \sum_{s=1}^{S} \pi^{0,s}_{kt}\right) \int_0^\infty L^0_{k,t-1}(z_{t-1}) g(z_t|z_{t-1}) dG(z_{t-1}) +}_{\text{Persistently Healthy}} \\
& \underbrace{\int_0^\infty \sum_{s=1}^{S} \sum_{l=1}^{K} \pi^{s,0}_{l,t-1} \mu^s_{kl,t-1}(z_{t-1}) L^s_{k,t-1}(z_{t-1}) g(z_t|z_{t-1}) dG(z_{t-1}) +}_{\text{Newly Recovered}} \\
& \underbrace{\int_0^\infty \sum_{s=1}^{S} \sum_{l=1}^{K} \pi^{s,-1}_{l,t-1} \mu^s_{kl,t-1}(z_{t-1}) L^s_{k,t-1}(z_{t-1}) g(z_t|z_{t-1}) dG(z_{t-1})}_{\text{Newborns Replacing Deceased}} . \qquad (6)
\end{aligned}
$$

[16] This assumption is plausible in our context because the population size in Sichuan province remained relatively stable over the two years in our sample. Although we assume that all newborns are healthy, they could alternatively follow any exogenous health distribution without affecting our analysis.

In the equation, $g(z_t|z_{t-1})$ is the density of z_t conditional on its previous realization z_{t-1}. The term $\mu^s_{kl,t-1}(z_{t-1})$ denotes the probability that a sick individual with severity $s > 0$ chooses city l for hospital care at time $t-1$, which satisfies $\sum_{l=1}^{K} \mu^s_{kl,t-1}(z_{t-1}) = 1$. As will be shown in the model solution (Section 5.5), this probability depends on the individual's productivity draw at time $t-1$.

The sick population originating from city k with severity $s > 0$ at time t, $L^s_{kt}(z_t)$, consists of two components: (i) newly sick individuals who transition from the healthy state at $t-1$, and (ii) existing sick individuals who remain ill. Formally, the law of motion is

$$L^s_{kt}(z_t) = \underbrace{\int_0^\infty \pi^{0,s}_{kt} L^0_{k,t-1}(z_{t-1}) g(z_t|z_{t-1}) dG(z_{t-1})}_{\text{Newly Sick}} + \underbrace{\int_0^\infty \sum_{l=1}^{K} \mu^s_{kl,t-1} \left(1 - \pi^{s,-1}_{l,t-1}(z_{t-1}) - \pi^{s,0}_{l,t-1}(z_{t-1})\right) L^s_{k,t-1}(z_{t-1}) g(z_t|z_{t-1}) dG(z_{t-1})}_{\text{Stay Sick}}. \tag{7}$$

The healthy and sick populations together constitute the total population of the city, which remains constant over time:

$$\int_0^\infty L^0_{kt}(z_t) + \sum_{s=1}^{S} L^s_{kt}(z_t) dG(z_t) = L_k \tag{8}$$

Congestion. The total number of patients seeking care in city l at time t is endogenously determined by the aggregation of incoming sick individuals from all origin cities:

$$p_{lt} = \int_0^\infty \sum_{s=1}^{S} \sum_{k=1}^{K} \mu^s_{klt}(z_t) L^s_{kt}(z_t) dG(z_t) \tag{9}$$

This endogenous congestion variable plays a crucial role in determining health outcomes, as it affects both recovery and mortality rates.

5.4 Equilibrium

Given a sequence of city-specific fundamentals—including wage rates w_{kt}, medical resources m_{kt}, and initial populations L_k—as well as a fixed monetary travel cost λ and bilateral travel disutility τ_{klt}, a sequential equilibrium consists of time paths for value functions $v^s_{kt}(z_t)$ and population distributions $L^s_{kt}(z_t)$ such that:

1. Individuals choose cities for hospital care by solving dynamic problems. In particular, sick individuals with severity $s > 0$ choose a city l to maximize their expected utility, as defined in Eq. (5), subject to the distribution of patient volumes across cities ($\{p_{lt}\}_{l=1}^K$).

2. Rational expectations are imposed such that individuals' beliefs about the law of motion for population and the distribution of patient volumes are consistent with the aggregate dynamics implied by individual decisions. That is, in equilibrium, Eqs. (6)-(9) must hold when evaluated using the endogenous choice probabilities $\mu^s_{klt}(z_t)$ derived from the individual-level optimization problems.

Steady State. With time-invariant city fundamentals and travel costs $\{\omega_k, m_k, L_k, \lambda, \tau_{kl}\}$, the steady state of the economy consists of constant value functions and population distribution $\{v^s_k(z), L^s_k(z)\}$ that satisfy the definition of a sequential equilibrium. In this steady state, we can analyze long-run implications of counterfactual changes in transportation networks or medical resources allocations.

5.5 Solution

To solve the model, we focus on the sick individual's dynamic problem described in Eq. (5). Since the idiosyncratic preference shocks ε_l follow a GEV-I distribution, we can derive closed-form expressions for city choice probabilities, conditional on the feasibility constraints of the choice set. Specifically, for a sick individual residing in city k, with severity $s(s > 0)$ and productivity level z_t, the probability of choosing city l for hospital

care is:

$$\mu^s_{klt}(z_t) = \begin{cases} \frac{\exp\left[v(\delta^s w_{kt} z_t - \mathbb{1}(l \neq k)\cdot\lambda) - \tau_{klt} + \beta W^s_{kl,t+1}(z_{t+1})\right]^{1/\kappa}}{\sum_{l' \in \mathbb{F}^s_{kt}(z_t)} \exp\left[u(\delta^s w_{kt} z_t - \mathbb{1}(l' \neq k)\cdot\lambda) - \tau_{kl't} + \beta W^s_{kl',t+1}(z_{t+1})\right]^{1/\kappa}} & \text{if } l \in \mathbb{F}^s_{kt}(z_t) \\ 0 & \text{otherwise} \end{cases}. \quad (10)$$

In the above equation, we use $W^s_{kl,t+1}(z_t)$ to denote the individual's expected continuation value conditional on choosing destination l:

$$W^s_{kl,t+1}(z_t) = \mathbf{E}\left[\pi^{s,0}_{lt} v^0_{k,t+1}(z_{t+1}) + \left(1 - \pi^{s,0}_{lt} - \pi^{s,-1}_{lt}\right) v^s_{k,t+1}(z_{t+1})\right], \quad (11)$$

where the expectation $\mathbf{E}$ is taken with respect to both preference shocks and productivity shocks in the future. $1/\kappa$ in Eq. (10) is the distance elasticity of seeking hospital care out-of-city: a higher κ implies less sensitivity to travel disutility and greater willingness to travel out-of-city for care.

Following the urban and trade literature (see, e.g., Ahlfeldt et al., 2015; Donaldson and Hornbeck, 2016; Donaldson, 2018), we interpret the denominator of Eq. (10),

$$\Phi^s_{kt}(z_t) = \sum_{l' \in \mathbb{F}^s_{kt}(z_t)} \exp\left[u(\delta^s w_{kt} z_t - \mathbb{1}(l' \neq k)\cdot\lambda) - \tau_{kl't} + \beta W^s_{kl',t+1}(z_{t+1})\right]^{1/\kappa}, \quad (12)$$

as the "medical access" for individuals residing in city k with severity $s (s > 0)$ and productivity z_t. Appendix B.2 and Appendix C.1 provide additional details and the full algorithm for solving the model in steady state.

5.6 Remarks

Before quantifying model parameters, we make two remarks. First, unlike predominantly static frameworks in prior works (Ho and Pakes, 2014; Hackmann, 2019), our dynamic model recognizes that patients face sequential treatment choices as their health and income status evolve stochastically over time. By incorporating forward-looking behavior, we capture how patients weigh not only immediate treatment outcomes but also the continuation value of accessing high-quality care under ongoing uncertainty. At the aggregate

level, the dynamic model allows congestion at treatment destinations to feed back into future treatment quality and subsequently, patients' location choices and health outcomes. These mechanisms are essential for understanding how infrastructure investments generate health benefits through multiple reinforcing channels over time, revealing equilibrium effects that static models would miss.

Second, we do not explicitly model the supply side of the market. We incorporate healthcare prices into patients' cost functions. Moreover, in balancing healthcare supply and demand, prices can be substituted by patients' waiting times. Waiting times, which are determined by the interaction of medical resources and patient volume in the destination city, affects the production of treatment quality. Although waiting times are not explicitly modeled in our framework due to the lack of data, Appendix B.4 details how they influence treatment quality and how we capture this channel using observed medical resources and patient volumes. This provides the microfoundation for our parameterization of treatment outcomes as functions of medical resources and patient volumes in Section 6.2.1. Therefore, in our model—as in Dingel et al. (2023)—prices do not play an active role in equilibrium; instead, the market clears through endogenous adjustments in treatment quality.

6 Quantification

Based on the model solution in steady state, we now quantify key parameters of the model. Each city is defined as a prefecture-level administrative region within Sichuan province, resulting in 21 cities in total, and each time period corresponds to one month.

6.1 Calibration

We first calibrate a list of parameters based on empirical evidence grounded in the literature.

Fixed Costs of out-of-city care (λ). This parameter governs the feasibility of medical travel: a larger λ implies fewer individuals can afford to seek care outside their home city.

We calibrate $\lambda = 0.0005$ to match the average cardinality of choice sets observed in the data. [17]

Time Discount Factor (β). We set the monthly discount factor at $\beta = 0.98^{1/12} \approx 0.9983$, corresponding to an annual real discount rate of 2%. This assumption is standard in dynamic models of intertemporal choice and reflects a moderate degree of impatience among patients.

Severity Level ($s(s > 0)$). Based on the hospital admission and discharge summary data, we categorize sick individuals into two severity levels ($S = 2$), distinguishing between less severe ($s = 1$) and severe ($s = 2$) patients.

Income Loss Due to Illness (δ^s). The parameter δ^s captures the proportional reduction in income associated with illness of severity level $s(s > 0)$. Using the hospital admission and discharge summary data, we find that patients—regardless of severity—spend approximately one-third of their working days in the hospital. Since this share does not vary significantly across severity types, we calibrate $\delta^s = \frac{1}{3}$ for all s, implying that being hospitalized results in a two-thirds reduction in income over the hospitalization period.

Disease Incidence Probabilities ($\pi^{0,s}$). We compute monthly disease incidence probabilities as the number of admission cases with CCVD conditions in our sample divided by the total population. The probabilities are stratified by severity level:[18] $\pi^{0,1} = \frac{0.00354}{24}$ and $\pi^{0,2} = \frac{0.00223}{24}$.

Wage Rates (w_{kt}). We impute w_{kt} using the city-level average income.

[17] We compute the average cardinality in the data as follows. Conditional on an origin city k, the cardinality of choice sets is the number of destination cities that we observe non-zero patient flows. The average cardinality is then the mean across all origins, calculated as $\sum_{l=1}^{K} \sum_{k=1}^{K} \mathbb{1}(p_{kl} > 0)/K$.

[18] The incidence probability of CCVDs varies widely across populations, depending on numerous factors including age, sex, ethnicity, and pre-existing health conditions. For example, the estimated incidence rate is about 92 cases per 100,000 persons per year in Japan (Liu et al., 2025), while over 1,000 per 100,000 in the UK (Conrad et al., 2024).

Persistence of Income Process (ρ_z^s). We do not estimate ρ_z directly from our data but instead adopt values from Fan et al. (2010), who estimate the income process in China. They find that the persistence parameter at the yearly level is 0.917, which corresponds to $\rho_z = 0.917^{1/12} \approx 0.99$ at the monthly level.

Standard Deviation of Productivity Shocks (σ_ν). According to Eq. (1), the variance of $\log z$ in the stationary distribution satisfies

$$\mathrm{Var}(\log z) = \rho_z^2 \,\mathrm{Var}(\log z) + \sigma_\nu^2,$$

which rearranges to

$$\sigma_\nu = \sqrt{(1-\rho_z^2)} \times \mathrm{Std}(\log z).$$

Thus, the standard deviation of productivity shock depends on the persistence of shocks, ρ_z, and the cross-sectional dispersion of $\log z$.

We measure the dispersion of $\log z$ using the dispersion of income, since individual income $\omega_{ikt} = w_{kt} z_{it}$, and any constant scaling of income does not affect the dispersion of log income:

$$\mathrm{Std}\big(\log(\tfrac{\omega_{ikt}}{w_{kt}})\big) = \mathrm{Std}\big(\log(\omega_{ikt}) - \log(w_{kt})\big) = \mathrm{Std}(\log(\omega_{ikt}))$$

for any $w_{kt} > 0$. From our individual-level data, we compute the standard deviation of log income as 1.62719. Plugging this value into the formula yields $\sigma_\nu = \sqrt{1-\rho_z^2} \times 1.62719$.

Utility Function. We assume a Constant Relative Risk Aversion (CRRA) utility function of the form

$$u(c) = \frac{c^{1-\varrho} - 1}{1-\varrho}, \quad c \geq 0,$$

and set $\varrho = 2$, following standard practice in the macroeconomics literature (see, e.g., Mendoza, 1991; Aguiar and Gopinath, 2007).

6.2 Indirect Inference

Conditional on the calibrated parameters, we then estimate the remaining key parameters using indirect inference. These include: (i) the inverse distance elasticity of medical travel (κ); (ii) the parameters governing the relationship between travel disutility (τ_{klt}) and travel distance; and (iii) the parameters linking recovery and mortality rates ($\{\pi_{lt}^{s,0}, \pi_{lt}^{s,-1}\}$) to medical resources and patient volume. We specify the functional form of τ_{klt} and $\{\pi_{lt}^{s,0}, \pi_{lt}^{s,-1}\}$ below.

6.2.1 Parameterization

Travel Disutility. We parameterize travel disutility as

$$\tau_{klt} = \begin{cases} \exp\left(\eta_0 + \eta_1 D_{klt}\right), & \text{if } l \neq k \\ 0 & \text{if } l = k \end{cases}, \tag{13}$$

where D_{klt} is the observed passenger travel time between cities k and l at time t that we take from Ma and Tang (2024). The parameters $\{\eta_0, \eta_1\}$ determine how travel disutility scales with intercity distance. They play a central role in quantifying the deterrent effect of travel time on medical travel decisions.

Recovery and Mortality Rates. The probability of recovery or death for a patient seeking care in city l at time t depends on both the quantity of medical resources m_{lt} and congestion of patient p_{lt} in l. We parameterize the relationships using a multinomial logistic specification:

$$\pi_{lt}^{s,0} = \frac{\exp\left[\gamma^{s,0}(m_{lt}, p_{lt})\right]}{1 + \exp\left[\gamma^{s,0}(m_{lt}, p_{lt})\right] + \exp\left[\gamma^{s,-1}(m_{lt}, p_{lt})\right]}, \tag{14}$$

$$\pi_{lt}^{s,-1} = \frac{\exp\left[\gamma^{s,-1}(m_{lt}, p_{lt})\right]}{1 + \exp\left[\gamma^{s,0}(m_{lt}, p_{lt})\right] + \exp\left[\gamma^{s,-1}(m_{lt}, p_{lt})\right]}, \tag{15}$$

where $\gamma^{s,0}(\cdot)$ and $\gamma^{s,-1}(\cdot)$ are log-linear in m_{lt} and p_{lt}:

$$\gamma^{s,0}(m_{lt}, p_{lt}) = \gamma_1^{sH} + \gamma_2^{sH} \log(m_{lt}) + \gamma_3^{sH} \log(p_{lt}), \tag{16}$$

$$\gamma^{s,-1}(m_{lt}, p_{lt}) = \gamma_1^{sD} + \gamma_2^{sD} \log(m_{lt}) + \gamma_3^{sD} \log(p_{lt}). \tag{17}$$

The above specifications allow medical resources and patient volume to impact treatment outcomes in a nonlinear way. We measure m_{lt} as the total number of beds in tertiary hospitals, which is observed directly from the hospital annual reports. The patient volume p_{lt} is endogenously determined in the model (Eq. (9)).

6.2.2 Parameter Identification and Estimation

κ and $\{\eta_0, \eta_1\}$

We identify these parameters through two auxiliary regressions derived from the model-implied gravity equation. Taking logs on both sides of Eq. (10), we arrive at the gravity equation of estimation:

$$\log \mu_{klt}^s(z_t) = \frac{\beta}{\kappa} W_{kl,t+1}^s(z_t) + \frac{1}{\kappa} \underbrace{u\left(\delta^s w_{kt} z_t - \mathbb{1}(l \neq k) \cdot \lambda\right)}_{\text{extensive margin}} - \frac{1}{\kappa} \underbrace{\exp\left(\eta_0 + \eta_1 D_{klt}\right)}_{\text{intensive margin}} - \Phi_{kt}^s(z_t). \tag{18}$$

In the above equation, the term $\delta^s w_{kt} z_{it} - \mathbb{1}(l \neq k)\lambda$ determines whether the individual seek care outside the home city—extensive margin of medical travel—since out-of-city care is feasible only when disposable income after paying the fixed cost remains positive. Conditional on this feasibility, the distance-related disutility between cities k and l, $\tau_{klt} = \exp(\eta_0 + \eta_1 D_{klt})$, affects the probability of choosing destination city l for hospital care—the intensive margin.

According to Eq. (18), variation in the extensive margin of medical travel induced by income conditional on disease severity helps identify κ.[19] Based on this observation, we

[19] In dynamic equilibrium, the value functions summarized by the W and the Φ affect the choice probability. As both terms are also functions of severity, we control for severity in the auxiliary regressions to alleviate concerns of omitted variable bias.

specify **the first auxiliary regression** in indirect inference, which is an individual-level linear probability model that relates the probability of seeking out-of-city care to income and severity:

$$\mathbf{1}(l \neq k|s)_i = \alpha_1^{\text{out}} + \alpha_2^{\text{out}} \log\left(\text{Income}_{ik}\right) + \alpha_3^{\text{out}} s_{ik} + \alpha_4^{\text{out}} \log\left(\text{Income}_{ik}\right) s_{ik} + \nu_{ik}, \quad (19)$$

where the dependent variable equals 1 if individual i in city k seeks out-of-city care and 0 otherwise. We use the coefficients $\{\alpha_1^{\text{out}}, \alpha_2^{\text{out}}, \alpha_3^{\text{out}}, \alpha_4^{\text{out}}\}$ as moment conditions to identify κ . The regression results are reported in Online Appendix Table A2.

Conditional on identifying κ, variation in the intensive margin with travel time (D_{klt}) helps identify η_0 and η_1. We therefore derive **the second auxiliary regression**, which is the reduced-form gravity equation that captures the relationship between patient flows to travel time:

$$\log \mu_{kl} = \alpha_1^{\text{pair}} + \alpha_2^{\text{pair}} \log\left(\text{Travel Time}_{kl}\right) + \alpha_3^{\text{pair}} \frac{m_l}{m_k} + \alpha_4^{\text{pair}} \frac{p_l}{p_k} + \text{FE}_l + \text{FE}_k + \nu_{kl}, \quad (20)$$

where μ_{kl} is the share of patients from city k seeking hospital care in city l. The regression is conducted at the city-pair level. We control for the relative medical resources and patient volume between the destination and the home cities, as well as the fixed effects for both cities. The coefficients $\{\alpha_1^{\text{pair}}, \alpha_2^{\text{pair}}, \alpha_3^{\text{pair}}, \alpha_4^{\text{pair}}\}$ serve as the moment conditions to identify $\{\eta_0, \eta_1\}$. The regression results are reported in Online Appendix Table A3.

$\gamma_{(\cdot)}^{sH}$ and $\gamma_{(\cdot)}^{sD}$

As shown in Eqs. (14)-(17), the parameters of $\gamma_{(\cdot)}^{sH}$ and $\gamma_{(\cdot)}^{sD}$ map the destination-specific treatment outcomes to medical resources (m_{lt}) and patient volume (p_{lt}) in the destination city. To capture these relationships empirically, we estimate **the third set of auxiliary regressions** at the individual level:

$$\mathbf{1}\left(\bar{s} = 0|s, l\right)_i = \alpha_1^{sH} + \alpha_2^{sH} \log\left(m_l\right) + \alpha_3^{sH} \log\left(p_l\right) + \nu_{il}^{sH} \quad (21)$$

$$\mathbf{1}\left(\bar{s} = -1|s, l\right)_i = \alpha_1^{sD} + \alpha_2^{sD} \log\left(m_l\right) + \alpha_3^{sD} \log\left(p_l\right) + \nu_{il}^{sD}. \quad (22)$$

In Eq. (21), the dependent variable equals one if patient i with severity $s(s > 0)$ recovers after receiving care in city l, and zero otherwise; in Eq. (22), it equals one if the patient passes away and zero otherwise. The regressions are estimated separately for each severity level $s(s > 0)$. We employ the coefficients $\{\alpha_1^{sH}, \alpha_2^{sH}, \alpha_3^{sH}\}$ and $\{\alpha_1^{sD}, \alpha_2^{sD}, \alpha_3^{sD}\}$ as the moment conditions in indirect inference. Corresponding regression results are reported in Online Appendix Table A4.

In total, we have 15 $(= 3 + 3 \times 2 + 3 \times 2)$ parameters to be estimated, collected in $\Theta = \{\kappa, \eta_0, \eta_1, \gamma_{(\cdot)}^{sH}, \gamma_{(\cdot)}^{sD}\}$, and we use 20 $(= 4 + 4 + 3 \times 2 + 3 \times 2)$ moment conditions for indirect inference. Let $\mathbf{A} = \{\alpha_{(\cdot)}^{\text{out}}, \alpha_{(\cdot)}^{\text{pair}}, \alpha_{(\cdot)}^{sH}, \alpha_{(\cdot)}^{sD}\}$ denote the vector of coefficients in the auxiliary regressions, and let $\widetilde{\mathbf{A}}_r(\Theta)$ denote the corresponding coefficients obtained from the r-th simulation, with $r = 1, \ldots, R$. The indirect inference estimator solves the following minimization problem:

$$\min_{\Theta} \left[\mathbf{A} - \frac{1}{R} \sum_{r=1}^{R} \widetilde{\mathbf{A}}_r(\Theta) \right]' \mathbf{W} \left[\mathbf{A} - \frac{1}{R} \sum_{r=1}^{R} \widetilde{\mathbf{A}}_r(\Theta) \right], \tag{23}$$

where $\mathbf{W}$ is the diagonal weighting matrix in which the ith diagonal equals the inverse of the squared standard errors of the ith elements in $\mathbf{A}$. In our implementation, we draw and simulate the same number of patients as in the reduced-form estimation, and set $R = 100$. Appendix C.3 provides more technical details on the estimation procedure.

6.2.3 Estimation Results

The estimation yields precise structural parameter estimates, as summarized in Table 2. Figure (A2) in the Online Appendix assesses the model's fit, showing that the targeted moments are matched closely overall.

The results indicate that CCVD patients are highly sensitive to travel disutility in medical travel. The estimate of $\kappa = 0.249$ implies a semi-elasticity of travel probability with respect to travel disutility (τ_{klt}) of $-1/\kappa \approx -4$ (see Eq. (18)). Under this elasticity, increasing τ_{klt} from the 25th to the 75th percentile of the 2010 transportation network would reduce the probability of out-of-city medical travel to essentially zero.[20] This

[20] On the 2010 transportation network, the 25th percentile of τ_{kl} across all city pairs is 3.15 and the

Table 2: Estimation Results

name	value	s.e.	note
κ	0.249^{***}	0.030	inverse travel elasticity
η_0	0.062	0.052	travel cost function, intercept
η_1	0.144^{***}	0.031	travel cost function, slope
γ_1^{1H}	1.735^{***}	0.063	recovery rate, intercept, $s=1$
γ_2^{1H}	0.133^{***}	0.022	recovery rate, slope on m_l, $s=1$
γ_3^{1H}	-0.288^{***}	0.050	recovery rate, slope on p_l, $s=1$
γ_1^{2H}	0.958^{***}	0.071	recovery rate, intercept, $s=2$
γ_2^{2H}	0.102^{***}	0.027	recovery rate, slope on m_l, $s=2$
γ_3^{2H}	-0.314^{***}	0.076	recovery rate, slope on p_l, $s=2$
γ_1^{1D}	-3.088^{**}	1.543	mortality rate, intercept, $s=1$
γ_2^{1D}	-2.488^{***}	0.868	mortality rate, slope on m_l, $s=1$
γ_3^{1D}	0.749	0.942	mortality rate, slope on p_l, $s=1$
γ_1^{2D}	-4.113^{***}	0.374	mortality rate, intercept, $s=2$
γ_2^{2D}	-1.044^{*}	0.612	mortality rate, slope on m_l, $s=2$
γ_3^{2D}	-1.102^{***}	0.150	mortality rate, slope on p_l, $s=2$

Notes: This table reports the results of the indirect inference. The asymptotic standard errors are computed using formulas reported in Online Appendix C.3. $1/\kappa$ is the elasticity of travel probability with respect to travel costs. $\{\eta_0, \eta_1\}$ are the parameters of the travel cost function. $\{\gamma_{(.)}^{sH}\}$ are the parameters governing the recovery rate as a function of medical resources (m_l) and patient volume (p_l). Similarly, $\{\gamma_{(.)}^{sD}\}$ are the parameters governing the mortality rate as a function of the same variables.

estimated elasticity is higher than the estimates of labor migration elasticity documented in the literature.[21] This is because we focus on CCVD patients. They face steep increases mental and physical discomfort from additional travel time, and many CCVD conditions progress rapidly, making their travel decisions very sensitive to travel time.

The remaining parameter estimates in Table 2 align well with intuition and established medical patterns. Travel disutility increases with travel time ($\eta_1 > 0$). Better medical resources raise recovery rates ($\gamma_2^{1H} > 0$ and $\gamma_2^{2H} > 0$) and reduce mortality rates ($\gamma_2^{1D} < 0$ and $\gamma_2^{2D} < 0$) for both less severe ($s = 1$) and severe ($s = 2$) CCVD cases. We also find negative impacts of patient congestion on recovery: conditional on medical resources,

75th percentile is 11.50, implying $\Delta \log \mu_{kl} \approx -(1/\kappa)(11.50 - 3.15) = -33.53$. This change in $\log \mu_{kl}$ corresponds to multiplying μ_{kl} by $e^{-33.53} \approx 2.8 \times 10^{-15}$, effectively driving the probability of out-of-city medical travel to zero.

[21] For example, Tombe and Zhu (2019) estimated the labor migration elasticity to be around 1.5 using the Chinese data; Stillwell et al. (2014)'s estimation based on the European data is between 1.4 and 2.2; Bryan and Morten (2019) estimated the elasticity to be 2.7 using Indonesian data.

higher patient volumes reduce the recovery rate ($\gamma_3^{1H} < 0$ and $\gamma_3^{2H} < 0$). The impacts on mortality are more nuanced. For less severe cases, higher patient volumes tend to increase mortality ($\gamma_3^{1D} > 0$), though the effect is statistically insignificant. In contrast, for severe CCVD cases, greater patient volumes are associated with lower mortality ($\gamma_3^{2D} < 0$), which may reflect learning-by-doing or agglomeration effects in specialized care, or the prioritization of treating the most severe patients in hospitals.

A comparison between the model estimates of $\gamma_{(\cdot)}^{sH}$ and $\gamma_{(\cdot)}^{sD}$ and the estimated coefficients ($\alpha_{(\cdot)}^{sH}$ and $\alpha_{(\cdot)}^{sD}$) in the third set of auxiliary regressions (Online Appendix Table A4) shows that their signs do not necessarily coincide, even though both sets of parameters relate treatment outcomes to medical resources and patient volumes. For instance, our model estimates indicate that greater patient congestion increases mortality rates for less severe cases ($\gamma_3^{1D} > 0$), whereas the auxiliary regression in Eq. (22) shows a negative correlation between patient volumes and mortality ($\alpha_3^{1D} < 0$). This discrepancy arises because the auxiliary regressions capture only the correlations between patient volumes and treatment outcomes in equilibrium, not the causal effect of congestion. Patient volume p_{lt} affects treatment quality, but it is also endogenously determined by individuals' city choices, which in turn depend on the treatment quality of each destination city. Our model explicitly accounts for this endogenous determination of p_{lt}, allowing us to separate the causal effects of patient volumes on treatment outcomes from the sorting of patients across cities. These causal effects are captured by γ_3^{sH} and γ_3^{sD}.

7 Quantitative Results

Based on these structural parameters, we can quantitatively evaluate the impacts of transportation networks on healthcare outcomes through counterfactual analysis.

We extend the counterfactual analysis to all cities in China. Implementing the analysis does not require individual-level data; instead, based on our model, we only need four sources of information: 1) the city-level population, 2) the city-level average income, 3) the city-level medical resources, and 4) the travel distance between cities. We obtain

population data from the *Chinese Population Census*, income data from the *China City Statistical Yearbooks*, the number of tertiary hospital beds from the *China Health Commission*, and transportation networks from Ma and Tang (2024). The counterfactual sample includes 263 cities, the largest common set across all four data sources. This sample accounts for roughly 96 percent of the national population and 98 percent of economic output between 2010 and 2018. Figure A3 in the Online Appendix maps the cities included in our counterfactual analysis.

Our analysis examines the impacts of changes in transportation networks between 2010 and 2018, during which the average travel time across all city pairs declined from 14.6 to 9.5 hours. We conduct two complementary exercises. In the first, which we refer to as the "2010 base", we solve the baseline steady state using all the information in 2010. We then compute a counterfactual steady state in which only the transportation network is updated to its 2018 configuration, while all other inputs—including city-level population and medical resources, as well as individuals' income levels—remain at their 2010 values. The differences between the counterfactual and the baseline equilibrium inform us about the impacts of better transportation connection if all the other conditions were held at the 2010 levels. In the second exercise, denoted as "2018 base", we first compute a steady state in which the transportation network is fixed at its 2010 configuration, while all other inputs take their 2018 values. We then compute a steady state using all the fundamental variables observed in 2018. The difference between these two equilibria evaluates the impact of the same transportation improvements conditional on the 2018 fundamentals. In the rest of the section, we present the results of both exercises side-by-side.

Expected Mortality. Throughout the counterfactual analyses, our key variable of interest is the expected mortality rate for CCVD patients. For sick individuals with productivity z residing in city k, we define their expected mortality rate, denoted by ϖ_{kz}, as

$$\varpi_{kz} = \frac{\sum_{s=1}^{S}\sum_{l=1}^{K}\mu_{kl}^{s}(z)\pi_{l}^{s,-1}L_{k}^{s}(z)}{\sum_{s=1}^{S}L_{k}^{s}(z)}. \tag{24}$$

The numerator in this equation is the expected number of deaths among all sick individuals with productivity z in city k, considering the equilibrium patient flows to all possible destination cities. The denominator is the total number of sick individuals in the city. This expected mortality rate ϖ_{kz} summarizes how transportation networks affect health outcomes. By altering travel times, changes in transportation networks modify patients' destination choices ($\mu_{kl}^{s}(z)$); these choice adjustments, in turn, affect both the mortality risks patients face at different treatment locations ($\pi_{l}^{s,-1}$) and the evolution of the sick population in city k ($L_{k}^{s}(z)$). Therefore, ϖ_{kz} serves as a sufficient statistic for evaluating the equilibrium impacts of counterfactual changes in transportation networks on mortality.

Based on the expected mortality rate ϖ_{kz} for each city-productivity group, we define the national average mortality rate for CCVD patients as

$$\overline{\varpi} = \sum_{k} \sum_{z} \xi_{kz} \varpi_{kz}, \tag{25}$$

where the weight assigned to group (k, z) is its share in the total patient population across cities,

$$\xi_{kz} = \frac{\sum_{s=1}^{S} L_{k}^{s}(z)}{\sum_{k',z',s'} L_{k'}^{s'}(z')}.$$

We further define the average expected mortality rate for city k (averaged over productivity levels) as

$$\overline{\varpi}_{k} = \frac{1}{\xi_{k}} \sum_{z} \xi_{kz} \, \varpi_{kz},$$

where $\xi_{k} = \sum_{z} \xi_{kz}$ is the marginal weight for city k. Similarly, the average expected mortality rate for productivity group z (averaged over all cities) is

$$\overline{\varpi}_{z} = \frac{1}{\xi_{z}} \sum_{k} \xi_{kz} \, \varpi_{kz},$$

where $\xi_{z} = \sum_{k} \xi_{kz}$ is the marginal weight for individual with productivity z. These measures allow us to investigate the health impacts of transportation networks nationally, by city, and by income group.

7.1 Aggregate Impacts

Improvements in connectivity substantially reduce the expected national mortality rate by enabling more patients to receive out-of-city care. Table 3 summarizes the results. The first row reports the results for the "2010 base". Upgrading the transportation network from its 2010 to 2018 configuration lowers the national expected mortality rate ($\overline{\varpi}$) by 3.08 percent, equivalent to around 9.88 thousand lives saved per year. Using a value of statistical life of 4.76 million RMB, this amounts to approximately 47 billion RMB in 2019 nominal terms.[22]

To benchmark this effect, we compare it with the health gain from expanding medical resources over the same period. From 2010 to 2018, the average number of tertiary hospital beds per city increases from 3,930 to 9,209.[23] Holding the transportation network and all other fundamentals at their 2010 levels, our simulation shows that increasing the number of tertiary hospital beds in each city from their 2010 to 2018 levels would save approximately 230.8 thousand lives per year.Thus, improved access to existing medical resources via better connectivity accounts for about $9.88/230.8 \approx 4.3\%$ of the mortality reduction achieved by expanding resources. Given that these health gains arise as unintended benefits of infrastructure investment, their magnitude is economically meaningful.

To understand the total mortality decline driven by transportation improvements, we further examine mortality changes for three patient groups. As shown in Table 3 Column (3), the largest decline comes from "induced travelers"—patients who switch from home-city care to out-of-city care due to improved connectivity.[24] By gaining access to cities with better treatment quality, mortality reduction of this group amounts to 9.01 thousand lives saved per year in the "2010 base". For "never-travelers"—patients who consistently receive care in their home city under both the 2010 and 2018 transportation networks—the mortality change depends on how local patient volumes shift. Cities with net patient outflows see reduced patient congestion and lower mortality, while those

[22] Our model is calibrated at a monthly frequency; yearly results are obtained by multiplying monthly impacts by 12. The value of statistical life is based on estimates from Cao et al. (2023).

[23] As shown in Appendix Figure A4, tertiary hospital bed capacity increases in all cities between the two years, and the increases are larger in cities that had fewer beds initially.

[24] Our simulation results suggest that transportation improvements increase out-of-city patient flows by roughly 75 percent.

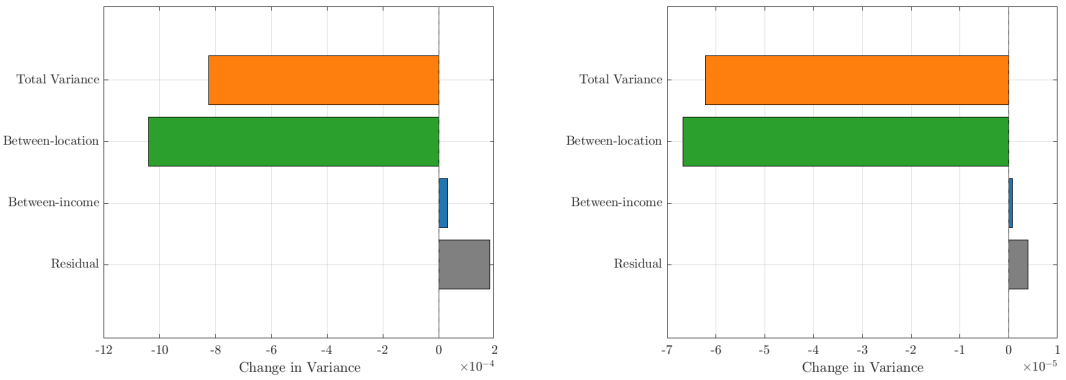
Table 3: The Aggregate Impacts of Transportation on Mortality

	%Δ Mortality	Δ Mortality (thousands)				Δ VSL
		Total	Induced-traveler	Never-traveler	Always-traveler	(billion ¥)
	(1)	(2)	(3)	(4)	(5)	(6)
Reduce τ, 2010 base	-3.08	-9.88	-9.01	-0.85	-0.01	47.01
Reduce τ, 2018 base	-2.01	-1.82	-2.17	0.35	0.00	8.65

Notes: This table reports the impacts of transportation network expansion on aggregate mortality. Each row represents the result of a counterfactual exercise. Column (1) reports the percentage changes in total mortality. Columns (2) to (5) report the changes in mortality level in one year. Patients who seek care out-of-city only when transportation networks are improved are labeled as "induced-traveler". Patients who always choose to receive hospital care in their home city in both the baseline and the counterfactual equilibria are referred to as "never-traveler". Patients who always seek hospital care out-of-city are "always-traveler". Column (6) reports the "value of statistical life" (VSL) of the yearly changes in total mortality (Column (2)) in the unit of billion Chinese RMB.

receiving net inflows face increased congestion and consequently higher mortality. On net, "never-travelers" experience a reduction of 0.85 thousand deaths per year (Column (4)). Similarly, for "always-travelers"—patients who always seek out-of-city care regardless of transportation improvements—the mortality change is determined by congestion changes in their destination cities. In the "2010 base", their mortality declines slightly by 0.01 thousand deaths per year (Column (5)).

As shown in the second row of Table 3, the health gains from improved connectivity are much smaller in the "2018 base": the same transportation improvements reduce $\overline{\varpi}$ by only 2.01 percent, corresponding to 1.82 thousand fewer deaths per year—just 18% of the effect estimated in the "2010 base." A key reason for this attenuation is the expansion of medical resources between 2010 and 2018. As the number of tertiary beds increases nationwide, especially in cities that were initially poorly endowed (see Appendix Figure A4), patients are more likely to receive high-quality care locally. Consequently, the demand for medical travel falls, moderating the mortality-reducing effects of better connectivity.



(a) 2010 Base (b) 2018 Base

Figure 6: Decomposition of the Changes in the Variance of Expected Mortality Rates

Notes: This figure decomposes the change in the variance of mortality rate before and after the improvement in transportation networks into three additive components: between-location variance, between-income variance, and a residual interaction term, according to the decomposition formula in Eq. (26). The "Total variance" bar represents the overall change in variance.

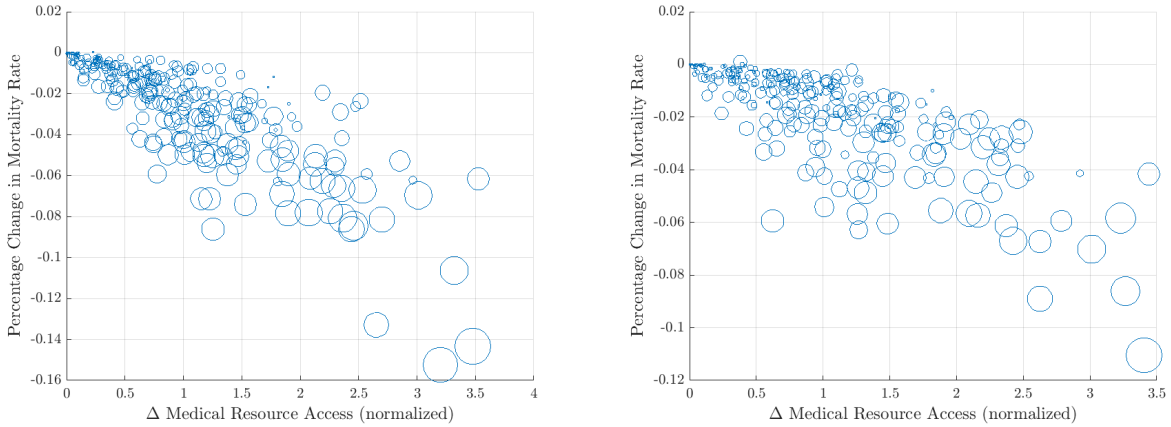
7.2 Distributional Effects

In addition to the aggregate impacts, we study how improved connectivity alters the distribution of health outcomes across cities and across income groups. We measure the inequality in mortality using the variance of expected mortality rates, where a higher variance reflects greater disparity in mortality. The overall variance can be additively decomposed into three terms:

$$\begin{aligned}\text{Var}\left(\varpi\right) &\equiv \sum_k \sum_z \xi_{kz}\left(\varpi_{kz} - \overline{\varpi}\right)^2 \\ &= \underbrace{\sum_k \xi_k(\overline{\varpi}_k - \overline{\varpi})^2}_{\text{Between-city variance}} + \underbrace{\sum_z \xi_z(\overline{\varpi}_z - \overline{\varpi})^2}_{\text{Between-income variance}} + \underbrace{\sum_k \sum_z \xi_{kz}\left(\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}\right)^2}_{\text{Residual (interaction) variance}}.\end{aligned} \tag{26}$$

The first term measures variation in mortality across cities; the second measures variation across income groups; and the third is a residual component capturing the interaction between city-level and income-level variation. Appendix D.1 provides the decomposition details.

Figure 6 reports the changes in total mortality variance and its three components



(a) 2010 Base (b) 2018 Base

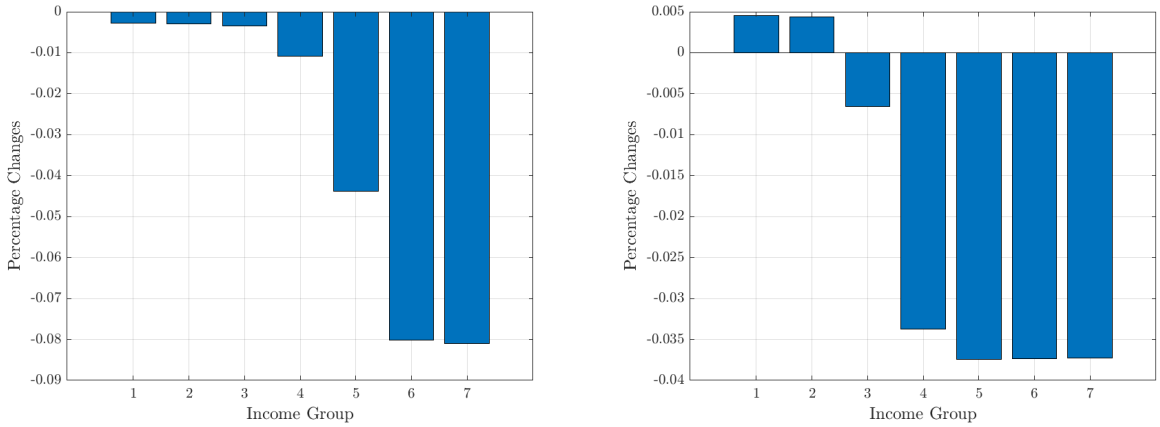
Figure 7: The Spatial Impacts of Transportation Networks on Mortality Rates

Notes: The horizontal axis is the normalized changes in medical access as defined in Eq. (12). The normalization enforces that the average changes across cities is equal to unity. The vertical axes in both panels are the percentage changes in mortality rate. The size of the circle represents the changes in out-of-city care probability: a larger circle indicates a larger increase in the probability of out-of-city care. Each dot represents a city.

when the transportation network improves from 2010 to 2018. The left panel reports results under the "2010 base" scenario, which shows that the total mortality variance declines by 8.2 basis units, or 5.4 percent. This overall reduction is driven primarily by a 10.4-basis-unit reduction in between-city variance, indicating spatial convergence in mortality. However, the between-income variance of mortality within cities rises slightly. Results under the "2018 base" scenario exhibit a similar pattern.

The reduction in between-city variance arises because the improved transportation network enables patients in initially underserved cities to access better medical resources outside their home cities. Figure 7 plots, for each city, the normalized change in medical access against the percentage change in mortality, with marker size indicating the increase in out-of-city care probability. The negative slope indicates that cities experiencing larger improvements in access also exhibit larger reductions in mortality, and the size gradient shows that these health gains are accompanied by greater out-of-city patient flows.

The between-income variance increases, because the health gains due to improved connectivity are strongly regressive: benefits are concentrated among higher-income individuals who are better able to overcome the financial barriers associated with seeking



(a) 2010 Base (b) 2018 Base

Figure 8: The Impacts of Transportation Network on Mortality Rate, by Income Groups

Notes: This figure shows the percentage changes in the expected mortality rates by income group due to the improvements in transportation networks from 2010 to 2018. The counterfactual simulation changes the transportation network to the year 2018 while keeping the other inputs the same as in the corresponding base year. The y-axis is $x'/x - 1$, where x' is the variable of interest in the counterfactual simulation, and x is the same variable in the corresponding base year.

out-of-city care.[25] Figure 8 reveals that the mortality reductions are far from evenly distributed across income groups.[26] Under the "2010 base" scenario, mortality among the highest-income group (Group 7) declined by more than 8 percent, well above the national average, while mortality for the poorest income groups remained virtually unchanged. Under the "2018 base" scenario, the disparity is even starker: mortality rates for low-income individuals increases slightly, likely because those who remain in cities attracting large inflows of medical "migrants" face increasingly congested hospital systems. These findings underscore the importance of complementing infrastructure investments with policies that improve healthcare affordability and capacity for low-income populations, to ensure that connectivity gains translate into equitable health outcomes.

[25]The evidence is consistent with findings in Dingel et al. (2023) that in the United States, patients with low socioeconomic status are less likely to travel longer distances to obtain high-quality healthcare.

[26]We discretize the income process into seven grid points using the methods introduced in Tauchen (1986). The choice of 7 grid points balances computational burden and approximation accuracy, consistent with common practice in the macroeconomics literature (e.g, Silos (2006); Heer and Maussner (2009))

8 Conclusion

This paper studies how the expansion of transportation infrastructure has affected healthcare access and health outcomes, focusing on mortality from CCVDs. Using comprehensive administrative records covering the universe of hospital admissions for CCVD patients in China's Sichuan province, we estimate a dynamic spatial model to quantify how medical resource availability, travel costs, and patient heterogeneity shape treatment-seeking behavior and subsequent health outcomes. We then combine the estimated model with national data on population, healthcare capacity, and transportation network development to conduct counterfactual analyses, assessing the extent to which transportation infrastructure improvements have contributed to changes in CCVD mortality nationwide.

Our analysis yields the following key insights. First, improvements in transportation infrastructure enable faster access to high-quality medical facilities from underserved areas, leading to reductions in mortality. We estimate that, conditional on a fixed distribution of medical resources in 2010, the expansion of China's transport network from 2010 to 2018 would have saved approximately 10,000 lives per year from CCVDs alone. Second, while better transport reduces spatial inequality in access to care, the benefits are distributed unevenly across income groups. High-income individuals, better able to overcome the fixed financial costs of out-of-city care, capture a disproportionate share of the gains. As a result, geographic convergence in health outcomes has coincided with widening income-related disparities.

These findings carry important policy implications. A long-standing policy debate concerns the potential misallocation of medical resources, stemming from the unequal distribution of healthcare capacity and the resulting disparities in access and outcomes (Finkelstein et al., 2021). Expansion of the transport network offers an efficient way to facilitate the "export" of high-quality hospital care to underserved areas while preserving scale economies in the healthcare sector (Trogdon, 2009; Dingel et al., 2023). We quantify the role of transportation networks in China, while emphasizing that transportation infrastructure and healthcare policy are complementary levers: without financial mechanisms that reduce the costs of traveling for care, improved connectivity may reinforce

existing socioeconomic inequalities. Targeted interventions, such as insurance reforms and means-tested travel subsidies, could help ensure that the health benefits of improved connectivity are more broadly shared.

References

Aguiar, M. and Gopinath, G. (2007). Emerging market business cycles: The cycle is the trend. *Journal of Political Economy*, 115(1):69–102.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189.

Allen, T. and Arkolakis, C. (2014). Trade and the topography of the spatial economy. *The Quarterly Journal of Economics*, 1085:1139.

Allen, T. and Arkolakis, C. (2022). The Welfare Effects of Transportation Infrastructure Improvements. *The Review of Economic Studies*.

Andersson, D., Berger, T., and Prawitz, E. (2023). Making a market: Infrastructure, integration, and the rise of innovation. *Review of Economics and Statistics*, 105(2):258–274.

Asher, S. and Novosad, P. (2020). Rural roads and local economic development. *American Economic Review*, 110(3):797–823.

Banerjee, A., Duflo, E., and Qian, N. (2020). On the road: Access to transportation infrastructure and economic growth in China. *Journal of Development Economics*, 145:102442.

Bryan, G. and Morten, M. (2019). The aggregate productivity effects of internal migration: Evidence from indonesia. *Journal of Political Economy*, 127(5):2229–2268.

Burns, L. R. and Huang, Y. (2017). History of China's healthcare system. In Burns, L. R. and Liu, G. G., editors, *China's Healthcare System and Reform*. Cambridge University Press.

Caliendo, L., Dvorkin, M., and Parro, F. (2019). Trade and labor market dynamics: General equilibrium analysis of the china trade shock. *Econometrica*, pages 741–835.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Cao, C., Song, X., Cai, W., Li, Y., Cong, J., Yu, X., Gao, M., and Wang, C. (2023). Estimating the value of a statistical life in china: A contingent valuation study in six representative cities. *Chinese Journal of Population, Resources and Environment*, 21(4):269–278.

Conrad, N., Molenberghs, G., Verbeke, G., Zaccardi, F., Lawson, C., Friday, J. M., Su, H., Jhund, P. S., Sattar, N., Rahimi, K., Cleland, J. G., Khunti, K., Budts, W., and McMurray, J. J. V. (2024). Trends in cardiovascular disease incidence among 22 million people in the UK over 20 years: Population based study. *BMJ*, 385.

Dingel, J. I., Gottlieb, J. D., Lozinski, M., and Mourot, P. (2023). Market size and trade in medical services. NBER Working Paper 31030, National Bureau of Economic Research.

Donaldson, D. (2018). Railroads of the Raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934.

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

Egger, P. H., Loumeau, G., and Loumeau, N. (2023). China's dazzling transport-infrastructure growth: Measurement and effects. *Journal of International Economics*, 142:103734.

Faber, B. (2014). Trade integration, market size, and industrialization: Evidence from China's national trunk highway system. *The Review of Economic Studies*, 81(3):1046–1070.

Fan, X., Song, Z., and Wang, Y. (2010). Estimating income processes in China. Working Paper 202, University of Zurich, Department of Economics, Center for Institutions, Policy and Culture in the Development Process.

Fang, H., Wang, L., and Yang, Y. (2020). Human mobility restrictions and the spread of the Novel Coronavirus (2019-nCoV) in China. *Journal of Public Economics*, 191:104272.

Fang, H., Wang, L., and Yang, Y. (2025). Competition and quality: Evidence from high-speed railways and airlines. *The Review of Economics and Statistics*, 107(2):494–509.

Fay, M., Lee, H. I., Mastruzzi, M., Han, S., and Cho, M. (2019). Hitting the trillion mark: A look at how much countries are spending on infrastructure. Policy Research Working Paper 8730, World Bank.

Finkelstein, A., Gentzkow, M., and Williams, H. (2021). Place-based drivers of mortality: Evidence from migration. *American Economic Review*, 111(8):2697–2735.

Hackmann, M. B. (2019). Incentivizing better quality of care: The role of Medicaid and competition in the nursing home industry. *American Economic Review*, 109(5):1684–1716.

Heer, B. and Maussner, A. (2009). *Dynamic General Equilibrium Modeling: Computational Methods and Applications*. Springer, 2nd edition.

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

Ho, K. and Pakes, A. (2014). Hospital choices, hospital prices, and financial incentives to physicians. *American Economic Review*, 104(12):3841–3884.

Kleinman, B., Liu, E., and Redding, S. J. (2023). Dynamic spatial general equilibrium. *Econometrica*, 91(2):385–424.

Li, J. (2014). The influence of state policy and proximity to medical services on health outcomes. *Journal of Urban Economics*, 80:97–109.

Lin, Y. (2017). Travel costs and urban specialization patterns: Evidence from China's high speed railway system. *Journal of Urban Economics*, 98:98 – 123.

Liu, J., Xu, A., Zhao, Z., Ren, B., Gao, Z., Fang, D., Hei, B., Sun, J., Bao, X., Ma, L., et al. (2025). Epidemiology and future trend predictions of ischemic stroke based on the global burden of disease study 1990–2021. *Communications Medicine*, 5(1):273.

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

Mendoza, E. G. (1991). Real business cycles in a small open economy. *American Economic Review*, 81(4):797–818.

Milcent, C. (2018). *Healthcare Reform in China: From Violence to Digital Healthcare*. Springer.

National Health Commission (2019). *China Health Statistical Yearbook 2019 (Chinese Edition)*. Peking Union Medical College Press.

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

Silos, P. (2006). Assessing Markov chain approximations: A minimal econometric approach. *Journal of Economic Dynamics & Control*, 30(6):1063–1079.

Stillwell, J., Daras, K., Bell, M., and Lomax, N. (2014). The image studio: A tool for internal migration analysis and modelling. *Applied Spatial Analysis and Policy*, 7(1):5–23.

Tauchen, G. (1986). Finite state Markov-chain approximations to univariate and vector autoregressions. *Economics Letters*, 20(2):177–181.

Tombe, T. and Zhu, X. (2019). Trade, migration, and productivity: A quantitative analysis of China. *American Economic Review*, 109(5):1843–72.

Trogdon, J. G. (2009). Demand for and regulation of cardiac services. *International Economic Review*, 50(4):1183–1204.

Zhang, Q. (2011). Optimization of medical equipment procurement process. *China Medical Device Information (in Chinese)*, 5:3–5.

Zhang, X. and Kanbur, R. (2009). Spatial inequality in education and health care in China. *China Economic Review*, 16:189–204.

Online Appendix [For Online Publication]

Contents

A Additional Tables and Figures 2
B Model Details 9
B.1 Choice Set 9
B.2 Model Solution 9
B.3 Markov Representation of Health Transitions 11
B.4 Microfoundation for Treatment Outcomes 11
C Numerical Details on Simulation and Estimation 14
C.1 Algorithm for Solving the Steady State 14
C.2 Details of the Simulations 15
C.3 Details of the Indirect Inference 17
D Supplements to Counterfactual Analysis 18
D.1 Variance Decomposition 18

A Additional Tables and Figures

Table A1: Summary Statistics

	Mean	SD
Panel A. Admission-level variables		
Female	0.457	0.498
Monthly income	8,353.027	6,450.316
Severe	0.387	0.487
Recovery	0.926	0.261
Mortality	0.065	0.247
Seeking out-of-city treatment	0.044	0.205
Number of admissions	611,575	
Panel B. Hospital-level variables		
Tertiary hospitals		
Number of beds	853.665	676.303
Number of observations	227	
Secondary hospitals		
Number of beds	217.880	186.449
Number of observations	652	
Primary or ungraded institutions		
Number of beds	71.740	90.731
Number of observations	1,656	
Panel C. City-level variables		
Number of tertiary hospital beds (10,000)	0.923	1.060
Number of admissions with CCVD (10,000)	1.586	1.562
Population (10,000)	466.056	291.599
Number of observations	42	
Panel D. City-pair-level variables		
Travel time by road (hours)	4.389	2.932
Travel time by railway (hours)	7.855	6.951
Minimum travel time (hours)	4.272	2.924
Probability of medical travel between the city pair	0.048	0.197
Number of observations	882	

Notes: "Severe" is a dummy variable equal to 1 if the admission is labeled as "critical" or "urgent" at admission, and 0 otherwise. In particular, "critical" admissions often involve life-threatening situations such as acute myocardial infarction, respiratory failure, or severe stroke; "urgent" admissions include acute exacerbations of chronic illnesses or sudden onset conditions like high fever. "Seeking out-of-city treatment" equals 1 if the patient's residence is outside the hospital's city, and 0 otherwise. Recovery equals 1 if the admitted patient is not readmitted within 30 days after discharge, and 0 otherwise. Mortality equals 1 if the patient dies during this admission or within 30 days after discharge, and 0 otherwise.

Table A2: Auxiliary Regression Results (I)

	Seeking out-of-city treatment
ln(Income)	0.008***
	(0.000)
Severe	-0.008***
	(0.002)
Severe × ln(Income)	0.002***
	(0.000)
Constant	-0.024***
	(0.002)
Observations	611,575

Notes: This table reports the regression results of Eq. (19). "Seeking out-of-city treatment" equals 1 if the patient's residence is outside the hospital's city, and 0 otherwise. "Severe" is a dummy variable equal to 1 if the admission is labeled as "critical" or "urgent" at admission, and 0 otherwise. In particular, "critical" admissions often involve life-threatening situations such as acute myocardial infarction, respiratory failure, or severe stroke; "urgent" admissions include acute exacerbations of chronic illnesses or sudden onset conditions like high fever. Standard errors clustered at the individual level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Auxiliary Regression Results (II)

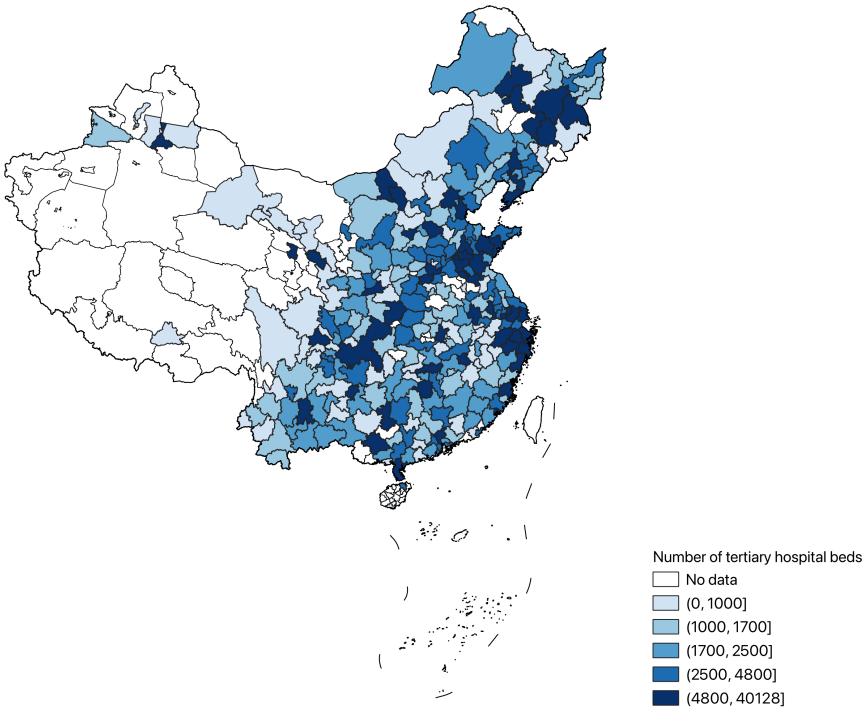
	ln(Share of admissions outside the home city)	
	(1)	(2)
ln(Minimum travel time)	-2.031***	-2.012***
	(0.278)	(0.279)
Diff. in tertiary hospital beds		0.042***
		(0.006)
Diff. in tertiary hospital patients		-0.017***
		(0.002)
Constant	-5.135***	-5.187***
	(0.312)	(0.318)
Origin-by-year FE	Yes	Yes
Destination-by-year FE	Yes	Yes
Observations	653	653

Notes: This table reports the regression results of Eq. (20). In Columns (2), we additionally control for the difference in numbers of tertiary hospital beds (and patients) between the patient's home city and the destination city, and this difference is measured as the ratio of the number in the destination city to that in the home city. Standard errors clustered at the home-city-by-destination-city level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

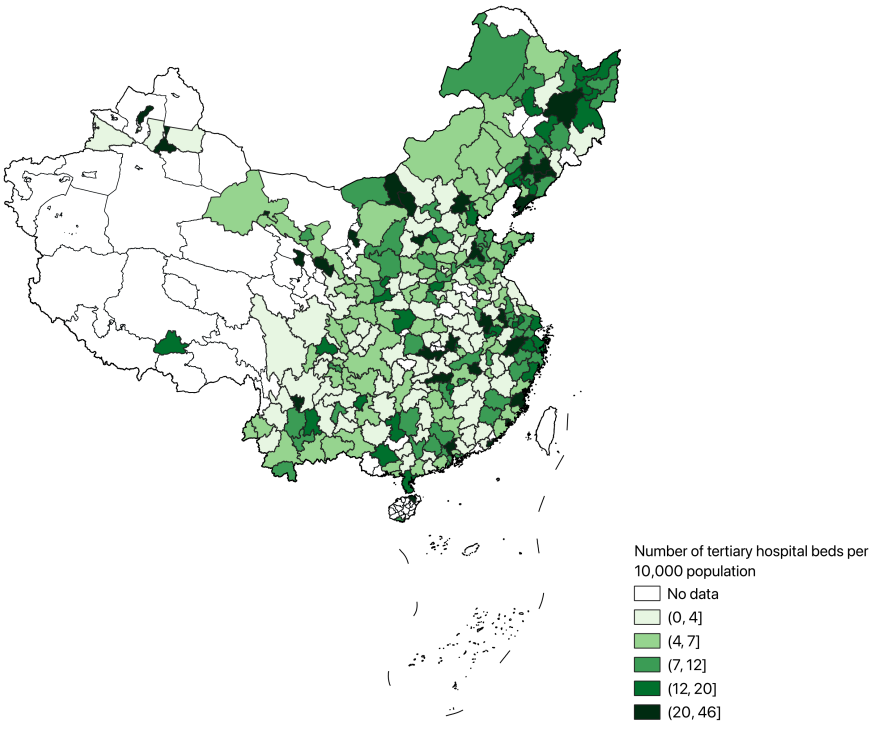
Table A4: Auxiliary Regression Results (III)

	Recovery		Mortality	
	(1)	(2)	(3)	(4)
ln(Number of tertiary beds)	0.005***	0.005***	0.006***	0.046***
	(0.002)	(0.002)	(0.001)	(0.002)
ln(Number of admissions with CCVD)	-0.013***	-0.009***	-0.012***	-0.071***
	(0.001)	(0.002)	(0.001)	(0.002)
Constant	0.879***	0.889***	-0.009**	-0.172***
	(0.006)	(0.007)	(0.004)	(0.007)
Severe	No	Yes	No	Yes
Observations	374,095	236,564	373,594	235,690

Notes: Columns (1)-(2) report the regression results of Eq. (21), and Columns (3)-(4) report the results of Eq. (22). Recovery equals 1 if the admitted patient is not readmitted within 30 days after discharge, and 0 otherwise. Mortality equals 1 if the patient dies during this admission or within 30 days after discharge, and 0 otherwise. Standard errors clustered at the individual level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.



(a) Number of Tertiary Hospital Beds by Cities



(b) Number of Tertiary Hospital Beds per 10,000 Population by Cities

Figure A1: Spatial Distribution of Medical Resources in 2010

Notes: The figure displays the number of tertiary hospital beds and that adjusted for population across cities in China in 2010. Data sources: 2011 Hospital Annual Report, China Health Commission. "No data" indicate cities that are excluded because at least one required variable is missing or cannot be reliably matched across sources. Administrative boundaries are shown at the prefecture level.

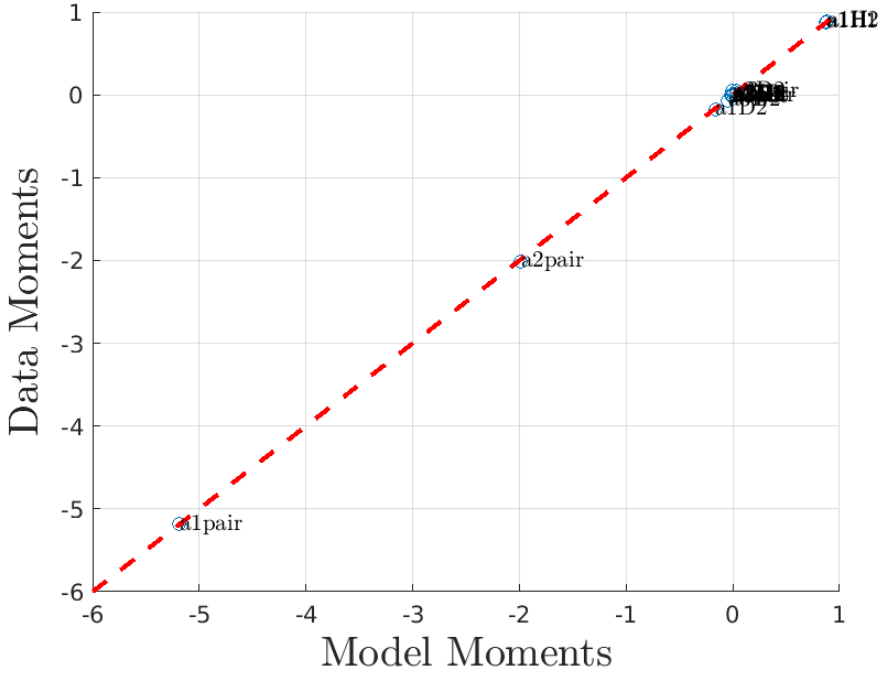


Figure A2: Goodness of Fit under Indirect Inference

Notes: This figure assesses goodness of fit under the indirect inference procedure. Each marker corresponds to a targeted moment condition used in estimation. The horizontal axis reports the model-implied moment computed from simulated data based on the estimated parameters, and the vertical axis reports the corresponding empirical moment from observed data. The red dashed line is the 45-degree line (perfect fit).

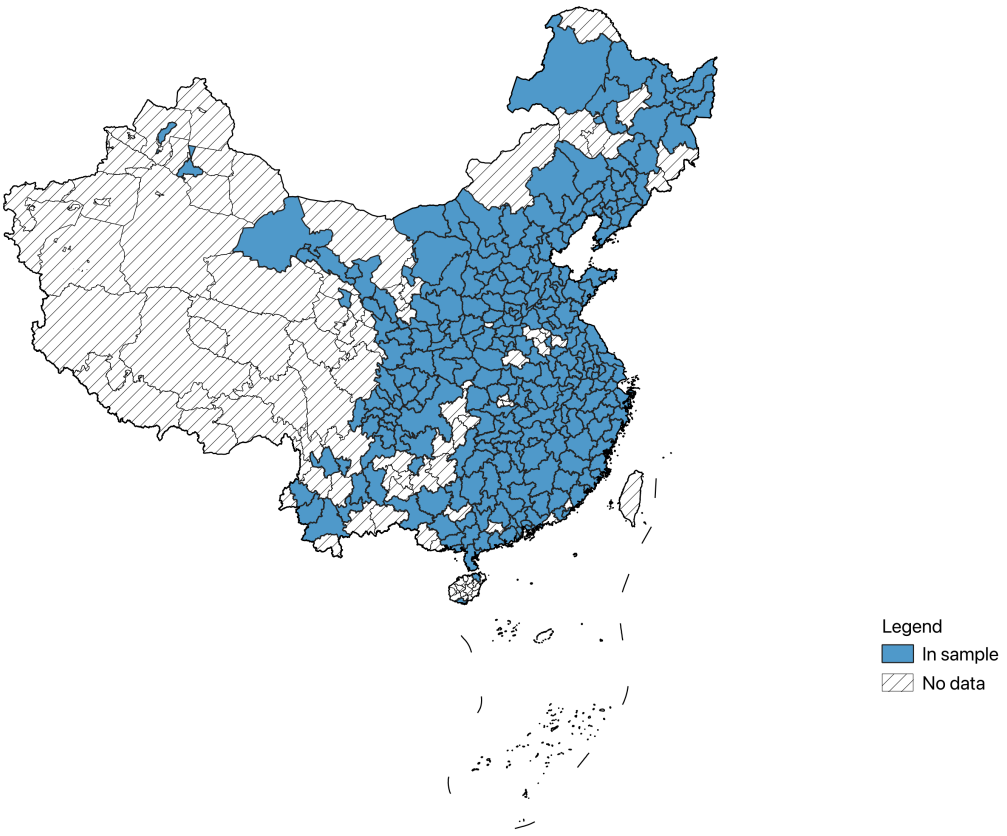


Figure A3: Geographic Coverage of the National City Sample

Notes: The map displays China's prefecture-level cities used in the national analysis sample. Shaded areas ("In sample") indicate cities for which we observe the full set of variables required for the national panel, constructed as the largest common set across the medical resource dataset and the *China City Statistical Yearbooks*. Hatched areas ("No data") indicate cities that are excluded because at least one required variable is missing or cannot be reliably matched across sources. Administrative boundaries are shown at the prefecture level.

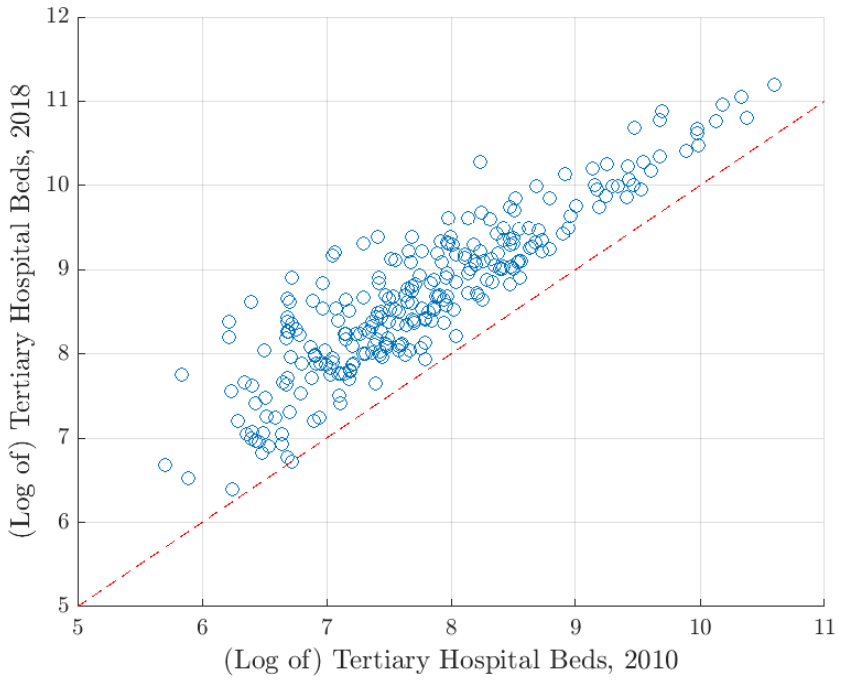


Figure A4: Changes in City-Level Tertiary Hospital Beds, 2010–2018

Notes: The figure plots, for each prefecture-level city in the national sample, the natural logarithm of tertiary-hospital bed counts in 2010 (x-axis) against the corresponding value in 2018 (y-axis). Each point represents one city. The red dashed line is the 45-degree line; points above (below) the line indicate cities in which tertiary-bed capacity increased (decreased) between 2010 and 2018. Bed counts are measured as the total number of beds in tertiary hospitals located in the city.

B Model Details

B.1 Choice Set

As discussed in the main text, the choice set of an individual, $\Omega^s_{kt}(z_{it})$, changes with home location, income, and health status. As is well-known in the literature of discrete choice, the cardinality of the choice set affects the expected value — a larger choice set implies a higher utility. Technically, the relationship arises because the expected value conditional on a location being chosen is strictly positive. This introduces a technical concern: because of this selection bias, sick individuals' value functions may be artificially inflated, leading to the paradoxical result that being sick appears more valuable than being healthy ($v^s_{kt}(z_{it}) > v^0_{kt}(z_t)$).

To address this, we subtract the conditional mean to eliminate the influence of extreme value shocks on value functions. In general, with K choices, the location parameter of $\max_{l=1}^{K}(\kappa\varepsilon_l)$ is $-\bar{\gamma}\kappa + \kappa \ln K$ and the scale parameter is κ. Therefore, the conditional mean is $\bar{\varepsilon} = \kappa \ln K$. As it will be clear later, the size of the choice set is not always K, and it depends on the home location, severity, and income of the patient; therefore, the conditional mean is not a constant but a function of $\{k, s, z\}$.[27]

The recursive formulation is then revised to be:

$$v^s_{kt}(z_{it}) = \max_l \Bigg\{ u(\delta^s w_{kt} z_{it} - \mathbb{1}(l \neq k) \cdot \lambda) - \tau_{klt} + \mathbf{E}\left[\beta\left[\pi^{Hs}_{lt} v^0_{kt}(z_{i,t+1}) + \left(1 - \pi^{Hs}_{lt} - \pi^{Ds}_{lt}\right) v^s_{kt}(z_{i,t+1})\right] + \kappa\varepsilon_l - \bar{\varepsilon}^s_{kt}(z_t)\right] \Bigg\}. \tag{A1}$$

The final term $\bar{\varepsilon}^s_{kt}(z_t)$ represents the inclusive value or expected maximum utility over all possible destination hospitals, integrating over the distribution of idiosyncratic shocks.

B.2 Model Solution

In this section, we derive the solution to a sick individual's expected value function, which is central for solving the model in steady state.

We define the expected value function for an individual, denoted by $V^s_{kt}(z_t)$, which captures the individual's maximum expected lifetime utility at time t. For a sick individual,

[27] Similar issues also exist in other dynamic discrete choice models, such as in ACM(2010), CDP(2019), and KLR(2023). In these cases, the conditional expectation of the extreme value shocks is innocuous because it only affects the level of the value function and none of the policy function. In our case, however, the relative value of sickness affects one's valuation of health and, subsequently, the value of treatment.

the expected value function is

$$V_{kt}^{s}(z_t) = \mathbf{E}_{\varepsilon}[v_{kt}^{s}(z_t)], \tag{A2}$$

where the expectation is taken with respect to preference shocks at time t. For a healthy individual, who simply consume current income and face future health and productivity shocks, the expected value function satisfies

$$V_{kt}^{0}(z_t) = u(w_{kt}z_t) + \beta \int_0^{\infty} \left[\left(1 - \sum_{s=1}^{S} \pi_{kt}^{0,s} \right) V_{k,t+1}^{0}(z_{t+1}) + \sum_{s=1}^{S} \pi_{kt}^{0,s} V_{k,t+1}^{s}(z_{t+1}) \right] dG(z_{t+1}|z_t). \tag{A3}$$

Given $V_{kt}^{s}(z_t)$, we can rewrite the expected continuation value, $W_{kl,t+1}^{s}(z_t)$ (defined in Eq. (11)) as follows:

$$\begin{aligned} W_{kl,t+1}^{s}(z_t) &= \int_0^{\infty} \mathbf{E}_{\varepsilon'} \left[\pi_{lt}^{Hs} v_{kt}^{0}(z_{t+1}) + \left(1 - \pi_{lt}^{Hs} - \pi_{lt}^{Ds} \right) v_{kt}^{s}(z_{t+1}) \right] dG(z_{t+1}|z_t) && \text{(A4)} \\ &= \int_0^{\infty} \left[\pi_{lt}^{s,0} V_{kt}^{0}(z_{t+1}) + \left(1 - \pi_{lt}^{s,0} - \pi_{lt}^{s,-1} \right) V_{kt}^{s}(z_{t+1}) \right] dG(z_{t+1}|z_t). && \text{(A5)} \end{aligned}$$

where the expectation $\mathbf{E}_{\varepsilon'}$ in the first line is only taken with respect to future preference shocks. $G(z_{t+1}|z_t)$ is the conditional cumulative distribution function for the productivity shock in the next period.

To derive the solution for $V_{kt}^{s}(z_t)$, we further define a term $\zeta_{klt}(z_t)$ as follows:

$$\zeta_{klt}(z_t) = u(\delta^s w_{kt} z_t - \mathbb{1}(l \neq k) \cdot \lambda) - \tau_{klt} + \beta W_{kl,t+1}^{s}(z_t) + \kappa \varepsilon_l - \bar{\varepsilon}_{kt}^{s}(z_t).$$

The preference shock ε_l follows a GEV-I distribution with location parameter $\bar{\gamma}$ and a scale parameter of 1. Since $\zeta_{klt}(z_t)$ is a linear transformation of ε_l, it also follows a GEV-I distribution but with a location parameter $u(\delta^s w_{kt} z_t - \mathbb{1}(l \neq k) \cdot \lambda) - \tau_{klt} + \beta W_{kl,t+1}^{s}(z_t) - \kappa\bar{\gamma} - \bar{\varepsilon}_{kt}^{s}(z_t)$ and a scale parameter of κ. Thus, the expected value function for a sick individual in Eq. (A2) is

$$\begin{aligned} V_{kt}^{s}(z_t) &= \kappa \log \left[\sum_{l \in \mathbb{F}_{kt}^{s}(z_t)} \exp \left(\frac{u(\delta^s w_{kt} z_t - \mathbb{1}(l \neq k) \cdot \lambda) - \tau_{klt} + \beta W_{kl,t+1}^{s}(z_t)}{\kappa} \right) \right] - \bar{\varepsilon}_{kt}^{s}(z_t) \\ &= \kappa \log \Phi_{kt}^{s}(z_t) - \kappa \log \bar{F}_{kt}^{s}(z_t). && \text{(A6)} \end{aligned}$$

Recall that $\bar{F}_{kt}^{s}(z_t)$ is the cardinality of set $\mathbb{F}_{kt}^{s}(z_t)$, and $\bar{\varepsilon}_{kt}^{s}(z_t) = \kappa \log \bar{F}_{kt}^{s}(z_t)$ corrects for the utility drift induced by the size of the choice set. When medical travel is infeasible ($\mathbb{F}_{kt}^{s}(z_t) = k$), $\bar{F}_{kt}^{s}(z_t) = 1$, and the correction term vanishes.

B.3 Markov Representation of Health Transitions

The entire model can be represented as a Markov process over $S+2$ health states. Transition probabilities depend not only on the individual's current state but also on their treatment choice. Let Π_{kl} denote the transition matrix for an individual residing in location k who receives treatment in location l. Then the state transitions evolve as follows:

	start = -1	0	1	$\cdots$	$S-1$	S
end = -1	1	0	π_{lt}^{D1}	$\cdots$	$\pi_{lt}^{D,S-1}$	$\pi_{lt}^{D,S}$
0	0	$1-\sum_{s=1}^{S}\pi_{kt}^{Is}$	π_{lt}^{H1}	$\cdots$	$\pi_{lt}^{H,S-1}$	$\pi_{lt}^{H,S}$
1	0	π_{kt}^{I1}	$1-\pi_{lt}^{D1}-\pi_{lt}^{H1}$	$\cdots$	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$S-1$	0	$\pi_{kt}^{I,S-1}$	0	$\cdots$	$1-\pi_{lt}^{D,S-1}-\pi_{lt}^{H,S-1}$	0
S	0	$\pi_{kt}^{I,S}$	0	$\cdots$	0	$1-\pi_{lt}^{D,S}-\pi_{lt}^{H,S}$

(A7)

This transition structure formalizes the stochastic evolution of health, governed both by environmental risk and by the individual's treatment choice.

B.4 Microfoundation for Treatment Outcomes

This appendix derives the multinomial logit specifications in Eqs. (14)-(17) from a model in which hospitals within a city produce treatment quality that stochastically determines patient outcomes.

We model treatment outcomes through a two-stage process. First, city l produces treatment quality q_l^s for patients with severity s, which depends on medical resources, patient volume, and waiting times. Second, individual patients experience idiosyncratic health shocks that, combined with treatment quality, determine whether they recover, die, or remain sick.

Medical Service Production and Waiting Times

City l produces treatment quality for patients with severity s according to:

$$\bar{q}_l^s = A^s \cdot \left(\frac{m_l}{p_l}\right)^{\alpha} \cdot p_l^{\beta}, \tag{A8}$$

where A^s is a baseline quality parameter for severity s, m_l denotes medical resources, and p_l denotes patient volume. This specification captures two key mechanisms behind treatment quality production. The first mechanism is resources per patient, captured by $(m_l/p_l)^{\alpha}$ with $\alpha > 0$. Cities with more resources per patient can provide higher-quality care through better equipment, more staff attention per patient, and enhanced diagnostic capabilities. The second mechanism is learning-by-doing, captured by the patient volume term p_l^{β}. When $\beta > 0$, treating more patients generates expertise as medical staff gain

experience with specific conditions. The net effect of patient volume is captured by $(\beta - \alpha)$, which reflects the balance between learning effects and resource dilution.

The effective treatment quality received by patients, denoted as q_l^s, is different from $\bar{q}_l^s$ due to waiting times. We model q_l^s as a mixture of treated and untreated states, where the waiting time determines the relative weight of each state. Upon arrival at city l, a patient with severity s enters a queue for specialized treatment. During the waiting period, the patient receives only baseline care, experiencing untreated quality q_0^s. Once treatment begins, the patient receives the full benefits of the medical resources and expertise in city l, experiencing treated quality $\bar{q}_l^s$. For tractability, we model the waiting time, W_l, as affecting the relative weights between the untreated quality and the treated quality as follows:

$$q_l^s = \bar{q}_l^s \cdot \exp(-\delta W_l) + q_0^s \cdot [1 - \exp(-\delta W_l)]. \tag{A9}$$

In the equation above, when waiting time approaches 0, the effective quality received by the patient approaches the treatment quality: $\lim_{W_l \to 0} q_{ls} = \bar{q}_l^s$. Conversely, the effective quality approaches the untreated quality as $W_l \to \infty$. When untreated quality is substantially lower than treated quality ($q_0^s \ll \bar{q}_l^s$), which is realistic for severe cardiovascular conditions requiring specialized tertiary care, we can approximate:

$$q_l^s \approx \bar{q}_l^s \cdot \exp(-\delta W_l). \tag{A10}$$

The parameter $\delta > 0$ governs the rate of quality degradation: larger δ indicates that delays are more costly.

We specify the waiting time as a log-linear function of medical resources and patient volume:

$$W_l = w_0 + w_m \log(m_l) + w_p \log(p_l), \tag{A11}$$

where $w_m < 0$ indicates that more resources reduce waiting times, and $w_p > 0$ indicates that more patients increase waiting times. This specification allows waiting times to respond proportionally to percentage changes in resources and volume, consistent with queuing theory where waiting times depend on capacity utilization rates. Substituting Eqs. (A11) and (A8) into Eq. (A10),

$$\begin{aligned} q_{ls} &= A^s \cdot \left(\frac{m_l}{p_l}\right)^{\alpha} \cdot p_l^{\beta} \cdot \exp[-\delta(w_0 + w_m \log m_l + w_p \log p_l)] \\ &= A^s \cdot m_l^{\alpha} \cdot p_l^{-\alpha} \cdot p_l^{\beta} \cdot \exp(-\delta w_0) \cdot \exp(-\delta w_m \log m_l) \cdot \exp(-\delta w_p \log p_l) \\ &= A^s \exp(-\delta w_0) \cdot m_l^{\alpha} \cdot m_l^{-\delta w_m} \cdot p_l^{\beta-\alpha} \cdot p_l^{-\delta w_p} \\ &= A^s \exp(-\delta w_0) \cdot m_l^{\alpha-\delta w_m} \cdot p_l^{\beta-\alpha-\delta w_p}. \end{aligned} \tag{A12}$$

Taking logarithms,

$$\log q_l^s = \log A^s - \delta w_0 + (\alpha - \delta w_m)\log(m_l) + (\beta - \alpha - \delta w_p)\log(p_l). \tag{A13}$$

The expression above highlights how medical resources and patient volume affect the effective treatment quality received by a patient. Better resources increase both treatment quality (α) and reduce waiting time (δw_m). Similarly, higher patient volume increases learning through β, dilutes the resources per patient through $-\alpha$, and increases waiting time through δw_p.

Patient-Level Outcomes

Individual patient i experiences idiosyncratic health shocks $\{\nu_i^0, \nu_i^{-1}, \nu_i^s\}$ with $s > 0$, which are independently drawn from a Type-I Extreme Value distribution. These shocks capture patient-specific factors that affect treatment response, such as genetic differences, comorbidity, and unmeasured baseline health status. Patient i recovers if treatment quality combined with the recovery shock exceeds the values for death and staying sick: $q_l^s + \nu_i^0 > \max\{q_l^s + \nu_i^{-1}, \nu_i^s\}$. Similarly, the patient dies if $q_l^s + \nu_i^{-1} > \max\{q_l^s + \nu_i^0, \nu_i^s\}$, and remains sick if $\nu_i^s > \max\{q_l^s + \nu_i^0, q_l^s + \nu_i^{-1}\}$.

The distribution assumption for patient health shocks yields multinomial logit probabilities of treatment outcomes. Specifically, the probability of recovery is:

$$\pi_l^{s,0} = \frac{\exp(\psi^{s,0} q_l^s)}{1 + \exp(\psi^{s,0} q_l^s) + \exp(\psi^{s,-1} q_l^s)}, \tag{A14}$$

where $\psi^{s,0} > 0$ measures how responsive recovery probabilities are to treatment quality for patients with severity s. The probability of death is:

$$\pi_l^{s,-1} = \frac{\exp(\psi^{s,-1} q_l^s)}{1 + \exp(\psi^{s,0} q_l^s) + \exp(\psi^{s,-1} q_l^s)}, \tag{A15}$$

where $\psi^{s,-1} < 0$ indicates that higher treatment quality reduces mortality for patients with severity s. The probability of remaining sick is $1 - \pi_l^{s,0} - \pi_l^{s,-1}$. We define the outcome-specific indices as:

$$\gamma^{s,0}(m_l, p_l) = \psi^{s,0} q_l^s \tag{A16}$$

$$\gamma^{s,-1}(m_l, p_l) = \psi^{s,-1} q_l^s. \tag{A17}$$

Substituting the log quality from Eq. (A13), we have

$$\begin{aligned}\gamma^{s,0}(m_l, p_l) &= \psi^{s,0}[\log A^s - \delta w_0 + (\alpha - \delta w_m)\log(m_l) + (\beta - \alpha - \delta w_p)\log(p_l)] \\ &= \psi^{s,0}(\log A^s - \delta w_0) + \psi^{s,0}(\alpha - \delta w_m)\log(m_l) + \psi^{s,0}(\beta - \alpha - \delta w_p)\log(p_l).\end{aligned}$$

Similarly,

$$\gamma^{s,-1}(m_l, p_l) = \psi^{s,-1}(\log A^s - \delta w_0) + \psi^{s,-1}(\alpha - \delta w_m)\log(m_l) + \psi^{s,-1}(\beta - \alpha - \delta w_p)\log(p_l).$$

Defining the reduced-form parameters

$$\gamma_1^{sH} = \psi^{s,0}(\log A^s - \delta w_0), \quad \gamma_2^{sH} = \psi^{s,0}(\alpha - \delta w_m), \quad \gamma_3^{sH} = \psi^{s,0}(\beta - \alpha - \delta w_p)$$
$$\gamma_1^{sD} = \psi^{s,-1}(\log A^s - \delta w_0), \quad \gamma_2^{sD} = \psi^{s,-1}(\alpha - \delta w_m), \quad \gamma_3^{sD} = \psi^{s,-1}(\beta - \alpha - \delta w_p),$$

we obtain the log-linear specifications

$$\gamma^{s,0}(m_l, p_l) = \gamma_1^{sH} + \gamma_2^{sH}\log(m_l) + \gamma_3^{sH}\log(p_l) \tag{A18}$$
$$\gamma^{s,-1}(m_l, p_l) = \gamma_1^{sD} + \gamma_2^{sD}\log(m_l) + \gamma_3^{sD}\log(p_l). \tag{A19}$$

Substituting these into Eqs. (A14) and (A15) yields

$$\pi_l^{s,0} = \frac{\exp[\gamma^{s,0}(m_l, p_l)]}{1 + \exp[\gamma^{s,0}(m_l, p_l)] + \exp[\gamma^{s,-1}(m_l, p_l)]},$$
$$\pi_l^{s,-1} = \frac{\exp[\gamma^{s,-1}(m_l, p_l)]}{1 + \exp[\gamma^{s,0}(m_l, p_l)] + \exp[\gamma^{s,-1}(m_l, p_l)]}.$$

The four equations above are exactly Eqs. (14)-(17) in the main text.

C Numerical Details on Simulation and Estimation

C.1 Algorithm for Solving the Steady State

Assume that we know the parameters $\{\gamma_{(\cdot)}^s\}$, the conditional CDF $G(\cdot)$, and the location fundamentals $\{\omega_k, m_k, \lambda\}$. Start with a guess of the expected value functions, $\{V_k^s(z)\}$, and a vector of patient distribution $\{p_l\}$. We iterate as follows:

1. Compute the recovery and the mortality rates of each location, $\{\pi_l^{Hs}, \pi_l^{Ds}\}$ using equations (14) and (15).
2. Compute the value of treatment, $\{W_{kl}^s(z)\}$, using equation (A5).
3. Compute the travel probability, $\{\mu_{kl}^s(z)\}$, using equation (10).
4. Compute the steady state $\{L_k^s(z)\}$ by iteratively calling equations (6) and (7) until convergence.
5. Update $\{p_l\}$ using equation (9).
6. Update $\{V_k^s\}$ using equation (A6).

7. Update $\{V_k^0\}$ using equation (A3).

Repeat the above steps until convergence. To numerically implement the above algorithm, we discretize the AR(1) income shock, $\log(z_{it})$, into 7 grid points following Tauchen (1986).

Technical Notes on Computation The value of option l should be computed as:

$$\begin{aligned}
&\exp\left[u(\delta^s w_{kt} z_t - \mathbb{1}(l \neq k)\cdot\lambda) - \tau_{klt} + \beta W^s_{kl,t+1}(z_t)\right]^{1/\kappa} \\
&= \exp\left[\frac{u(\delta^s w_{kt} z_t - \mathbb{1}(l \neq k)\cdot\lambda) - \tau_{klt} + \beta W^s_{kl,t+1}(z_t)}{\kappa}\right] \\
&= \exp\left[\frac{u(\delta^s w_{kt} z_t - \mathbb{1}(l \neq k)\cdot\lambda) - \tau_{klt}}{\kappa}\right] \times \exp\left[\frac{\beta W^s_{kl,t+1}(z_t)}{\kappa}\right].
\end{aligned}$$

The denominator in the choice probabilities, $\Phi^s_{kt}(z)$, conditional on the origin k, should be computed as the inner product of the following two vectors:

$$\Phi^s_{kt}(z) = \left\langle \left\{\exp\left[\frac{u(\delta^s w_{kt} z_t - \mathbb{1}(l' \neq k)\cdot\lambda) - \tau_{klt}}{\kappa}\right]\right\}_{l'\in\mathbb{F}^s_{kt}(z_t)}, \left\{\exp\left[\frac{\beta W^s_{kl,t+1}(z_t)}{\kappa}\right]\right\}_{l'\in\mathbb{F}^s_{kt}(z_t)} \right\rangle.$$

The advantage of this formulation is that in both cases, one term does not depend on $W^s_{kl,t+1}(z)$, and therefore can be computed prior to the fixed point iteration, saving computational time.

C.2 Details of the Simulations

Simulation Procedure The data sample we observe contains N cases spanning over 24 months. In the case of Cerebro-cardio, $N = 610,642$, and in the case of cancer, $N = 199,664$. We draw N_m new patients each month to simulate the sample in the model. New patients are identified by their home location, severity, and initial income shocks. The location and severity distributions follow the observed distributions in the data. The income shock follows the stationary distribution of z_t derived from the Tauchen discretization.

We simulate the model for $T+24$ months and discard the first T months of simulation to eliminate the potential biases introduced by the initial distribution of patients. In a simulation of month t, the pool of patients includes the remaining patients from the previous period $t-1$ and the new patients drawn in period t. For each patient in the pool at location k, severity $s(s > 0)$, and income shock z, we use the steady state policy function, $\mu^s_{kl}(z)$, to simulate their treatment location l. We then use the steady state $\{\pi_l^{Hs}, \pi_l^{Ds}\}$ to simulate their treatment outcome. In the last step, we remove the recovered and deceased patients from the pool and move on to the next period with the remaining patients.

Simulation Length and Sample Size The number of new patients to be simulated each month, N_m, and the length of the pre-simulation, T, depend on the parameters of interest $\gamma_{(\cdot)}^{Hs}$ and $\gamma_{(\cdot)}^{Ds}$ through the recovery and the mortality rates. As a result, we need to determine these two parameters for every guess of the input parameters, Θ.

Denote the average stay rate for severity $s(s > 0)$ as $\bar{\pi}^s = \frac{1}{K}\sum_{l=1}^{K}\left(1 - \pi_l^{Hs} - \pi_l^{Ds}\right)$. It is straightforward to see that at period t of the simulation, the pool patients at severity $s(s > 0)$ equals to

$$\frac{\pi^{Is}}{\sum_{s'=1}^{S}\pi^{Is'}}\left[N_m + \bar{\pi}^s N_m + (\bar{\pi}^s)^2 N_m + \cdots + (\bar{\pi}^s)^{t-1} N_m\right] = \widetilde{\pi}^{Is} N_m \frac{1 - (\bar{\pi}^s)^t}{1 - \bar{\pi}^s}.$$

In the expression above, $\widetilde{\pi}^{Is} = \frac{\pi^{Is}}{\sum_{s=1}^{S}\pi^{Is}}$ is the fraction of new patients with severity $s(s > 0)$, and therefore, $\widetilde{\pi}^{Is} N_m$ is the number of new patients in period t. The term $\frac{1-(\bar{\pi}^s)^t}{1-\bar{\pi}^s} \geq 1$ captures the relative weight of remaining old patients in the cross-sectional pool at period t. If patients exit the pool quickly so that $\bar{\pi}^s \to 0$, then, $\frac{1-(\bar{\pi}^s)^t}{1-\bar{\pi}^s} \to 1$, so the cross-section pool contains mostly new patients. Conversely, a higher $\bar{\pi}^s$ implies that a patient requires more time to recover; therefore, the cross-section pool of patients would contain a higher fraction of remaining patients. In a simulation, a higher $\bar{\pi}^s$ subsequently implies that more periods are needed to reach a stable cross-section pool of patients.

We first pick the simulation length T so that the number of cases starting from period T is stable, defined as the relative variations in the pool size between T and $T + 1$ is smaller than a pre-set threshold, ε^N:

$$\begin{aligned}
\frac{(\bar{\pi}^s)^T - (\bar{\pi}^s)^{T+1}}{1 - (\bar{\pi}^s)^T} &< \varepsilon^N \\
(1 + \varepsilon^N)(\bar{\pi}^s)^T - (\bar{\pi}^s)^{T+1} &< \varepsilon^N \\
(\bar{\pi}^s)^T\left[(1 + \varepsilon^N) - \bar{\pi}^s\right] &< \varepsilon^N \\
T\log(\bar{\pi}^s) &< \log\left[\frac{\varepsilon^N}{(1 + \varepsilon^N) - \bar{\pi}^s}\right] \\
T(s) &\approx \frac{\log\left[\frac{\varepsilon^N}{(1+\varepsilon^N)-\bar{\pi}^s}\right]}{\log(\bar{\pi}^s)}.
\end{aligned}$$

The value of T depends on severity $s(s > 0)$ through $\bar{\pi}^s$. We set T to be the nearest integer to $\max_{s\in S}\{T(s)\}$. The pre-simulation length ensures that after period T, the number of patients each month is approximately constant at $\sum_{s=1}^{S}\widetilde{\pi}^{Is} N_m \frac{1-(\bar{\pi}^s)^T}{1-\bar{\pi}^s}$. In practice, we set $\varepsilon^N = 0.0001$.

Given that we observe $\frac{N}{24}$ total cases each month in the data, we set N_m to the integer

closest to:

$$\frac{N}{24} = N_m \sum_{s=1}^{S} \widetilde{\pi}^{Is} \frac{1-(\bar{\pi}^s)^T}{1-\bar{\pi}^s}$$
$$N_m = \frac{N}{24} \left[\sum_{s=1}^{S} \widetilde{\pi}^{Is} \frac{1-(\bar{\pi}^s)^T}{1-\bar{\pi}^s} \right]^{-1}.$$

Bounding the Treatment Probability During the estimation process, the minimization algorithm might evaluate certain parameters of $\{\gamma_{(\cdot)}^{Hs}\}$ and $\{\gamma_{(\cdot)}^{Ds}\}$ that lead to extreme values of $\{\pi_l^{Hs}, \pi_l^{Ds}\}$. If $\pi_l^{Hs} \to 0$ or $\pi_l^{Ds} \to 1$, computing the stationary population distribution across $\{(k, z, s)\}$ could be time consuming for the reasons discussed above. As the limit cases of zero recovery rate and 100 percent mortality rate are empirically irrelevant, we set the lower bound of the recovery rate to be 1 percent and the upper bound of the mortality rate to be 99 percent. We verified that these bounds are not binding in the final estimates.

C.3 Details of the Indirect Inference

Estimating the structural parameters via indirect inference involves numerically minimizing a high-dimensional objective function as defined in Eq. (23). To minimize the objective function, we apply a mixture of particle swarm optimization (PSO) with pattern search (PS). We apply PSO first to search over a wide range of the parameter space, and then switch to PS after convergence in PSO to speed up the optimization process.

We compute the asymptotic standard errors following the methods outlined in Cameron and Trivedi (2005). In particular:

$$\widehat{\text{Var}}(\Theta) = \left(\widehat{\mathbf{G}}'\mathbf{W}\widehat{\mathbf{G}}\right)^{-1} \widehat{\mathbf{G}}'\mathbf{W}\widehat{\mathbf{\Sigma}}\mathbf{W}\widehat{\mathbf{G}} \left(\widehat{\mathbf{G}}'\mathbf{W}\widehat{\mathbf{G}}\right)^{-1}. \tag{A20}$$

In the equation above, $\widehat{\mathbf{G}}$ is the estimated gradient matrix, in which the ith row and the jth column is the partial derivative of the ith element of vector $\mathbf{A}$ with respect to the jth parameter, evaluated at the estimated Θ. We numerically compute the gradient matrix using a two-sided difference approximation. $\mathbf{W}$ is the weighting matrix, and $\widehat{\mathbf{\Sigma}}$ is an estimate of the variance-covariance matrix of the moment conditions. In our text, $\widehat{\mathbf{\Sigma}}$ is a diagonal matrix that contains the squared standard errors of the coefficients in the auxiliary regressions. Note that as we set $\mathbf{W} = \left(\widehat{\mathbf{\Sigma}}\right)^{-1}$, equation (A20) simplifies to:

$$\widehat{\text{Var}}(\Theta) = \left(\widehat{\mathbf{G}}'\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{G}}\right)^{-1}, \tag{A21}$$

which we use to estimate the standard errors as reported in the main text.

D Supplements to Counterfactual Analysis

D.1 Variance Decomposition

In this section, we provide the details on decomposing the variance of the expected mortality rates. As we carry out all the analysis in steady state, we omit the time subscript for expositional ease. Recall that we define the expected mortality rate of individuals in location k, income shock z as:

$$\varpi_{kz} = \frac{\sum_{s=1}^{S}\sum_{l=1}^{K}\mu_{kl}^{s}(z)\pi_{l}^{Ds}L_{k}^{s}(z)}{\sum_{s=1}^{S}L_{k}^{s}(z)}. \tag{A22}$$

The numerator in the equation represents the total expected mortality in location k, type z, considering the expected patient flows to all possible treatment locations. The denominator is the total patient population in location k, type z. We are interested in understanding the variance of the expected mortality rate across locations and income types.

To carry out the variance decomposition, first denote the national average mortality rate as:

$$\overline{\varpi} = \sum_{j=1}^{K}\omega_{kz}\varpi_{kz}, \tag{A23}$$

where the weight for cell (k,z) is $\omega_{kz} = \sum_{s=1}^{S}L_{k}^{s}(z)/\sum_{k',z',s'}L_{k'}^{s'}(z)$.

The variance of the expected mortality rate can then be decomposed into three terms:

$$\mathrm{Var}\left(\varpi\right) = \sum_{k,z}\omega_{kz}\left(\varpi_{kz}-\overline{\varpi}\right)^{2} = \mathrm{Var}_{K} + \mathrm{Var}_{Z} + \mathrm{Var}_{KZ}.$$

1. The first term,

$$\mathrm{Var}_{K} = \sum_{k=1}^{K}\omega_{k}\left(\overline{\varpi}_{k}-\overline{\varpi}\right)^{2}, \tag{A24}$$

   is the between-location variance. In this expression, $\omega_k = \sum_z \omega_{kz}$ is the marginal weight of location k, and $\overline{\varpi}_k = \frac{1}{\omega_k}\sum_z \omega_{kz}\varpi_{kz}$ is the average mortality in location k across income groups.

2. The second term is the between-income group variance:

$$\mathrm{Var}_{Z} = \sum_{z=1}^{Z}\omega_{z}\left(\overline{\varpi}_{z}-\overline{\varpi}\right)^{2}, \tag{A25}$$

   where $\omega_z = \sum_k \omega_{kz}$ is the marginal weight for type z, and $\overline{\varpi}_z = \frac{1}{\omega_z}\sum_k \omega_{kz}\varpi_{kz}$ is

the average mortality by income group z.

3. The last term is the residual:

$$\text{Var}_{KZ} = \sum_k \sum_z \omega_{kz} \left(\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi} \right)^2 . \tag{A26}$$

The details of the decomposition are as follows. First note that:

$$\varpi_{kz} - \overline{\varpi} = (\overline{\varpi}_k - \overline{\varpi}) + (\overline{\varpi}_z - \overline{\varpi}) + (\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}) .$$

Squaring both sides:

$$\begin{aligned}(\varpi_{kz} - \overline{\varpi})^2 = (\overline{\varpi}_k - \overline{\varpi})^2 + (\overline{\varpi}_z - \overline{\varpi})^2 + (\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi})^2 \\ + 2(\overline{\varpi}_k - \overline{\varpi})(\overline{\varpi}_z - \overline{\varpi}) \\ + 2(\overline{\varpi}_k - \overline{\varpi})(\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}) \\ + 2(\overline{\varpi}_z - \overline{\varpi})(\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}).\end{aligned}$$

The first line contains the three main terms in the decomposition, while the rest are the cross-terms. Notice that summing over all k, z with weights ω_{kz}, the cross-terms vanish.

The First Cross Term To be specific, the first term, summed over (k, z):

$$\sum_k \sum_z \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi} \right) \left(\overline{\varpi}_z - \overline{\varpi} \right)$$

equals zero. To see this, begin by rewriting the sum as:

$$\sum_k \sum_z \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi} \right) \left(\overline{\varpi}_z - \overline{\varpi} \right) = \sum_k \left(\overline{\varpi}_k - \overline{\varpi} \right) \sum_z \omega_{kz} \left(\overline{\varpi}_z - \overline{\varpi} \right)$$

Now consider the inner sum:

$$\sum_z \omega_{kz} \left(\overline{\varpi}_z - \overline{\varpi} \right) = \sum_z \omega_{kz} \overline{\varpi}_z - \overline{\varpi} \sum_z \omega_{kz} = \sum_z \omega_{kz} \overline{\varpi}_z - \overline{\varpi} \cdot \omega_k .$$

Thus, the full expression becomes:

$$\begin{aligned}\sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi} \right) \left(\overline{\varpi}_z - \overline{\varpi} \right) &= \sum_k \left(\overline{\varpi}_k - \overline{\varpi} \right) \left(\sum_z \omega_{kz} \overline{\varpi}_z - \overline{\varpi} \cdot \omega_k \right) \\ &= \sum_k \left(\overline{\varpi}_k - \overline{\varpi} \right) \sum_z \omega_{kz} \overline{\varpi}_z - \overline{\varpi} \sum_k \left(\overline{\varpi}_k - \overline{\varpi} \right) \omega_k .\end{aligned}$$

The second term is zero because:

$$\sum_k (\overline{\varpi}_k - \overline{\varpi})\,\omega_k = \sum_k \omega_k \overline{\varpi}_k - \overline{\varpi} \sum_k \omega_k = \overline{\varpi} - \overline{\varpi} = 0.$$

Now focus on the first term:

$$\sum_k (\overline{\varpi}_k - \overline{\varpi}) \sum_z \omega_{kz} \overline{\varpi}_z = \sum_k \sum_z \omega_{kz} (\overline{\varpi}_k - \overline{\varpi})\, \overline{\varpi}_z = \sum_z \overline{\varpi}_z \sum_k \omega_{kz} (\overline{\varpi}_k - \overline{\varpi}).$$

Next, consider:

$$\sum_k \omega_{kz} (\overline{\varpi}_k - \overline{\varpi}) = \sum_k \omega_{kz} \overline{\varpi}_k - \overline{\varpi} \sum_k \omega_{kz} = \sum_k \omega_{kz} \overline{\varpi}_k - \overline{\varpi} \cdot \omega_z.$$

Then the total becomes:

$$\sum_z \overline{\varpi}_z \left(\sum_k \omega_{kz} \overline{\varpi}_k - \overline{\varpi} \cdot \omega_z \right) = \sum_z \sum_k \omega_{kz} \overline{\varpi}_k \overline{\varpi}_z - \overline{\varpi} \sum_z \omega_z \overline{\varpi}_z.$$

Since:

$$\sum_z \omega_z \overline{\varpi}_z = \overline{\varpi} \quad \text{and} \quad \sum_{k,z} \omega_{kz} \overline{\varpi}_k \overline{\varpi}_z = \sum_k \overline{\varpi}_k \sum_z \omega_{kz} \overline{\varpi}_z,$$

this expression reduces to:

$$\sum_{k,z} \omega_{kz} \overline{\varpi}_k \overline{\varpi}_z - \overline{\varpi}^2.$$

Now expand the first term:

$$\begin{aligned} \sum_{k,z} \omega_{kz}\, \overline{\varpi}_k\, \overline{\varpi}_z &= \sum_{k,z} \omega_{kz} \left(\frac{1}{\omega_k} \sum_{z'} \omega_{kz'} \varpi_{kz'} \right) \left(\frac{1}{\omega_z} \sum_{k'} \omega_{k'z} \varpi_{k'z} \right) \\ &= \sum_{k,z} \frac{\omega_{kz}}{\omega_k \omega_z} \left(\sum_{z'} \omega_{kz'} \varpi_{kz'} \right) \left(\sum_{k'} \omega_{k'z} \varpi_{k'z} \right). \end{aligned}$$

Now observe that the factor $\sum_{z'} \omega_{kz'} \varpi_{kz'}$ is independent of z; similarly the factor $\sum_{k'} \omega_{k'z} \varpi_{k'z}$ is independent of k. So we can reorder the sums:

$$\begin{aligned} \sum_{k,z} \omega_{kz}\, \overline{\varpi}_k\, \overline{\varpi}_z &= \left(\sum_k \frac{1}{\omega_k} \sum_{z'} \omega_{kz'} \varpi_{kz'} \sum_z \omega_{kz} \cdot \frac{1}{\omega_z} \sum_{k'} \omega_{k'z} \varpi_{k'z} \right) \\ &= \left(\sum_k \sum_{z'} \omega_{kz'} \varpi_{kz'} \right) \left(\sum_z \sum_{k'} \omega_{k'z} \varpi_{k'z} \right) = \overline{\varpi} \cdot \overline{\varpi} = \overline{\varpi}^2. \end{aligned}$$

Thus:

$$\sum_{k,z} \omega_{kz}\, \overline{\varpi}_k\, \overline{\varpi}_z = \overline{\varpi}^2.$$

Hence, the first cross-term cancels:

$$\sum_{k,z} \omega_{kz}(\overline{\varpi}_k - \overline{\varpi})(\overline{\varpi}_z - \overline{\varpi}) = 0.$$

The Second Cross Term We begin with the expression:

$$\sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \left(\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}\right).$$

Distribute the first factor and expand the second:

$$\begin{aligned} &= \sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \varpi_{kz} - \sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \overline{\varpi}_k \\ &\quad - \sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \overline{\varpi}_z + \sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \overline{\varpi}. \end{aligned}$$

Now evaluate each term separately. The first term is:

$$\sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \varpi_{kz} = \sum_{k} \left(\overline{\varpi}_k - \overline{\varpi}\right) \sum_{z} \omega_{kz} \varpi_{kz} = \sum_{k} \left(\overline{\varpi}_k - \overline{\varpi}\right) \cdot \omega_k \cdot \overline{\varpi}_k.$$

The second term is:

$$\sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \overline{\varpi}_k = \sum_{k} \left(\overline{\varpi}_k - \overline{\varpi}\right) \overline{\varpi}_k \sum_{z} \omega_{kz} = \sum_{k} \left(\overline{\varpi}_k - \overline{\varpi}\right) \overline{\varpi}_k \cdot \omega_k.$$

Thus, the two terms cancel out, and the first line equals zero. The third term is:

$$\sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \overline{\varpi}_z = \sum_{z} \overline{\varpi}_z \sum_{k} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) = \sum_{z} \overline{\varpi}_z \left(\sum_{k} \omega_{kz} \overline{\varpi}_k - \overline{\varpi} \cdot \omega_z \right).$$

Now note that:

$$\sum_{z} \overline{\varpi}_z \sum_{k} \omega_{kz} \overline{\varpi}_k = \sum_{k,z} \omega_{kz} \overline{\varpi}_k \overline{\varpi}_z, \quad \sum_{z} \overline{\varpi}_z \cdot \overline{\varpi} \cdot \omega_z = \overline{\varpi} \sum_{z} \omega_z \overline{\varpi}_z = \overline{\varpi}^2.$$

From earlier derivation:

$$\sum_{k,z} \omega_{kz} \overline{\varpi}_k \overline{\varpi}_z = \overline{\varpi}^2.$$

So Term 3 equals $\overline{\varpi}^2 - \overline{\varpi}^2 = 0$. The last term is also zero:

$$\sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \overline{\varpi} = \overline{\varpi} \sum_k \left(\overline{\varpi}_k - \overline{\varpi}\right) \sum_z \omega_{kz} = \overline{\varpi} \sum_k \omega_k \left(\overline{\varpi}_k - \overline{\varpi}\right) = \overline{\varpi} \left(\sum_k \omega_k \overline{\varpi}_k - \overline{\varpi} \right) = 0.$$

Each of the four terms in the expansion cancels. Therefore,

$$\sum_{k,z} \omega_{kz} \left(\overline{\varpi}_k - \overline{\varpi}\right) \left(\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}\right) = 0.$$

The Third Cross Term The third term is also zero by symmetric reasoning as above.

$$\sum_{k,z} \omega_{kz} (\overline{\varpi}_z - \overline{\varpi}) (\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}) = 0.$$

Therefore, to sum up, all the cross terms are zero once summed over (k, z):

$$\begin{aligned}
\sum_{k,z} \omega_{kz} (\overline{\varpi}_k - \overline{\varpi}) (\overline{\varpi}_z - \overline{\varpi}) &= 0, \\
\sum_{k,z} \omega_{kz} (\overline{\varpi}_k - \overline{\varpi}) (\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}) &= 0, \\
\sum_{k,z} \omega_{kz} (\overline{\varpi}_z - \overline{\varpi}) (\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}) &= 0.
\end{aligned}$$

Thus, we obtain the exact additive decomposition:

$$\mathrm{Var}(\varpi) = \underbrace{\sum_k \omega_k (\overline{\varpi}_k - \overline{\varpi})^2}_{\text{Between-location variance}} + \underbrace{\sum_z \omega_z (\overline{\varpi}_z - \overline{\varpi})^2}_{\text{Between-income variance}} + \underbrace{\sum_{k,z} \omega_{kz} \left(\varpi_{kz} - \overline{\varpi}_k - \overline{\varpi}_z + \overline{\varpi}\right)^2}_{\text{Residual (interaction) variance}}.$$