Estimating an Equilibrium Model of Health Insurance and Healthcare Expense

Junjian Yi* Shaoyang Zhao† Hang Zou‡

March 16, 2021

Abstract

We formulate a general equilibrium model to disentangle demand-side responses from supply-side responses in accounting for the equilibrium effect of health insurance on healthcare expense. We estimate the model using a policy change in public health insurance in China. Using model estimates, we simulate and decompose the equilibrium effect of the policy change. We find that both demand and supply factors significantly contribute to the equilibrium effect, but their relative importance differs by area. Supply-side responses account for 71.62% of the increase in healthcare expense in urban areas, and demand-side responses for 61.55% in rural areas.

Keywords: Public health insurance; healthcare expense; equilibrium analysis; demand-side responses; supply-side responses; individual welfare

JEL Codes: I11, I13, I38, L13

*Department of Economics, National University of Singapore; e-mail: junjian.yi@gmail.com.
†School of Economics, Sichuan University; e-mail: zhaoshaoyang@scu.edu.cn.
‡Department of Economics, National University of Singapore; e-mail: hang.zou@u.nus.edu.

1 Introduction

The dramatic rise in healthcare expense concerns both policymakers and the public worldwide. From 1978 to 2019, total healthcare expense per capita in real terms rose fourfold, from $2,758 to $11,582, and the share of healthcare expense of GDP rose more than twofold, from 8.2% to 17.7%, in the U.S. (CMS, 2019). During the same period, with the fast economic growth, total healthcare expense per capita in real terms rose 61-fold, from $11 to $681, and the share out of GDP rose more than twofold, from 3.0% to 6.6%, in China (NBS, 2020). If the growth of healthcare expense keeps outpacing the growth of GDP in the next decades, the government's budget will be strained, which undermines its ability to finance other vital services such as education, defense, and transportation. In the long run, the economic growth would be stifled.

The economics and public health literature has widely investigated factors that may drive the dramatic rise in healthcare expense from multiple perspectives, such as income growth (Newhouse, 1977; Hall and Jones, 2007); an aging population (Crivelli et al., 2006; Lopreite and Mauro, 2017); progress in medical technology (Bundorf et al., 2009; Smith et al., 2009); and health insurance (Newhouse, 1992; Finkelstein, 2007). Of these factors, economists have been particularly interested in health insurance since Arrow (1963). When government provides health insurance, the price of healthcare services decreases, demand increases, and total expense increases.

Early studies conclude that the two public health insurance schemes—Medicare and Medicaid—are the main drivers of the dramatic rise in healthcare expense in the U.S. (Feldstein, 1971, 1977). However, this conclusion is not supported by two famous experiments: the RAND and Oregon health insurance experiments. Estimates based on the RAND experiment show that the price elasticities of healthcare demand are as low as -0.2 (Manning et al., 1987). Those based on the Oregon experiment are even smaller (Finkelstein et al., 2012). The small estimates suggest that public health insurance cannot explain the several-fold increase in total healthcare expense in the U.S. and elsewhere.

We suggest that estimates based on the two experiments may not fully illustrate the relationship between public health insurance and healthcare expense. The two experiments include only a few thousand subjects, and thus they induce a partial equilibrium effect. By contrast, public health insurance schemes, such as Medicare

and Medicaid, induce general equilibrium effects because they change healthcare demand nationwide, which incentivize hospitals to adjust their investment and pricing strategies. For example, Finkelstein (2007) finds that hospitals adjusted their total number of doctors, nurses, and beds in response to Medicare. To fully understand the relationship between public health insurance and healthcare expense, we must study both individuals' responses from the demand side and hospitals' responses from the supply side within a unified framework.

We provide the first general equilibrium model in the literature to endogenize both patient and hospital responses simultaneously. On the demand side, individuals face a discrete choice of hospital care. Each individual maximizes her utility, which is determined by the out-of-pocket (OOP) expense, hospital quality, and hospital tier, by visiting a hospital or staying at home when facing a health shock. On the supply side, hospitals compete in a two-stage game. In the first stage, all hospitals choose healthcare quality; in the second stage, all hospitals simultaneously set the total expense and claimable expense per visit in the market.[1] We adopt subgame perfect Nash equilibrium (SPNE) as the equilibrium notion and derive the optimality conditions for hospital choices.

We estimate the model using a policy change regarding public health insurance in China. The Chinese government changed the premiums and plans for both urban and rural areas in 2010, and more than three-fourths of enrollees switched to health insurance plans with higher reimbursement rates. The policy change provides a unique opportunity for us to carry out the identification for estimating a general equilibrium model. On the one hand, the change in reimbursement rate changes the price of healthcare, which helps us identify individuals' preference parameters in hospital choice on the demand side. On the other, the policy change may have differential impacts across hospitals in different tiers since reimbursement rates differed across hospitals in different tiers. The differential impacts, which are rarely explored in the literature, would help us to empirically identify our structural model of hospital behavior on the supply side.

We employ three datasets—enrollment data, inpatient claims data, and annual hospital report data—from Chengdu, the capital city of Sichuan in China. The three datasets complement each other and enable us to estimate the general equilibrium model of the healthcare market. The data reveal four stylized facts associated with

[1]See Eq. (1) in Section 2.2 for a discussion of total expense and claimable expense in detail.

the policy change. First, rural enrollees visited hospitals more frequently. Second, urban patients switched from low-tier to high-tier hospitals. Third, although the total expense per visit increased significantly, the OOP expense per visit remained at a roughly similar level. Fourth, hospitals hired more doctors and nurses and bought more beds and advanced medical equipment.

To explain the responses from both demand and supply sides, we estimate our model in three steps. First, we recover an individual's preference for hospital care by estimating the discrete choice model. We take advantage of the policy change to address the endogeneity of the OOP expense in hospital choice. Second, we estimate the marginal cost function and marginal psychological cost function. Third, we estimate the fixed cost function using the generalized method of moments (GMM). Our estimation results show that most parameters are precisely estimated, and the signs of parameter estimates are consistent with the theory.

Based on our model estimates, we simulate the change in healthcare expense with the policy change. We then disentangle demand-side factors from supply-side factors in accounting for the equilibrium effect of health insurance on healthcare expense. Our decomposition differs from that in the literature. For example, Brot-Goldberg et al. (2017) decompose the difference in *observed* healthcare expense before and after the policy change. By contrast, we decompose the difference in *simulated* healthcare expense with and without the policy based on our general equilibrium model. In our model, we simultaneously endogenize responses from both the demand and supply sides. This model-based approach enables us to decompose the equilibrium effect of insurance on healthcare expense into more channels than approaches adopted in the literature (Brot-Goldberg et al., 2017). In particular, we are able to disentangle the channels on the demand side from those on the supply side. This is a first for the literature. Furthermore, the model-based decomposition is free from confounding factors.

We decompose the change in healthcare expense into five terms: "patient sorting," "quantity increase," "quality adjustment," "price adjustment," and "cross." The first two terms reflect responses from the demand side, the next two reflect responses from the supply side, and the last the covariance between demand- and supply-side responses. We find that both demand and supply factors significantly contribute to the equilibrium effect of health insurance on healthcare expense. However, their relative importance differs by area. In urban areas, quality and price adjustments

from the supply side mainly explain the equilibrium effect, accounting for 33.24% and 38.38% of the increase in healthcare expense, respectively. By contrast, in rural areas, the quantity increase from the demand side mainly explains the equilibrium effect, accounting for 62.57% of the increase in healthcare expense. Our finding on the difference between urban and rural areas is consistent with the finding in the RAND health insurance experiment. Manning et al. (1987) show that the price elasticity of healthcare demand is higher when the reimbursement rate is lower. In our study, the reimbursement rate is lower in rural areas than in urban ones before the policy change. When the government increases the reimbursement rate, rural residents more actively respond to the policy change.

We further use our model estimates to evaluate the welfare consequence of the policy change. We define the welfare for each individual as the consumer surplus minus the premium (Small and Rosen, 1981). We find that the policy change enhanced the welfare of urban residents but deteriorated that of rural ones. The difference in the change in the premium explains the difference in the change in the welfare between urban and rural residents. When individuals switched to the new plan induced by the policy, the premium decreases in urban areas and increases in rural ones.

Our paper contributes to health economics—in particular, to the literature on health insurance and healthcare expense (Feldstein, 1971, 1977; Manning et al., 1987; Newhouse, 1992; Finkelstein, 2007; Finkelstein et al., 2012; Chandra et al., 2014; Brot-Goldberg et al., 2017). The literature focuses on the endogenous response to health insurance from either the demand or supply side. We develop the first general equilibrium model to endogenize demand- and supply-side responses simultaneously. Based on our model estimates, we are able to simulate the equilibrium effect of health insurance on healthcare expense, through which we decompose the equilibrium effect into multiple channels from demand and supply responses. Our decomposition analysis complements that of Brot-Goldberg et al. (2017), who decompose the observed spending reduction associated with the deductible increase into (i) consumer price shopping, (ii) quantity reductions, and (iii) quantity substitutions. They find that all spending reductions are achieved through reductions in quantity.

Our result shows that quantifying the relative importance of demand and supply responses is crucial for designing policies to contain the rising healthcare expense. In this regard, our analysis complements that of Finkelstein et al. (2016). They separate the roles of demand and supply factors to explain the geographic variation

in U.S. health care utilization by exploiting the migration of Medicare patients. Their empirical strategy accounts for demand differentials driven by both observable and unobservable patient characteristics, but holds supply-side characteristics constant. They find that 40-50% of geographic variation in utilization is attributed to demand factors, with the remainder to supply factors.

Our paper also contributes to empirical industrial organization; in particular, to the literature that treats product characteristics as endogenous (Fan, 2013; Crawford et al., 2019; Hackmann, 2019). Fan (2013) is the closest to ours; she formulates a structural model to understand the effect of ownership consolidation in U.S. newspaper markets, taking into account both price adjustment and characteristics adjustment. She finds that overlooking the characteristics adjustment can lead to considerable differences in the estimated effects of mergers. We extend this framework to the healthcare setting, in which hospitals' pricing strategy is more complicated because hospitals are partially altruistic and patients pay only a proportion of their healthcare expense with health insurance. We further use the policy change to strengthen the identification.

2 Institutional Setting

This section describes the institutional background of the healthcare system, the urban and rural resident basic medical insurance scheme (URRBMI), and the policy change in the URRBMI from 2009 to 2010 in Chengdu. As the capital city of Sichuan, Chengdu is one of the most developed regions in southwest China, with a GDP per capita of ¥41,253 in 2010 (CBS, 2011).[2] The healthcare system and public health insurance in Chengdu are representative of the rest of China.

2.1 Healthcare System

In China, most hospitals are public hospitals. For example, they provided 91.9% of outpatient service and 91.6% of inpatient service in China in 2010.[3] Public hospitals are classified into three tiers. The classification is mainly based on hospitals' size and

[2]In 2010, the GDP per capita was ¥29,992 in China (NBS, 2011) and ¥21,182 in Sichuan (SBS, 2011). ¥1 ≈ $0.15 in 2010.

[3]Data source: National Bureau of Statistics (2011a).

quality measures, such as the number of inpatient beds and the quality of management and medical care (Ministry of Health, 1989). Tier-3 hospitals provide the most sophisticated acute care and specialist services. They also play a dominant role in medical education and research. Tier-2 hospitals are responsible for comprehensive healthcare services and medical training for health workers in tier-1 facilities. Tier-1 hospitals mainly provide primary care and preventive care services. In addition to hospitals in these three tiers, the township health center provides healthcare services at the township level. We refer to the township health center as the tier-0 hospital. Hospitals of different tiers are designed to treat diseases of different severity and complexity. However, there is no gatekeeping referral system to triage patients with different medical needs to hospitals in different tiers. Appendix A details China's tiered healthcare system.

In public hospitals, all physicians are employees. Their total income consists of a fixed salary and a commission of revenues generated by treating patients. The latter usually accounts for as high as three-quarters of physicians' total income (Milcent, 2018). Therefore, throughout our analysis, we assume that hospitals and physicians have the same incentives to maximize their revenues.

A public hospital's revenue depends on the patient's total healthcare expense, which in turn depends on the number of medical services and the fee for each medical service. For example, to treat a patient with low back pain, the hospital could provide consultation; image tests such as X-rays, CT scans, and MRI; lab tests such as blood and urine tests; physical therapy; surgery; and inpatient services such as nursing. For each patient, the hospital chooses the type and amount of medical services and the government sets the fee for each service. In brief, public hospitals are paid under a fee-for-service scheme, which is different from a bundled payment scheme, such as the diagnosis-related-group system adopted in the U.S.

2.2 Urban and Rural Resident Basic Medical Insurance

History of Public Health Insurances in China

Under the public health insurance, the total healthcare expense is paid by both the government and the patient. The current public health insurance system in China consists of two insurance schemes: urban employee basic medical insurance (UEBMI) and urban and rural residents basic medical insurance (URRBMI). The UEBMI covers

all individuals who have formal jobs and hold urban hukou;[4] these individuals account for about one-fourth of the Chinese population. The URRBMI covers the remaining three-fourths of the population.[5] Our study focuses on the URRBMI. We next briefly review the history of the URRBMI; Appendix B provides details on the public health insurance system in China.

The Chinese government combines two public health insurance schemes—the New Rural Cooperative Medical Scheme (NRCMS) and the Urban Resident Basic Medical Insurance (URBMI)—to form the URRBMI. The NRCMS serves all rural residents. In the late 1950s, the Chinese government established the rural cooperative medical scheme (RCMS) to provide health insurance for rural residents. The RCMS was operated at the commune level.[6] The health workers, who were called "barefoot doctors," provided primary and preventive care for rural residents for free. However, communes have been completely dismantled and the RCMS virtually collapsed after 1979, when China launched its economic reforms. From 1979 to 2003, no public health insurance was provided for rural residents. During this period, rural residents' OOP expense increased dramatically, and thus they were vulnerable to health risks. Finally, in 2003, the government started to roll out a new public health insurance scheme—the NRCMS—for rural residents across the nation. By 2007, 96% of all rural residents were covered by the NRCMS. Premiums for the scheme are highly subsidized by the government. The scheme mainly covers inpatient services. The reimbursement rate varies across hospitals of different tiers, with a lower rate for higher-tier hospitals (Milcent, 2018).

As for urban residents without formal jobs, no public health insurance was provided before 2007. The Chinese government started the URBMI for these residents in 2007. The scheme covers inpatient and critical outpatient services. Similar to the NRCMS, the reimbursement rate under the URBMI varies across hospitals in different tiers, with a lower rate for higher-tier hospitals. However, the URBMI differed from the NRCMS in multiple dimensions. For example, both the premium and reimbursement rate are higher under the URBMI than under the NRCMS.

[4]The hukou system is a household registration system in China. The system categorizes each Chinese citizen as either a rural hukou holder or an urban hukou holder. A person's hukou is inherited from their parents, and changes in hukou are rare. The hukou is used to determine one's eligibility for social services and welfare based on his registered place of residence (Milcent, 2018).

[5]Data source: http://www.nhsa.gov.cn/art/2020/6/24/art_7_3268.html.

[6]A commune is a large rural collective work unit. Individuals in a commune work together and obtain their earnings based on their labor contribution (Milcent, 2018).

The differences between the URBMI and the NRCMS raised concerns about not only the inequality between rural and urban residents but also the efficiency of government administrative management. Hence, the Chinese government started to combine the URBMI and the NRCMS to form a new public health insurance scheme—the URRBMI—in 2008.[7]

The URRBMI in Chengdu

The URRBMI was implemented in Chengdu in 2009. It covers urban residents without formal jobs and all rural residents.[8] The enrollment rate in the URRBMI has been above 95% since 2009.[9]

Three plans were provided under the URRBMI in 2009: Low, Medium, and High. The High (Low) plan charged the highest (lowest) premium and had the highest (lowest) reimbursement rate. Residents could choose one of the three plans at the beginning of 2009. Premiums and reimbursement rates differed across plans. Table 1 tabulates the premiums, government subsidies, and enrollment rates across plans and areas. Chengdu consisted of 19 county-level administrative units—9 districts, 4 county-level cities, and 6 counties.[10] Under the URRBMI, these 19 county-level administrative units are classified into urban and rural areas.[11]

Premiums were the same in both urban and rural areas in 2009. Specifically, the premiums were ¥20, ¥120, and ¥220 under Low, Medium, and High plans, respectively. The government subsidized premiums by ¥80 for each plan. Enrollment rates for the three plans differed across areas. Eighty-six percent of urban residents enrolled in the Medium plan, and 75% of rural residents enrolled in the Low plan.

Reimbursement Structure

We next describe the reimbursement structure under the URRBMI. Not all healthcare expense are claimable in China. To fully understand the reimbursement struc-

[7]The private health insurance industry is less developed in China. For example, less than 5% of Chinese had private health insurance in 2016 (Xiang, 2020).

[8]The remaining urban residents with formal jobs are covered by the UEBMI.

[9]Data source: https://www.sc.gov.cn/10462/10464/10465/10595/2011/8/2/10172757.shtml.

[10]The administrative divisions of China consist of 5 levels: central, province, prefecture, county, and township. Chengdu is a prefecture-level administrative unit. The 9 districts are Jinjiang, Qingyang, Jinniu, Wuhou, Chenghua, Longquanyi, Qingbaijiang, Xindu, and Wenjiang; the 4 county-level cities are Pengzhou, Qionglai, Chongzhou, and Dujiangyan; the 6 counties are Shuangliu, Jintang, Pi, Dayi, Pujiang, and Xinjin.

[11]The urban areas include 6 districts—Jinjiang, Qingyang, Jinniu, Wuhou, Chenghua, and Wenjiang—and Shuangliu county. The rest belong to rural areas.

Table 1: Changes in Insurance Policies by Plan and Area

	Plan		Premium (¥)		Subsidy (¥)		Enrollment Rate	
Area	2009	2010	2009	2010	2009	2010	2009	2010
Urban	Low	N.A.	20	N.A.	80	N.A.	2.61%	N.A.
	Medium	N.A.	120	N.A.	80	N.A.	85.68%	N.A.
	High	High	220	100	80	220	11.71%	100%
Rural	Low	N.A.	20	N.A.	80	N.A.	74.68%	N.A.
	Medium	Medium	120	40	80	180	24.01%	93.71%
	High	High	220	140	80	180	1.31%	6.29%

ture, we first introduce the formula to calculate the OOP expense (p), which is set by the government:

$$p = \begin{cases} e - (C - Dud)r & \text{if} \quad C \geq Dud, \\ e & \text{if} \quad C < Dud, \end{cases} \tag{1}$$

where e denotes the total expense, C the claimable expense, Dud the deductible, and r the reimbursement rate. The National Healthcare Security Administration (NHSA) issued a list of drugs and medical services in 2000.[12] Only the drugs and medical services on this list are claimable. The deductible varied across hospitals of different tiers but not across plans. Specifically, the deductibles were ¥500, ¥200, ¥100, and ¥50 when patients visited tier-3, tier-2, tier-1, and tier-0 hospitals, respectively. Based on this formula, the reimbursement rate (r) is defined with respect to the "feasible" claimable expense, which in turn is defined as the claimable expense minus the deductible ($C - Dud$).

Reimbursement rates varied not only across insurance plans but also across hospitals in different tiers (Table 2). Given the hospital tier, the reimbursement rate is highest (lowest) for the High (Low) plan. Assume an individual visited a tier-3 hospital; the URRBMI would reimburse ¥65, ¥50, and ¥35 for every ¥100 feasible claimable expense under the High, Medium, and Low plans, respectively. Given the plan, the reimbursement rate decreases with hospital tier. Assume an individual who enrolled in the High plan; the URRBMI would reimburse ¥65, ¥80, ¥85, and ¥90

[12]This list is adopted to balance the basic medical needs of enrollees and healthcare expense containment for the government. The drugs and medical services on the list are mostly essential, and their prices are low (Milcent, 2018). For example, expensively branded and imported drugs account for less than 5% of all items on the list. The list is updated from year to year.

for every ¥100 feasible claimable expense when she visited a tier-3, tier-2, tier-1, and tier-0 hospital, respectively.

Table 2: Reimbursement Rates by Plan and Hospital Tier

Insurance Plans	Hospital Tiers			
	Tier-0	Tier-1	Tier-2	Tier-3
Low	65%	60%	55%	35%
Medium	90%	80%	65%	50%
High	90%	85%	80%	65%

2.3 Policy Change

To increase the accessibility of healthcare services and alleviate the financial burden for enrollees, the government changed the plans and premiums by areas in 2010 (Table 1). First, only the High plan was available in urban areas, and Medium and High plans in rural areas in 2010. Second, premiums were changed across plans and areas. Specifically, the premium for the High plan decreased from ¥220 to ¥100 in urban areas; in rural areas, the premium for the High plan decreased from ¥220 to ¥140, and for the Medium plan from ¥120 to ¥40. Third, the premium subsidy was increased. Specifically, it increased from ¥80 to ¥220 for the High plan in urban areas; in rural areas, it increased from ¥80 to ¥180 for both Medium and High plans. Correspondingly, residents changed their plans. Most urban residents switched from the Medium to the High plan, and rural residents from the Low to the Medium plan. Despite these changes, we note that the reimbursement structure (Eq. (1) and reimbursement rates (Table 2)) did not change.

The policy change provides a unique opportunity for us to identify and estimate a general equilibrium model of the healthcare market. On the one hand, the change in reimbursement rate changes the price of healthcare, which helps us identify individuals' preference parameters in hospital choice on the demand side. On the other, the policy change may have differential impacts across hospitals of different tiers, and these differential impacts, which are rarely explored in the literature, would help us to empirically identify our structural model of hospital behavior on the supply side. The reason is that reimbursement rates differed across hospital tiers. Table 2 shows that when urban residents switched from the Medium to the High plan, the reim-

bursement rate increased by 15, 15, 5, and 0 percentage points if they visited a tier-3, 2, 1, and 0 hospital, respectively. When rural residents switched from the Low to the Medium plan, the reimbursement rate increased by 15, 10, 20, and 25 percentage points if they visited a tier-3, 2, 1, and 0 hospital, respectively.

3 Data and Descriptive Patterns

This section introduces the three datasets used in our empirical analysis and describes patients' and hospitals' responses to the policy change.

3.1 Data

Our analysis draws on three datasets. The first is URRBMI enrollment data, the second is URRBMI claims data, and the third is annual hospital report data. These three datasets complement each other and provide a unique opportunity to estimate a general equilibrium model of the healthcare market. Annual hospital report data complement enrollment and claim data by providing information on hospital characteristics, so we are able to estimate the healthcare demand function. Based on the estimated demand function, we then estimate the hospital cost functions. The policy change in the URRBMI helps us address the identification issue.

The enrollment data are a balanced panel from 2009 to 2010. The dataset includes a 10% random sample of all enrollees in the URRBMI in Chengdu. In total, there are 376,988 unique enrollees. The dataset contains each enrollee's basic information, including age, gender, where she lives, and the plan she chooses.

The claims data cover all inpatient visits in Chengdu from 2009 to 2010 and contain the admission and discharge dates, the name of the hospital visited, diagnosis codes, and healthcare expense. The information on healthcare expense is comprehensive and includes total expense, claimble expense, and expense broken down by drugs, tests, and medical services. Each enrollee is assigned a unique ID in both the enrollment and claims data, and this ID is the same for the same person in both datasets. Enrollment and claims data are provided by the Chengdu Healthcare Security Administration.

Annual hospital report data from 2009 to 2010 are provided by the Chengdu Municipal Health Commission. The data provide comprehensive information on hospital

characteristics such as hospital name, tier, and ownership and the number of doctors, nurses, inpatient beds, and advanced medical equipment.[13] The data also cover information on annual hospital operation including hospital revenue and expense; number of outpatient, inpatient, and emergency visits; emergency room (ER) mortality rate per 1,000 ER visits; inpatient mortality rate per 1,000 admissions; rate of adverse drug reactions per 1,000 inpatient visits; and number of medical malpractice lawsuits.

Our estimation sample includes two types of switchers induced by the policy change in public health insurance from 2009 to 2010. Enrollment data show that 55,422 (85.7%) urban residents switched from the Medium to the High plan, and 221,758 (70.9%) rural residents switched from the Low to the Medium plan;[14] thus, our sample size is 277,180. In our estimation sample, the average age was 46 in 2009, and half were male. Although enrollment in the URRBMI is voluntary, the enrollment rate was over 95% in our sample period. Therefore, we regard these two types of switches across health insurance plans as exogenous.

3.2 Descriptive Patterns

We now describe patients' and hospitals' responses to the policy change. Figure 1 plots the average expense for inpatient care per enrollee by month. We note that the average expense increased over time in both urban and rural areas. On average, the monthly expense increased from ¥25.9 in 2009 to ¥44.7 in 2010 in urban areas, and from ¥12.5 to ¥23.9 in rural areas.[15] The average expense per enrollee (TS) is determined by three components—the number of inpatient visits per enrollee (D), the fraction of patients who visit hospital h (s_h), and the expense per visit to hospital h (e_h):

$$TS = \left(\sum_{h \in \mathcal{H}} e_h \cdot s_h \right) \cdot D. \tag{2}$$

[13] This refers to medical equipment valued at more than ¥10,000.

[14] Urban residents who switched from the Medium to the High plan accounted for 97% of all urban switchers induced by the policy change. Rural residents who switched from the Low to the Medium plan accounted for 95.1% of all rural switchers induced by the policy change.

[15] We note that the part of the increase is due to the time trend. To remove the time trend, we first use the data in 2009 to regress the average expense on a constant and a time trend. Using the two estimated coefficients, we then predict the average expense in 2010. The predicted values of average expense are ¥39.2 in urban areas and ¥15.1 in rural areas. We will carefully consider the time trend in our estimation below.

The change in the average expense per enrollee is determined by changes in these three components, which we examine one by one.

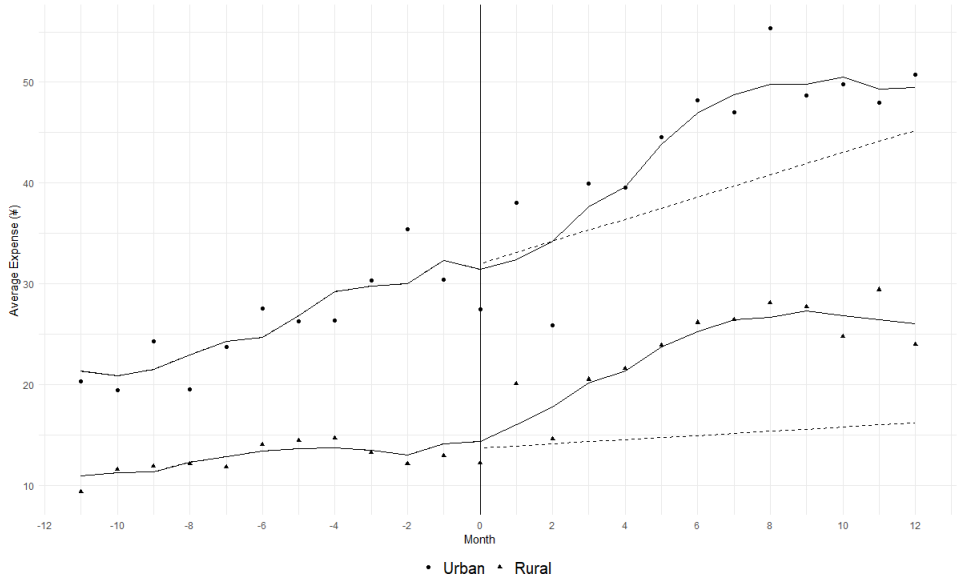

Figure 1: Total Expense per Enrollee

Note: The horizontal line indexes the calendar month with respect to December 2009, the last month before the policy change. The vertical line shows the average expense on inpatient care per enrollee. The dots (triangles) are for urban (rural) enrollees. Dots and triangles represent the values based on raw data; solid lines represent the 5-month moving average; and dashed lines represent the predicted values without the policy change. Using the data in 2009, we first regress the values based on raw data on a constant and a time trend. Using the two estimated coefficients, we then predict the value of the average expense in 2010.

Figure 2 plots the average number of visits per enrollee by month (D). We note that the number of visits increased over time in both urban and rural areas. To remove the time trend, we predict the number of visits without the policy in 2010. Using the data in 2009, we first regress the number of visits on a constant and a time trend. Based on these two estimates, we then predict the number of visits in 2010. Figure 2(a) shows that the observed number of visits is slightly small than the predicted one in urban areas. By contrast, Figure 2(b) shows that the observed number of visits is significantly larger than the predicted number in rural areas. Thus, we conclude that the number of visits increased in rural areas but not in urban ones.

We then examine the share of visits by hospital tiers (s_h). Figure 3 plots the

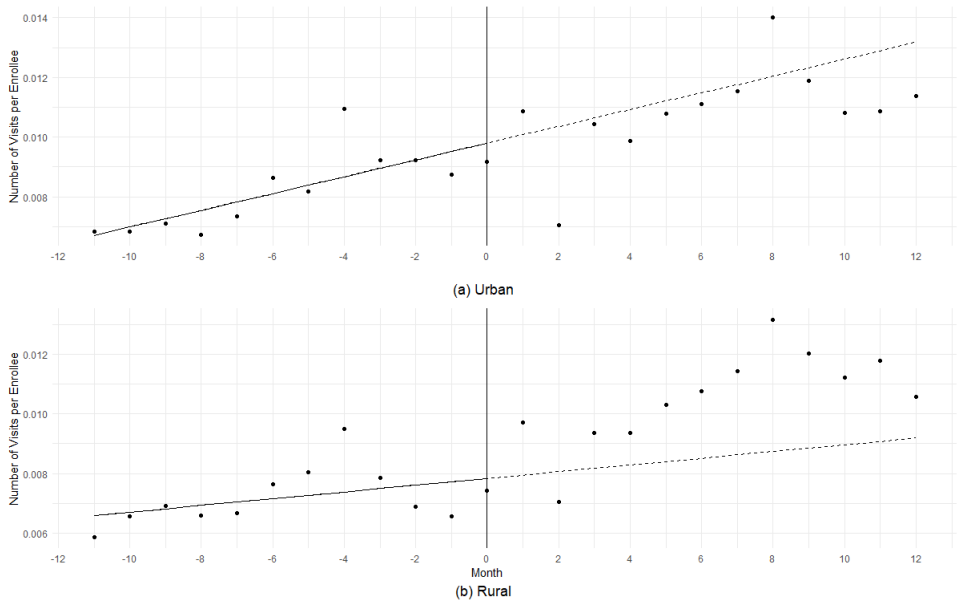

Figure 2: Number of Inpatient Visits per Enrollee

Note: Panels (a) and (b) represent urban and rural enrollees, respectively. The horizontal line indexes the calendar month with respect to December 2009, the last month before the policy change. The vertical line shows the average number of inpatient visits per enrollee. Dots represent the values based on raw data. Solid lines represent the predicted values in 2009. Using the data in 2009, we regress the number of visits on a constant and a time trend. Based on the two estimated coefficients, we predict the average number of inpatient visits in 2009. The dashed line represents the predicted values in 2010 without the policy change using the two estimated coefficients in 2009.

fraction of patients who visit high-tier/low-tier hospitals.[16] Figure 3(a) shows that the average share of urban patients who visited high-tier hospitals increased from 41.1% in 2009 to 43.9% in 2010, but the share who visited low-tier hospitals decreased from 58.9% to 56.1%. This result indicates that urban patients switched from low-tier to high-tier hospitals in response to the policy change. By contrast, Figure 3(b) does not show such a pattern of switching for rural patients.

We finally check the total expense (e_h) and OOP expense per visit. Figure 4 plots the monthly total and OOP expenses by areas and hospital tiers. An interesting pattern emerges. On the one hand, the total expense increased across areas and hospital tiers in all subfigures. For example, the total expense per visit increased from ¥3,982 in 2009 to ¥5,013 in 2010 for high-tier hospitals in urban areas. On the other, the OOP expense more or less remained at the same level. The only exception

[16]We define tier-3 and 2 (tier-1 and 0) hospitals as high-tier (low-tier) hospitals.

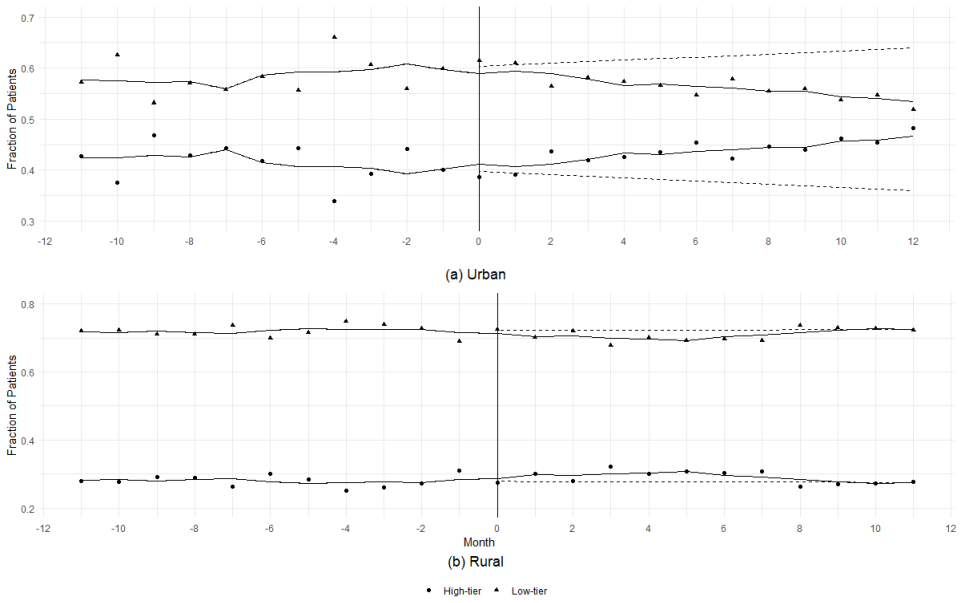

Figure 3: Fraction of Patients Visiting High-tier/Low-tier Hospitals

Note: Panels (a) and (b) represent urban and rural patients, respectively. We define tier-3 and 2 (tier-1 and 0) hospitals as high-tier (low-tier) hospitals. The horizontal line indexes the calendar month with respect to December 2009, the last month before the policy change. The vertical line shows the fraction of patients visiting high-tier/low-tier hospitals. The dots (triangles) are for high-tier (low-tier) hospital visits. Dots and triangles represent the values based on raw data; solid lines represent the 5-month moving average; and dashed lines represent the predicted values without the policy change. Using the data in 2009, we regress the values based on raw data on a constant and a time trend. Using the two estimated coefficients, we then predict the fraction of visits in 2010.

is for visits to low-tier hospitals in rural areas. The OOP expense decreased from ¥339 in 2009 to ¥239 in 2010, despite the increase in the total expense.

The pattern revealed in Figure 4 is interesting. We would expect the number of visits and the total expense to increase when the reimbursement rate (price) of healthcare service increases (decreases). But it is difficult to understand why the OOP expense per visit remained stable in a partial equilibrium framework. Finkelstein (2007) contends that market-wide changes in demand may alter hospitals' incentives. Thus, this pattern suggests that the supply side of the healthcare market—hospitals—may have endogenously responded to the policy change.

We then check hospitals' decisions to employ doctors and nurses and to invest in inpatient beds and advanced medical equipment. Table 3 shows that in response to the policy change, hospitals employed more doctors and nurses and bought more beds

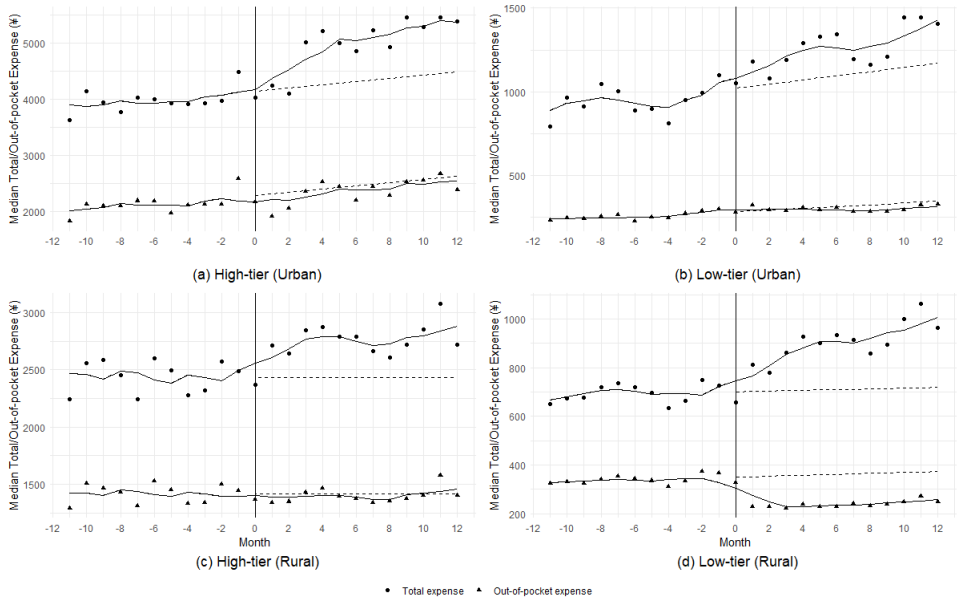

Figure 4: Total and OOP Expense per Visit

Note: Panels (a), (b), (c), and (d) show the total and OOP expenses of high-tier hospitals in urban areas, low-tier hospitals in urban areas, high-tier hospitals in rural areas, and low-tier hospitals in rural areas, respectively. We define the tier-3 and 2 (tier-1 and 0) hospitals as high-tier (low-tier) hospitals. There are four subfigures in the figure. In each subfigure, the horizontal line indexes the calendar month with respect to December 2009, the last month before the policy change. The vertical line shows the median total and OOP expenses. The dots (triangles) are for total (OOP) expenses based on raw data; solid lines represent the five-month moving average; and dashed lines represent the predicted values without the policy change. Using the data in 2009, we regress the values based on raw data on a constant and a time trend. Using the two estimated coefficients, we then predict the total and OOP expenses in 2010.

and advanced medical equipment in both urban and rural areas. The responses were larger for high-tier hospitals than low-tier ones.

In summary, our data reveal four stylized facts associated with the policy change:

1. Rural enrollees visited hospitals more.

2. Urban patients switched from low- to high-tier hospitals.

3. Although total healthcare expense per visit increased, the OOP expense per visit more or less remained the same.

4. Hospitals employed more doctors and nurses and bought more beds and advanced medical equipment.

Table 3: Hospitals' Employment and Investment Decisions

		# Doctors	# Nurses	# Beds	# Equipment
		(1)	(2)	(3)	(4)
Hospital Tiers	Year	Urban			
High	2009	185.20	265.59	482.12	507.92
	2010	199.80	304.53	520.31	617.71
Low	2009	21.92	22.08	73.59	25.90
	2010	23.76	26.37	76.00	28.08
		Rural			
High	2009	81.50	95.25	204.48	128.52
	2010	87.79	110.29	237.15	158.65
Low	2009	8.84	7.24	39.21	9.71
	2010	9.01	8.61	44.17	11.50

Note: Data source: annual hospital report data from 2009 to 2010. Sample size: 380 hospitals. We define tier-3 and 2 (tier-1 and 0) hospitals as high-tier (low-tier) hospitals. Equipment refers to medical equipment valued at more than ¥10,000.

4 Model

In this section, we formulate a general equilibrium model of the healthcare market. On the demand side, individuals face a discrete choice of hospital care. On the supply side, hospitals compete in a two-stage game. In the first stage, hospitals choose healthcare quality; in the second, they set the total expense and claimable expense per visit.

4.1 Individual Behavior

Discrete Choice for Hospital Care

We start with the individual's demand for hospital care. Motivated by the fact that most individuals have inpatient visits at most once per year in our sample, we employ a static discrete choice model of demand for hospital care with a time horizon of one year.[17] Individual i's indirect utility from visiting hospital h at year t, u_{iht}, is determined by the OOP expense (p_{ht}), observed hospital quality (q_{ht}), hospital tier

[17]Ninety-eight percent and 97% of individuals have inpatient visits at most once in 2009 and 2010, respectively.

(k_{ht}), and unobserved hospital quality (ξ_{ht}):

$$u\left(p_{ht}, q_{ht}, \xi_{ht}, \epsilon_{iht}; \mathbf{\Theta}\right) = \alpha_i p_{ht} + \beta_i q_{ht} + \lambda_i k_{ht} + \xi_{ht} + \epsilon_{iht}, \tag{3}$$

where ϵ_{iht} is a logit error. The vector of parameters $\mathbf{\Theta} = \{\alpha_i, \beta_i, \lambda_i\}$ represents the individual's preferences for OOP expense, observed hospital quality, and hospital tiers, which differ by age, gender, and residential areas:

$$\begin{pmatrix} \alpha_i \\ \beta_i \\ \lambda_i \end{pmatrix} = \begin{pmatrix} \bar{\alpha} \\ \bar{\beta} \\ \bar{\lambda} \end{pmatrix} + \text{age}_i \cdot \begin{pmatrix} \alpha_{\text{age}} \\ \beta_{\text{age}} \\ \lambda_{\text{age}} \end{pmatrix} + \text{male}_i \cdot \begin{pmatrix} \alpha_{\text{male}} \\ \beta_{\text{male}} \\ \lambda_{\text{male}} \end{pmatrix} + \text{urban}_i \cdot \begin{pmatrix} \alpha_{\text{urban}} \\ \beta_{\text{urban}} \\ \lambda_{\text{urban}} \end{pmatrix}.$$

The mean utility for visiting hospital h is

$$\delta_{ht}(p_{ht}, q_{ht}, k_{ht}, \xi_{ht}; \mathbf{\Theta}_1) = \bar{\alpha} p_{ht} + \bar{\beta} q_{ht} + \bar{\lambda} k_{ht} + \xi_{ht}, \tag{4}$$

where $\mathbf{\Theta}_1$ denotes all parameters in Eq. (4). The individual-specific utility is

$$\begin{aligned} &\mu_{iht}\left(p_{ht}, q_{ht}, k_{ht}, \text{age}_i, \text{male}_i, \text{urban}_i; \mathbf{\Theta}_2\right) \\ =& \left(\alpha_{\text{age}} \cdot \text{age}_i \cdot + \alpha_{\text{male}} \cdot \text{male}_i + \alpha_{\text{urban}} \cdot \text{urban}_i\right) \cdot p_{ht} \\ +& \left(\beta_{\text{age}} \cdot \text{age}_i + \beta_{\text{male}} \cdot \text{male}_i + \beta_{\text{urban}} \cdot \text{urban}_i\right) \cdot q_{ht} \\ +& \left(\lambda_{\text{age}} \cdot \text{age}_i + \lambda_{\text{male}} \cdot \text{male}_i + \lambda_{\text{urban}} \cdot \text{urban}_i\right) \cdot k_{ht}, \end{aligned} \tag{5}$$

where $\mathbf{\Theta}_2$ denotes all parameters in Eq. (5). Instead of visiting any hospital, individuals can stay at home and not have hospital care, from which their utility is assumed to be

$$u_{i0t} = \epsilon_{i0t}.$$

Thus, the individual's problem is to maximize their utility by visiting hospital h or staying at home:

$$u_{iht} \geq u_{ijt} \quad \forall j \neq h.$$

In equilibrium, the probability of visiting hospital h for individual i is given by

$$P_{iht} = \frac{\exp\left(\delta_{ht} + \mu_{iht}\right)}{1 + \sum\limits_{j \in \mathfrak{I}_{mt}} \exp\left(\delta_{jt} + \mu_{ijt}\right)}, \tag{6}$$

where $\mathfrak{I}_{mt}$ denotes all hospitals in market m in year t.[18]

Out-of-pocket Expense (p_{ht})

We calculate the average OOP expense for each hospital using Eq. (1). Our claims data record the total expense (e) and claimable expense (C). Our enrollment data record the health insurance plan for each individual, from which we know the deductible (Dud) and reimbursement rate (r). We assume each enrollee forms an expectation on the OOP expense for each hospital before making a hospital choice. We average the total and claimable expenses across all visits for each hospital in year t, and use Eq. (1) to compute the average OOP expense as a proxy for the expected OOP expense when individual i plans to visit hospital h in that year (p_{ht} in Eq. (3)).

Panels (a) and (b) in Table 4 summarize the total and OOP expenses by hospital tiers. We find that hospitals of higher tiers charged larger total and OOP expenses. This could mean that patients who visited higher-tier hospitals had more serious illnesses. In addition, the total expenses increased from 2009 to 2010 across hospitals of all tiers. By contrast, the OOP expense marginally increased for tier-3 hospitals but decreased for lower-tier hospitals.

Observed Hospital Quality (q_{ht})

We construct the observed hospital quality index (q_{ht} in Eq. (3)) using factor analysis:

$$\boldsymbol{x}_{ht} = \boldsymbol{\kappa} + \boldsymbol{\varsigma} \cdot \bar{q}_{ht} + \boldsymbol{\varrho}_{ht}, \tag{7}$$

where $\boldsymbol{x}_{ht}$ is a vector of four observed hospital characteristics: number of doctors (x^1_{ht}), nurses (x^2_{ht}), beds (x^3_{ht}), and advanced medical equipment (x^4_{ht}); $\boldsymbol{\kappa} \equiv (\kappa^1, \kappa^2, \kappa^3, \kappa^4)$ is the mean of $\boldsymbol{x}_{ht}$; $\boldsymbol{\varsigma}$ represents a vector of factor loadings; $\bar{q}_{ht}$ is the common factor; and $\boldsymbol{\varrho}_{ht}$ denotes a vector of unique factors (Lee, 2011). We estimate the vector of factor loadings ($\boldsymbol{\varsigma}$).

The common factor $\bar{q}_{ht}$ in Eq. (7) is

$$\bar{q}_{ht} = \vartheta_1 \cdot \frac{x^1_{ht} - \kappa^1}{\sigma^1} + \vartheta_2 \cdot \frac{x^2_{ht} - \kappa^2}{\sigma^2} + \vartheta_3 \cdot \frac{x^3_{ht} - \kappa^3}{\sigma^3} + \vartheta_4 \cdot \frac{x^4_{ht} - \kappa^4}{\sigma^4}, \tag{8}$$

[18] We define 12 different markets in our empirical analysis, because Chengdu is geographically divided into 12 regions. As discussed above, no referral system has been adopted to triage patients to different hospitals in China. Thus, in principle, residents can choose any hospital in Chengdu. However, the geographical distance between regions prevents residents from visiting hospitals across regions. By our definition, only 7.44% of all inpatient visits occurred when patients and hospitals belonged to different markets.

Table 4: Summary of Total, OOP, and Claimable Expenses, Claimable Ratio and Observed Quality Index

	(a) Total Expense		(b) OOP Expense	
Hospital Tiers	2009	2010	2009	2010
Tier-3	9.326 (8.712)	11.696 (11.432)	6.128 (5.749)	6.869 (6.905)
Tier-2	3.626 (3.358)	4.176 (3.936)	2.087 (1.994)	2.014 (1.891)
Tier-1	2.252 (1.898)	2.616 (2.237)	1.149 (0.939)	1.090 (0.894)
Tier-0	0.936 (0.856)	1.100 (1.027)	0.417 (0.370)	0.294 (0.263)
Hospital Tiers	(c) Claimable Expense		(d) Claimable Ratio	
Tier-3	6.898 (6.426)	7.926 (7.578)	0.742 (0.748)	0.681 (0.695)
Tier-2	2.825 (2.669)	3.262 (3.053)	0.786 (0.788)	0.790 (0.785)
Tier-1	1.746 (1.458)	1.958 (1.792)	0.786 (0.801)	0.769 (0.791)
Tier-0	0.796 (0.711)	0.946 (0.902)	0.860 (0.863)	0.869 (0.880)
Hospital Tiers	(e) Quality Index			
Tier-3	4.100 (2.458)	4.746 (2.879)		
Tier-2	0.654 (0.485)	0.703 (0.515)		
Tier-1	0.116 (0.088)	0.119 (0.097)		
Tier-0	0.063 (0.049)	0.069 (0.054)		

Note: This table presents the means of total, OOP, and claimable expenses; claimable ratio; and quality index. See Eq. (1) for the definition of claimable expense. The claimable ratio is defined as the claimable expense divided by total expense. The medians are in parentheses. The unit for Panels (a), (b), and (c) is ¥1,000.

where $\boldsymbol{\sigma} \equiv (\sigma^1, \sigma^2, \sigma^3, \sigma^4)$ is the standard deviation of $\boldsymbol{x}_{ht}$, and $\boldsymbol{\vartheta} \equiv (\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)$ is a vector of scoring coefficients, which is the inverse of the correlation matrix of observed hospital characteristics multiplied by the estimated vector of factor loadings ($\boldsymbol{\varsigma}$). Rearranging terms in Eq. (8), we calculate the observed hospital quality index (q_{ht} in Eq. (3)) as follows:

$$\underbrace{\bar{q}_{ht} + \vartheta_1 \cdot \frac{\kappa^1}{\sigma^1} + \vartheta_2 \cdot \frac{\kappa^2}{\sigma^2} + \vartheta_3 \cdot \frac{\kappa^3}{\sigma^3} + \vartheta_4 \cdot \frac{\kappa^4}{\sigma^4}}_{q_{ht}} = \underbrace{\frac{\vartheta_1}{\sigma^1}}_{\phi_1} \cdot x^1_{ht} + \underbrace{\frac{\vartheta_2}{\sigma^2}}_{\phi_2} \cdot x^2_{ht} + \underbrace{\frac{\vartheta_3}{\sigma^3}}_{\phi_3} \cdot x^3_{ht} + \underbrace{\frac{\vartheta_4}{\sigma^4}}_{\phi_4} \cdot x^4_{ht}. \tag{9}$$

Panel (e) in Table 4 summarizes the observed hospital quality index by tiers. We find that hospitals in higher tiers had better quality. In addition, the quality improved from 2009 to 2010 across hospitals in all tiers.

4.2 Hospital Behavior

On the supply side, hospitals play a two-stage game: Hospitals choose the observed quality (q_h) in the first stage and then simultaneously set the total expense (e_h) and claimable expense (C_h) in the second stage.

Stage 1

In the first stage, after observing the realization of the fixed-cost shock (ν_h), hospitals choose the observed quality (q_h), which incurs the fixed cost (fc_h).[19] Hospitals maximize the profit minus the fixed cost:

$$\max_{q_h} \quad \mathbb{E}_{(\boldsymbol{\xi},\boldsymbol{\chi},\boldsymbol{\iota})}\pi_h^{\mathrm{II}}\left(\boldsymbol{e}^*\left(\boldsymbol{q}\right),\boldsymbol{C}^*\left(\boldsymbol{q}\right);\boldsymbol{q}\right) - fc(q_h,\nu_h;\boldsymbol{\tau}), \tag{10}$$

where $\mathbb{E}_{(\boldsymbol{\xi},\boldsymbol{\chi},\boldsymbol{\iota})}\pi_h^{\mathrm{II}}$ denotes hospitals' expected variable profit conditional on their expectation on demand shock ($\boldsymbol{\xi}$) and marginal cost shocks ($\boldsymbol{\chi}$ and $\boldsymbol{\iota}$) in the market, and $e_h^*\left(\boldsymbol{q}\right)$ and $C_h^*\left(\boldsymbol{q}\right)$ represent the equilibrium total and claimable expenses in the second stage.[20] The fixed cost (fc_h) is a function of the observed hospital quality (q_h) and unobserved fixed cost shock (ν_h). The vector of parameters ($\boldsymbol{\tau}$) is to be estimated.

Stage 2

Given the hospital's choice on quality in the first stage, in the second stage, hospitals observe the realization of demand shocks (ξ_h) and marginal cost shocks (χ_h and ι_h), then compete with one another by simultaneously setting the total expense (e_h) and claimable expense (C_h). In the second stage, we assume that hospitals are partially altruistic. They maximize their variable profits (π_h^{II}) minus the psychological cost (c_h):

$$\max_{e_h,C_h} \quad \underbrace{\left(e_h - mc_h\right)\cdot S_h}_{\pi_h^{\mathrm{II}}} - c_h, \tag{11}$$

where the total expense (e_h) minus the marginal cost (mc_h) represents the hospital's markup, and S_h is the total demand for hospital h.

We assume that the psychological cost (c_h) is an increasing function of the claimable ratio (R_h), which is defined as the claimable expense divided by total expense ($R_h =$

[19] We suppress the time subscript t for presentational simplicity.

[20] Equilibrium total and claimable expenses also depend on other variables such as individual characteristics, which are omitted here for ease of exposition.

$\frac{C_h}{e_h}$). As stated in Section 2, only the drugs and medical services on the list issued by the NHSA are claimable. Some procedures or drugs are not on the list, such as imported brand drugs. However, these drugs are essential for patient health in some cases. When hospitals prescribe these drugs, on the one hand, the market share and profit decrease. On the other, patient health increases, and the psychological cost decreases. This creates a trade-off in setting the optimal claimable ratio. Without this trade-off, the claimable ratio is always 1—contradicting the data, in which the average claimable ratio is about 0.8 across all hospitals.

Panels (c) and (d) in Table 4 summarize the claimable expense and claimable ratio by tiers. We find that the claimable expense was larger but the claimable ratio was lower for hospitals in higher tiers. From 2009 to 2010, the claimable expense increased across hospitals in all tiers.

4.3 Optimality Conditions

In the SPNE, all hospitals simultaneously choose the observed quality (q), total expense (e), and claimable expense (C), which constitute a Nash equilibrium in every subgame. We now derive the optimality conditions for the observed quality, total expense, and claimable expense, which help us identify the hospital's cost structure.[21] We derive the optimality conditions by backward induction.

We start with the second stage, in which all hospitals choose total and claimable expenses simultaneously. We have two first-order conditions by differentiating the objective function in the second stage (Eq. (11)) with respect to the total expense (e_h) and claimable expense (C_h):

$$S_h + (e_h - mc_h) \cdot \frac{\partial S_h}{\partial e_h} = 0, \tag{12}$$

$$(e_h - mc_h) \cdot \frac{\partial S_h}{\partial C_h} - \frac{\partial c_h}{\partial R_h} \cdot \frac{\partial R_h}{\partial C_h} = 0. \tag{13}$$

The first equation shows that the hospital's markup increases but the demand decreases when the total expense increases. On the left-hand side of Eq. (12), the first term is positive and the second term ($\frac{\partial S_h}{\partial e_h}$) is negative. When the claimable expense

[21] Following the literature, we assume the existence of a pure-strategy SPNE for the two-stage game. Finding a set of sufficient conditions for the existence of a Nash equilibrium is beyond the scope of this paper (Fan, 2013; Eizenberg, 2014).

increases, the second equation shows that the demand increases but the psychological cost also increases because the hospital cares about patient health. For a given level of total expense, the OOP expense decreases with the claimable expense. Thus, the demand increases with the claimable expense ($\frac{\partial S_h}{\partial C_h} > 0$), and the first term on the left-hand side of Eq. (13) is positive; the second term is positive because both $\frac{\partial c_h}{\partial R_h}$ and $\frac{R_h}{C_h}$ are positive.

By Eq. (12), we have

$$mc_h = e_h + S_h / \frac{\partial S_h}{\partial e_h}. \tag{14}$$

We then let the hospital's marginal cost linearly depend on the observed hospital quality (q_h), hospital location (l_h), year, and an error term for marginal cost shock (χ_h):[22]

$$mc_h = \gamma_0 + \gamma_1 \cdot q_h + \gamma_2 \cdot l_h + \gamma_3 \cdot \mathbf{1}\left(\text{Year} = 2010\right) \cdot l_h + \chi_h, \tag{15}$$

where l_h is a dummy variable indicating that the hospital is located in an urban area, and $\mathbf{1}\left(\text{Year} = 2010\right)$ is a dummy variable indicating that the year is 2010. The location dummy (l_h) is used to capture the difference in the marginal cost between urban and rural areas.

By Eq. (13), we have

$$\frac{\partial c_h}{\partial R_h} = \left(e_h - mc_h\right) \cdot \frac{\partial S_h}{\partial R_h}. \tag{16}$$

We then let the hospital's marginal psychological cost linearly depend on the hospital tier (k_h), hospital location, year, and an error term for marginal psychological cost shock (ι_h):

$$\ln\left(\frac{\partial c_h}{\partial R_h}\right) = \omega_0 + \omega_1 \cdot k_h + \omega_2 \cdot l_h + \omega_3 \cdot \mathbf{1}\left(\text{Year} = 2010\right) \cdot l_h + \iota_h. \tag{17}$$

It might be better to assume that the marginal psychological cost is a function of the severity of patient illness, for which our data do not record information. We use the hospital tier as a proxy variable for patient severity, because higher-tier hospitals are designated to treat patients with more severe conditions in China's healthcare system.

In the first stage, all hospitals choose their quality simultaneously. We have one first-order condition by differentiating the objective function in the first stage (Eq.

[22] We use the year dummy to take care of the time trend.

(10)) with respect to the quality (q_h):

$$\frac{\partial \mathbb{E}\pi_h^{\mathrm{II}}}{\partial q_h} - \frac{\partial f c_h}{\partial q_h} = 0, \tag{18}$$

where

$$\frac{\partial \mathbb{E}\pi_h^{\mathrm{II}}}{\partial q_h} = \mathbb{E}\left[\left(\frac{\partial e_h^*}{\partial q_h} - \frac{\partial m c_h}{\partial q_h}\right) \cdot S_h + (e_h - m c_h) \cdot \left(\frac{\partial S_h}{\partial q_h} + \sum_{j \in \mathfrak{I}_{m(h)}} \frac{\partial S_h}{\partial e_j} \cdot \frac{\partial e_j^*}{\partial q_h} + \sum_{j \in \mathfrak{I}_{m(h)}} \frac{\partial S_h}{\partial C_j} \cdot \frac{\partial C_j^*}{\partial q_h}\right)\right], \tag{19}$$

and $\mathfrak{I}_{m(h)}$ refers to all hospitals in market m that h belongs to.

Eq. (18) shows that the hospital's profit increases but the fixed cost also increases when the quality increases. On the left-hand side of Eq. (18), both the first and second terms are positive. Eq. (19) shows that the effect of quality on the hospital's expected profit can be decomposed into two components: the effect on markup and the effect on demand. The first term on the right-hand side of Eq. (19) captures the effect on the profit through the effect on the markup. When the quality increases, both total expense and marginal cost increase, and thus the markup changes. The second term captures the effect on the profit through the effect on the demand. It can be further decomposed into two channels: a direct channel $\left((e_h - m c_h) \cdot \frac{\partial S_h}{\partial q_h}\right)$ and an indirect channel $\left((e_h - m c_h) \cdot \left(\sum_{j \in \mathfrak{I}_{m(h)}} \frac{\partial S_h}{\partial e_j} \cdot \frac{\partial e_j^*}{\partial q_h} + \sum_{j \in \mathfrak{I}_{m(h)}} \frac{\partial S_h}{\partial C_j} \cdot \frac{\partial C_j^*}{\partial q_h}\right)\right)$. The direct channel shows how a change in quality affects demand for the hospital directly. The indirect channel shows how a change in quality affects demand through an impact on the equilibrium total and claimable expense for all hospitals in the market.

Finally, we parameterize the fixed cost function as follows:

$$f c\left(q_h, \nu_h; \boldsymbol{\tau}\right) = e^{\tau_0 + \tau_1 \cdot \ln(q_h) + \nu_h}, \tag{20}$$

where ν_h represents an error term for fixed cost shock.

5 Estimation

In this section, we describe the estimation methods and report the estimation results.

5.1 Methods

We have 22 parameters to be estimated: 12 parameters in the discrete choice of demand for hospital care (Eqs. (4) and (5)) and 10 parameters in the hospital cost structure. The latter include 4 parameters in the marginal cost function (Eq. (15)); 4 in the marginal psychological cost function (Eq. (17)); and 2 in the fixed cost function (Eq. (20)). We sequentially estimate our general equilibrium model in three steps (Eizenberg, 2014).[23]

Step 1

In the first step, we estimate the discrete choice model of demand for hospital care. Estimating Eq. (3) is difficult because ξ_{ht} is unobserved, which correlates with p_{ht}. To address this concern, we use a nested fixed point algorithm to estimate the model (Berry et al., 1995). This algorithm consists of an outer loop to search over the parameter space of $\mathbf{\Theta}_2$ by maximizing the log-likelihood function:

$$LL\left(\mathbf{\Theta}_2\right) = \sum_i \sum_t \sum_h 1_{iht} \cdot \ln\left[\frac{\exp\left(\delta_{ht} + \mu_{iht}\right)}{1 + \sum\limits_{j \in \Im_{mt}} \exp\left(\delta_{jt} + \mu_{ijt}\right)}\right], \tag{21}$$

where 1_{iht} is an indicator that hospital h is chosen by individual i in year t, and an inner loop to invert out the mean utility δ_{ht} for any given conjectured value of $\mathbf{\Theta}_2$ using a contraction mapping:

$$\delta_{ht}^{(N+1)} = \delta_{ht}^{(N)} + \ln\left(S_{ht}\right) - \ln\left(\sum_i \frac{\exp\left(\delta_{ht}^{(N)} + \mu_{iht}\right)}{1 + \sum\limits_{j \in \Im_{mt}} \exp\left(\delta_{jt}^{(N)} + \mu_{ijt}\right)}\right), \tag{22}$$

where $\delta_{ht}^{(N)}$ is the mean utility at the Nth iteration. This algorithm is computationally demanding because inverting out the mean utility in the inner loop is time consuming (Lee and Seo, 2015). To accelerate convergence of the inner loop, we adopt the SQUAREM algorithm developed by Varadhan and Roland (2008).[24] Appendix C

[23] Two methods are commonly used to estimate the general equilibrium model in the literature: One is the GMM to estimate demand and supply jointly (Berry and Jia, 2010); the other is to estimate demand and supply sequentially (Gowrisankaran et al., 2015; Barwick et al., 2020).

[24] Reynaerts et al. (2012) compare computational performance for various fixed point acceleration algorithms and find that the SQUAREM algorithm works well on BLP contraction mapping in Eq. (22).

details the algorithm.

We estimate the mean utility (δ_{ht}) in Eq. (4) and the parameters in the individual-specific utility function ($\mathbf{\Theta}_2$) in Eq. (5) using the nested fixed point algorithm. We proceed to estimating $\mathbf{\Theta}_1$ in Eq. (4). We use a two-stage least squares (2SLS) estimator because the OOP expense (p_{ht}) correlates with the unobserved hospital quality (ξ_{ht}). In addition to the standard instruments used by Berry et al. (1995)—the number of doctors, nurses, beds, and advanced medical equipment in hospitals other than h in market m—we explore a new instrument, the reimbursement rate, which is described in Section 2.2. The new instrument strengthens the identification of individuals' preference parameters in the discrete choice model of demand. On the one hand, the reimbursement rate correlates with the OOP expense by definition (Eq. (1)). On the other, the hospital's unobserved quality does not correlate with the reimbursement rate, because the change in the latter is induced by the policy change, which is not predicted by hospitals.

Step 2

We proceed with the second step, in which we estimate the marginal cost function (Eq. (15)) and marginal psychological cost function (Eq. (17)) in the second stage of the game for hospital competition. We first calculate $\frac{\partial S_h}{\partial e_h}$ and $\frac{\partial S_h}{\partial R_h}$ for all hospitals after estimation of the discrete choice model of demand for hospital care. We then calculate the marginal cost (mc_h) and marginal psychological cost ($\frac{\partial c_h}{\partial R_h}$) by Eqs. (14) and (16), respectively. Finally, we estimate Eqs. (15) and (17) using the ordinary least squares (OLS) estimator. OLS estimates are consistent because we assume that hospitals choose quality before the realization of marginal cost shocks, as discussed in Section 4.2.

Step 3

In the final step, we estimate the parameters in the fixed cost function (Eq. (20)). We first differentiate Eq. (20) with respect to quality:

$$\frac{\partial fc_h}{\partial q_h} = \frac{\tau_1}{q_h} \cdot e^{\tau_0 + \tau_1 \cdot \ln(q_h) + \nu_h}. \tag{23}$$

We have

$$\ln\left(\frac{\partial fc_h}{\partial q_h}\right) = \ln\left(\frac{\tau_1}{q_h}\right) + \tau_0 + \tau_1 \cdot \ln(q_h) + \nu_h. \tag{24}$$

Three challenges complicate estimating Eq. (24). First, we have to calculate $\frac{\partial fc_h}{\partial q_h}$. By the first-order condition of Eq. (18), $\frac{\partial fc_h}{\partial q_h} = \frac{\partial \mathbb{E}\pi_h^{\mathrm{II}}}{\partial q_h}$. We use Eq. (19) to calculate $\frac{\partial \mathbb{E}\pi_h^{\mathrm{II}}}{\partial q_h}$.[25] Calculating $\frac{\partial \mathbb{E}\pi_h^{\mathrm{II}}}{\partial q_h}$ is difficult, because we have to calculate $\frac{\partial e_j^*}{\partial q_h}$ and $\frac{\partial C_j^*}{\partial q_h}$ for all hospitals in the market in equilibrium. We use the theorem of the system of implicit functions to perform the calculation, the details of which are presented in Appendix D. Second, hospital quality (q_h) correlates with the unobserved fixed cost shock (ν_h). Third, Eq. (24) is nonlinear with respect to parameter τ_1.

We employ the GMM to perform the estimation. We use two types of instruments, which are denoted by $\boldsymbol{Z}$. The first includes the average age, fraction of males, and their squares for individuals in market m that hospital h belongs to. These instruments are standard in the literature (Fan, 2013). The second includes proxy variables for the price of q_h, such as the average annual wage rate for medical personnel and the average years of education of individuals in market m that hospital h belongs to. The second type of instrument substantiates our identification for parameters in the fixed cost function. The moment condition for our GMM estimation is

$$\mathbb{E}\left(\boldsymbol{\nu}\boldsymbol{Z}\right) = 0. \tag{25}$$

5.2 Results

Table 5 presents parameter estimates. Most parameters are precisely estimated, and the signs of parameter estimates are consistent with the theory. Individuals' utility decreases with the OOP expense ($\bar{\alpha} < 0$). Male and urban residents are more sensitive to the OOP expense, since both α_{male} and α_{urban} are statistically significantly negative. By contrast, the disutility associated with the OOP expense is smaller for elders ($\alpha_{\text{age}} > 0$), because they are less healthy and their demand for hospital care is less elastic. Individuals value higher hospital quality ($\bar{\beta} > 0$). Male and urban residents are more sensitive to hospital quality, since both β_{male} and β_{urban} are statistically significantly positive. By contrast, elders are less sensitive to hospital quality ($\beta_{\text{age}} < 0$). The reason might be that elders are less likely to switch across hospitals because they are more likely to suffer from chronic diseases. Individuals' utility increases

[25] For computational simplicity, we assume that all hospitals of a given tier in urban (rural) areas form expectations for variable profits based on the means of $\boldsymbol{\xi}$, $\boldsymbol{\chi}$, and $\boldsymbol{\iota}$ in a given tier in urban (rural) areas.

Table 5: Estimation Results

	Parameters	Estimates	Standard Errors
(a) Mean Utility:			
OOP Expense (¥1,000)	$\bar{\alpha}$	-8.8706***	2.0848
Observed Quality	$\bar{\beta}$	6.1156**	2.4252
Hospital Tier	$\bar{\lambda}$	5.0835***	1.5788
(b) Individual-specific Utility:			
(b1) Interactions With OOP Expense:			
Age/100	α_{age}	2.4287***	0.0516
Male	α_{male}	-0.1673***	0.0165
Urban	α_{urban}	-0.5556***	0.0336
(b2) Interactions with Observed Quality:			
Age/100	β_{age}	-1.1171***	0.0350
Male	β_{male}	0.1103***	0.0106
Urban	β_{urban}	0.2248***	0.0152
(b3) Interactions with Hospital Tier:			
Age/100	λ_{age}	-0.8537***	0.0523
Male	λ_{male}	-0.0021	0.0168
Urban	λ_{urban}	-1.2378***	0.0562
(c) Marginal Cost:			
Constant	γ_0	1.1829***	0.0694
Observed Quality	γ_1	1.3898***	0.4066
Location Dummy	γ_2	0.9864***	0.3076
Location Dummy * Year-2010 Dummy	γ_3	0.6394	0.4229
(d) Marginal Psychological Cost:			
Constant	ω_0	3.9571***	0.0431
Hospital Tier	ω_1	0.5248***	0.0638
Location Dummy	ω_2	0.3247**	0.1593
Location Dummy * Year-2010 Dummy	ω_3	0.5967***	0.1984
(e) Fixed Cost:			
Constant	τ_0	-3.3269***	0.1067
Ln(Observed Quality)	τ_1	0.7866***	0.0859

Note: Standard errors of parameters reported in Panel (b) are calculated by taking the square root of the diagonal elements in the negative inverse of the Hessian matrix. We use the automatic differentiation routine to calculate the Hessian matrix (Revels et al., 2016). Standard errors of all other parameters are calculated by bootstrapping.
*** Significant at the 1% level
** Significant at the 5% level

with hospital tier. Both elders and urban residents are less sensitive to hospital tier, since both λ_{age} and λ_{urban} are statistically significantly negative. We find no gender difference in the preference for hospital tier.

Turning to parameter estimates in the hospital cost structure, the marginal cost increases with the quality ($\gamma_1 > 0$). The positive sign of γ_2 indicates that hospitals located in urban areas have higher marginal costs than those located in rural ones. The marginal psychological cost increases with hospital tier ($\omega_1 > 0$), because the medical condition is on average more serious for patients who visit higher-tier hospitals. Physicians in higher-tier hospitals are more likely to prescribe drugs and provide

services outside the list issued by the NHSA for their patients. Thus, the claimable ratio is lower in higher-tier hospitals. Compared with those in rural areas, hospitals in urban areas have higher marginal psychological cost because they are more likely to be higher-tier ($\omega_2 > 0$). The fixed cost increases with quality ($\tau_1 > 0$). The elasticity of the fixed cost with respect to quality is precisely estimated and around 0.8.

6 Simulation and Decomposition

We decompose the difference in the simulated healthcare expense with and without the policy as follows:

$$\begin{aligned} TS^2 - TS^1 = & \underbrace{\sum_h e_h^1 \cdot \left(D^2 \cdot s_h^2 - D^2 \cdot s_h^1\right)}_{\text{Patient Sorting}} + \underbrace{\sum_h e_h^1 \cdot \left(D^2 \cdot s_h^1 - D^1 \cdot s_h^1\right)}_{\text{Quantity increase}} \\ & + \underbrace{\sum_h \left(e_h^2 - e_h^*\right) \cdot D^1 \cdot s_h^1}_{\text{Quality adjustment}} + \underbrace{\sum_h \left(e_h^* - e_h^1\right) \cdot D^1 \cdot s_h^1}_{\text{Price adjustment}} \\ & + \underbrace{\sum_h \left(e_h^2 - e_h^1\right) \cdot \left(D^2 \cdot s_h^2 - D^1 \cdot s_h^1\right)}_{\text{Cross}}, \end{aligned} \tag{26}$$

where the superscript 2 (1) denotes the simulated case with (without) the policy; TS, e, D, and s are defined in Eq. (2). Note that e^* denotes the simulated average total expense set by the hospital with the policy but holding the quality at the level without the policy.

We decompose the change in healthcare expense into five terms in Eq. (26). The first term, "patient sorting," reflects the change in healthcare expense due to patients' switching across hospitals, holding the total number of hospital visits constant. The second term, "quantity increase," reflects the change due to the change in the total number of hospital visits, assuming that patients do not switch across hospitals. Both terms reflect responses from the demand side. The third term, "quality adjustment," reflects the change in healthcare expense due to the change in hospital quality. The fourth term, "price adjustment," reflects the change due to the change in average total expense set by the hospital. Both terms reflect responses from the supply side. The final term, "cross," reflects the covariance between demand- and supply-side

responses.

We must simulate our model to conduct the decomposition. However, it is computationally infeasible to fully simulate our model for the 12 markets in Chengdu, because we would have to simulate the SPNE for hospital competition in each market. With a large number of hospitals, the computation is time consuming because we would have to iteratively find the optimal quality for all hospitals in the first stage—and for any guess regarding quality, we would have to simulate the total and claimable expenses at the new equilibrium for all hospitals in the second stage.

Therefore, we make two major assumptions to simplify our simulation. First, we assume two markets only. One is an urban market, and the other is a rural market. Second, in the urban market, we assume four hospitals that are at the tiers of 3, 2, 1, and 0; in the rural market, we assume three hospitals that are at the tiers of 2, 1, and 0 [26]. Focusing on a small number of hospitals in two markets greatly reduces our computational burden. At the same time, the decomposition results based on this simplified simulation illustrate the importance of each channel through which the health insurance affects healthcare expense. When simulating the model, 20,000 individuals are randomly drawn from our data in each area. Appendix E details our simulation.

Table 6 presents simulation and decomposition results for the urban market. Columns (1) and (2), respectively, show simulation results for individuals' and hospitals' endogenous choices without and with the policy. The simulated individual and hospital responses to the policy change are consistent with the stylized facts presented in Section 3.2. Column (3) shows the decomposition result. We first examine the demand side. The total number of visits per enrollee increases from 0.021 to 0.022, which accounts for 7.37% of the increase in total expense per enrollee. We also observe that patients switch from low-tier to high-tier hospitals. For example, the share of patients who visit tier-3 hospitals increases from 34.01% to 37.74%. Patient sorting accounts for 14.26% of the increase in total expense. In total, patient responses from the two channels on the demand side account for 21.63% of the increase in total expense.

We then examine the supply side. We first observe that quality increases for hospitals across all tiers. The quality adjustment accounts for 33.24% of the increase in total expense. In particular, the quality adjustment from tier-3 hospitals accounts

[26] No tier-3 hospitals in rural areas are in our data.

Table 6: Simulation and Decomposition Results for Urban Areas

	w/o policy (1)	w/ policy (2)		Decomposition (3)
			Total change	100.00%
Demand side			Demand side	21.63%
Total number of visits per enrollee	0.021	0.022	(a) Quantity increase	7.37%
Market share			(b) Sorting	14.26%
Tier-3	34.01%	37.74%		
Tier-2	30.46%	28.16%		
Tier-1	12.02%	11.70%		
Tier-0	23.51%	22.40%		
Supply side			Supply side	71.62%
Hospital quality			(c) Quality adjustment	33.24%
Average	0.541	0.662		
Tier-3	1.435	1.674	(c1) Tier-3	18.58%
Tier-2	0.540	0.705	(c2) Tier-2	11.48%
Tier-1	0.120	0.161	(c3) Tier-1	1.13%
Tier-0	0.068	0.106	(c4) Tier-0	2.05%
Total expense per visit (¥1,000)			(d) Price adjustment	38.38%
Average	4.485	5.043		
Tier-3	6.251	7.353	(d1) Tier-3	43.05%
Tier-2	5.073	5.425	(d2) Tier-2	6.14%
Tier-1	3.182	3.633	(d3) Tier-1	7.78%
Tier-0	1.835	1.407	(d4) Tier-0	-18.59%
Claimable expense per visit (¥1,000)				
Average	3.733	3.980		
Tier-3	4.830	5.501		
Tier-2	4.445	4.498		
Tier-1	2.673	3.107		
Tier-0	1.765	1.222		
			(e) Cross	6.75%
Market Equilibrium				
Average total expense per enrollee (¥)	95.902	108.910		
Average claimable expense per enrollee (¥)	79.814	85.951		
Average OOP expense per enrollee (¥)	49.141	49.945		

Note: Columns (1) and (2) show the simulation results without and with the policy, respectively. Column (3) shows the decomposition result of the difference in the simulated total expense per enrollee with and without the policy based on Eq. (26).

for 18.58% of the increase in total expense. Second, we find that the average total expense per visit increases from ¥4,485 to ¥5,043, and the average claimable expense per visit increases from ¥3,733 to ¥3,980. The price adjustment, including changes in both total expense and claimable expense per visit, accounts for 38.38% of the increase in total expense per enrollee. In total, hospital responses from the two channels on the supply side account for 71.62% of the increase in total expense per enrollee.

We observe that tier-0 hospitals increase their quality, which is similar to higher-

tier hospitals. But different from higher-tier hospitals, tier-0 hospitals decrease the total expense per visit. This is because the reimbursement rate remains unchanged for tier-0 hospitals but increases for higher-tier hospitals when urban residents switch from the Medium to the High plan induced by the policy change (Table 1). To compete with higher-tier hospitals, tier-0 hospitals decrease the total expense. This fact explains why the price adjustment by tier-0 hospitals accounts for -18.59% of the increase in total expense per enrollee.

In equilibrium, the average total and claimable expense per enrollee increases from ¥95.9 to ¥108.9 and from ¥79.8 to ¥86.0, respectively. The average OOP expense per enrollee remains almost unchanged.

Table 7 presents the simulation and decomposition results for the rural market. We first examine the demand side. The total number of visits per enrollee increases from 0.008 to 0.012, which accounts for 62.57% of the increase in total expense per enrollee. Different from urban patients, a small proportion of rural patients switch from high-tier to low-tier hospitals. This is because the reimbursement rate increases more for visiting low-tier than high-tier hospitals when rural residents switch from the Low to the Medium plan induced by the policy change (Table 1). Therefore, the policy change disincentivizes rural residents to switch from low-tier to high-tier hospitals. In total, patient responses from the two channels on the demand side account for 61.55% of the increase in total expense.

We next examine the supply side. We first observe that quality increases for hospitals across all tiers, which accounts for 8.79% of the increase in total expense. In particular, the quality adjustment from tier-2 hospitals accounts for 7.39% of the increase in total expense. Second, we find that the average total expense per visit increases from ¥1,985 to ¥2,331, and the average claimable expense per visit increases from ¥1,610 to ¥1,879. The price adjustment, including changes in both total expense and claimable expense per visit, accounts for 19.02% of the increase in total expense per enrollee. In total, hospital responses from the two channels on the supply side account for 27.81% of the increase in total expense per enrollee.

In equilibrium, the average total and claimable expense per enrollee increase from ¥16.7 to ¥27.7 and from ¥13.5 to ¥22.3, respectively. The average OOP expense per enrollee merely increases from ¥9.5 to ¥12.6.

In summary, supply-side responses mainly account for the increase in total expense per enrollee in the urban market, while demand-side responses mainly account for the

Table 7: Simulation and Decomposition Results for Rural Areas

	w/o policy (1)	w/ policy (2)		Decomposition (3)
			Total change	100.00%
Demand side			Demand side	61.55%
Total number of visits per enrollee	0.008	0.012	(a) Quantity increase	62.57%
Market share			(b) Sorting	-1.02%
Tier-2	45.79%	44.15%		
Tier-1	18.22%	20.06%		
Tier-0	35.99%	35.79%		
Supply side			Supply side	27.81%
Hospital quality			(c) Quality adjustment	8.79%
Average	0.151	0.218		
Tier-2	0.305	0.456	(c1) Tier-2	7.39%
Tier-1	0.091	0.118	(c2) Tier-1	0.52%
Tier-0	0.058	0.081	(c3) Tier-0	0.88%
Total expense per visit (¥1,000)			(d) Price adjustment	19.02%
Average	1.985	2.331		
Tier-2	2.809	3.392	(d1) Tier-2	13.09%
Tier-1	2.083	2.236	(d2) Tier-1	1.63%
Tier-0	0.887	1.074	(d3) Tier-0	4.30%
Claimable expense per visit (¥1,000)				
Average	1.610	1.879		
Tier-2	2.242	2.715		
Tier-1	1.660	1.731		
Tier-0	0.780	0.932		
			Cross	10.64%
Market equilibrium				
Average total expense per enrollee (¥)	16.706	27.663		
Average claimable expense per enrollee (¥)	13.546	22.304		
Average OOP expense per enrollee (¥)	9.507	12.621		

Note: Columns (1) and (2) show the simulation results without and with the policy, respectively. Column (3) shows the decomposition result of the difference in the simulated total expense per enrollee with and without the policy based on Eq. (26).

increase in the rural market. Our finding on the difference between the urban and rural market is consistent with the finding in the RAND health insurance experiment. Manning et al. (1987) show that the price elasticity of healthcare demand is higher when the reimbursement rate is lower. In our study, the reimbursement rate is lower in rural areas than in urban ones before the policy change. When the government increases the reimbursement rate, rural residents more actively respond to the policy change.

7 Welfare Analysis

In this section, we investigate the welfare effect of the policy change. Following Small and Rosen (1981), the expected consumer surplus for individual i is

$$\begin{aligned} E\left(CS_i\right) &= \frac{1}{\alpha_i} E\left(\max_h \left(u_{ih}\right)\right) \\ &= \frac{1}{\alpha_i}\left[\log \sum_h \exp\left(\delta_h + \mu_{ih}\right) + \text{Euler Constant}\right]. \end{aligned} \tag{27}$$

We then define the expected welfare for individual i as follows:

$$EW_i = E(CS_i) - P_i, \tag{28}$$

where P_i denotes the premium for the insurance plan the individual has chosen. The effect of the policy change on individual i's welfare is

$$\Delta_i = EW_i^{2010} - EW_i^{2009}. \tag{29}$$

Figure 5 presents our welfare analysis results. We have three findings. First, comparing Subfigures (a)-(b) with (c)-(d), the policy change enhanced the welfare of urban residents but deteriorated that of rural ones. Specifically, the welfare for male residents in urban areas increased between ¥25.64 and ¥33.41, while that in rural areas decreased between ¥15.75 and ¥12.83. The welfare for female residents in urban areas increased between ¥26.28 and ¥38.01, while that in rural areas decreased between ¥15.28 and ¥11.83.

The difference in the change in the premium explains the difference in the change in the welfare between urban and rural residents. According to Eqs. (28) and (29), the change in the welfare (Δ) is the sum of the change in the consumer surplus ($\Delta\left(CS\right)$) and the change in the insurance premium (ΔP). In urban areas, the consumer surplus increased between ¥5.64 and ¥18.01 for different age groups, and the premium decreased by ¥20 from ¥120 to ¥100 for all urban residents. The net welfare increased for urban residents. In rural areas, although the consumer surplus increased between ¥4.25 and ¥8.17 for different age groups, the premium increased by ¥20 from ¥20 to ¥40 for all rural residents. The net welfare decreased.

Second, in each subfigure, welfare increased more or decreased less for elders than

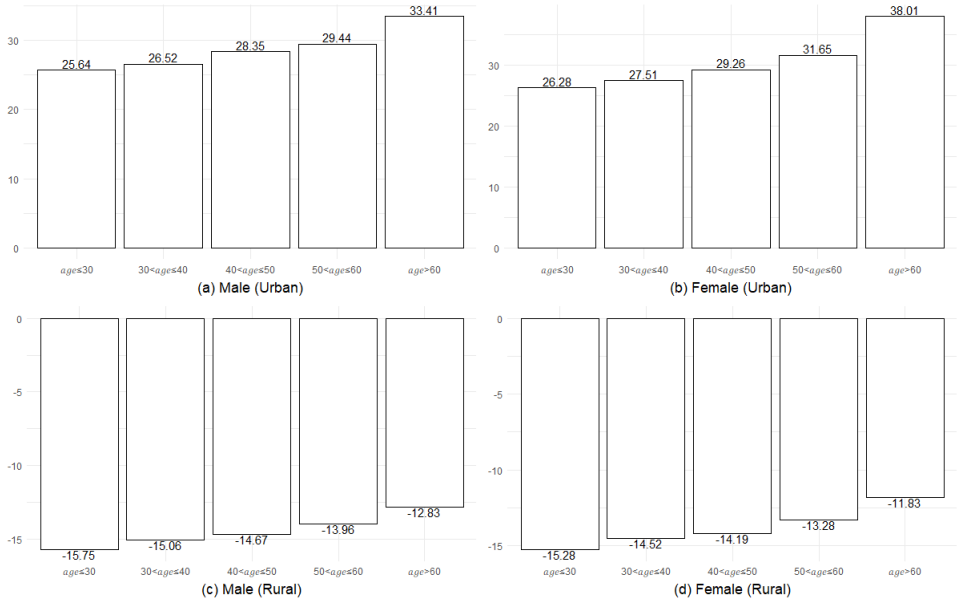

Figure 5: The Change in Expected Welfare (¥)

Note: Subfigures (a), (b), (c), and (d) show the change in expected welfare for urban male, urban female, rural male, and rural female residents, respectively.

youngers. The reason is that elders visit hospitals more frequently, and correspondingly, their consumer surplus increased more. Third, comparing Subfigures (a)-(c) with (b)-(d), female residents benefited more from the policy change than male residents. The reason might be that compared with males, females are less sensitive to the OOP expense, as suggested by the negative estimate of α_{male} in Table 5.

8 Conclusion

In this paper, we formulate and estimate a structural model to understand the equilibrium effect of health insurance on healthcare expense by examining a change in the public health insurance in China. Based on model estimates, our simulation analysis quantitatively disentangles demand-side responses from supply-side responses in accounting for the equilibrium effect. We find that both demand and supply factors significantly contribute to the equilibrium effect of health insurance on healthcare expense. However, the relative importance of demand and supply responses differs by areas. Quality and price adjustments from the supply side mainly explain the equilibrium effect in urban areas; the increase in number of visits from the demand

side mainly explains the equilibrium effect in rural areas. We further use the model estimates to evaluate the welfare consequence of the policy change, and find that the policy change enhanced the welfare of urban residents but deteriorated that of rural ones.

A couple of limitations to this paper need to be acknowledged. First, throughout the paper, we assume that hospitals and physicians have same incentives, since all physicians are employees of hospitals in China. Thus, the results may not be directly generalizable to other countries in which physicians are not employees of hospitals. Second, we model the interaction between public insurance and hospitals, and thus the model is less suitable for understanding the interaction between commercial insurance and hospitals. For example, in the U.S., the set of commercial insurers, referred to as managed care organizations (MCOs), bargain with hospitals over inclusion in their networks and negotiate the price each hospital will be paid by each MCO for hospital care (Gowrisankaran et al., 2015; Ho and Lee, 2019). Third, we study the short-run equilibrium effect of health insurance on healthcare expense because of data availability. In the long run, market-wide changes in demand may induce new hospitals to incur costs associated with market entry (Finkelstein, 2007), which may alter the structure of healthcare market. We relegate this interesting extension to future research.

References

Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *American Economic Review*, 53(5):941–973.

Barwick, P. J., Cao, S., and Li, S. (2020). Local Protectionism, Market Structure, and Social Welfare: China's Automobile Market. *American Economic Journal: Economic Policy*, Forthcoming.

Berry, S. and Jia, P. (2010). Tracing the Woes: An Empirical Analysis of the Airline Industry. *American Economic Journal: Microeconomics*, 2(3):1–43.

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

Brot-Goldberg, Z. C., Chandra, A., Handel, B. R., and Kolstad, J. T. (2017). What does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics. *Quarterly Journal of Economics*, 132(3):1261–1318.

Bundorf, M. K., Royalty, A., and Baker, L. C. (2009). Health Care Cost Growth among the Privately Insured. *Health Affairs*, 28(5):1294–1304.

Chandra, A., Gruber, J., and McKnight, R. (2014). The Impact of Patient Cost-Sharing on Low-Income Populations: Evidence from Massachusetts. *Journal of Health Economics*, 33:57–66.

Chengdu Bureau of Statistics (2011). *Chengdu Statistical Yearbook 2011 (Chinese Edition)*. China Statistics Press.

Crawford, G. S., Shcherbakov, O., and Shum, M. (2019). Quality Overprovision in Cable Television Markets. *American Economic Review*, 109(3):956–95.

Crivelli, L., Filippini, M., and Mosca, I. (2006). Federalism and Regional Health Care Expenditures: An Empirical Analysis for the Swiss Cantons. *Health Economics*, 15(5):535–541.

Eizenberg, A. (2014). Upstream Innovation and Product Variety in the U.S. Home PC Market. *Review of Economic Studies*, 81(3):1003–1045.

Fan, Y. (2013). Ownership Consolidation and Product Characteristics: A Study of the US Daily Newspaper Market. *American Economic Review*, 103(5):1598–1628.

Feldstein, M. (1977). Quality Change and the Demand for Hospital Care. *Econometrica*, 45(7):1681–1702.

Feldstein, M. S. (1971). Hospital Cost Inflation: A Study of Nonprofit Price Dynamics. *American Economic Review*, 61(5):853–872.

Finkelstein, A. (2007). The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare. *Quarterly Journal of Economics*, 122(1):1–37.

Finkelstein, A., Gentzkow, M., and Williams, H. (2016). Sources of Geographic Variation in Health Care: Evidence from Patient Migration. *Quarterly Journal of Economics*, 131(4):1681–1726.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Group, O. H. S. (2012). The Oregon Health Insurance Experiment: Evidence from the First Year. *Quarterly Journal of Economics*, 127(3):1057–1106.

Gowrisankaran, G., Nevo, A., and Town, R. (2015). Mergers When Prices Are Negotiated: Evidence from the Hospital Industry. *American Economic Review*, 105(1):172–203.

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.

Hall, R. E. and Jones, C. I. (2007). The Value of Life and the Rise in Health Spending. *Quarterly Journal of Economics*, 122(1):39–72.

Ho, K. and Lee, R. S. (2019). Equilibrium Provider Networks: Bargaining and Exclusion in Health Care Markets. *American Economic Review*, 109(2):473–522.

Lee, J. and Seo, K. (2015). A Computationally Fast Estimator for Random Coefficients Logit Demand Models Using Aggregate Data. *RAND Journal of Economics*, 46(1):86–102.

Lee, S.-Y. (2011). *Handbook of Latent Variable and Related Models*. Elsevier.

Lopreite, M. and Mauro, M. (2017). The Effects of Population Ageing on Health Care Expenditure: A Bayesian VAR Analysis Using Data from Italy. *Health Policy*, 121(6):663–674.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., and Leibowitz, A. (1987). Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. *American Economic Review*, 77(3):251–277.

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

Ministry of Health (1989). *Public Hospital Classification Standard*. Beijing, China.

National Bureau of Statistics (2011a). *China Health Statistical Yearbook 2011 (Chinese Edition)*. China Statistics Press.

National Bureau of Statistics (2011b). *China Statistical Yearbook 2011 (Chinese Edition)*. China Statistics Press.

National Bureau of Statistics (2020). *China Statistical Yearbook 2020 (Chinese Edition)*. China Statistics Press.

Newhouse, J. P. (1977). Medical-Care Expenditure: A Cross-National Survey. *Journal of Human Resources*, 12(1):115–125.

Newhouse, J. P. (1992). Medical Care Costs: How Much Welfare Loss? *Journal of Economic Perspectives*, 6(3):3–21.

Revels, J., Lubin, M., and Papamarkou, T. (2016). Forward-Mode Automatic Differentiation in Julia. *arXiv:1607.07892 [cs.MS]*.

Reynaerts, J., Varadha, R., and Nash, J. C. (2012). Enhencing the Convergence Properties of the BLP (1995) Contraction Mapping. *Unpublished Manuscript*.

Sichuan Provincial Bureau of Statistics (2011). *Sichuan Statistical Yearbook 2011 (Chinese Edition)*. China Statistics Press.

Small, K. A. and Rosen, H. S. (1981). Applied Welfare Economics with Discrete Choice Models. *Econometrica*, 49(1):105–130.

Smith, S., Newhouse, J. P., and Freeland, M. S. (2009). Income, Insurance, and Technology: Why Does Health Spending Outpace Economic Growth? *Health Affairs*, 28(5):1276–1284.

The Center for Medicare and Medicaid Services (2019). National Health Expenditures. `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical`.

Varadhan, R. and Roland, C. (2008). Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353.

Xiang, J. (2020). Physicians as Persuaders: Evidence from Hospitals in China. *Unpublished Manuscript*.

Online Appendix for "Estimating an Equilibrium Model of Health Insurance and Healthcare Expense"

Junjian Yi* Shaoyang Zhao† Hang Zou‡

March 16, 2021

Contents

*Department of Economics, National University of Singapore; e-mail: junjian.yi@gmail.com.

†School of Economics, Sichuan University; e-mail: zhaoshaoyang@scu.edu.cn.

‡Department of Economics, National University of Singapore; e-mail: hang.zou@u.nus.edu.

A The Chinese Three-Tier Hospital System

China has been developing its healthcare system since 1949. One of the major accomplishments is the three-tier public hospital system. The classification of hospitals is based on weighted scores that measure the number of beds, medical personnel, medical equipment, level of service provision, medical technology, and quality of management and medical care (Ministry of Health, 1989). Different tiers are designed to provide different levels of healthcare services. Tier-1 hospitals generally have fewer than 100 beds and are tasked with providing primary care and preventive care. Tier-2 hospitals are equipped with 100 to 500 beds and are responsible for more comprehensive healthcare services and medical training for health workers in tier-1 facilities. With a bed capacity of over 500, tier-3 hospitals provide the most sophisticated acute care and specialist services. They also play a dominant role in medical education and research, and serve as medical hubs for multiple regions (Song et al., 2020).

Before 1979, the government owned, funded, and operated all healthcare facilities (Blumenthal and Hsiao, 2005, 2015). With adequate government funding, the three-tier system achieved great success in improving the population's health and life expectancy. From 1952 to 1982, infant mortality decreased from 200 to 34 per 1,000 live births, life expectancy rose from 35 to 68 years, and long-lasting scourges such as schistosomiasis were largely eliminated (Blumenthal and Hsiao, 2005, 2015).

Besides strong financial support from the government, the referral system was another important reason for the success of the three-tier hospital system before 1979. During this period, patients entered the healthcare system through visiting tier-1 hospitals and then transferred to higher-tier hospitals if they required more comprehensive or intensive care (Song et al., 2020).

However, after the economic reforms initiated in December 1978, China reduced the role of government in all economic and social sectors, thus guiding the healthcare system onto a different track (Blumenthal and Hsiao, 2015; Yip and Hsiao, 2008). The three-tier system

remained intact. However, the government funding of hospitals fell dramatically. Between 1978 and 1999, the central government's share of national healthcare spending decreased from 32% to 15% (Blumenthal and Hsiao, 2005).

Reduction in the government's fiscal support forced hospitals to rely more on the sale of healthcare services to cover their expenses. Hence, public hospitals came to behave like for-profit entities. Although the government imposed strict regulations on routine healthcare services, such as standard diagnostic tests, it permitted hospitals to earn profits from new drugs and advanced technology, with profit margins of 15% or more (Blumenthal and Hsiao, 2005; Hesketh and Zhu, 1997). In addition, the government altered the way it compensated physicians by including bonuses determined by the revenue physicians generate for their hospitals. The result was the overprescription of expensive pharmaceuticals and overuse of high-tech services, such as imaging, which led to a rapid increase in healthcare expense.

At the same time, the medical referral system that funnelled patients from lower-tier to higher-tier hospitals collapsed. Patients had the autonomy to visit any hospitals instead of entering the healthcare system from tier-1 facilities, and thus caused allocative inefficiency in the healthcare market.

The Chinese government recognized these unintended consequences. In 2009, it introduced a nationwide comprehensive health reform that aimed to provide more affordable and equitable access to healthcare services for all citizens by 2020. Specifically, the reform has five goals. First, expand public health insurance to cover more than 90% of the Chinese population, including improved basic medical insurance for urban employees and residents, the new rural cooperative Medicare scheme for rural residents, and the Medicaid system for the poor. Second, establish a national essential drug system to meet the basic need for treatment and ensure an affordable drug supply. Third, provide more public financing and infrastructure support to improve the medical care and public health service system at the grassroots level. Fourth, promote the basic public health service. Fifth, launch a pilot reform of public hospitals. More details can be found in Chen (2009).

The 2009 healthcare reform has made great progress in expanding insurance coverage (Song et al., 2020). By 2012, 95% of the population was covered by government-provided insurance schemes (Blumenthal and Hsiao, 2015). However, much work remains on healthcare service delivery. Although the government played a bigger role in the production and distribution of healthcare services, public hospitals still received limited funding from the government. Nearly 50% of hospital revenue still comes from the sale of healthcare services (Milcent, 2018). Although the government tried to contain the increase in healthcare expense by introducing payment methods other than fee-for-service, such as a diagnosis-related-group scheme and global budgeting, the healthcare expense per capita rose dramatically from ¥1,807 in 2011 to ¥4,703 in 2019 (National Bureau of Statistics, 2020). Facing the ever-growing healthcare expense, efficient healthcare delivery at an affordable cost has become the primary goal for the government to ensure the most effective development of China's healthcare system.

B The Chinese Public Health Insurance System

The Chinese public health insurance system consists of two insurance schemes: the urban employee basic medical insurance (UEBMI) and the urban and rural residents basic medical insurance (URRBMI). In particular, URRBMI was rolled out by integrating two insurance schemes: the new rural cooperative medical scheme (NRCMS) and urban residents basic medical insurance (URBMI). These schemes not only cover different groups of individuals, but also function independently and differ in financing and reimbursement. This section provides details on the UEBMI, NRCMS, URBMI, and URRBMI sequentially.

B.1 Urban Employee Basic Medical Insurance

The UEBMI evolved from two insurance schemes in the planned economy era: the government insurance scheme (GIS) and the labor insurance scheme (LIS). Established in 1952, the GIS was mainly designed for civil servants. Enrollees needed to visit designated hospitals to obtain free outpatient and inpatient services. Founded in 1951, the LIS covered employees in industrial enterprises. Financed by employers, the LIS not only guaranteed free healthcare services for its enrollees, but also reimbursed half of the healthcare spending of enrollees' family members.

Due to the absence of appropriate cost-reduction mechanisms in the planned economy era, overutilization of healthcare services became prevalent, resulting in a rapid increase in healthcare expense (Dong, 2009). Between 1980 and 1984, the annual increase in the healthcare expense was 10.7% (Liu and Wang, 1991). Struggling with the hefty financial burden induced by these two insurance schemes, the government decided to reform.

In 1994, the government carried out pilot trials of the UEBMI in Zhejiang, a city in Jiangsu province, and Jiujiang, a city in Jiangxi province. After a 4-year trial period, the UEBMI was launched nationwide in 1998. All urban employers and employees are required to enroll in the scheme, and share the responsibility for premium contributions. The total

premium equals 8% of the employee's monthly payroll, which is contributed to the scheme, with the employee contributing 2% and the employer providing the remaining 6%.

The total contribution is split into two accounts: an individual account and a social account. The employee's entire contribution plus 30% of the employer's contribution is allocated to the individual account, and the remaining employer's contribution is allocated to the social account. These two accounts function independently. In general, the individual account can be used to pay for outpatient expense, while the social account pays for the cost associated with inpatient visits. On average, 80% of medical costs for inpatient visits can be reimbursed by the UEBMI in 2014. As for outpatient services, the reimbursement rate varies across provinces, with an average between 50% and 80% (Sun et al., 2017).

B.2 New Rural Cooperative Medical Scheme

The NRCMS evolved from the rural cooperative medical scheme (RCMS) established in the late 1950s. The RCMS was a commune-based scheme.[1] It was mainly financed by commune welfare funds, which were part of the communes' collective revenue. Enrolled commune workers obtained free primary and preventive healthcare services from "barefoot doctors," rural health workers who served a particular commune. By the mid-1970s, the RCMS covered more than 90% of the rural population (Xu et al., 2009). The RCMS not only succeeded in funding healthcare for rural people, but also significantly improved their health (Dong, 2009). Between the early 1950s and early 1980s, the malaria rate decreased from 5.55% of the entire Chinese population to only 0.3%. The success of the RCMS was highly praised by the international community and recognized as a benchmark model for other developing countries (Dong, 2009; Xu et al., 2009).

However, after the economic reforms initiated in December 1978, the government completely dismantled communes and the RCMS lost its main source of funding. As a result, the RCMS nearly collapsed and only 5% of village maintained the scheme by the late 1980s (Xu

[1] A commune is a large rural collective work unit. Individuals in a commune work together and obtain their earnings based on their labor contribution (Dong, 2009).

et al., 2009). Without the RCMS, rural residents had no way to pool risks for healthcare expense, and were more vulnerable to health risks. Hence, the call for a new healthcare system for rural residents grew louder.

In 2003, large-scale pilot trials of the NRCMS were carried out in hundreds of counties in four provinces. The pilot period lasted for 3 years and the NRCMS was officially rolled out during the 11th Five-year Plan (2006–2010). The scheme is managed at the central level and operated at the county level. Enrollment in the NRCMS is on a household basis. As a government-led scheme, NRCMS is largely financed by subsidies from the central and local governments and, to a lesser extent, through premiums paid by households. In 2012, the total fund was ¥300 per capita, with 80% coming from government subsidies.

Although the scheme is voluntary in principle, the economic incentive to enroll is strong for rural residents due to high government subsidies. By 2014, the NRCMS had covered 98.9% of the rural population (Sun et al., 2017). The scheme mainly covers inpatient services. The reimbursement rate varies between 30% and 80%, and differs across different tiers of hospitals, with a lower rate for higher-tier hospitals (Xu et al., 2009). Different cost-sharing arrangements are partially designed to efficiently allocate patients with different medical needs to different healthcare facilities (Lai et al., 2018). Less severe patients may be more price elastic and end up choosing lower-tier hospitals.

B.3 Urban Residents Basic Medical Insurance

When urban employees and rural residents were covered by the UEBMI and NRCMS, respectively, urban residents without formal-sector jobs were completely left out of the healthcare safety net prior to 2007 (Lin et al., 2009). To achieve universal health insurance coverage, a large-scale pilot trial of the URBMI was launched in 79 cities following the guidelines in 2007 State Council Document No. 20. After a 2-year pilot period, the URBMI was eventually launched nationwide following the 2009 healthcare reform, closing the insurance coverage gap for urban residents without formal employment.

The URBMI is managed by the government and largely pooled at the prefecture level. Similar to the NRCMS, funding for the URBMI comes from two sources: government subsidies and individual contributions. In general, the government subsidy is relatively high and individuals pay less than 1% of their average disposable income for premiums (Si, 2020). Premiums also vary across areas. In 2010, the government subsidy per person was ¥180 on average, and individuals contributed around ¥20−¥170 in the central and western provinces and ¥40 − ¥250 in the eastern provinces (Yip et al., 2012).

Although the scheme is on a voluntary basis, the enrollment rate is high due to high government subsidies. By 2014, the URBMI covered more than 95% of all urban residents without formal employment (Sun et al., 2017). The scheme covers inpatient and critical outpatient services. Similar to the NRCMS, the reimbursement rate also varies across different levels of facilities, with a lower rate for higher-level facilities. On average, 48% of medical costs associated with inpatient and critical outpatient services were reimbursed by the URBMI in 2011 (Yu, 2015).

B.4 Urban and Rural Residents Basic Medical Insurance

The above three public health insurance schemes have covered over 95% of the total population, which is a great step for China in accomplishing universal health insurance coverage. However, the Chinese government still faces some challenges. Although these schemes increase individuals' utilization of healthcare services (Zhou et al., 2014; Sun et al., 2020; Liu and Zhao, 2014), they play a limited role in alleviating individuals' financial burdens, especially for inpatient visits (Huang and Gan, 2017; Sun et al., 2020; Liu and Zhao, 2014). In addition, regional inequities in healthcare resources and services still exist. Between 2005 and 2017, the number of medical personnel—including licensed doctors and registered nurses—and beds per 1,000 population in urban areas increased more than in rural areas (Tao et al., 2020). Hence, the next step toward universal health insurance coverage will be providing more affordable and equitable healthcare services.

The Chinese government first focused on the URBMI and NRCMS. To improve the fairness and cost-effectiveness of the URBMI and NRCMS, pilot trials to integrate the URBMI and NRCMS as the URRBMI were carried out in 2008. In 2016, the Chinese government issued the "Guiding Opinions on Integrating the Urban and Rural Residents Basic Medical Insurance System." The document required that the URBMI and NRCMS should be integrated based on the principle of six unifications.

First, unification of coverage. The URRBMI should cover rural residents and urban residents without formal employment. Second, the unification of financing policy. As with the URBMI and NRCMS, the URRBMI will still be financed by government subsidy and individual premium payment. Third, the unification of benefits package. The government will gradually integrate covered services and reimbursement schemes in the URBMI and NRCMS to ensure equitable healthcare delivery. Fourth, the unification of lists of drugs and services. The government will unify the essential medical drugs and services catalogue and clarify the reimbursement rule for each item in the catalogue. Fifth, the unification of the management of designated hospitals. Sixth, the unification of funds management. After release of the document, provincial governments started to introduce the URRBMI comprehensively and the nationwide launch was implemented in 2019.

C The SQUAREM Algorithm

The SQUAREM algorithm was developed by Varadhan and Roland (2008). The intuition is to form an estimate of the Jacobian through two residuals—$\boldsymbol{r}^N$ and $\boldsymbol{v}^N$. The residual $\boldsymbol{r}^N$ is determined by the difference between the current iteration $\boldsymbol{\delta}^N$ and next iteration $g\left(\boldsymbol{\delta}^N\right)$. The residual $\boldsymbol{v}^N$ is determined by the change in the residual $\boldsymbol{r}^N$ from this iteration to the next one. The residual and the curvature can also be used to construct a step-size ζ^N (Conlon and Gortmaker, 2020). The exact algorithm is described below:

$$\boldsymbol{\delta}^{N+1} = \boldsymbol{\delta}^N - 2\zeta^N \boldsymbol{r}^N + \left(\zeta^N\right)^2 \boldsymbol{v}^N,$$

where

$$\zeta^N = -\frac{||\boldsymbol{r}^N||}{||\boldsymbol{v}^N||},$$

$$\boldsymbol{r}^N = g\left(\boldsymbol{\delta}^N\right) - \boldsymbol{\delta}^N,$$

$$\boldsymbol{v}^N = g\left(g\left(\boldsymbol{\delta}^N\right)\right) - 2g\left(\boldsymbol{\delta}^N\right) + \boldsymbol{\delta}^N,$$

and $g\left(\cdot\right)$ indicates BLP contraction mapping. In practice, we start with BLP contraction mapping. After $||g\left(\boldsymbol{\delta}\right) - \boldsymbol{\delta}|| \leq n$, we switch to the SQUAREM algorithm. n is a user-defined threshold, which can be problem dependent. In our analysis, we use $n = 5$.

D The System of Implicit Functions

We consider a system of equations

$$\begin{cases} F_1(x_1,\ldots,x_n,y_1,\ldots,y_m)=c_1 \\ \ldots\ldots\ldots\ldots.. \\ F_m(x_1,\ldots,x_n,y_1,\ldots,y_m)=c_m \end{cases}$$

and suppose a point $(x_1^*,\ldots,x_n^*,y_1^*,\ldots,y_m^*)\in R^{n+m}$ is a solution.

Theorem 1 *If the determinant of Jacobian matrix*

$$\begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \cdots & \cdots & \cdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix}$$

evaluated at $(x_1^*,\ldots,x_n^*,y_1^*,\ldots,y_m^*)$ *is nonzero, then there exist functions*

$$y_1(x_1,\ldots,x_n),\ldots,y_m(x_1,\ldots,x_n)$$

defined on a ball about $(x_1^*,\ldots,x_n^*)$ *satisfying the conditions*

$$F_i(x_1,\ldots,x_n,y_1(x_1,\ldots,x_n),\ldots,y_m(x_1,\ldots,x_n))\equiv c_i, i=1,\ldots,m,$$

and $y_i(x_1^*,\ldots,x_n^*)=y_i^*, i=1,\ldots,m.$

Furthermore, the derivatives $\frac{\partial y_i}{\partial x_k}$ *can be solved from the system of linear equations*

$$\begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \cdots & \cdots & \cdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial y_1}{\partial x_k} \\ \cdots \\ \frac{\partial y_m}{\partial x_k} \end{pmatrix} = - \begin{pmatrix} \frac{\partial F_1}{\partial x_k} \\ \cdots \\ \frac{\partial F_m}{\partial x_k} \end{pmatrix} \tag{1}$$

Back to our case, given a market m, hospital h has the following two first-order conditions:

$$S_h + (e_h - mc_h) \cdot \frac{\partial S_h}{\partial e_h} = 0,$$

$$(e_h - mc_h) \cdot \frac{\partial S_h}{\partial R_h} - \frac{\partial c_h}{\partial R_h} = 0.^2$$

We denote

$$F_u = S_h + (e_h - mc_h) \cdot \frac{\partial S_h}{\partial e_h},$$

$$F_v = (e_h - mc_h) \cdot \frac{\partial S_h}{\partial R_h} - \frac{\partial c_h}{\partial R_h}.$$

The whole Jacobian matrix has $2 \cdot ||H_m||$ rows and $2 \cdot ||H_m||$ columns, where $||H_m||$ is the number of hospitals in market m. In order to obtain the u-th row of the Jacobian matrix, we need to compute

$$\frac{\partial F_u}{\partial e_h} = 2 \cdot \frac{\partial S_h}{\partial e_h} + (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial e_h^2},$$

$$\frac{\partial F_u}{\partial e_j} = \frac{\partial S_h}{\partial e_j} + (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial e_h \partial e_j} \quad \forall j \neq h,$$

$$\frac{\partial F_u}{\partial R_h} = \frac{\partial S_h}{\partial R_h} + (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial e_h \partial R_h},$$

$$\frac{\partial F_u}{\partial R_j} = \frac{\partial S_h}{\partial R_j} + (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial e_h \partial R_j} \quad \forall j \neq h.$$

Similarly, in order to obtain the v-th row of the Jacobian matrix, we need to compute

$$\frac{\partial F_v}{\partial e_h} = \frac{\partial S_h}{\partial R_h} + (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial R_h \partial e_h},$$

$$\frac{\partial F_v}{\partial e_j} = (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial R_h \partial e_j} \quad \forall j \neq h,$$

$$\frac{\partial F_v}{\partial R_h} = (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial R_h^2},$$

[2] In practice, we work with the claimable ratio for convenience due to the equivalence between the optimization of claimable expense and that of the claimable ratio.

$$\frac{\partial F_v}{\partial R_j} = (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial R_h \partial R_j} \quad \forall j \neq h.$$

Besides those values, we still need to obtain the right-hand-side vector in equation (1). However, in our case, it becomes a $||H_m|| \times ||H_m||$ matrix. Specifically, for the u-th row of the matrix, we need to calculate

$$\frac{\partial F_u}{\partial q_h} = \frac{\partial S_h}{\partial q_h} + \left[-\frac{\partial mc_h}{\partial q_h} \cdot \frac{S_h}{e_h} + (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial e_h \partial q_h} \right],$$

$$\frac{\partial F_u}{\partial q_j} = \frac{\partial S_h}{\partial q_j} + (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial e_h \partial q_j} \quad \forall j \neq h.$$

Similarly, for the v-th row of the matrix, we need to calculate

$$\frac{\partial F_v}{\partial q_h} = -\frac{\partial mc_h}{\partial q_h} \cdot \frac{\partial S_h}{\partial R_h} + (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial R_h \partial q_h},$$

$$\frac{\partial F_v}{\partial q_j} = (e_h - mc_h) \cdot \frac{\partial^2 S_h}{\partial R_h \partial q_j} \quad \forall j \neq h.$$

After all of the calculations, we can obtain $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{q}}$ and $\frac{\partial \boldsymbol{R}}{\partial \boldsymbol{q}}$ by either left multiplying the inversion of the Jacobian matrix on both sides of equation (1) or Cramer's rule.

E Details on Decomposition

To perform the decomposition in Eq. (26), we need to simulate equilibrium market outcomes for three scenarios: (i) with the policy; (ii) without the policy; (iii) with the policy but holding the quality at the level without the policy. However, simulating SPNE with a large number of players is computationally challenging. To render the analysis feasible, we make the following simplification assumptions:

- We construct two "pseudo" markets: urban and rural.
- Players are tier groups rather than hospitals. Based on the data, there are four players in the urban market—tier-3, tier-2, tier-1, and tier-0—and three players in the rural market—tier-2, tier-1, and tier-0. We also include an outside option in both markets.
- All unobservables are obtained by taking the mean of the empirical values of those unobserved shocks across all hospitals in those tier groups. The values in 2009 are used to simulate market outcomes without the policy, and those in 2010 are used to simulate market outcomes with the policy.
- In the first stage, players' expectations on unobservables are obtained by taking the median of the empirical values of those unobserved shocks across all hospitals in those tier groups.
- We randomly draw 20,000 individuals from our data in each area.

As can be seen from these assumptions, we use aggregate information on the markets in urban and rural areas. Since this aggregation captures the key features of markets in urban and rural areas, the decomposition results are still informative for understanding the relative importance of supply- and demand-side responses.

References

Blumenthal, D. and Hsiao, W. (2005). Privatization and Its Discontents: The Evolving Chinese Health Care System. *New England Journal of Medicine*, 353(11):1165–1170.

Blumenthal, D. and Hsiao, W. (2015). Lessons from the East—China's Rapidly Evolving Health Care System. *New England Journal of Medicine*, 372(14):1281–1285.

Chen, Z. (2009). Launch of the Health-Care Reform Plan in China. *Lancet*, 373(9672):1322–1324.

Conlon, C. and Gortmaker, J. (2020). Best Practices for Differentiated Products Demand Estimation with PyBLP. *RAND Journal of Economics*, 54(4):1108–1161.

Dong, K. (2009). Medical Insurance System Evolution in China. *China Economic Review*, 20(4):591–597.

Hesketh, T. and Zhu, W. X. (1997). Health in China: The Healthcare Market. *BMJ*, 314(7094):1616.

Huang, F. and Gan, L. (2017). The Impacts of China's Urban Employee Basic Medical Insurance on Healthcare Expenditures and Health Outcomes. *Health Economics*, 26(2):149–163.

Lai, S., Shen, C., Xu, Y., Yang, X., Si, Y., Gao, J., Zhou, Z., and Chen, G. (2018). The Distribution of Benefits under China's New Rural Cooperative Medical System: Evidence from Western Rural China. *International Journal for Equity in Health*, 17(1):1–14.

Lin, W., Liu, G. G., and Chen, G. (2009). The Urban Resident Basic Medical Insurance: A Landmark Reform towards Universal Coverage in China. *Health Economics*, 18(S2):S83–S96.

Liu, H. and Zhao, Z. (2014). Does Health Insurance Matter? evidence from China's Urban Resident Basic Medical Insurance. *Journal of Comparative Economics*, 42(4):1007–1020.

Liu, X. and Wang, J. (1991). An Introduction to China's Health Care System. *Journal of Public Health Policy*, 12(1):104–116.

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

Ministry of Health (1989). *Public Hospital Classification Standard*. Beijing, China.

National Bureau of Statistics (2020). *China Statistical Yearbook 2020 (Chinese Edition)*. China Statistics Press.

Si, W. (2020). Public Health Insurance and the Labor Market: Evidence from China's Urban Resident Basic Medical Insurance. *Health Economics*, 30(2):403–431.

Song, C., Yang, N., Yi, J., and Yuan, Y. (2020). Information Provision, Patient Sorting, and Healthcare Quality. *Unpublished Manuscript*.

Sun, J., Lyu, X., and Yang, F. (2020). The Effect of New Rural Cooperative Medical Scheme on the Socioeconomic Inequality in Inpatient Service Utilization among the Elderly in China. *Risk Management and Healthcare Policy*, 13:1383.

Sun, Y., Gregersen, H., and Yuan, W. (2017). Chinese Health Care System and Clinical Epidemiology. *Clinical Epidemiology*, 9:167.

Tao, W., Zeng, Z., Dang, H., Li, P., Chuong, L., Yue, D., Wen, J., Zhao, R., Li, W., and Kominski, G. (2020). Towards Universal Health Coverage: Achievements and Challenges of 10 Years of Healthcare Reform in China. *BMJ Global Health*, 5(3):e002087.

Varadhan, R. and Roland, C. (2008). Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353.

Xu, K., Saksena, P., Fu, X. Z. H., Lei, H., Chen, N., and Carrin, G. (2009). Health Care Financing in Rural China: New Rural Cooperative Medical Scheme. *World Health Organization Publ., Geneva*.

Yip, W. and Hsiao, W. C. (2008). The Chinese Health System at a Crossroads. *Health Affairs*, 27(2):460–468.

Yip, W. C.-M., Hsiao, W. C., Chen, W., Hu, S., Ma, J., and Maynard, A. (2012). Early Appraisal of China's Huge and Complex Health-Care Reforms. *Lancet*, 379(9818):833–842.

Yu, H. (2015). Universal Health Insurance Coverage for 1.3 Billion People: What Accounts for China's Success? *Health Policy*, 119(9):1145–1152.

Zhou, Z., Zhou, Z., Gao, J., Yang, X., Xue, Q., Chen, G., et al. (2014). The Effect of Urban Basic Medical Insurance on Health Service Utilisation in Shaanxi Province, China: A Comparison of Two Schemes. *PLoS One*, 9(4):e94909.