Understanding Deductible and Reimbursement Maximum: A Study of Rural China's Tiered Medical System

Julie Shi[2], Xi Wang[2], and Castiel Chen Zhuang[*1]

[1]Department of Economics, University of Washington

[2]Department of Economics, School of Economics, Peking University

October 20, 2021

Abstract

A high-deductible coverage is shown to reduce inappropriate health care, while little is known about its effect on social welfare under a reimbursement limit. This paper utilizes a large claims level dataset from rural China and estimates preferences of inpatients under tier-dependent nonlinear cost-sharing schemes. Patients with high health risks prefer generous hospitals for financial protection, but it is countervailed by a potential mistrust of quality—this partly explains why patients with common diseases or minor illnesses may bypass primary care; moral hazard increases with health risks and willingness to pay in general but is modest. Increasing all hospitals' deductibles by 1,000 yuan improves social welfare by 2 percent and encourages more patients with lower health risks to visit

[*]Corresponding author. Email: cczhuang@uw.edu. We are grateful to Shi Chen, Rachel Heath, Vanessa Oddo, Yuya Takahashi, Xu Tan, Jing Tao, and seminar participants in the Department of Economics at University of Washington for their helpful comments. The authors are responsible for all errors.

lower-tiered hospitals and save the medical resources in higher-tiered hospitals for advanced treatments; the current reimbursement maximum is close to the optimal, and the additional insurer cost and efficiency loss of increasing the limit to promote policy acceptance do not outweigh the positive effects of higher deductibles.

Keywords: hierarchical medical system, hospital choice, moral hazard.

JEL Classification: D12, D81, D82, G22, I13.

1 Introduction

Health expenditure has grown rapidly all over the world, increasing from 4.6% of gross domestic product (GDP) in 1970 to 10.0% of GDP in 2018 (Stadhouders et al., 2019; World Health Organization, 2020). Meanwhile, medical waste accounts for a great portion of medical costs. In the United States, for example, around thirty percent of health-care spending may be considered waste (Shrank et al., 2019). Researchers and policymakers have long focused on how to contain the escalating medical costs and reduce waste. One common approach is to rely on consumer incentives to control for moral hazard by applying high cost-sharing to treatments and services that are not cost-effective. For instance, Medicare Part D sets different reimbursement tiers for drugs, with generics occupying the lowest tier and having the least out-of-pocket (OOP) payment (Duggan et al., 2008); employees are increasingly encouraged to participate in high-deductible health plans (HDHPs) that provide consumers with incentives to control cost (Agarwal et al., 2017; Mazurenko et al., 2019).

It is well documented by randomized controlled experiments and quasi-experiments that consumers respond to demand-side incentives, while the literature focuses mainly on cost containment (Newhouse and the Insurance Experiment Group, 1993; Finkelstein et al., 2012) and how spending reduction is achieved (Brot-Goldberg et al., 2017). A more important aspect, which has been less discussed, is how the efficiency of a medical system is influenced. Particularly, what remains as a question is whether the expenditure reduction incentivized by high cost-sharing is achieved by consuming more high-value care, which would improve welfare, or through reduction of necessary services, which may in contrast deteriorate population health. However, as medical systems are complex, and it is difficult to clearly distinguish between high-value care and low-value care, the evidence on this aspect is limited. As the costs of medical waste due to overtreatment or low-value care are estimated to add up to 75.7–101.2 billion US$ in the United States alone,[1] this issue is of great policy relevance. Recent articles therefore have started to conduct welfare analyses of healthcare programs by imposing assumptions about the structures of those programs, such as Finkelstein

[1]Estimates are based on Colla et al. (2015), Carter et al. (2017), French et al. (2017), Langer-Gould et al. (2013), Mannocci et al. (2016), Mulcahy et al. (2018), National Academies of Sciences, Engineering, and Medicine (2018), Reid et al. (2016), and Schwartz et al. (2014).

et al. (2019).

In this article, we investigate how cost-sharing structures affect patients' choices on hospitals that provide medical services at different efficiency levels and subsequent spending, taking advantage of the hierarchical delivery system in China. The Chinese hospitals are graded into multiple tiers. Those in higher tiers are typically responsible for treating more complicated illnesses, and employing higher labor and capital costs. Thus, high-tiered hospitals are less cost-effective than low-tiered hospitals in treating the same common disease. Patients can freely choose hospitals without referral. For cost containment, a health insurance applies higher cost-sharing for services received in high-tiered hospitals. The coinsurance rates at different tiers of hospitals also vary by year complying to the annual budget. Taking the unique setting of rural China's medical delivery system, we construct a structural framework to analyze how insurance policies (need not be observed in data) on patient cost-sharing affect rural residents' decision on hospital choice and their consecutive medical spending, and simulate alternative cost-sharing structures and measure welfare impacts.

To identify how patients respond to incentives, we adopt structural modeling by leveraging the exogenous variation in hospital options and isolated variation along the dimension of coinsurance level. Our entails two decision stages. First, patients make a discrete choice over hospitals under uncertainty about health risk; then, they make a continuous spending choice upon realizing their health states. As a usual rational decision, when choosing a hospital, a patient forms an expectation under uncertainty and takes into consideration his/her own opinion about financial risk (e.g., he/she would decide if a higher coinsurance rate is riskier than a potentially lower service quality associated with a lower price), the expected service utility from the second stage, the hospital level fixed effects, and the taste shock. In the second stage, the patient incorporates health status, moral hazard type, and the cost structure to choose expenditure optimally. Note that, this model allows for heterogeneity in risk preferences, moral hazard types, and health states, so that we can obtain the richest possible understanding of how patients select hospitals and make utilization decisions in a consecutive manner. When building the model, we consult the well-established insurance choice literature including Cardon and Hendel (2001), Carlin and Town (2009), Bundorf et al. (2012), Einav et al. (2013), Handel (2013), and

Azevedo and Gottlieb (2017). Their settings allow for variation in coinsurance level among insurance plans, and our setting allows for variation in cost-sharing across hospitals under the same plan.

Our estimation results reveal that there is substantial heterogeneity in willingness to pay for a more generous hospital. While this willingness to pay is mainly driven by a high value of financial risk protection among patients with large spending, some patients with small spending may subjectively associate more generous hospitals with lower quality and higher long-term risks (mistrust of service quality) in more generous but lower-tiered hospitals and become unwilling to pay.[2] These patients, regardless of their health states, may choose to bypass lower-tired hospitals which are usually more generous in cost-sharing. On the other hand, moral hazard is considered modest in the sense that it could explain at most 6 percent of the total spending. As a result, the expected reduction in OOP spending contributes to willingness to pay more than the expected increase in utility from overconsumption does. In addition to the above-mentioned compositions of willingness to pay, we find that patients are willing to pay for a higher-tiered hospital even when its other observable characteristics match those of a lower-tiered hospital. This could be because higher-tiered hospitals in our context typically have higher social reputation, and thus higher perceived quality.

There could be efficiency loss due to the current policy. Owing to low deductibles and the potential mistrust of quality in more generous hospitals, patients with low willingness to pay, who also are more likely to be at low spending risk, tend to choose less generous hospitals. These hospitals tend to be higher-tiered hospitals, which are supposed to deal with more complicated diseases. Mistrust not only distorts the allocation of resources, but can also reduce willingness to pay (and thus consumer and social welfare) directly. Therefore, we believe that delaying patients' exposure to reimbursement by having higher deductibles can alleviate the negative impact of mistrust[3] and promote a more efficient allocation of medical resources at the lower end

[2] As suggested by Avdic et al. (2019), subjective quality of hospital can affect choices of patients. Although we do not have a direct measure of subjective quality, our risk attitude parameter may indirectly reveal the association between satisfaction and generosity.

[3] There remains a question for us to empirically investigate if lower prices are indeed associated with lower quality in different tiers of hospitals. Anecdotal evidence suggests that some patients with common diseases or minor illnesses do not require treatment in a hospital but still get hospitalized

of the willingness-to-pay distribution.[4] This is in line with the purpose of HDHPs—encouraging patients to make higher-value choices. Furthermore, at the higher end of the willingness-to-pay distribution, it is possible that the reimbursement cap is causing the distortion of resource allocation. Raising the cap, for example, might increase the willingness to pay and thus promote the welfare of very risk-averse patients who value financial risk protection more, while pushing some to higher-tiered hospitals that become relatively more attractive than before, leading to a lower increase in the overall willingness to pay; at the same time, part of the patients who are mistrustful of quality associated with generosity may experience a decrease in willingness to pay, and these patients might be pushed to higher-tiered hospitals that become relatively less unattractive than before, leading to a smaller decrease in the average willingness to pay and consumer welfare. Since the current cap is not binding (i.e., none of our patients exhausted the reimbursement limit), we do not expect insurer/government costs to change much in response to a slight change of the cap.

Our model focuses on the financial dimension of the policy design associated with the multi-tiered hospital system and permits a rich space of potential (counterfactual) contracts. We utilize our structural model estimates to investigate three types of alternative policies. First, we experiment with a few high-deductible policies, and find that increasing the deductibles within a certain degree can lead to increased social welfare. However, increasing the deductible gaps between hospitals of different tiers could lead to unexpected efficiency loss. As a result, the scenario in which social welfare is increased the most is to increase deductibles moderately for all hospitals without increasing the gaps between them. Second, we experiment with policies with alternative caps, and it turns out that the current reimbursement cap is close to the optimal level—largely raising or lowering the reimbursement caps all lead to efficiency loss. Raising caps can potentially improve efficiency but the effect is quite modest. Nevertheless, we find that raising or even removing the caps has limited impact on

by their physicians as the number of inpatients is one of their key performance indicators. For these patients who do not look for reimbursement initially, discounted prices may indeed lead to mistrust of quality.

[4]Here we assume that the (subjective) service quality can vary by reimbursement generosity within the same hospital. Thus, for those who associate higher quality with lower generosity, the uncompensated portion (e.g., before reaching a deductible) of a service has the highest quality.

welfare and insurer/government cost, making it a candidate policy tool to compensate for high deductibles to improve policy acceptance. Third, we test the combination of the two policies. It turns out that, accompanying higher deductibles with a slight increase in reimbursement caps can further improve social welfare, and the welfare gain is higher than the sum of those obtained separately, indicating synergy.

This paper complements and contributes to a few literature branches. First, it builds on the prior studies about determinants of hospital choice including Burns and Wholey (1992), Roh et al. (2008), Brown and Theoharides (2009), Escarce and Kapur (2009), Ho and Pakes (2011, 2014a), Sanders et al. (2015), Baker et al. (2016), Mak (2018), Avdic et al. (2019), and Zhu et al. (2019). They emphasize hospital features (such as distance and quality) and patient characteristics (such as health) as the determinants. Our work connects to these earlier studies in a few aspects: (i) we include these factors in a stylized manner, which improves the flexibility of our model and allows us to focus more on how choice is affected by cost-sharing structures; (ii) similar to Gaynor and Vogt (2003) and Ho and Pakes (2011, 2014a), we estimate a discrete choice model of hospital demand; (iii) in addition, we introduce moral hazard and subjective risk attitude as the factors affecting hospital choice, by taking the consecutive utilization choice after hospital choice into account. We emphasize that cost-sharing structure is another important feature affecting hospital choices, while incorporating other determinants (such as distance, moral hazard, and subjective risk attitude) in our flexible model.

Second, it is related to the literature on the value-based insurance design (VBID) that aims to encourage the use of higher-value services by aligning cost with value (Perez et al., 2019). As suggested by Agarwal et al. (2018) and Ma et al. (2019), the VBID can promote primary care by modifying cost-sharing without an increase in total health spending. There has been a rising branch of literature on the design of an optimal health insurance menu (Einav et al., 2010; Bundorf et al., 2012; Geruso, 2017; Ho and Lee, 2019). However, very few discussions are given to the design of an appropriate hospital menu within an insurance contract. We add to this literature by considering a design based on different tiers of hospitals. If hospitals of lower tiers provide services with higher social values, patients should incentivized to choose lower-tiered hospitals.

Third, this work connects the literature on (ex post) moral hazard in healthcare. The moral hazard issues induced by cost-sharing are well documented theoretically and empirically, by Pauly (1968), Manning and Marquis (1996), Cutler and Zeckhauser (2000), Aron-Dine et al. (2015), Keane and Stavrunova (2016), Hudson et al. (2017), Brot-Goldberg et al. (2017), and many others. These papers not only attempt to verify the tendency of overconsumption when patients do not pay the full costs, but also try to further understand the nature of consumer response (e.g., how consumer reacts to non-linear contracts). However, while offering compelling evidence, most of these reduced form studies provide little guidance for forecasting health-care spending under situations not directly observed in the data. As suggested by Einav and Finkelstein (2018), to complement the limitations of prospective policy analysis in guiding the optimal design of healthcare system to address moral hazard, we need economic models that rely on deeper economic primitives. Following Einav et al. (2013), Bajari et al. (2014), Kowalski (2015), Einav et al. (2015, 2017), and Lu et al. (2019), we therefore rely on a more sophisticated economic model of individual behavior and investigate an optimal policy design.

Fourth, our project adds new insights to the literature on cost containment in general. The multi-tiered medical system has been adopted widely around the world, and a common challenge is that people tend to bypass primary care (Liang et al., 2020a). In some countries (such as the United States), bypassing primary care is mainly limited by the community doctor "gatekeepers" system[5] instead of through imposing economic measures (Zhou et al., 2021). In other countries and economic regions, however, most initiatives have not led to imposition of gatekeeping regulations. This opens a window for regulations targeting the demand side. Myriad cost control policies have been discussed in the literature, e.g., global budget (Bazzoli et al., 2004), price controls (Nguyen, 1996; Iizuka, 2007; Duggan et al., 2008; Kaiser et al., 2014; Gothe et al., 2015; Fu et al., 2018), payment reforms (Ho and Pakes, 2014b; Huckfeldt et al., 2014; Lemak et al., 2015), and consumer cost-sharing (Joyce et al., 2002; Bundorf, 2016; Brot-Goldberg et al., 2017). A systemic literature review evaluating these healthcare cost containment policies is carried out by Stadhouders et al. (2019).[6] Our

[5]In the United States, the health maintenance organization (HMO) plans require patients' choices through a primary care physician's referral.

[6]While higher cost-sharing is reported to be effective, some studies have reported that it is

paper considers both the healthcare system and patient cost-sharing, and suggests a potential tool for cost control. If healthcare facilities are sorted by the efficiency of services they provide, applying different cost-sharing structures to different facilities may reduce medical waste and improve welfare.

The rest of the paper proceeds as follows. In Section 2, we introduce the hierarchical medical system in China as well as the insurance plan that covers all the patients in our study. Section 3 first illustrates our theoretical framework, and then presents the empirical implementation of our model. Section 4 describes our data and the variation it provides. Section 5 shows the model estimates and calculates willingness to pay and social surplus. Section 6 evaluates welfare and distributional outcomes under alternative pricing policies. Finally, we note concluding remarks in Section 7.

2 Institutional Background

2.1 The Hierarchical Medical System in Rural China

Hospitals in China are categorized into three main tiers—tertiary hospitals, secondary hospitals, and primary care facilities (Wang et al., 2014). In rural areas, three names of healthcare institutions are commonly used: village clinics, township health centers (THCs), and county-level hospitals. Village clinics and THCs typically correspond to primary care facilities in China's general medical hierarchy (Liu et al., 2018), while county-level hospitals are mostly secondary, with a few being tertiary. We focus on two levels of hospitals available in our inpatient data—the county-level hospitals and THCs.

Within a county, a referral is not necessary for an inpatient visit, and patients can freely choose between THCs (typically with higher reimbursement rates) and county hospitals (to receive lower but non-zero reimbursement rates) for inpatient care, as shown in Figure 1.

associated with adverse outcomes, especially among vulnerable populations such as elderly and poor patients (Zeber et al., 2007; Hartung et al., 2008; Trivedi et al., 2010).

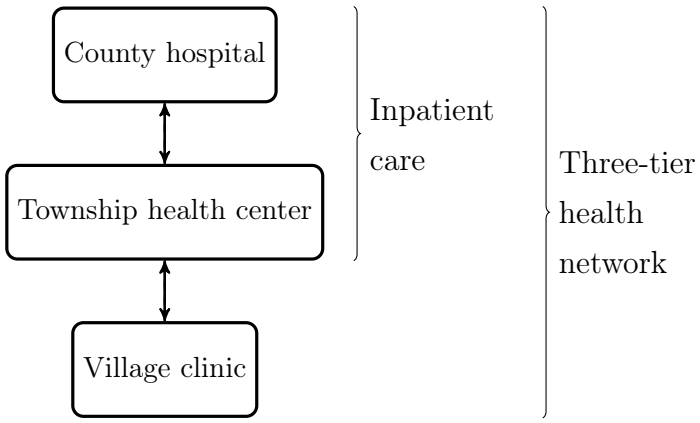


Figure 1: The Hierarchical Medical System in the Study Area

2.2 Research County and Healthcare System

We draw data from a county in the southwestern part of China. As discussed by Lu et al. (2019), this county is comparable to the median county of China, providing us with the external validity to generalize empirical results regarding inpatient responses to healthcare policies in rural China.

The area of the study county is comparable to New York City, while the population is comparable to Oakland in California, as of the study period. To better visualize the geographic distribution of the local health institutions and residential areas in our sample, we draw a map in Figure 2. As shown, health care is accessible across the study area, but some neighborhoods (e.g., those located in the southeast) have slightly better access to healthcare than others (e.g., those located in the northwest). Among the 29 healthcare institutions, 15 are THCs, while 14 are county hospitals (one of them is not available until 2013). More characteristics of the healthcare system are described in Table 1.

In our target population, patients are covered by a single health insurance program, the New Rural Cooperative Medical Scheme (NRCMS), aiming to achieve universal access to healthcare and reduce financial burden for all rural residents in China. The NRCMS was initiated in July 2003 in 310 pilot counties[7] and then was rapidly expanded to more than 85.0% of counties nationwide within 4 years (Bai and Wu,

[7]It is the third level (below provincial and prefecture levels) of administrative division in China.

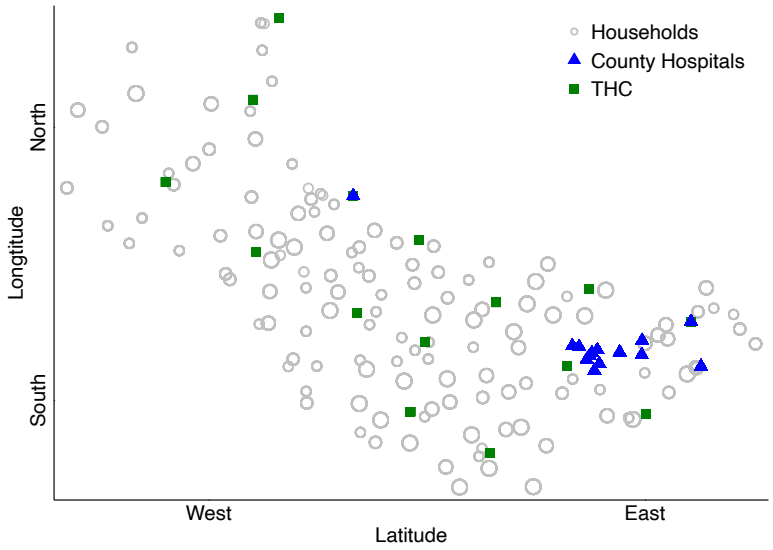


Figure 2: Geography of the Study County

Notes: Each circle represents a community. The size of a circle indicates the number of hospital visits from the corresponding neighborhood, with a larger size suggesting more visits.

2014). By the start of our study period (2012), the program covered about 805 million individuals, or 98.3% of rural residents, and the participation rate kept increasing in 2013 and 2014 and reached 98.9% (China Health and Family Planning Statistical Yearbook, 2015). More details about the program are covered by Chen et al. (2019).

Under the NRCMS, inpatients in the study county receive an average reimbursement

Table 1: Basic Characteristics of the Study County's Healthcare System

	Overall	2012	2013	2014
Proportion of county hospitals	0.477	0.464	0.483	0.483
	(0.502)	(0.508)	(0.509)	(0.509)
Number of physicians	27.558	26.800	27.615	28.231
	(46.487)	(44.331)	(48.444)	(48.322)
Number of beds	84.208	86.208	81.385	84.462
	(102.205)	(105.555)	(101.229)	(103.906)
N	86	28	29	29

Notes: This table presents the summary statistics for the study healthcare institutions. Standard deviations are in the parentheses under the means.

rate of 71.3% in THCs and an average rate of 55.6% in county hospitals, between 2012 and 2014. The average reimbursement rates received in these two tiers of healthcare facilities in each year from 2012 to 2014 are declining, as shown in Figure 3.

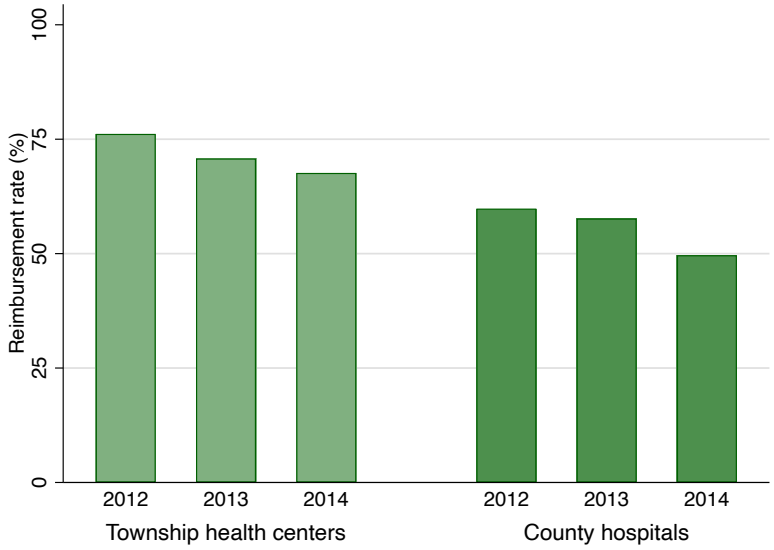


Figure 3: Average Reimbursement Rates in the Study County

3 Model

3.1 Theoretical Framework

In this section, we present a stylized framework of hospital choice and health care utilization. This theoretical model provides the main ingredients in our empirical specification and counterfactual analyses.

3.1.1 Demand and Patient Incentives

We consider a two-stage demand model resembling that of Einav et al. (2013). In the first stage, a forward-looking utility-maximizing patient chooses a health-care provider without knowing her exact health status.[8] The patient forms an expectation

[8]For patients with chronic conditions, we assume that they still do not know exactly how they progress until they pay the next visit, but they may have a smaller uncertainty (can be as close to zero as possible).

regarding her health realization based on all available information. In the second stage, the patient learns about her health state (after being admitted to a hospital) and decides how much to spend on health care.

Patients are characterized by type θ: $\{F, \omega, \psi\}$, where F represents a patient's belief about her subsequent health status λ; ω is a "moral hazard" parameter (measured by the additional spending that would be induced by moving from no insurance to full coverage); $\psi \in \mathbb{R}$ is the "risk attitude" parameter—it measures both how much patients dislike financial uncertainty in OOP costs and how well patients think a hospital can handle this uncertainty (assuming that they are not additively separable). Population is then defined by the distribution $G(\theta)$.

A patient chooses a hopsital from a set of health-care providers denoted by $J = \{1, 2, ..., j, ..., N_J\}$ in the first stage. Specifically, we denote $j = 0$ as the hospital that charges the full cost, which is excluded from the empirical choice set.[9] After choosing a hospital, patients realize their health status λ and decide the dollar amount $m \in \mathbb{R}_+$ of health care utilization. Health care utilization provides patients with benefits—we denote the money-metric valuation of benefits as $b(m, \lambda, \omega)$. It also costs patients money—we denote the OOP cost as $c(m, j)$. A utility-maximizing patient should trade off the benefits and OOP costs to find the optimal spending $m^*(\lambda, \omega, j) = \arg\max_m \{b(m, \lambda, \omega) - c(m, j)\}$. We define the indirect benefit by substituting m^*, i.e., $b_j^*(\lambda, \omega) = b(m^*(\lambda, \omega, j), \lambda, \omega)$; similarly, the indirect OOP cost is $c_j^*(\lambda, \omega, j) = c(m^*(\lambda, \omega, j), j)$, while the indirect payoff from utilization in hospital j is $x_j^*(\lambda, \omega, j) = b_j^*(\lambda, \omega) - c_j^*(\lambda, \omega, j)$.

The patient's utility function is given by $v(m, y) = v_\psi(y + l_j + b(m, \lambda, \omega))$, where v_ψ is strictly increasing and its shape depends on the value of ψ; $y = \hat{y} - c(m, j) - p$ is the "residual income" defined by subtracting the OOP cost of health care c, and other costs (such as transportation costs) p, from the initial income $\hat{y}$; l_j is the money-metric valuation of a hospital level.[10] Due to the uncertainty in health, the patient forms an

[9] It can also denote "not going to any hospital"; in such case, the patient bears the full consequence of not getting treated, and we assume this consequence can be exactly measured by the full cost of the treatment. In our sample, almost none of the patients go to an uncontracted hospital (outside of the county); we also only consider patients who choose to get treated whenever they have a disease.

[10] We can regard p as a "price" or opportunity cost patients have to pay for each hospital's admission before utilize health care.

expected utility, given by $U(j,p,\theta) = \mathbb{E}[v_\psi(\hat{y} - p - c_j^*(\lambda,\omega,j) + l_j + b_j^*(\lambda,\omega))|\lambda \sim F]$, when choosing a hospital. We can also write the expected utility as

$$U(j,p,\theta) = \mathbb{E}[v_\psi(\hat{y} - p + l_j + x_j^*(\lambda,\omega,j))|\lambda \sim F]. \tag{1}$$

We assume the socially optimal utilization to be the same as the privately optimal one (i.e., without reimbursement). Denote $m^*(\lambda,\omega,0)$ as the socially optimal (uninsured) spending. Due to moral hazard, $m^*(\lambda,\omega,j) \geq m^*(\lambda,\omega,0)$. Let's name the difference between the two amounts "moral hazard spending". A patient's resulted benefit from this moral hazard-induced utilization can be decomposed into two parts:

$$\begin{aligned}\Delta x_j^*(\lambda,\omega,j) =& x_j^*(\lambda,\omega,j) - x_0^*(\lambda,\omega,j) \\ =& \underbrace{[b_j^*(\lambda,\omega) - b_0^*(\lambda,\omega)]}_{\text{induced benefit from extra spending}} - \underbrace{[c_j^*(\lambda,\omega,j) - c_0^*(\lambda,\omega,j)]}_{\text{induced OOP cost}}\end{aligned} \tag{2}$$

Note that, $b_0^*(\lambda,\omega) = b(m^*(\lambda,\omega,0),\lambda,\omega)$ is the indirect benefit of uninsured behavior, while $c_0^*(\lambda,\omega,j) = c(m^*(\lambda,\omega,0),j)$ is the indirect OOP cost at insured prices.

To measure a patient's welfare gain from insurance plans, we calculate her willingness to pay (WTP) for the decreased cost sharing, holding all else constant except for the hospital level, we assume that v_ψ is of the constant absolute risk aversion (CARA) form. Then, we define $WTP = p - p_0$ such that

$$U(j,p,\theta) = U(0,p_0,\theta). \tag{3}$$

It can be shown that

$$\begin{aligned}& WTP(j,\theta,c_j) \\ =& \underbrace{\mathbb{E}_\lambda[c_0^*(\lambda,\omega,0) - c_0^*(\lambda,\omega,j)]}_{\text{mean reduced OOP cost at uninsured spending}} + \underbrace{\mathbb{E}_\lambda[\Delta x_j^*(\lambda,\omega,j)]}_{\text{mean payoff from moral hazard spending}} \\ &+ \underbrace{RP(0,\theta) - RP(j,\theta)}_{\text{value of risk change}} + \underbrace{l_j - l_0}_{\text{value of upgrade/downgrade}}\end{aligned} \tag{4}$$

where $RP(j,\theta) = U(j,p,\theta) - v_\psi^{-1}(U(j,p,\theta))$ is the lottery-like risk premium that does not depend on $\hat{y} - p$ (as it will be canceled out under CARA). The above willingness to pay is comprised of four terms. The first term captures the transfer of health-care cost liability from the patient to the insurer associated with hospital j, which occurs

even without moral hazard. The next two terms are relevant to social welfare, and they depend on patient preferences: the second term suggests that patients value the ability to consume more health care when they have (lower) coinsurance; the third term tells how patients value the ability to smooth consumption across health states and how they rate hospital j's ability to help them do so. These first three terms are mentioned in a similar fashion by Azevedo and Gottlieb (2017). Their third term (risk-sharing value) does not consider a possible increase in (subjective) risk as consumers are not tied to any specific service provider in their setting. In addition, we have a fourth term that measures how an average patient values hospital j's level (a summary of all characteristics), and it is independent of individual preferences.

3.1.2 Supply Regulation

We denote the insurer cost as $k(m,j) = m - c(m,j) + l_j - l_0$ (supposing the cost/saving of upgrading/downgrading a hospital is equal to an average patient's value). Thus, the reduced OOP cost in Equation (4) is the increased insurer cost. Define $k_j^*(\lambda,\omega,j) = k(m^*(\lambda,\omega,j),j)$ and $k_0^*(\lambda,\omega,j) = k(m^*(\lambda,\omega,0),j)$.[11] Then, the social surplus (SS) of choosing hospital j against 0 is the difference between $WTP(j,\theta,c_j)$ and the expected insured cost $\mathbb{E}_\lambda[k_j^*(\lambda,\omega,j)]$, written as:

$$\begin{aligned} SS(j,\theta) = & \underbrace{RP(0,\theta) - RP(j,\theta)}_{\text{value of risk change}} \\ & - \underbrace{\mathbb{E}_\lambda\left[k_j^*(\lambda,\omega,j) - k_0^*(\lambda,\omega,j) - \Delta x_j^*(\lambda,\omega,j)\right]}_{\text{social cost of moral hazard}}. \end{aligned} \tag{5}$$

In Equation (5), the social cost is independent of uncertain payoffs as we assume that the insurer is risk neutral. The socially optimal hospital will trade off the value of subjective risk change and social cost of moral hazard: $j^{\text{eff}}(\theta) = \arg\max_{j\in J} SS(j,\theta)$. Given the vector of prices $\mathbf{p} = \{p_j\}_{j\in J}$, the vector of cost-sharing structures $\mathbf{c} = \{c_j\}_{j\in J}$ associated with all potential hospitals, and the vector of hospital fixed values $\mathbf{l} = \{l_j\}_{j\in J}$, patients choose the privately optimal hospital by trading off their private utility and prices (opportunity costs): $j^*(\theta,\mathbf{p},\mathbf{c}) = \arg\max_{j\in J}\{WTP(j,\theta,c_j) - p_j\}$.

[11] Note that, since $c_0^*(\lambda,\omega,0) = m^*(\lambda,\omega,0)$, and $c_0^*(\lambda,\omega,j) = c(m^*(\lambda,\omega,0),j)$, we have $k_0^*(\lambda,\omega,j) = c_0^*(\lambda,\omega,0) - c_0^*(\lambda,\omega,j) + l_j - l_0$.

The regulator can design the cost structure of the healthcare system to align privately optimal $j^*(\theta, \mathbf{p}, \mathbf{c})$ and socially optimal $j^{\text{eff}}(\theta)$ allocations as closely as possible. The equilibrium social welfare can be written as:

$$W(\mathbf{p}, \mathbf{c}) = \int SS(j^*(\theta, \mathbf{p}, \mathbf{c}), \theta) dG(\theta). \tag{6}$$

In our counterfactual analyses, we will explore how the regulator should provide a vertical menu of hospitals with different coinsurance policies. For policymakers, there is a trade-off between risk-smoothing and moral hazard. As more risk-smoothing is welfare-improving, higher coverage also promotes more unnecessary expenditures, and therefore higher social costs. A thorough policy comparison is beyond the scope of this paper, but we implement several policy experiments after model estimation.

3.2 Empirical Model

3.2.1 Parameterization

Second Stage: Utilization Decision. Following Einav et al. (2013) and Lu et al. (2019), we assume that the benefit of health-care spending m is quadratic in its difference from the health risk λ (the amount of spending necessary to treat one's disease). That is,

$$b(m_{it}, \lambda_{it}, \omega_i) = (m_{it} - \lambda_{it}) - \frac{1}{2\omega_i}(m_{it} - \lambda_{it})^2 \tag{7}$$

where the price sensitivity ω_i affects the curvature of the benefit from health-care spending. When choosing the optimal total spending, the patient i takes the OOP cost $c_{jt}(m)$ into consideration. That is,

$$m^*_{jt}(\lambda_{it}, \omega_i) = \arg\max_m \left\{ b(m, \lambda_{it}, \omega_i) - c_{jt}(m) \right\}. \tag{8}$$

The first order condition is

$$m^*_{jt}(\lambda_{it}, \omega_i) = \omega_i(1 - c'_{jt}(m^*_{jt}(\lambda_{it}, \omega_i))) + \lambda_{it}. \tag{9}$$

Note that, without any coverage, the patient would spend λ_{it} exactly; however, with full coverage, the patient would spend $\lambda_{it} + \omega_i$. This suggests that ω_i is the overconsumption induced by moving from no insurance to full coverage, while λ_{it} reflects the patient's underlying fundamental need for health care.

Let's also denote $b^*_{jt}(\lambda_{it}, \omega_i)$ as the benefit of optimal utilization and $c^*_{jt}(\lambda_{it}, \omega_i)$ as the associated OOP cost, when substituting for m^*. Given the optimal decision in the second stage, patients only face uncertainty about payoffs through the uncertainty in $b^*_{jt}(\lambda_{it}, \omega_i) - c^*_{jt}(\lambda_{it}, \omega_i)$ in the first stage.

First Stage: Hospital Choice. Before choosing a hospital, the patient receives a private signal about her latent health status $\lambda_{it} \sim F^{\lambda}_{it}$. She therefore chooses a health-care provider j from set J_{it} (all the hospitals available[12] to the patient) to maximize the objective function below:

$$v_{ijt}\left(F^{\lambda}_{it}(\cdot), \omega_i, \psi_i\right) = \int \frac{-\exp\left(-\psi_i u^*_{ijt}(\lambda, \omega_i)\right)}{\psi_i} dF^{\lambda}_{it}(\lambda), \quad \psi_i \neq 0. \tag{10}$$

Here, in line with our theoretical model, preferences are assumed to exhibit CARA and the coefficient of absolute risk aversion is ψ_i.[13] It is important to note that, this risk attitude parameter can reflect two effects that are countervailing: (i) attitudes toward financial uncertainty, and (ii) subjective perceptions about service quality and its relation with financial uncertainty. With moral hazard, the von Neumann Morgenstern (vNM) utility function is defined over the payoff from health spending (in the second stage) and some hospital and individual characteristics. By extending Equation (1), we define this payoff by

$$u^*_{ijt}(\lambda, \omega_i) = \beta_{0,j} + \beta_1\left(b^*_{jt}(\lambda, \omega_i) - c^*_{jt}(\lambda, \omega_i)\right) + \beta_2 D_{ijt} + Z'_{jt}\beta_3 + \sigma_\epsilon \epsilon_{ijt} \tag{11}$$

where $\beta_{0,j}$ is the fixed effect of provider j's level/tier, D_{ijt} measures the travel distance between patient i's home address and health-care provider j, Z_{jt} contains observed measures of hospital features, and ϵ_{ijt} is the idiosyncratic taste shock that follows an i.i.d. type-I extreme value distribution, with the magnitude σ_ϵ to be estimated. As a result,

$$j^*\left(F^{\lambda}_{it}(\cdot), \omega_i, \psi_i\right) = \arg\max_{j \in J_{it}} v_{ijt}\left(F^{\lambda}_{it}(\cdot), \omega_i, \psi_i\right). \tag{12}$$

[12] Availability is defined by two aspects: (1) the patient's disease can be treated in the hospitals, and (2) the hospitals are nearer to the chosen hospital (not necessarily to the patient's home) than other hospitals are. In practice, we chose the nearest 2–3 hospitals (including the chosen one).

[13] We allow for a negative value of ψ_i, which is identified from the cases where patients with higher uncertainty in spending choose a less generous hospital holding all other characteristics similar. We may interpret it as a belief that less generous hospitals/services are more capable of controlling uncertain risks, rather than risk-taking. This relaxation improves our model fit greatly.

Health Information. Suppose patients believe that health risks are drawn from a right-truncated lognormal distribution of health states,[14]

$$\log \lambda_{it} \sim N(\mu_{\lambda,it}, \sigma^2_{\lambda,it})\mathbf{1}\{0 < \lambda \leq \overline{\lambda}\}, \tag{13}$$

with support $(0, \infty)$. $\sigma_{\lambda,it}$ indicates the precision of the patient's information about her subsequent health and is assumed to be time-varying.

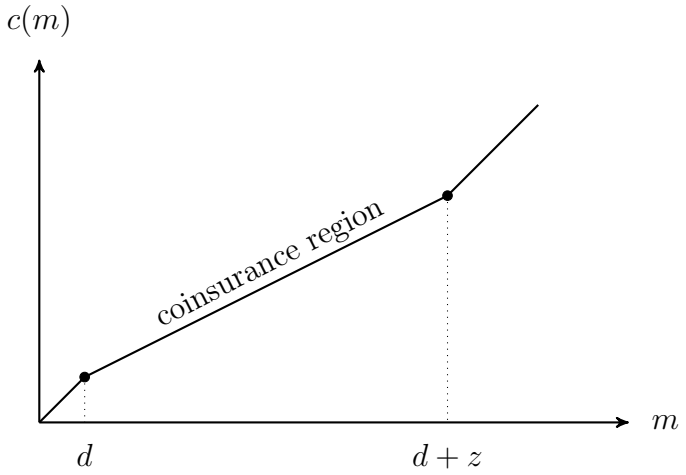


Figure 4: Out-of-Pocket Cost Function, $a = 0.5$

Reimbursement Scheme. In our context, the OOP cost function is nonlinear—or more precisely, piecewise linear (see Figure 4 for an illustration). The marginal OOP cost function $c'(m)$ in Equation (9) is thus piecewise constant. In the region $m \leq d$ or $m \geq d + z$ where d denotes deductible and z denotes the maximum reimbursable spending, we have $c'(m) = 1$; in the region $d < m < d + z$, $c'(m) = 1 - a \in (0, 1)$. As a result, the optimal m^* would be a step function of λ, also depending on ω. We next derive cutoff values on the health state that determine which OOP region a patient will find a specific solution optimal.

[14]The truncation helps us avoid the explosion of numerical integration and the violation of an implicit assumption in our model. That is, we assume that patients can afford all potential OOP cost realizations.

If $z > \omega a/2$ (large coinsurance region), then

$$m^* = \begin{cases} \lambda & \text{if } \lambda \leq d - \omega a/2 \\ \lambda + \omega a & \text{if } d - \omega a/2 < \lambda \leq d + z - \omega a \\ d + z & \text{if } d + z - \omega a < \lambda \leq d + z \\ \lambda & \text{if } \lambda > d + z \end{cases}. \tag{14}$$

If $0 < z \leq \omega a/2$ (small but non-zero coinsurance region), then

$$m^* = \begin{cases} \lambda & \text{if } \lambda \leq d - \omega a/2 \\ \lambda + \omega a & \text{if } d - \omega a/2 < \lambda \leq d + z - \sqrt{2\omega a z} \\ d + z & \text{if } d + z - \sqrt{2\omega a z} < \lambda \leq d + z \\ \lambda & \text{if } \lambda > d + z \end{cases}. \tag{15}$$

If $z = 0$ (no coinsurance region), then $m^* = \lambda$. All hospitals in our empirical setting have large coinsurance regions. Derivations can be provided upon request.[15]

3.2.2 Identification

Our goal is to recover the joint distribution across patients of willingness to pay, risk attitude, and the social cost of moral hazard associated with different hospitals. Variation in these objects comes from variation in either patient preferences (the risk attitude and moral-hazard parameters) or in the distribution of health states. Major concerns include (1) distinguishing preferences (ω_i) from private information about health ($\mu_{\lambda,it}$), (2) distinguishing taste for reimbursed spending (β_1) from risk attitude (ψ_i), and (3) identifying heterogeneity in the risk attitude (ψ_i) and moral hazard (ω_i) parameters.

First, when observing a positive correlation between reimbursement generosity and total health-care spending (conditional on observable characteristics) in the data, we can explain it as either the effect of private health information affecting hospital choice

[15] We can see from Equation (14) that there is bunching at the convex kink point $d+z$, as discussed by Einav et al. (2017). However, unlike the "donut hole" in the context of prescription drug insurance for the elderly in Medicare Part D, our kink point is quite high in the budget set, and empirically almost none of our patients are near there. Thus, our identification will not rely on bunching.

(selection) or lower OOP prices driving utilization (moral hazard). To distinguish one explanation to the other, we need variation in hospital menus J_{it} (sets of available or accessible hospitals). When hospital choices vary with menus, the degree of moral hazard can be identified by the extent to which patients facing more generous hospital menus also have higher health-care spending. On the other hand, when observing patients who face similar menus making different hospital choices, we can identify the amount of private information about health and the magnitude of the idiosyncratic shock σ_ϵ: conditional on observables and the predicted effects of moral hazard, if patients who inexplicably choose more generous hospital inexplicably spend more on health care, this variation in hospital choice will be attributed to private information about health; otherwise, we attribute any residual unexplained variation in hospital choice to the idiosyncratic shock.

Second, in our model, both risk parameter ψ_i and taste for reimbursed spending β_1 affect hospital choice but not spending. To distinguish between them, we can utilize cases in which observably different patients face similar hospital menus. Risk attitude is then identified by how a patient associates uncertainty in reimbursed spending with reimbursement rate, holding the expected OOP cost fixed. The taste for reimbursed spending is identified by the rate at which patients trade off other hospital characteristics (e.g., distance) with expected OOP cost, holding uncertainty in OOP cost fixed.

Third, we rely on the panel nature of our data to identify unobserved heterogeneity in the risk attitude and moral hazard parameters. By observing the same patients making choices under different circumstances, we can apply the previous arguments patient by patient to obtain patient-specific estimates. To ask less of the data, we assume that the distribution of unobserved heterogeneity is multivariate normal. The variance and covariance of the unobserved components of patient types are identified by the extent to which different patients consistently act in different ways.

3.2.3 Estimation

Structural Settings. First, we need to make some assumptions about the structure of the parameters of individual health status distributions $F_{it}^{\lambda}(\cdot)$: let's assume a fixed-

effect structure on $\mu_{\lambda,it}$ and that $\sigma_{\lambda,it}$ can be projected on time-varying patient characteristics. That is,

$$\mu_{\lambda,it} = \overline{\mu_{\lambda,i}} + (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\beta_\mu \tag{16}$$

$$\sigma_{\lambda,it} = \mathbf{x}^\sigma_{it}\beta_\sigma \tag{17}$$

where $\overline{\mathbf{x}}_i$ is a vector of within-individual averages of $\mathbf{x}_{it}$, including an individual health risk predictor[16] calculated based on age, gender, education, marital status, and the International Classification of Diseases (ICD) 10 codes; $\overline{\mu_{\lambda,i}}$ is the average (over time) of $\mu_{\lambda,it}$ drawn from the jointly right-truncated normal distribution described below.

$$\begin{pmatrix} \overline{\mu_{\lambda,i}} \\ \log\omega_i \\ \psi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \overline{\mathbf{x}}_i\beta_\mu \\ \overline{\mathbf{x}}^\omega_i\beta_\omega \\ \overline{\mathbf{x}}^\psi_i\beta_\psi \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma^2_\mu & \sigma_{\mu,\omega} & \sigma_{\mu,\psi} \\ \sigma_{\mu,\omega} & \sigma^2_\omega & \sigma_{\omega,\psi} \\ \sigma_{\mu,\psi} & \sigma_{\omega,\psi} & \sigma^2_\psi \end{pmatrix}}_{\Sigma_1} \right) \mathbf{1}\{0 < \omega \leq \overline{\lambda}\} \tag{18}$$

There are both observed (via mean) and unobserved (via covariance) heterogeneity in each parameter. Covariates $\mathbf{x}^\sigma_{it}$, $\mathbf{x}^\omega_{it}$ and $\mathbf{x}^\psi_{it}$ include a standardized risk predictor[17] and a constant.

The parameters to be estimated are the 4 vectors of mean shifters $(\beta_\mu, \beta_\sigma, \beta_\omega, \beta_\psi)$, 6 variance and covariance parameters $(\sigma_\mu, \sigma_\omega, \sigma_\psi, \sigma_{\mu,\omega}, \sigma_{\mu,\psi}, \sigma_{\omega,\psi})$, and 5 (vectors of) taste/magnitude parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma_\epsilon)$.

Algorithm. We resort to a maximum likelihood approach.

Denote the full set of parameters to be estimated as θ, which describes the joint distribution of $\alpha_{it} = \{\mu_{\lambda,it}, \omega_i, \psi_i\}$ (i.e., health state, risk attitude, and moral hazard). For each guess of θ, we simulate the distribution of α_{it} using Gaussian quadrature, yielding simulated points $\alpha_{its}(\theta) = \{\mu_{\lambda,its}, \omega_{is}, \psi_{is}\}$ as well as weights W_s. Given a simulation draw s, we calculate the conditional probability density at the observed health-care spending and the probability of observed hospital choices.

[16] More details about the calculation of risk predictor can be found in Appendix A.2.

[17] The risk predictor is shifted to make the smallest value 0, and then scaled down to make the largest value 1, leading to the standardized risk predictor between 0 and 1.

First, we construct the distribution of spending for each patient-visit implied by the model and guess of θ. Our model predicts that $m^* = \omega_{is}(1 - c'_{jt}(m^*)) + \lambda$. By inverting the expression, the corresponding health state realization is $\lambda_{ijts} = m_{it} - \omega_{is}(1 - c'_{jt}(m_{it}))$. Then, the density of m_{it} is given by the density of λ_{ijts}, so the probability density of total health-care spending conditional on hospital, guess of parameters, and patient observables is given by

$$f_m(m_{it}|c_{jt}, \alpha_{its}, \theta, \mathbf{x}_{it}) = \frac{\Phi'\left(\frac{\log \lambda_{ijts} - \mu_{\lambda,its}}{\sigma_{\lambda,it}}\right)}{\Phi\left(\frac{\log \overline{\lambda} - \mu_{\lambda,its}}{\sigma_{\lambda,it}}\right)} \tag{19}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Second, we calculate the probability of each hospital choice. Given θ and α_{its}, we simulate the distribution of health states by $\lambda_{ijtsd} = \exp(\mu_{\lambda,its} + \sigma_{\lambda,it} Z_d)$ where Z_d is a vector of points approximating a standard normal distribution, and we denote W_d as the associated Gaussian quadrature weights. Then, we calculate the optimal health-care spending m_{ijtsd} associated with each potential health state realization based on formula (14):[18]

$$m^*_{ijtsd} = \begin{cases} \lambda_{ijtsd} + \omega_{is} a_{jt} & \text{if } d_{jt} - \frac{\omega_{is} a_{jt}}{2} < \lambda_{ijtsd} \leq d_{jt} + z_{jt} - \omega_{is} a_{jt} \\ d_{jt} + z_{jt} & \text{if } d_{jt} + z_{jt} - \omega_{is} a_{jt} < \lambda_{ijtsd} \leq d_{jt} + z_{jt} \\ \lambda_{ijtsd} & \text{otherwise} \end{cases} . \tag{20}$$

Now, we have the distributions of privately optimal total spending m^*_{ijtsd} for each patient-hospital-visit and draw of α_{its} to calculate the patient's expected utility from choosing each potential hospital. Then, the numerical approximation to Equation (10) is constructed using the quadrature weights W_d mentioned above:

$$v_{ijts} = \frac{-\sum_{d=1}^{N_d} W_d \cdot \exp\left(-\psi_{is} u^*_{ijts}(\lambda_{ijtsd}, \omega_{is})\right)}{\psi_{is}}, \quad \psi_{is} \neq 0, \tag{21}$$

where N_d is the number of support points and the payoff u^* is calculated as in Equation (11). In practice, we estimate the model in certainty-equivalent (CE) units of v_{ijts} to avoid numerical issues when dealing with double-exponentiation:

$$v^{CE}_{ijts} = \overline{u}_{ijts} - \frac{1}{\psi_{is}} \log\left(\sum_{d=1}^{N_d} W_d \cdot \exp\left(-\psi_{is}(u^*_{ijts}(\lambda_{ijtsd}, \omega_{is}) - \overline{u}_{ijts})\right)\right), \quad \psi_{is} \neq 0 \tag{22}$$

[18] Note that, $z_{j,t}$ depends on $z_{j,t-1}$ if both t and $t-1$ are in the same year.

where $\overline{u}_{ijts} = \mathbb{E}_d[u^*_{ijts}(\lambda_{ijtsd}, \omega_{is})]$.

The choice probabilities conditional on α_{its} are given by the standard logit formula

$$L_{ijts} = \frac{\exp\left(v^{CE}_{ijts}/\sigma_\epsilon\right)}{\sum_{j\in J_{it}} \exp\left(v^{CE}_{ijts}/\sigma_\epsilon\right)}. \tag{23}$$

Third, we write the numerical approximation to the likelihood of the sequence of choices and spending amounts for a given patient:

$$L_i = \sum_{s=1}^{N_s} W_s \prod_{t=1}^{T} \prod_{j\in J_{it}} \left(f_m(m_{it}|c_{jt}, \alpha_{its}, \theta, \mathbf{x}_{it}) L_{ijts}\right)^{d_{ijt}} \tag{24}$$

where N_s is the number of support points in the first step, and $d_{ijt} = 1$ if patient i chose hospital j in visit t and 0 otherwise. The simulated log-likelihood function for parameters θ is then

$$LL(\theta) = \sum_{i=1}^{N} \log(L_i). \tag{25}$$

Recovering Individual Types. We assume that individual types $\alpha_{it}(\theta) = \{\mu_{\lambda,it}, \omega_i, \psi_i\}$ are distributed multivariate normal with observable heterogeneity in the mean vector based on Equation (18). The above algorithm will provide us with $\hat{\theta}$, an estimate of θ, which helps us back out individual types using a sequence of observed hospital choices and medical expenses, denoted as $\mathbf{y}$. Denote the population distribution of types as $g(\alpha|\hat{\theta})$, the probability of observed outcomes as $p(\mathbf{y}|\hat{\theta})$, and the conditional probability of observed outcomes $p(\mathbf{y}|\alpha)$ (the "conditioning of individual tastes"). Then, according to Bayes' rule, the density of α conditional on parameters and observed outcomes $h(\alpha|\hat{\theta}, \mathbf{y})$ (the posterior distribution of α) can be written as

$$h(\alpha|\hat{\theta}, \mathbf{y}) = \frac{p(\mathbf{y}|\alpha) \cdot g(\alpha|\hat{\theta})}{p(\mathbf{y}|\hat{\theta})}. \tag{26}$$

The numerical approximation to each patient's posterior distribution of unobserved heterogeneity is therefore

$$h_{is}(\alpha|\hat{\theta}, \mathbf{y}_i) = \frac{L_{is} \cdot W_s}{L_i}, \tag{27}$$

where $L_{is} = \prod_{t=1}^{T} \prod_{j\in J_{it}} \left(f_m(m_{it}|c_{jt}, \alpha_{its}, \theta, \mathbf{x}_{it}) L_{ijts}\right)^{d_{ijt}}$ and $\sum_{s=1}^{N_s} h_{is}(\alpha|\hat{\theta}, \mathbf{y}_i) = 1$. Each patient's expected types with respect to the posterior distribution of unobserved

heterogeneity are hence

$$\mathbb{E}\overline{\lambda}_{it} = \sum_{s=1}^{N_s} h_{is}(\alpha|\hat{\theta}, \mathbf{y}_i) e^{\mu_{\lambda,its}+\frac{1}{2}\sigma^2_{\lambda,it}} \cdot \frac{\Phi\left(\frac{\log\overline{\lambda}-\mu_{\lambda,its}-\sigma^2_{\lambda,it}}{\sigma_{\lambda,it}}\right)}{\Phi\left(\frac{\log\overline{\lambda}-\mu_{\lambda,its}}{\sigma_{\lambda,it}}\right)}, \tag{28}$$

$$\mathbb{E}\omega_i = \sum_{s=1}^{N_s} h_{is}(\alpha|\hat{\theta}, \mathbf{y}_i)\omega_{is}, \tag{29}$$

$$\mathbb{E}\psi_i = \sum_{s=1}^{N_s} h_{is}(\alpha|\hat{\theta}, \mathbf{y}_i)\psi_{is}. \tag{30}$$

4 Data

4.1 Summary Statistics

We use a set of unique medical claims data for all the enrollees of the NRCMS program from a county-level city located in the southwestern part of China. The data record all inpatient service visits at the local health institutions. Detailed information concerning each visit is contained in the data, including the date of visit, diagnosis (the ICD 10 code), medical organization visited, total medical expenditure, and the amount of insurance reimbursement received. We also find patient demographics in the data, such as birthdate, gender, marital status, and education level.

Between 2012 and 2014, our data record 113,662 inpatient service visits by 66,316 patients. In 2014 alone, 39,743 inpatient visits are made by 29,647 rural residents, indicating that at least 9.3% of the rural residents (if we assume that all of them are covered by the NRCMS) in the study county get hospitalized at least once. We select a sample from the data with complete information of our interest (e.g., home address), which leads to a one-third reduction in the sample size. The summary statistics of our main variables of interest are reported in Table 2. As presented in the table, our study sample contains nearly eighty thousand inpatient visits by forty-seven thousand patients with complete information.[19]

[19] In Table A3, we show the summary statistics of the main variables in the full sample. The means are comparable to the ones in the first column of Table 2.

Table 2: Descriptive Statistics for the Estimation Sample

	2012–2014	2012	2013	2014
Patient level				
Male	0.429	0.425	0.427	0.425
	(0.495)	(0.494)	(0.495)	(0.494)
Number of visits per patient	1.708	1.349	1.349	1.343
	(1.534)	(0.897)	(0.919)	(0.892)
Total patients	46,577	18,521	19,551	20,971
Patient-visit level				
Age	56.247	55.105	55.837	57.644
	(18.508)	(19.445)	(18.432)	(17.618)
—Age 18–60	0.494	0.499	0.508	0.476
	(0.500)	(0.500)	(0.500)	(0.499)
Years of schooling	5.246	5.116	5.294	5.318
	(3.716)	(3.747)	(3.712)	(3.690)
—Middle school or more	0.290	0.280	0.297	0.292
	(0.454)	(0.449)	(0.457)	(0.455)
Married	0.731	0.716	0.735	0.741
	(0.443)	(0.451)	(0.442)	(0.438)
Proportion of county hospital visits	0.564	0.563	0.576	0.554
	(0.496)	(0.496)	(0.494)	(0.497)
Relative health risk[§]	1.000	1.000	1.000	1.000
	(1.077)	(1.094)	(1.092)	(1.048)
Total medical spending (thousand)	3.150	2.869	3.236	3.318
	(5.114)	(4.786)	(5.400)	(5.109)
Deductible paid (thousand)	0.239	0.232	0.242	0.243
	(0.145)	(0.147)	(0.145)	(0.144)
Reimbursement rate received	0.624	0.670	0.632	0.577
	(0.159)	(0.182)	(0.134)	(0.143)
Total visits	79,531	24,984	26,384	28,163

Notes: This table presents the summary statistics for the estimation sample. Standard deviations are in the parentheses under the means. Medical spending and deductible are both in thousands of RMB yuan; years of schooling are based on the highest education level attended; age is calculated as the calendar year age in the year getting treated. [§]Relative health risk is measured by the rescaled risk score explained in Appendix A.2.

About 42.9% of patients in our sample are male, and an average person visits a hospital almost twice during our study period (less than once a year). Among these visits, about 56.4% of them occur in county hospitals, and the remainder are in township health centers. Approximately half of the visits are paid by patients aged 18 to 60 years, and the average age of patients at the time of their visits is 56 years. The average patient has completed only five years of schooling (i.e., mostly finishes elementary school)—this is not surprising given the fact that our respondents are predominantly elderly rural residents—only 29.0% of them have been to middle school or above. Interestingly, around 73.1% of the visits are paid by married patients, while the remainder are paid by patients who are single, divorced, or widowed. Given the fact that half of the sample are under the age of 18 or over the age of 60, the married proportion seems high. The rescaled risk score is obtained from the Johns Hopkins ACG system (v. 12.1), and more details can be found in Appendix A.2.[20] For each inpatient visit, patients spend on average 3.2 thousand RMB yuan in our study sample, and the deductible is about 0.2 thousand RMB yuan. On average, after paying for the deductible, patients can reimburse 62.4% of the rest of the total spending.[21]

Patient characteristics are stable across the three study years. On the other hand, the healthcare system of the study county experiences a transformation toward lower reimbursement rates and higher deductibles. This provides variation in cost-sharing policies for the identification of patient preferences.

4.2 Variation in Reimbursement Rates and Hospital Menus

It's important for this research to check two key features of our setting—the plausibly exogenous variation in hospital menus and isolated variation along the dimension of

[20] 9.5% of the diagnoses suggest a serious disease in the department of general surgery [e.g., severe acute pancreatitis (ICD 10: K85)], neurosurgery [e.g., acoustic neuroma (ICD 10: D33.3), urinary surgery [e.g., muscle-invasive bladder cancer (ICD 10: C67)], orthopedics [e.g., spinal tuberculosis (ICD 10: A18.0 and M49.0)], or hematology [e.g., myelodysplastic syndrome (ICD 10: D46)].

[21] For the purposes of our empirical model, we estimate the reimbursement rate that best fits the relationship between OOP spending and total spending observed in the claims data. There is no maximum for OOP but for reimbursement. Thus, we limit the range of total spending when estimating the reimbursement rate. A more detailed procedure is described in Appendix A.1.

coinsurance level. To better illustrate the variation, we focus on the most popular disease among our patients—acute exacerbations of chronic obstructive pulmonary disease (AECOPD, ICD 10: J40–J44).[22] This disease can be treated in either a THC or a county hospital. Since most (around 75%) of the AECOPD patients are treated in THCs, we focus on patients who choose a THC in this discussion.

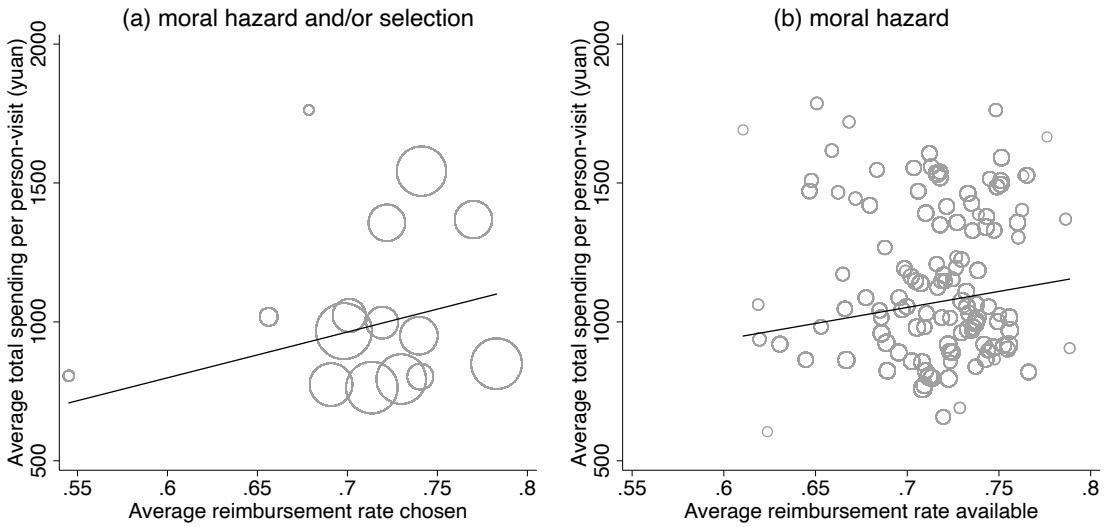


Figure 5: Average Spending by Generosity Chosen and Available

Notes: The figure shows the relationship between average total spending per person-visit and reimbursement rate for individuals that choose to treat AECOPD in THCs between 2012 and 2014. In the left panel, each dot represents one of the 15 THCs treating this disease. In the right panel, individuals are grouped by their community (or hospital menu), and each dot represents a unique value of average reimbursement rate available. The size of each dot indicates the number of individuals represented.

As shown in Figure 5, conditional on AECOPD and THC, health-care spending is positively correlated with reimbursement generosity. In the left panel, individuals are grouped by their chosen hospital, and the plot shows the average spending per person-visit in each of the THCs, weighting each hospital by admission. Consistent with our expectation, individuals who chose more generous THCs have higher spending, indicating moral hazard, adverse selection, or both. To distinguish between moral hazard and selection, the right panel, grouping individuals by their community (and

[22] See Liang et al. (2020b) for a more detailed classification of AECOPD into 4 sub-diseases.

thus their corresponding hospital menu), plots the average rate of reimbursement in the hospitals available/accessible to that community against the average spending of individuals living in the community. We notice that individuals with access to more generous THCs, thus arguably more likely to choose a THC with a higher reimbursement rate, have larger spending. This suggests the presence of moral hazard as well as the coexistence of adverse selection on unobservables.

Our structural model is identified in a similar way. A key identifying assumption is that, conditional on observables, hospital menus are not related to unobservables of individuals that could affect health-care spending. This can be threaten by township governments or community leaders trying to set hospital generosity in response to unobservable information about residents that would drive spending. For example, if more generous hospitals are open for communities with unobservably healthier residents, the extent of moral hazard can be underestimated. To see if this is the case, we seek to explain hospital menu generosity by individual health risk predictor and other observables. We argue that, if hospital menus are not responding to our risk predictor, it is unlikely that they are responding to unobservables, because we should have better information on these patients (collected when they are admitted to the hospitals) than township governments and community leaders do before their admission. To hold hospital level and disease effects fixed, we again focus on AECOPD patients treated in THCs. As shown by Table A2, conditional on community features and hospital features chosen (such as number of beds and number of doctors), we do not find any correlation between hospital menu generosity and the individual health indicator.

5 Results

5.1 Parameter Estimates

The parameter estimates of our structural model are shown in Table 3. Based on the results, we estimate an average moral hazard parameter ω of 0.4 thousand RMB yuan (less than 0.1 thousand US$ in 2014). This is smaller than Einav et al. (2013)'s estimate of the average ω, 1.3 thousand US$. This difference is due to several reasons.

Table 3: Parameter Estimates

Variable	Parameter	Robust Std. Err.
County hospital fixed effect, β_0	4.9225	0.0204
Net benefit from utilization stage, β_1	4.4367	0.0062
Distance (km), β_2	-0.1000*	–
Number of doctors, β_3	-0.0151	0.0003
Number of beds, β_3	-0.0413	0.0001
Taste shock's scale, σ_ϵ	-5.6434	0.0013
Health state mean × risk predictor, β_μ	1.0129	0.0002
Health state mean intercept, β_μ	-0.0092	0.0001
Health state mean's std. dev., σ_μ	0.0330	0.0000
Health state std. dev. × standardized risk predictor§, β_σ	0.1104	0.0003
Health state std. dev. intercept, β_σ	0.2258	0.0003
Risk attitude × standardized risk predictor§, β_ψ	2.5258	0.0231
Risk attitude intercept, β_ψ	2.3387	0.0238
Risk attitude std. dev., σ_ψ	4.5979	0.0164
Log moral hazard × standardized risk predictor, β_ω	10.3156	0.0049
Log moral hazard intercept, β_ω	-8.0730	0.0048
Log moral hazard std. dev., σ_ω	4.1135	0.0002
Corr. b/w health and log moral hazard, $\rho_{\mu,\omega}$	0.5519	0.0003
Corr. b/w health and risk attitude, $\rho_{\mu,\psi}$	0.6168	0.0063
Corr. b/w log moral hazard and risk attitude, $\rho_{\omega,\psi}$	0.9753	0.0010

Notes: Parameter estimates are all significant at the 1% level; robust standard errors are calculated based on the numerically approximated gradient and Hessian of the likelihood function; the model is estimated on an unbalanced panel of 46,577 individuals over three years. * By normalization. § The risk predictor is shifted to make the smallest value 0, and then scaled down to make the largest value 1, leading to the standardized risk predictor between 0 and 1.

First, our ω represents the extra total spending per visit,[23] while Einav et al. (2013)'s ω is the extra total spending per year. Second, we focus on inpatient care in rural China, where the average price level is much lower and resources are more limited; our average per-visit spending is around 3.2 thousand RMB yuan (about 0.5 thousand US$ in 2014), while Einav et al. (2013)'s average annual spending is more than ten times our level. Therefore, the estimated ω is still quite significant in our case. Note that, ω is the additional total spending induced by moving a patient from no coverage to full reimbursement. Thus, this estimate implies that moving from a hospital with half the prices (after a deductible and before reaching a cap) to one with zero costs is expected to increase inpatient-care spending by six percent of the mean spending. Interestingly, our model suggests that moral hazard is idiosyncratically more serious among people who privately expect that they are less healthy, as $\rho_{\mu,\omega} > 0$.

We find that patients in rural China have a wide range of risk attitudes, with the mean (median) coefficient of absolute risk aversion being -0.2 (0.4). We may translate it to an amount of money, say $\$X$, such that individuals are indifferent between (i) a payoff of zero and (ii) an equal-odds gamble between gaining \$100.0 and losing $\$X$. Based on our calculations, the mean (median) value of such indifferent value ($\$X$) is \$101.6 (\$96.6). The fact that some patients are willing to lose more money than their potential gain does not necessarily suggests that they love gambling in our context. Rather, they may perceive high cost-sharing as an indicator of high service quality (and are willing to pay \$3.4 for a chance to enjoy a higher quality). In our data, we observe that some patients would prefer a less generous hospitals given similar hospital characteristics, and these behaviors or preferences could not be explained by a hospital tier fixed effect. We reflect these preferences by allowing for negative coefficients, but we may not use the traditional term "risk taking" even though it is shown widespread among rural Chinese patients in various forms (Carlsson et al., 2012; Jin et al., 2017).[24] Our estimation suggests that risk aversion also increases idiosyncratically with private information about higher spending expectation, as $\rho_{\mu,\psi} > 0$, which is intuitive. Finally, we do find that more risk averse people (who may care less about service quality) are idiosyncratically more prone to moral hazard. For the

[23] Some patients can have multiple visits per year.

[24] We also notice that having no restriction on the coefficient value can improve model fit greatly, although it may complicate the economic meaning of this coefficient.

unconditional joint distribution of the three dimensions of patient type, please refer to Figure A2.

Our estimates illustrate the trade-off between travel distance, OOP costs, and the access to a county hospital. An average patient would be willing to travel an additional distance of 49.2 kilometers (km) to switch from a THC to a county hospital (perhaps due to its higher social reputation), and an additional distance of 44.4 km for a unit increase in net payoff of utilization. Interestingly, we notice that more doctors or beds are associated with less willingness to travel.

5.2 Model Fit

We evaluate model fit from two perspectives, corresponding to the hospital choice stage and the utilization stage respectively.

First, we can compare the observed and predicted market shares for each hospital. According to Figure 6, the model prediction is quite good at the hospital tier level. To inspect the flexibility of the model with respect to the choice of a specific hospital within a tier, we also show the market shares at the hospital level in Figure A3. It turns out to be reasonably good as well.

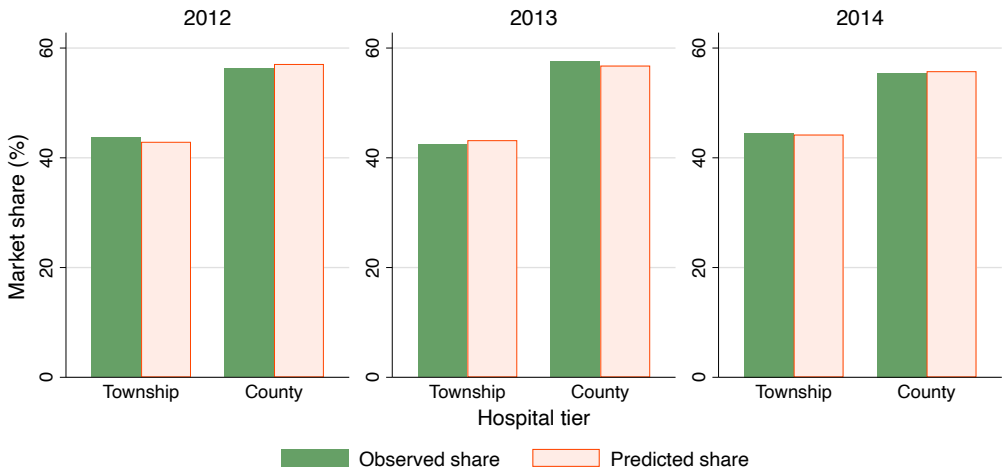


Figure 6: Model Fit: Hospital Tier Choices

Notes: The figure shows the observed and predicted market shares at the hospital tier level calculated based on Table 3. An observation is a person-visit in each year.

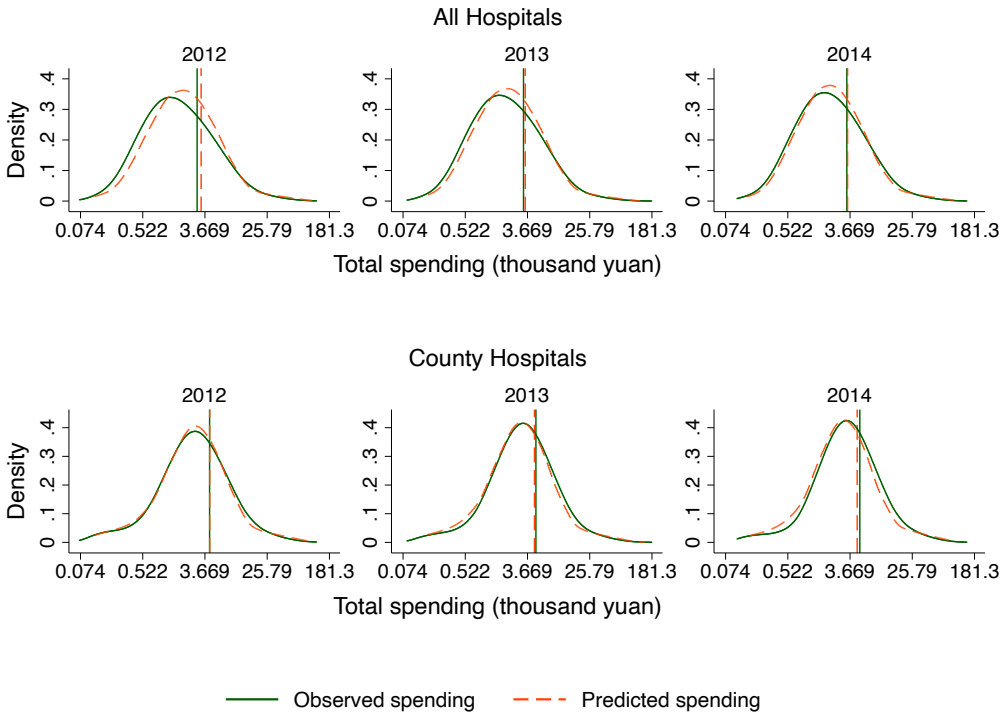


Figure 7: Model Fit: Inpatient-Care Spending

Notes: The kernel density plots of the observed and predicted distributions of total inpatient-care spending are on a log scale. An observation is a person-visit in each year. Predicted distributions are calculated based on Table 3.

Second, we can compare the observed and predicted distributions of patients' total inpatient-care spending per visit each year. The expected spending of each patient is used to construct the predicted spending distribution in the population of patients. We show the kernel density plots of spending on a log scale in Figure 7. If we pool THCs and county hospitals together, our model tends to overestimate the spending slightly in 2012. The predicted mean matches the observed mean well in 2013 and 2014, however. This could be partly because our estimation procedure pools all the three years of data together, while there might be some heterogeneity in 2012. If we focus on only county hospitals, nevertheless, the fit is much better, although there might be a slight underestimation of spending in 2014. On average, the inpatient-care spending is predicted to be 3,186 RMB yuan across patient-visit observations in our data, which is close to the observed average (3,150 RMB yuan). By implementing a Kolmogorov-Smirnov test, we cannot reject the equality of the observed and predicted

distributions of spending.[25]

5.3 Willingness to Pay

In this subsection, we construct each patient's willingness to pay for different levels of hospitals and reimbursement rates according to our previous parameter estimation. To map our empirical model to the theoretical framework, a few simplifications are needed. First, we limit our focus to the AECOPD patients who only have one inpatient visit between 2012 and 2014 (N = 4,612). This allows us to assign a single type $\alpha_i = \{F_i^\lambda, \omega_i, \psi_i\}$ to each patient, where F_i^λ is a right-truncated lognormal distribution described by $\{\mu_{\lambda,i}, \sigma_{\lambda,i}, \overline{\lambda}\}$.[26] Second, we assume that the idiosyncratic shock is utility-irrelevant.[27] Next, we hold all non-financial features fixed to limit our attention to the cost-sharing dimension. Last, we assume that all patients have the same per unit opportunity cost of travel, which is 22.5 RMB yuan per km.[28]

The reference hospital $j = 0$ is a county hospital that does not reimburse any cost, which is not observed in our data. Based on Equation (21), we calculate the utility of choosing hospital $j > 0$ in CE units, denoted as v_{ij}^{CE} and then calculate willingness to pay as $WTP_{ij} = v_{ij}^{CE} - v_{i0}^{CE}$. We decompose WTP into four terms according to Equation (4)—a "transfer" term that represents the mean reduced OOP cost holding patient behavior constant, a "moral hazard" term that describes the mean net payoff from moral hazard spending, a "risk attitude" term that shows how much a patient values the reduction of financial uncertainty (financial risk protection) over the mis-

[25] To avoid repeated values or ties in the test of continuous distributions, we obtain 499 quantiles for the observed distribution as well as the predicted one. Then, the Kolmogorov-Smirnov (K-S) test is implemented using the 998 quantiles. Our combined K-S statistic is 0.0641, and its corresponding p-value is 0.256.

[26] This is done by integrating over everyone's posterior distribution of types described by Equations (28)–(30).

[27] We consider the remaining choice determinants in ϵ as monkey-on-the-shoulder tastes (Akerlof and Shiller, 2015) or mistakes (Handel and Kolstad, 2015), and thus omit this term in our utility calculation.

[28] Since all patients live in the same county-level city, they are limited to very few means of transportation. Thus, for simplicity, we assume all of them to have the same per unit opportunity cost, backed out from the coefficient β_1 in Table 3 (i.e., $0.1 \times 1{,}000/4.4367$).

trust of quality associated with low prices, and finally a fixed "tier change value" term that reflects the monetary value of a hospital level change to an average patient.

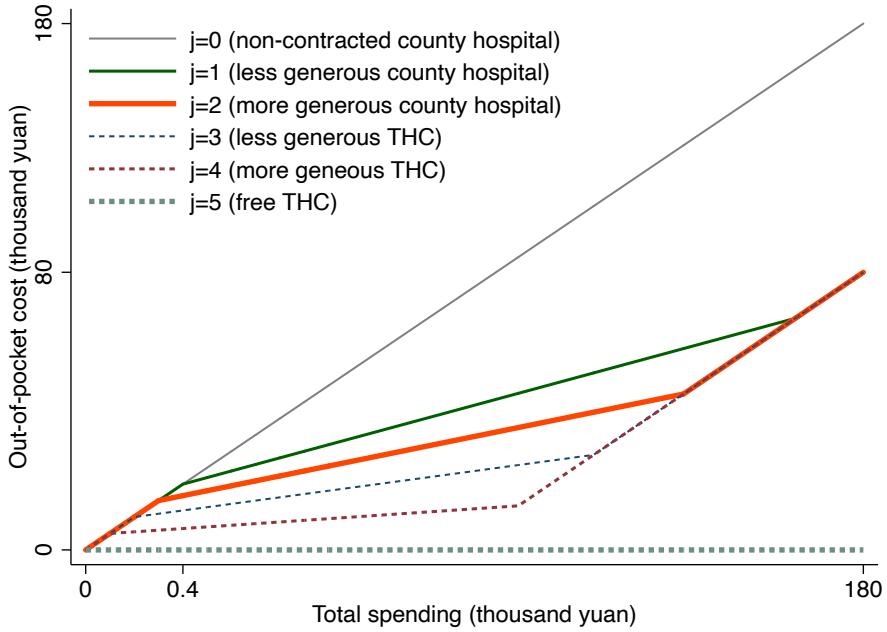


Figure 8: A Simplified Tiered Medical System

Notes: This graph shows a subset of hospitals representing a tiered medical system, including county hospitals $j = 0, 1, 2$ and THCs $j = 3, 4, 5$. All hospitals have the same reimbursement maximum, 100 thousand RMB yuan per year (except for $j = 0$ where the maximum is 0 and $j = 5$ with an infinite maximum). The exact deductibles and coinsurance rates are 400 yuan, 40%, for $j = 1$; 300 yuan, 30% for $j = 2$; 200 yuan, 20% for $j = 3$; 100 yuan, 10% for $j = 4$; 0 yuan, 0% for $j = 5$. The coinsurance rate for $j = 0$ is 100%. The graph is not to scale.

For tractability, we summarize our tiered hospital system under NRCMS into a list of four focal hospitals ($j = 1, ..., 4$) with the same reimbursement maximum (100 thousand RMB yuan per year). In addition, we consider a free hospital. Their deductibles are 0.4, 0.3, 0.2, 0.1, and 0.0 thousand RMB yuan, while reimbursement rates are 60%, 70%, 80%, 90%, and 100% (full coverage), respectively. Moreover, $j = 0, 1, 2$ are county hospitals, while $j = 3, 4, 5$ are THCs. Figure 8 shows the OOP cost functions of these four focal hospitals, the null county hospital, and the counterfactual free THC.

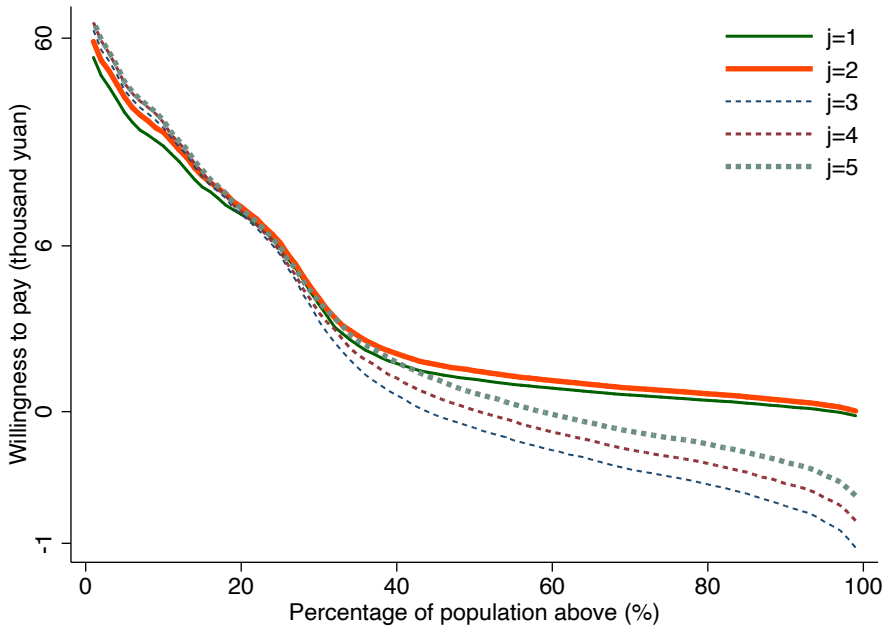


Figure 9: Marginal Willingness to Pay

Notes: This graph illustrates the distribution of the marginal willingness to pay across AECOPD patients in each hospital. It includes 5 connected scatter plots, with respect to 99 percentiles of individuals ordered by the willingness-to-pay value. They are marginal with respect to a non-contracted county hospital with the same features $j = 0$ as the reference point. The vertical axis is on a log scale.

We present the distributions of willingness to pay among these AECOPD patients in Figure 9. We sort patients according to their values of willingness to pay on the horizontal axis, and those with lower willingness to pay are on the right as in a demand curve. Some patients, especially those at the lower end of the willingness-to-pay distribution, seem to perceive generosity as an indicator of lower quality more than a financial risk protection, and thus are not willing to visit a more generous hospital (unless being compensated).[29] In order to encourage 99 percent of the population to go to THC $j = 3$ with a deductible of 200 RMB yuan and a coinsurance rate

[29]We notice a positive correlation between the reimbursement rate and total number of visits during 2012–2014 in our data. It seems to suggest that patients who visit more generous hospitals also tend to get hospitalized more frequently (due to less efficient treatments). This makes the association between low quality and high reimbursement one of the plausible explanations for "risk taking".

of 20 percent instead of a non-contracted county hospital ($j = 0$) holding other characteristics fixed, a travel subsidy that is worth 1 thousand RMB yuan should be given (or the THC needs to be 45 km closer). On the other hand, the patients with the top 1 percent willingness to pay are willing to pay 60 thousand RMB yuan for the full coverage in a THC (or to travel about 2.7 thousand km). Clearly, the range of willingness to pay is wide among these AECOPD patients. Slightly more than half of them prefer county hospitals ($j = 1, 2$) to THCs ($j = 3, 4, 5$) holding other characteristics fixed; interestingly, these are also the people with lower willingness to pay for any coverage, suggesting a tendency to bypass primary care. Some of them spent quite little (see Figure A6), suggesting common diseases or minor illnesses.

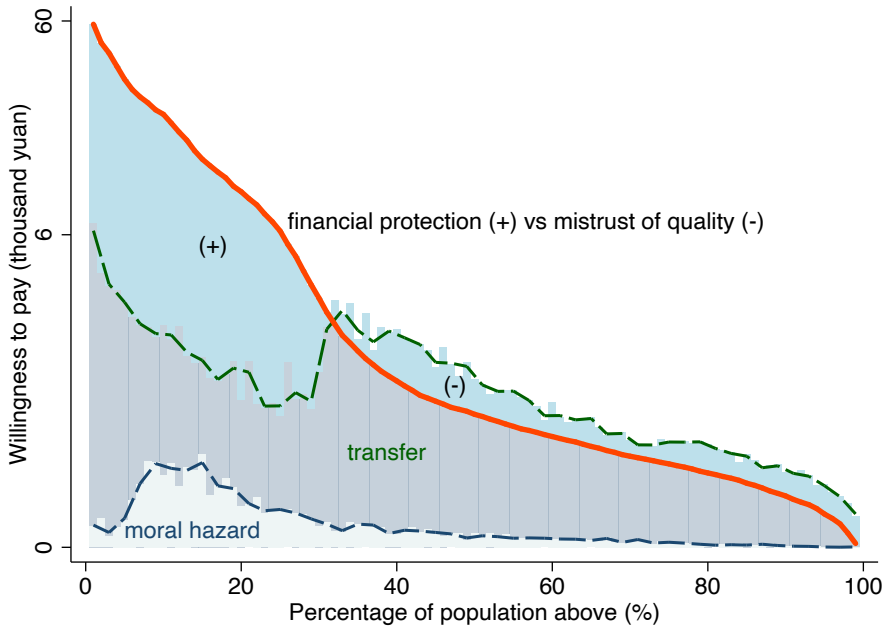


Figure 10: Decompose Marginal Willingness to Pay

Notes: This graph illustrates the distribution of the decomposition of willingness to pay across AECOPD patients in hospital $j = 2$. The willingness to pay is marginal with respect to a non-contracted county hospital $j = 0$ with the same features as the reference point. The vertical axis is on a log scale.

We further decompose the marginal WTP for $j = 2$ as Figure 10 shows. Note that, since both $j = 0$ and $j = 2$ are county hospitals, the "tier change value" is zero (assuming that all non-financial characteristics are the same) and thus we are left

with three components. As we can see, for most (more than 60%) of the AECOPD patients, willingness to pay mainly comes from the "transfer" term, and the net payoff from moral hazard spending only represents a very small portion of the willingness to pay, while mistrust of quality can lead to a lower willingness to pay (by -0.1 to -1 thousand RMB yuan), perhaps due to subjective perceptions about how more generous hospitals may not handle their health risks as efficient. For those with high (top 35%) willingness to pay, the value of financial protection finally outweighs the mistrust (of quality associated with generosity) and explains most of the willingness to pay, although "transfer" and the net payoff from moral hazard spending are also relatively high compared to those with low willingness to pay. At the top 1% percentile of the willingness-to-pay distribution, in addition to paying 6 thousand RMB yuan (for transportation) to avoid paying nearly 6 thousand RMB in expected OOP costs, patients are also willing to pay an additional 50–55 thousand RMB yuan to reduce financial uncertainty by 70% (i.e., the reimbursement rate in $j = 2$). This suggests that social surplus could be improved by allocating more of these patients with high valuation of financial risk protection to hospitals with higher reimbursement rates.

It's important to recall that, we determine patients' privately optimal choices given transportation subsidies/costs here, while these choices may not be socially optimal. To discuss socially optimal choices, we shall calculate the social surplus generated by allocating a patient to a given hospital based on Section 3.1.2.

5.4 Social Surplus

We can now calculate the social surplus $SS_{ij} = WTP_{ij} - \overline{k}_{ij}$, where $\overline{k}_{ij}$ is the expected insurer or government cost with respect to the distribution of λ_i, for every AECOPD patient covered in the previous subsection.

According to Equation (5), we may decompose SS into two parts, as illustrated by Figure A4. From the figure, we can see that the social cost of moral hazard is relatively small compared to willingness to pay especially for those with very high willingness to pay. As a result, social welfare gains from more generous hospitals are mainly driven by patients with the highest willingness to pay. This is driven by the shape of risk

attitude (see Figure A5) as well as the shape of risk itself.[30]

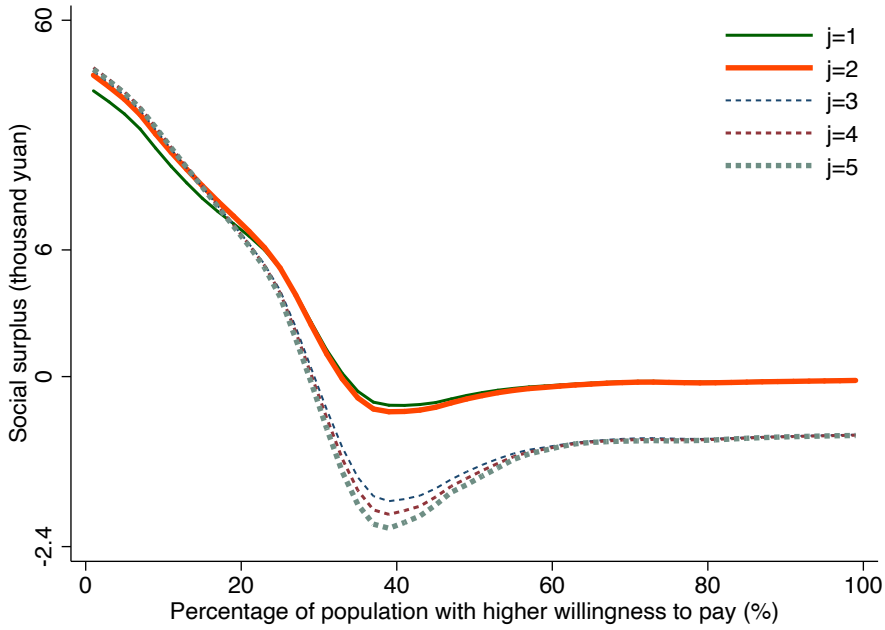


Figure 11: Social Surplus

Notes: This graph illustrates the distribution of the social surplus across AECOPD patients in each contracted hospital relative to the non-contracted county hospital ($j = 0$). It includes 5 local polynomial smoothed lines based on 99 percentiles of individuals ordered by the willingness-to-pay value. The vertical axis is on a log scale.

Eventually, we show the marginal social surplus generated by allocating patients to each hospital relative to the non-contracted (null) county hospital in Figure 11, by subtracting Figure A4b from Figure A4a. Since patients can be screened by their willingness to pay, this is relevant for the optimal design of a health insurance program.

As we can see, for 65–70% of the population, social surplus curves for all contracted hospitals lay below zero, indicating that a non-contracted county hospital is the best

[30]On the one hand, patients with high willingness to pay are typically more risk-averse and thus value financial risk protection more. On the other hand, patients with high willingness to pay tend to have poorer expected health, and thus are more likely to realize health states above the reimbursement maximum, leaving them the largest uncertainty about OOP costs.

hospital (from a social welfare perspective) when willingness to pay is low. This is because cost transfer at the lower end does not generate enough willingness to pay due to mistrust of quality. We can also find that, none of these hospitals are strictly the best. That is, the upper envelope of these social surplus curves is composed of multiple hospitals. At low levels of willingness to pay, the county hospital with the least generosity ($j = 0$) is the best (from a social welfare perspective); as willingness to pay increases, the more generous county hospitals ($j = 1$ and then $j = 2$) become the best; at very high levels of willingness to pay, the more generous THCs ($j = 4$ and $j = 5$) become the best. Clearly, cost transfer is beneficial to the society only when consumers value it enough, and vertical differentiation is necessary for maximizing social welfare.

Nevertheless, the socially efficient hospital is an average or overall concept and is not necessarily the best for every patient. From Figure 9, we can notice that it is not even the best for an average patient sometimes.[31] To further investigate the heterogeneity in privately versus socially optimal hospitals across patients, Figure A7 shows the distribution of efficient hospitals at every percentile of the average willingness-to-pay distribution. On average, when we assume zero additional (e.g., transportation) subsidies or costs, the non-contracted county hospital ($j = 0$) is only privately efficient for 0.8% of AECOPD patients, but is socially efficient for 49.0% of them; the most generous county hospital ($j = 2$) is privately efficient for 78.3% of them, but is only socially efficient for 25.7% of them; the THC with full insurance ($j = 5$) is privately efficient for 20.9% of them, but is only socially efficient for 6.7% of them; the less generous county hospital and THCs ($j = 1, 3, 4$) are never privately efficient, but are socially efficient for 13.2%, 1.4%, and 4.0% of these patients, respectively. Therefore, new policies can be designed to guide patients to choose the socially efficient hospitals over the privately efficient ones to achieve a larger social welfare target, which will be explored in the next section.

[31] For example, at low levels of willingness to pay, the average socially efficient hospital is the county hospital with the least generosity ($j = 0$), while the privately efficient hospital for an average patient is the most generous county hospital ($j = 2$).

6 Counterfactual Policies

The main objective of designing alternative (counterfactual) policies is to see how deductible and reimbursement maximum work and if there is a more socially efficient way to allocate patients to resources.[32] We consider three types of policies: (1) higher deductibles, (2) reimbursement cap adjustments, and (3) the combinations of the previous two.

6.1 High Deductibles

Based on the discussions in Section 5.3 and especially the evidence shown by Figure 10, we can see that, patients with low willingness to pay (who tend to be healthier) do not value the financial protection aspects in more generous hospitals as much as the disutility from their subjective perceptions about the low quality associated with price discounts. Then, based on Section 5.4, the more socially optimal allocation of resources would be to have these patients choose less generous hospitals (even the non-contracted ones). Under low deductibles, however, they start to receive price discounts too early, which does not generate more social welfare but leads them to have more moral hazard spending in more generous hospitals. This leads to a further social welfare loss. We may thus increase the deductibles and reserve the benefit of cost transfer to less "mistrustful" and more risk-averse patients with higher willingness to pay. By doing so, we improve the social welfare by the amount of disutility from mistrust plus the social cost of moral hazard. The question is, how high should deductibles be set at and should different hospitals with different reimbursement generosity also increase deductibles differently?

The NRCMS policy features low deductibles for all hospitals ($\leq$ 0.4 thousand RMB yuan), and lower deductibles for lower-tiered hospitals, but the differences between them are small in absolute value compared to the differences in average spending. We hence experiment with six alternative high-deductible policies: (i) increase the deductibles of all hospitals slightly (by 0.5 thousand RMB yuan); (ii) increase the

[32] We do not consider the capacity constraint of each hospital, and thus do not check any potential overcrowding issue.

deductibles of all hospitals moderately (by 1 thousand RMB yuan); (iii) increase the deductibles of all hospitals greatly (by 2 thousand RMB yuan); (iv) increase the deductibles and the gaps between hospitals slightly (multiply deductibles by 3); (v) increase the deductibles and the gaps moderately (multiply deductibles by 5); and (vi) increase the deductibles and the gaps greatly (multiply deductibles by 10).

Table 4: Outcomes of Alternative High-Deductible Policies

Policy	% of Current SS	% of Current WTP	% of County Hospital Visits	Average Insurer Cost
Current deductibles	100.00	100.00	56.53	1.827
(i) +0.5 thousand yuan	101.60	98.00	56.47	1.533
(ii) +1 thousand yuan	102.22	95.86	56.42	1.296
(iii) +2 thousand yuan	99.98	90.50	56.31	0.974
(iv) 3 times	101.02	97.88	57.05	1.563
(v) 5 times	101.74	95.93	57.57	1.336
(vi) 10 times	99.31	89.47	58.86	0.928

Notes: The table summarizes outcomes under the six high-deductible policies we consider as well as the current outcome, among the 79,531 individuals. Average insurer cost is in thousands of RMB yuan.

Table 4 provides outcomes under the current (original) deductibles, as well as those under each of the six alternative policies with higher deductibles. Due to mistrust of quality associated with price generosity at the lower end of the spending distribution, increasing the gaps may encourage more patients to switch to higher-tiered (county) hospitals as they become relatively more attractive. If we increase deductibles without increasing the gaps, those mistrustful patients may maintain their hospital choice while more risk-averse patients can switch to lower-tired hospitals, leading to slightly more patients (especially those with small spending) to switch from county hospitals to THCs. Among more risk-averse patients, higher deductibles reduce the willingness to pay as they take away the value of "transfer" and moral hazard spending, but this is partially compensated by higher reimbursement rates in THCs when they are encouraged to switch from county hospitals; on the other hand, "mistrust" associated with price discounts can be mitigated. Since patients tend to be more mistrustful than risk-averse at the lower end of the spending distribution, and the "transfer" term does not contribute to social surplus, the overall social welfare can be increased under

higher deductibles by the delayed exposure to mistrust and moral hazard. Of course, as we continue to increase deductibles, the loss of patient welfare will eventually outweigh the gain from insurer cost saving.

In our population, it seems that policy (ii), increasing the deductibles of all hospitals moderately (by 1 thousand RMB yuan) without increasing their gaps, is a more efficient and logical choice. It encourages more patients with low health risks/needs to choose lower-tiered hospitals and save the medical resources in higher-tiered hospitals to those with higher risks/needs, and at the same time increases social welfare and reduces insurer/government costs. Consumer surplus (the sum of willingness to pay) is slightly lower, but the allocation of resources becomes more efficient from the societal perspective.

6.2 Reimbursement Maximum Adjustments

From Sections 5.3 and 5.4, we learn that patients with very high willingness to pay tend to be quite risk-averse and thus value the financial protection aspects very much, while their degrees of moral hazard are modest. Adjusting the shape of risk/uncertainty for them could lead to efficiency gains. However, since there are both mistrustful and risk-averse patients at the higher end of spending distribution (implied by the non-monotonic trend of the willingness to pay explained by "transfer" in Figure 10, as well as Figure A6), and due to the fact that reimbursement maximum is working only on the higher end of the spending distribution, we may not have a definite answer to how we should adjust the cap. We experiment with fifteen alternative policies in Table 5.

The current reimbursement cap is set at 100 thousand RMB yuan per year for every hospital, which seems to be close to the optimal level. First, changing the caps within a certain range (e.g., -10 to 10 thousand RMB yuan) does not affect the average insurer cost significantly. This is partly because none of the patients in our data use up the reimbursement limit and most of them are quite far away from it. Second, changing the cap too much in either direction (e.g., $\pm$50 thousand RMB yuan) seems to affect both consumer welfare and social welfare negatively. Changing the maximum can alter the shape of risks facing patients and lead to redistribution

Table 5: Outcomes of Alternative Reimbursement Caps

Policy	% of Current SS	% of Current WTP	% of County Hospital Visits	Average Insurer Cost
Current cap	100.00	100.00	56.53	1.827
(i) No cap in THCs	98.50	98.81	56.61	1.828
(ii) No cap in county hospitals	99.63	99.72	56.49	1.828
(iii) No cap in all hospitals	98.84	99.10	56.51	1.829
(iv) +10k in THCs	99.25	99.42	56.62	1.829
(v) +10k in county hospitals	99.91	99.94	56.48	1.828
(vi) +10k in all hospitals	99.97	100.00	56.53	1.830
(vii) +5k in THCs	99.28	99.43	56.60	1.828
(viii) +5k in county hospitals	100.04	100.03	56.48	1.828
(ix) +5k in all hospitals	100.09	100.09	56.53	1.829
(x) -5k in THCs	99.58	99.67	56.48	1.827
(xi) -5k in county hospitals	99.10	99.28	56.61	1.827
(xii) -10k in THCs	99.03	99.24	56.47	1.829
(xiii) -10k in county hospitals	98.60	98.86	56.63	1.827
(xiv) -50k in THCs	93.33	94.71	56.50	1.829
(xv) -50k in county hospitals	86.63	88.93	56.57	1.791

Notes: The table summarizes outcomes under the fifteen cap-adjustment policies we consider as well as the current outcome, among the 79,531 individuals. Average insurer cost is in thousands of RMB yuan.

of hospital choices and reevaluation of financial protection and mistrust. Since the relationship between spending and how patients value financial protection is neither linear nor monotonic, there could be a certain level of reimbursement that is socially optimal. Third, we find that increasing the maximum of all hospitals or just county hospitals slightly (by 5 thousand RMB yuan) can slightly improve both consumer welfare and social welfare—how much patients appreciate this financial protection aspect outweighs how much they are mistrustful of it. Fourth, it is interesting to note that, having the maximum in THCs higher than that in county hospitals can further reduce welfare. This is probably because it worsens mistrust of quality in THCs and reduces willingness to pay, and at the same time reallocates more patients

to less generous county hospitals in which the value of financial protection is lower and further reduces willingness to pay. Fifth, when we lower the maximum in county hospitals greatly (by 50 thousand RMB yuan), it starts to become binding for some patients, and insurer costs can be reduced. However, due to the large reduction in risk protection value and considerable increase in moral hazard spending (and social cost associated with it) from more risk-averse patients who switch to THCs,[33] both patient welfare and social welfare drop significantly, and the latter drops more.

Based on the above discussions, although the current reimbursement maximum is already close to the optimal level, increasing the maximum by a small amount seems to be a potential policy tool to reduce policy resistance and improve acceptance without large negative impacts.

6.3 Combination Policies

We have discussed high deductibles and reimbursement cap adjustments separately. There is a concern that when we implement two sets of policies together, unexpected effects could arise. In this section, we are particularly interested in compensating higher deductibles (+0.5 to +1 thousand RMB yuan) by higher reimbursement caps (+5 thousand to unlimited). Table 6 lists the outcomes under these combination policies.

First, there seems to be a "synergy" effect. For example, increasing deductibles in all hospitals by 1 thousand RMB yuan alone can lead to a 2.22% increase in social welfare as shown by Table 4, and raising the reimbursement maximum in all hospitals by 5 thousand RMB yuan alone can lead to a 0.09% increase in social welfare as shown by Table 5; however, if we combine these two policies, the social welfare increase is 2.37%, which is larger than 2.22% + 0.09% = 2.31%. Similar agglomeration effects can be found in other combination policies in Table 6.

Second, this table shows that, the positive effects of high deductibles can outweigh the negative impacts of completely removing the reimbursement caps (i.e., allowing unlimited reimbursement). Thus, there is plenty of wiggle room for reimbursement

[33] This cannot be fully compensated by the reduction of mistrust disutility among mistrustful patients who switch to county hospitals.

Table 6: Outcomes of Combination Policies

Policy	% of Current SS	% of Current WTP	% of County Hospital Visits	Average Insurer Cost
Current policy	100.00	100.00	56.53	1.827
(i) deductibles +1k & caps +5k	102.37	96.00	56.42	1.297
(ii) deductibles +1k & caps +10k	102.23	95.89	56.42	1.297
(iii) deductibles +1k & no cap	101.18	95.05	56.41	1.297
(iv) deductibles +0.5k & caps +5k	101.74	98.13	56.48	1.534
(v) deductibles +0.5k & caps +10k	101.67	98.08	56.48	1.535
(vi) deductibles +0.5k & no cap	100.64	97.26	56.46	1.535

Notes: The table summarizes outcomes under the six combination policies we consider as well as the current outcome, among the 79,531 individuals. Average insurer cost is in thousands of RMB yuan.

maximum adjustments if policymakers intend to reduce resistance of high-deductible policies by a higher reimbursement limit.

7 Concluding Remarks

This paper takes an initiative to understand how deductible and reimbursement cap work and explore how patients can be incentivized to make more socially optimal choices of hospital and spending in a free-access tiered medical system. We utilize a framework with multi-dimensional consumer heterogeneity, hospital menus that feature nonlinear pricing schemes, and endogenous health care utilization through moral hazard. We distinguish the components of willingness to pay that generate social surplus from those affecting only allocations and thus only redistributive. We present the difficulty of aligning the social incentive to mitigate residual uncertainty and the private incentive to maximize transfer, due to mistrust as well as moral hazard.

There is rich variability in consumer preferences, and vertical differentiation is needed to improve allocative efficiency of medical resources in our context. Patients with

lower willingness to pay tend to be mistrustful of quality associated with generosity and thus a lower coverage should be offered; on the other hand, high willingness-to-pay patients value financial protection enough to make a higher coverage efficient. The current policy, nevertheless, assigns lower coverage in higher-tiered hospitals, which further encourages patients with common diseases and minor illnesses to bypass primary care, as they tend to have lower willingness to pay. We propose to delay exposure to cost sharing by introducing higher deductibles, to mitigate the negative impact of mistrust, encourage primary care, and save insurer cost. Our counterfactual analysis suggests that a moderate increase of the deductibles in all hospitals (by 1 thousand RMB yuan) can achieve a 2-percentage point increase in social welfare, and significantly lower insurer cost by almost 30 percentage points (from 1.8 to 1.3 thousand RMB yuan). Patient welfare is lower due to having to pay more out-of-pocket, and thus policymakers may need to consider compensating tools to improve policy acceptance among patients. The compensating tool we consider is an increase of reimbursement limit. We find that there is plenty of leeway. Since moral hazard is modest compared to how much patients value financial protection at the higher end of the spending distribution and the reimbursement cap is not binding for most patients, removing the maximum (allowing unlimited reimbursement) would not completely take away the efficiency improvement from moderately higher deductibles.

It is important to be mindful that there are a few limitations that need be taken into consideration when interpreting the above conclusions. First, since we only observe patients who make a visit to a hospital (either a THC or a county hospital) within our study area, we do not model how patients decide whether to go to a hospital to treat their diseases when needed.[34] Thus, our counterfactual policies do not measure the welfare loss of patients when they are discouraged from seeking health care. In this sense, the welfare gains due to cost saving by high-deductible policies mainly reflect higher-value choices made by patients, rather than reduced needed care,[35] by assuming that they would continue to seek health care. Second, we do not consider

[34] We also do not model how patients decide whether to travel to hospitals outside the study area. However, those cases are rare, and they are most likely to pay the full cost themselves when they do so.

[35] Of course, they also reflect reduced unnecessarily care included by moral hazard, but this tends to be negligible at the lower end of the spending distribution.

protection by limited liability such as bankruptcy protection (Gross and Notowidigdo, 2011) and liquidity constraints (Ericson and Sydnor, 2018), which could potentially affect the shape of risks facing our patients. It would be interesting to explore how these distortions can affect consumer behaviors and our conclusions in future work. Third, we do not consider externalities of health care utilization, such as crowding out because of limited capability, by assuming that the socially optimal level is the one chosen by patients without insurance. If there are positive externalities, the socially desirable level could include some additional health utilization induced by insurance. It could be challenging to evaluate externalities and determine the truly socially optimal level of health care utilization, but it should be considered a direction of future research. Fourth, to simplify our estimation of moral hazard, we assume health care to be a homogenous good conditional on the hospital chosen. However, the reality can be multidimensional and complex, and it could be important to extend our parsimonious model to capture more behavioral characteristics as a next step. Last, future research should try to separate mistrust of quality from risk aversion when studying consumers' health-care provider decisions.

References

Rajender Agarwal, Olena Mazurenko, and Nir Menachemi. High-deductible health plans reduce health care cost and utilization, including use of needed preventive services. *Health Affairs*, 36(10):1762–1768, 2017.

Rajender Agarwal, Ashutosh Gupta, and A. Mark Fendrick. Value-based insurance design improves medication adherence without an increase in total health care spending. *Health Affairs*, 37(7):1057–1064, 2018.

George A. Akerlof and Robert J. Shiller. *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton University Press, 2015.

Aviva Aron-Dine, Liran Einav, Amy Finkelstein, and Mark Cullen. Moral hazard in health insurance: do dynamic incentives matter? *Review of Economics and Statistics*, 97(4):725–741, 2015.

Daniel Avdic, Giuseppe Moscelli, Adam Pilny, and Ieva Sriubaite. Subjective and objective quality and choice of hospital: Evidence from maternal care services in Germany. *Journal of Health Economics*, 68:102229, 2019.

Eduardo M. Azevedo and Daniel Gottlieb. Perfect competition in markets with adverse selection. *Econometrica*, 85(1):67–105, 2017.

Chong-En Bai and Binzhen Wu. Health insurance and consumption: Evidence from China's New Cooperative Medical Scheme. *Journal of Comparative Economics*, 42 (2):450–469, 2014.

Patrick Bajari, Christina Dalton, Han Hong, and Ahmed Khwaja. Moral hazard, adverse selection, and health expenditures: A semiparametric analysis. *RAND Journal of Economics*, 45(4):747–763, 2014.

Laurence C. Baker, M. Kate Bundorf, and Daniel P. Kessler. The effect of hospital/physician integration on hospital choice. *Journal of Health Economics*, 50:1–8, 2016.

Gloria J. Bazzoli, Richard C. Lindrooth, Romana Hasnain-Wynia, and Jack Needleman. The Balanced Budget Act of 1997 and US hospital operations. *INQUIRY:*

The Journal of Health Care Organization, Provision, and Financing, 41(4):401–417, 2004.

Zarek C. Brot-Goldberg, Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad. 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, 2017.

Philip H. Brown and Caroline Theoharides. Health-seeking behavior and hospital choice in China's New Cooperative Medical System. *Health Economics*, 18(S2): S47–S64, 2009.

M. Kate Bundorf. Consumer-directed health plans: A review of the evidence. *Journal of Risk and Insurance*, 83(1):9–41, 2016.

M. Kate Bundorf, Jonathan Levin, and Neale Mahoney. Pricing and welfare in health plan choice. *American Economic Review*, 102(7):3214–3248, 2012.

Lawton R. Burns and Douglas R. Wholey. The impact of physician characteristics in conditional choice models for hospital care. *Journal of Health Economics*, 11(1): 43–62, 1992.

James H. Cardon and Igal Hendel. Asymmetric informationin health insurance: evidencefrom the National Medical Expenditure Survey. *RAND Journal of Economics*, 32(3):408–427, 2001.

Caroline Carlin and Robert Town. Adverse selection, welfare, and optimal pricing of employer sponsored health plans. Working Paper, University of Minnesota, 2009.

Fredrik Carlsson, Haoran He, Peter Martinsson, Ping Qin, and Matthias Sutter. Household decision making in rural China: Using experiments to estimate the influences of spouses. *Journal of Economic Behavior & Organization*, 84(2):525–536, 2012.

Elizabeth A. Carter, Pamela E. Morin, and Keith D. Lind. Costs and trends in utilization of low-value services among older adults with commercial insurance or Medicare Advantage. *Medical Care*, 55(11):931–939, 2017.

Yi Chen, Julie Shi, and Castiel Chen Zhuang. Income-dependent impacts of health insurance on medical expenditures: Theory and evidence from China. *China Economic Review*, 53:290–310, 2019.

China Health and Family Planning Statistical Yearbook. *[In Chinese]*. China Union Medical College Press, 2015.

Carrie H. Colla, Nancy E. Morden, Thomas D. Sequist, William L. Schpero, and Meredith B. Rosenthal. Choosing wisely: prevalence and correlates of low-value health care services in the United States. *Journal of General Internal Medicine*, 30 (2):221–228, 2015.

David M. Cutler and Richard J. Zeckhauser. The anatomy of health insurance. In *Handbook of Health Economics*, volume 1, pages 563–643. Elsevier, 2000.

Mark Duggan, Patrick Healy, and Fiona Scott Morton. Providing prescription drug coverage to the elderly: America's experiment with Medicare Part D. *Journal of Economic Perspectives*, 22(4):69–92, 2008.

Liran Einav and Amy Finkelstein. Moral hazard in health insurance: What we know and how we know it. *Journal of the European Economic Association*, 16(4):957–982, 2018.

Liran Einav, Amy Finkelstein, and Jonathan Levin. Beyond testing: Empirical models of insurance markets. *Annual Review of Economics*, 2:311–336, 2010.

Liran Einav, Amy Finkelstein, Stephen P. Ryan, Paul Schrimpf, and Mark R. Cullen. Selection on moral hazard in health insurance. *American Economic Review*, 103 (1):178–219, 2013.

Liran Einav, Amy Finkelstein, and Paul Schrimpf. The response of drug expenditure to nonlinear contract design: Evidence from Medicare Part D. *Quarterly Journal of Economics*, 130(2):841–899, 2015.

Liran Einav, Amy Finkelstein, and Paul Schrimpf. Bunching at the kink: implications for spending responses to health insurance contracts. *Journal of Public Economics*, 146:27–40, 2017.

Keith Marzilli Ericson and Justin R. Sydnor. Liquidity constraints and the value of insurance. Working Paper 24993, NBER, 2018.

José J. Escarce and Kanika Kapur. Do patients bypass rural hospitals?: Determinants of inpatient hospital choice in rural California. *Journal of Health Care for the Poor and Underserved*, 20(3):625–644, 2009.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. The Oregon health insurance experiment: evidence from the first year. *Quarterly Journal of Economics*, 127(3):1057–1106, 2012.

Amy Finkelstein, Nathaniel Hendren, and Erzo F. P. Luttmer. The value of Medicaid: Interpreting results from the Oregon Health Insurance Experiment. *Journal of Political Economy*, 127(6):2836–2874, 2019.

Eric B. French, Jeremy McCauley, Maria Aragon, Pieter Bakx, Martin Chalkley, Stacey H. Chen, Bent J. Christensen, Hongwei Chuang, Aurelie Côté-Sergent, Mariacristina De Nardi, Elliott Fan, Damien Échevin, Pierre-Yves Geoffard, Christelle Gastaldi-Ménager, Mette Gørtz, Yoko Ibuka, John B. Jones, Malene Kallestrup-Lamb, Martin Karlsson, Tobias J. Klein, Grégoire de Lagasnerie, Pierre-Carl Michaud, Owen O'Donnell, Nigel Rice, Jonathan S. Skinner, Eddy van Doorslaer, Nicolas R. Ziebarth, and Elaine Kelly. End-of-life medical spending in last twelve months of life is lower than previously reported. *Health Affairs*, 36(7):1211–1217, 2017.

Hongqiao Fu, Ling Li, and Winnie Yip. Intended and unintended impacts of price changes for drugs and medical services: evidence from China. *Social Science & Medicine*, 211:114–122, 2018.

Martin Gaynor and William B. Vogt. Competition among hospitals. *RAND Journal of Economics*, 34(4):764–785, 2003.

Michael Geruso. Demand heterogeneity in insurance markets: Implications for equity and efficiency. *Quantitative Economics*, 8(3):929–975, 2017.

H. Gothe, I. Schall, K. Saverno, M. Mitrovic, A. Luzak, D. Brixner, and U. Siebert. The impact of generic substitution on health and economic outcomes: a systematic review. *Applied Health Economics and Health Policy*, 13(1):21–33, 2015.

Tal Gross and Matthew J. Notowidigdo. Health insurance and the consumer bankruptcy decision: Evidence from expansions of Medicaid. *Journal of Public Economics*, 95(7–8):767–778, 2011.

Benjamin R. Handel. Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, 103(7):2643–2682, 2013.

Benjamin R. Handel and Jonathan T. Kolstad. Health insurance for "humans": Information frictions, plan choice and consumer welfare. *American Economic Review*, 105(8):2449–2500, August 2015. doi: 10.1257/aer.20131126. URL https://www.aeaweb.org/articles?id=10.1257/aer.20131126.

Daniel M. Hartung, Matthew J. Carlson, Dale F. Kraemer, Dean G. Haxby, Kathy L. Ketchum, and Merwyn R. Greenlick. Impact of a Medicaid copayment policy on prescription drug and health services utilization in a fee-for-service Medicaid population. *Medical Care*, 46(6):565–572, 2008.

Kate Ho and Robin Lee. Health insurance menu design: Managing the spending coverage tradeoff. In Presentation. Conference Celebrating the Scholarly Career of Mark Satterthwaite, Kellogg School of Management, 2019.

Katherine Ho and Ariel Pakes. Do physician incentives affect hospital choice?: A progress report. *International Journal of Industrial Organization*, 29(3):317–322, 2011.

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

Katherine Ho and Ariel Pakes. Physician payment reform and hospital referrals. *American Economic Review*, 104(5):200–205, 2014b.

Peter J. Huckfeldt, Neeraj Sood, José J. Escarce, David C. Grabowski, and Joseph P. Newhouse. Effects of Medicare payment reform: Evidence from the home health

interim and prospective payment systems. *Journal of Health Economics*, 34:1–18, 2014.

Paul Hudson, W. J. Wouter Botzen, Jeffrey Czajkowski, and Heidi Kreibich. Moral hazard in natural disaster insurance markets: empirical evidence from Germany and the United States. *Land Economics*, 93(2):179–208, 2017.

Toshiaki Iizuka. Experts' agency problems: evidence from the prescription drug market in Japan. *RAND Journal of Economics*, 38(3):844–862, 2007.

Jianjun Jin, Rui He, Haozhou Gong, Xia Xu, and Chunyang He. Farmers' risk preferences in rural China: Measurements and determinants. *International Journal of Environmental Research and Public Health*, 14(7):713, 2017.

Geoffrey F. Joyce, José J. Escarce, Matthew D. Solomon, and Dana P. Goldman. Employer drug benefit plans and spending on prescription drugs. *JAMA*, 288(14): 1733–1739, 2002.

Ulrich Kaiser, Susan J. Mendez, Thomas Rønde, and Hannes Ullrich. Regulation of pharmaceutical prices: evidence from a reference price reform in Denmark. *Journal of Health Economics*, 36:174–187, 2014.

Michael Keane and Olena Stavrunova. Adverse selection, moral hazard and the demand for Medigap insurance. *Journal of Econometrics*, 190(1):62–78, 2016.

Amanda E. Kowalski. Estimating the tradeoff between risk protection and moral hazard with a nonlinear budget set model of health insurance. *International Journal of Industrial Organization*, 43:122–135, 2015.

Annette M. Langer-Gould, Wayne E. Anderson, Melissa J. Armstrong, Adam B. Cohen, Matthew A. Eccher, Donald J. Iverson, Sonja B. Potrebic, Amanda Becker, Rod Larson, Alicia Gedan, Thomas S.D. Getchius, and Gary S. Gronseth. The American Academy of Neurology's top five choosing wisely recommendations. *Neurology*, 81(11):1004–1011, 2013.

Christy Harris Lemak, Tammie A. Nahra, Genna R. Cohen, Natalie D. Erb, Michael L. Paustian, David Share, and Richard A. Hirth. Michigan's fee-for-value

physician incentive program reduces spending and improves quality in primary care. *Health Affairs*, 34(4):645–652, 2015.

Li-Lin Liang, Nicole Huang, Yi-Jung Shen, Annie Yu-An Chen, and Yiing-Jenq Chou. Do patients bypass primary care for common health problems under a free-access system? experience of Taiwan. *BMC Health Services Research*, 20(1):1050, 2020a.

Lirong Liang, Yunxiao Shang, Wuxiang Xie, Julie Shi, Zhaohui Tong, and Mohammad S. Jalali. Trends in hospitalization expenditures for acute exacerbations of COPD in Beijing from 2009 to 2017. *International Journal of Chronic Obstructive Pulmonary Disease*, 15:1165, 2020b.

Yun Liu, Qingxia Kong, Shasha Yuan, and Joris van de Klundert. Factors influencing choice of health system access level in China: a systematic review. *PLoS One*, 13 (8):e0201887, 2018.

Yi Lu, Julie Shi, and Wanyu Yang. Expenditure response to health insurance policies: Evidence from kinks in rural China. *Journal of Public Economics*, 12(4):104049, 2019.

Qinli Ma, Gosia Sylwestrzak, Manish Oza, Lorraine Garneau, and DeVries andrea R. Evaluation of value-based insurance design for primary care. *The American Journal of Managed Care*, 25(5):221–227, 2019.

Henry Y. Mak. Managing imperfect competition by pay for performance and reference pricing. *Journal of Health Economics*, 57:131–146, 2018.

Willard G. Manning and M. Susan Marquis. Health insurance: the tradeoff between risk pooling and moral hazard. *Journal of Health Economics*, 15(5):609–639, 1996.

Alice Mannocci, Gabriella De Carli, Virginia Di Bari, Rosella Saulle, Brigid Unim, Nicola Nicolotti, Lorenzo Carbonari, Vincenzo Puro, and Giuseppe La Torre. How much do needlestick injuries cost? A systematic review of the economic evaluations of needlestick and sharps injuries among healthcare personnel. *Infection Control & Hospital Epidemiology*, 37(6):635–646, 2016.

Olena Mazurenko, Melinda J.B. Buntin, and Nir Menachemi. High-deductible health plans and prevention. *Annual Review of Public Health*, 40:411–421, 2019.

Andrew W. Mulcahy, Jakub P. Hlávka, and Spencer R. Case. Biosimilar cost savings in the United States: Initial experience and future potential. *RAND Health Quarterly*, 7(4):3, 2018.

National Academies of Sciences, Engineering, and Medicine. *Making Medicines Affordable: A National Imperative*. National Academies Press, 2018.

Joseph P. Newhouse and the Insurance Experiment Group. *Free for All?: Lessons from the RAND Health Insurance Experiment*. Harvard University Press, 1993.

Xuan Nguyen Nguyen. Physician volume response to price controls. *Health Policy*, 35(2):189–204, 1996.

Mark V. Pauly. The economics of moral hazard: Comment. *American Economic Review*, 58(3):531–537, 1968.

Susan L. Perez, Melissa Gosdin, Jessie Kemmick Pintor, and Patrick S. Romano. Consumers' perceptions and choices related to three value-based insurance design approaches. *Health Affairs*, 38(3):456–463, 2019.

Rachel O. Reid, Brendan Rabideau, and Neeraj Sood. Low-value health care services in a commercially insured population. *JAMA Internal Medicine*, 176(10):1567–1571, 2016.

Chul-Young Roh, Keon-Hyung Lee, and Myron D. Fottler. Determinants of hospital choice of rural hospital patients: the impact of networks, service scopes, and market competition. *Journal of Medical Systems*, 32(4):343–353, 2008.

Scott R. Sanders, Lance D. Erickson, Vaughn RA Call, Matthew L. McKnight, and Dawson W. Hedges. Rural health care bypass behavior: how community and spatial characteristics affect primary health care selection. *Journal of Rural Health*, 31(2): 146–156, 2015.

Aaron L. Schwartz, Bruce E. Landon, Adam G. Elshaug, Michael E. Chernew, and J. Michael McWilliams. Measuring low-value care in Medicare. *JAMA Internal Medicine*, 174(7):1067–1076, 2014.

William H. Shrank, Teresa L. Rogstad, and Natasha Parekh. Waste in the us health care system: estimated costs and potential for savings. *JAMA*, 322(15):1501–1509, 2019.

Niek Stadhouders, Florien Kruse, Marit Tanke, Xander Koolman, and Patrick Jeurissen. Effective healthcare cost-containment policies: a systematic review. *Health Policy*, 123(1):71–79, 2019.

Amal N. Trivedi, Husein Moloo, and Vincent Mor. Increased ambulatory care copayments and hospitalizations among the elderly. *New England Journal of Medicine*, 362(4):320–328, 2010.

Jin Wang, Pan Wang, Xinghe Wang, Yingdong Zheng, and Yonghong Xiao. Use and prescription of antibiotics in primary health care settings in China. *JAMA Internal Medicine*, 174(12):1914–1920, 2014.

World Health Organization. *Global Spending on Health 2020: Weathering the Storm*. World Health Organization, 2020.

John E. Zeber, Kyle L. Grazier, Marcia Valenstein, Frederic C. Blow, and Paula M. Lantz. Effect of a medication copayment increase in veterans with schizophrenia. *American Journal of Managed Care*, 13(6):335, 2007.

Zhongliang Zhou, Yaxin Zhao, Chi Shen, Sha Lai, Rashed Nawaz, and Jianmin Gao. Evaluating the effect of hierarchical medical system on health seeking behavior: A difference-in-differences analysis in China. *Social Science & Medicine*, 268:113372, 2021.

Jingrong Zhu, Jinlin Li, Zengbo Zhang, Hao Li, and Lingfei Cai. Exploring determinants of health provider choice and heterogeneity in preference among outpatients in Beijing: A labelled discrete choice experiment. *BMJ Open*, 9(4):e023363, 2019.

Appendices

A Additional Materials

A.1 Estimation of Hospital Cost-Sharing Rules

The cost-sharing function of each hospital is a crucial input to our empirical model. Although we describe hospitals using only the deductibles and reimbursement rates, hospitals are characterized by a much more complex set of payment rules. To model moral hazard structurally, we assume that health care is a homogenous good over which a patient chooses only the quantity to consume in our parsimonious framework and model this decision as being based in part on out-of-pocket cost. A univariate

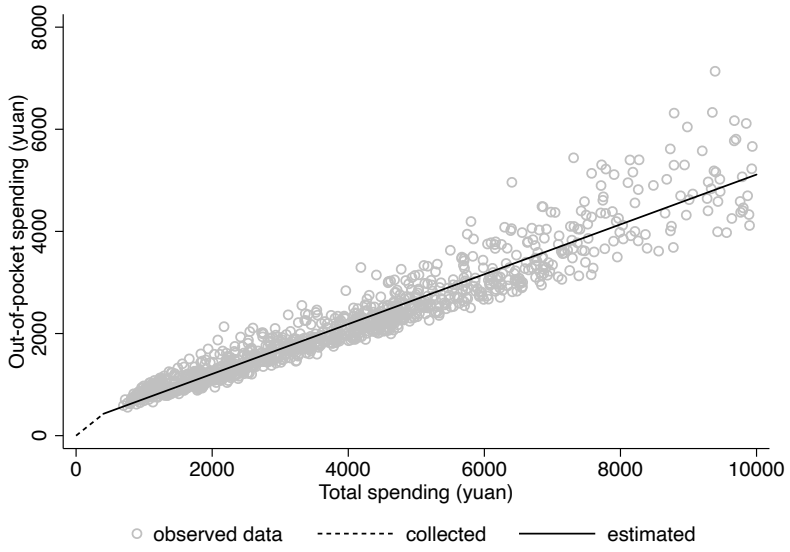


Figure A1: An Example of Hospital Cost-Sharing Rules Estimation

Notes: The plot shows the observed data (each dot represents a person-visit) used to estimate the cost-sharing rules for individuals who went to hospital 1 to treat diseases of the respiratory system (ICD 10: J00–J99) in 2014. For a better graphical illustration, we look at those who spent less than 10 thousand RMB yuan. The solid line depicts the estimated cost-sharing function of the hospital, minimizing the sum of squared errors between observed and predicted out-of-pocket spending. The estimated reimbursement rate is 48%, suggesting a coinsurance rate of 52%.

function that maps total spending into out-of-pocket cost is thus required as an input to our empirical model.

The out-of-pocket cost function in our application is defined by three parameters: a deductible, a reimbursement rate, and a reimbursement maximum. We take the true deductibles (mostly publicly available from each hospital) as given because they correspond very well to our observed data. As far as we learn from local officials, the reimbursement maximum is 100 thousand RMB yuan per patient-year[36] during our study period. Cases with an annual reimbursement of over 100 thousand RMB yuan do not occur in our data. Then, we are left with the reimbursement rate to estimate.

As shown by Figure A1, we can estimate the cost-sharing rules of each hospital in each year by disease category. For example, we estimated that the coinsurance rate for diseases of the respiratory system (ICD 10: J00–J99) is about 52% in hospital 1 in 2014, after paying 0.4 thousand RMB yuan as deductible. Since the reimbursement maximum is 100 thousand RMB yuan per year, patients in this hospital would have to face the full cost after spending more than 208 thousand RMB yuan per year.

A.2 Calculation of Individual Health Risk Predictors

The calculation of health risk predictors takes two steps. First, we resort to the Johns Hopkins ACG system (v. 12.1), which is widely applied in the literature such as Carlin and Town (2009), Handel (2013), Handel and Kolstad (2015), and Brot-Goldberg et al. (2017). By entering patient information, such as diagnosis (ICD 10 code), age, gender, the place of service (inpatient care), as well as the total spending in RMB yuan, into the software, we get the unscaled predicted total cost risk coefficient for everyone in each year (mean: 1.443; range: 0.000 to 14.861). Then, the rescaled risk score is obtained by dividing the unscaled predicted total cost risk coefficient by the mean.

Next, we adjust the risk score by running a linear regression. Before running the regression, we take a natural log of the rescaled risk score[37] to deal with its high

[36]Thus, the per patient-visit reimbursement maximum is 100 thousand RMB yuan minus the reimbursement amount accumulated from the previous visits within the year.

[37]To avoid the natural log of zero, we shift the risk coefficient by 0.05 first.

skewness. Then, we regress the natural log of actual total spending in thousands of RMB yuan on the log rescaled risk score, its interactions with each of the percentile indicators, the education level indicators, the indicator for a married person, and the hospital dummies, besides the integer age and gender indicators, and the ICD 10 code indicators. Finally, we predict the log spending using this linear model, and the predicted values are our re-scaled risk predictors (range: -2.526 to 4.896). The main reasons for this adjustment are two-fold. On the one hand, the Johns Hopkins ACG system is mainly based on the United States (although it has also been implemented internationally in the United Kingdom, Europe, Singapore, Vietnam, and Australia according to the sales staff), while our data is from rural China, and thus adjusting the risk coefficient may improve the accuracy of the cost prediction in our context, which can then improve our model fit. On the other hand, the risk coefficient from the ACG system does not contain information on a patient's educational attainment, marital status, and the hospital chosen; therefore, by running this additional regression, we can incorporate additional information that we expect to play a role in determining health status.

Table A1: Spending Distributions by Risk Quartile

	Percentile of total spending (in thousands of RMB yuan)						
Risk quartile	1st	10th	25th	50th	75th	90th	99th
1	0.168	0.276	0.431	0.597	0.751	0.884	1.339
2	0.585	0.834	0.985	1.184	1.438	1.686	2.658
3	1.022	1.635	1.956	2.419	3.025	3.554	5.109
4	1.940	3.814	4.413	5.708	8.286	15.208	42.007

Notes: This table is based on the estimation sample from 2012 to 2014, the same as the first column of Table 2.

The health risk predictors are different from log total spending, although they are highly (positively) correlated. To show how different but correlated they are, we summarize the total spending distributions by quartile of the risk predictor. As shown in Table A1, a patient in a higher risk quartile does not necessarily have a higher total spending than a patient in a lower risk quartile.

A.3 Variation in Hospital Menu Generosity

Hospital menu generosity is measured by the weighted average of the reimbursement rates in the hospitals available to each community each year. It is calculated for each disease, as the hospitals available for treating each disease can be different. The

Table A2: Hospital Menu Generosity and Individual Health

	All	2012	2013	2014
Individual Health				
Risk predictor	0.0002	0.0015	0.0011	-0.0017
	(0.0009)	(0.0010)	(0.0012)	(0.0014)
Community Characteristics				
Age 18–60	0.0016	-0.0013	0.0031*	0.0029**
	(0.0010)	(0.0011)	(0.0016)	(0.0012)
Male	-0.0006	0.0003	-0.0019	0.0000
	(0.0007)	(0.0007)	(0.0012)	(0.0010)
Years of schooling ≥ 9	0.0007	0.0015	0.0004	0.0001
	(0.0012)	(0.0013)	(0.0016)	(0.0014)
Married	0.0018***	0.0022**	0.0021*	0.0013
	(0.0007)	(0.0009)	(0.0011)	(0.0009)
Longitude	-0.1660***	-0.2135***	-0.1840***	-0.1082**
	(0.0405)	(0.0497)	(0.0477)	(0.0531)
Latitude	-0.1037***	-0.1482***	-0.0662*	-0.1094***
	(0.0301)	(0.0244)	(0.0381)	(0.0416)
Menu Characteristics				
Number of doctors	-0.0021***	-0.0020***	-0.0025***	-0.0023***
	(0.0005)	(0.0006)	(0.0007)	(0.0008)
Number of beds	0.0006***	0.0005**	0.0009***	0.0006
	(0.0002)	(0.0003)	(0.0003)	(0.0004)
Year fixed effects	Yes	No	No	No
Dependent variable's mean	0.7022	0.7627	0.6943	0.6517
R-squared	0.8430	0.5859	0.4535	0.4491
Number of observations	12,867	4,368	3,858	4,641

Notes: The dependent variable is hospital menu generosity, as measured by average reimbursement rate conditional on choosing a THC to treat AECOPD. Robust standard errors clustered at the community level are in parentheses; ***p<0.01, **p<0.05, *p<0.1.

weights are the proportion of patients going to each hospital from each community. By using the weights, we incorporate the likelihood that an individual would choose a generous hospital when presented with such a menu, as if the individual had been acting like the average individual in the community.

To investigate what explain the hospital menu generosity, we regress the average reimbursement rates on individual health risk predictors (calculated in Appendix A.2), community characteristics (such as age, gender, education, and marriage rate), and the menu characteristics (such as the average number of doctors/beds). All models in Table A2 fail to reject the null hypothesis that risk predictors are not correlated with the generosity of hospital menu, conditional on community and menu characteristics. Hospital menus are consistently more generous when there are fewer doctors available, and may be more generous in the southwest, or when there are more hospital beds. None of these relationships seem to be inconsistent with our understanding of how community benefits are decided. Nevertheless, there is no strong evidence that the communities try to set hospital generosity based on unobservable information that could drive inpatient-care spending.

B Additional Tables and Figures

Table A3: Descriptive Statistics for the Raw Full Sample

	N	Mean	SD	Min	Max
Patient level					
Male	66,298	0.434	0.496	0	1
Number of visits per patient	66,316	1.714	1.532	1	35
Patient-visit level					
Age	113,662	52.152	22.530	0	126
Years of schooling	100,519	5.270	3.736	0	18
Married	100,519	0.725	0.447	0	1
Proportion of county hospital visits	113,662	0.593	0.491	0	1
Total medical spending (thousand)	113,662	3.065	5.115	0.060	263.540
Deductible (thousand)	113,662	0.248	0.145	0	0.400
Reimbursement rate received	113,662	0.616	0.110	0.407	0.835

Notes: For patient level variables, N refers to the total number of patients; for patient-visit level variables, N refers to the total number of visits. Other variable details are the same as in Table 2.

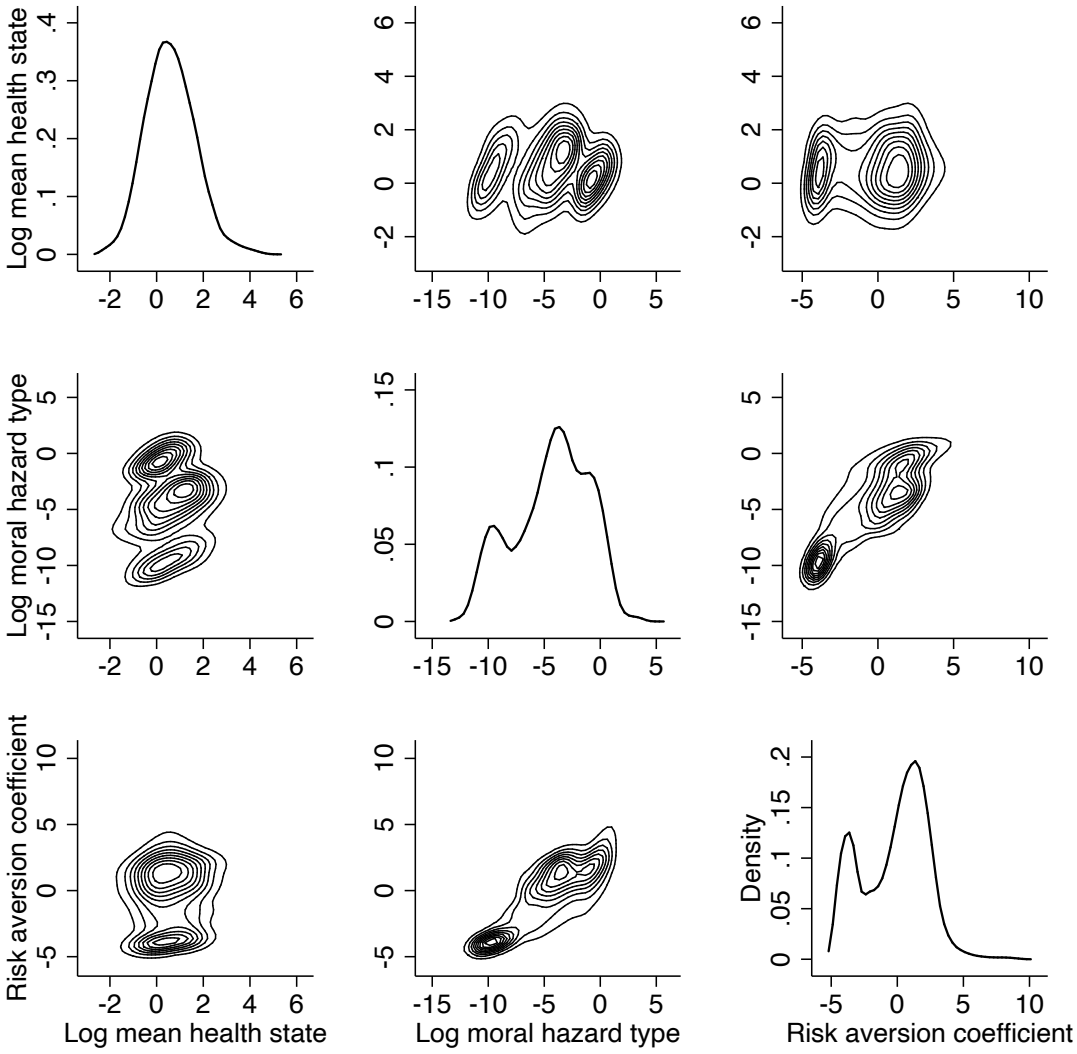


Figure A2: Joint Distribution of Individual Types

Notes: This figure presents the joint distribution of individual types implied by the estimates in Table 3. The diagonals are the one-way distributions of each parameter across individuals (with the vertical axis being the density), while the off-diagonals show bivariate distributions (with both axes being the values).

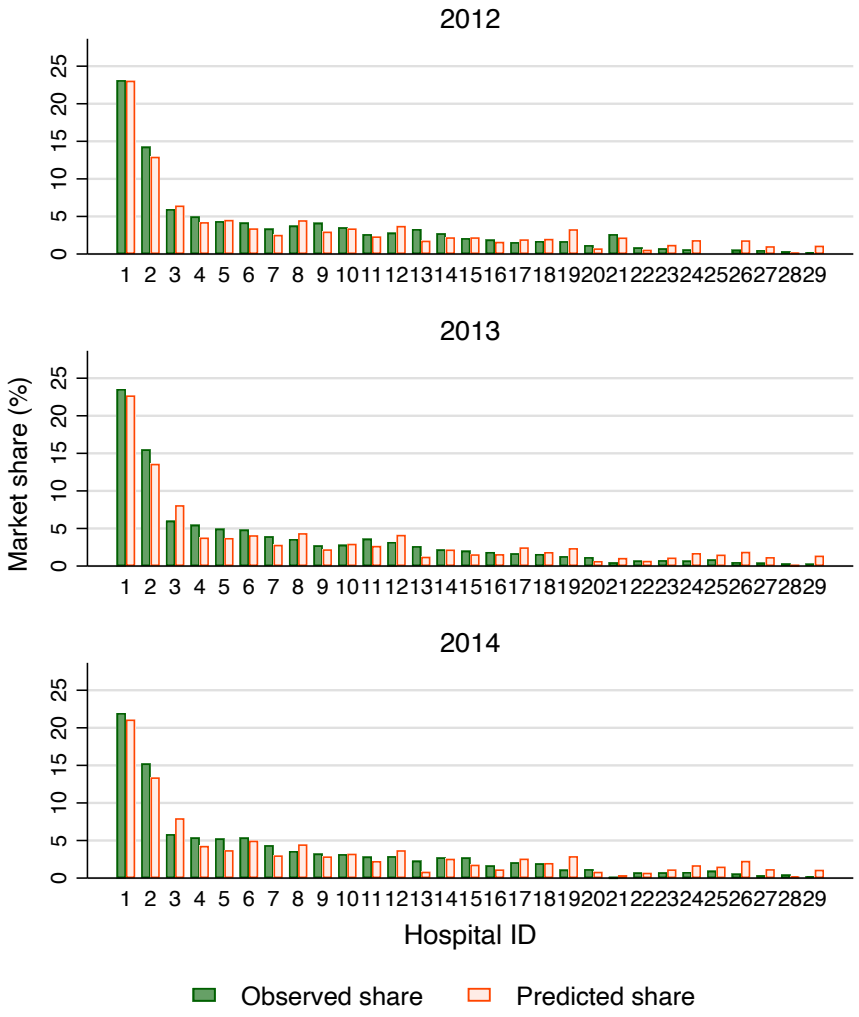
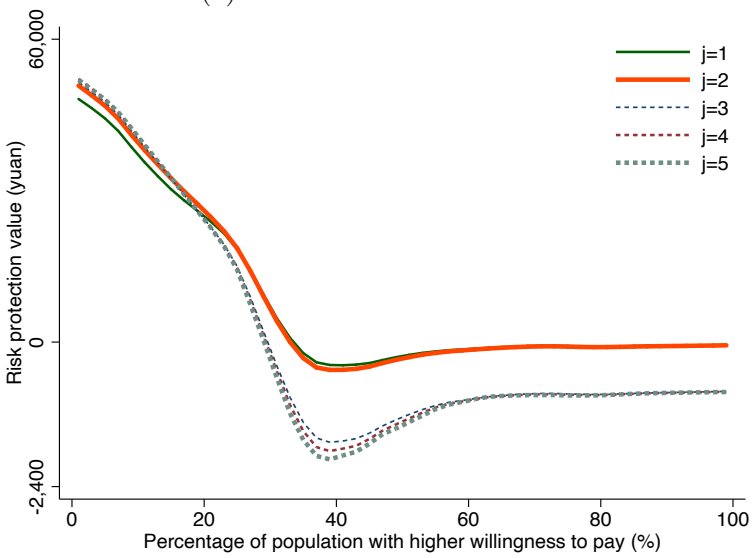


Figure A3: Model Fit: Hospital Choices

Notes: The figure shows the observed and predicted market shares at the hospital level. An observation is a person-visit in each year. Predicted shares are calculated based on Table 3.

(a) Value of Risk Protection



(b) Social Cost of Moral Hazard

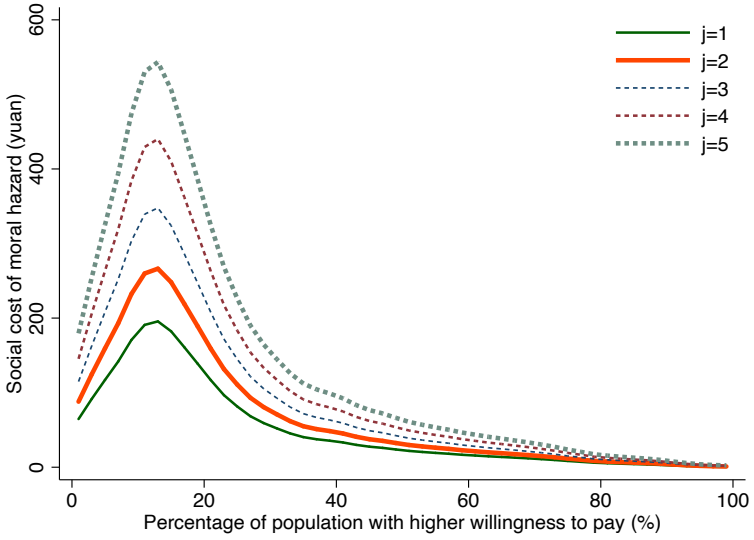


Figure A4: Decompose Social Surplus

Notes: The graph shows the distribution of (a) the value of risk protection and (b) the marginal social cost of moral hazard across AECOPD patients in each focal hospital, relative to the null county hospital ($j = 0$). Each panel includes 5 local polynomial smoothed lines based on 99 percentiles of individuals ordered by the willingness-to-pay value. The vertical axis of panel (a) is on a log scale.

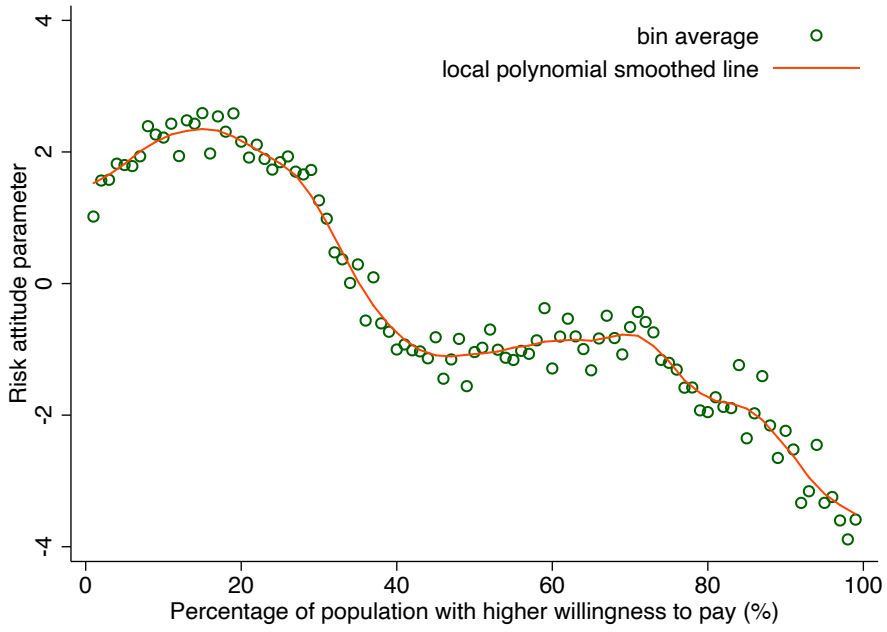


Figure A5: Risk Attitude Parameter by Willingness to Pay

Notes: This graph illustrates the distribution of the risk attitude parameter across AECOPD patients by willingness to pay. It consists of 99 binned scatters and a local polynomial smoothed line based on these scatters.

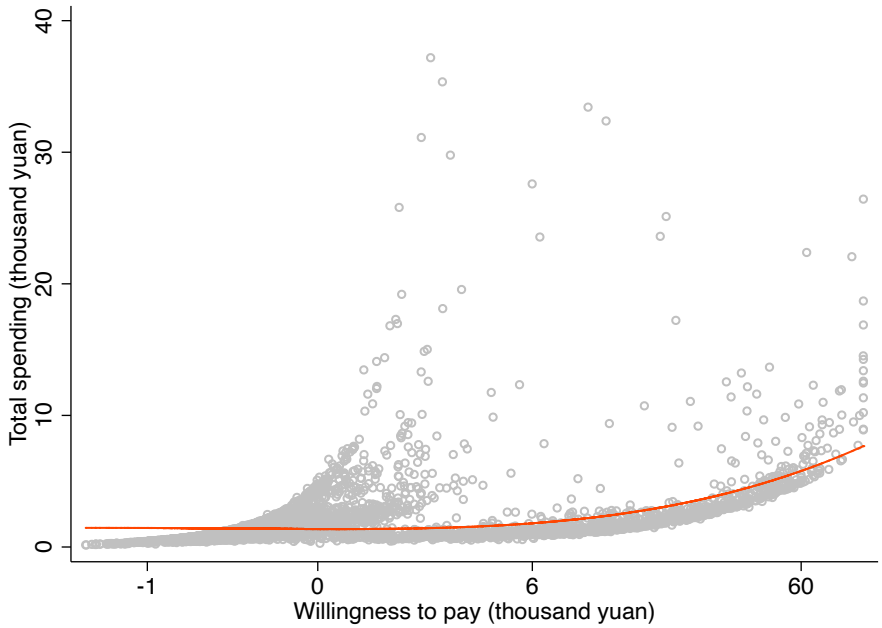
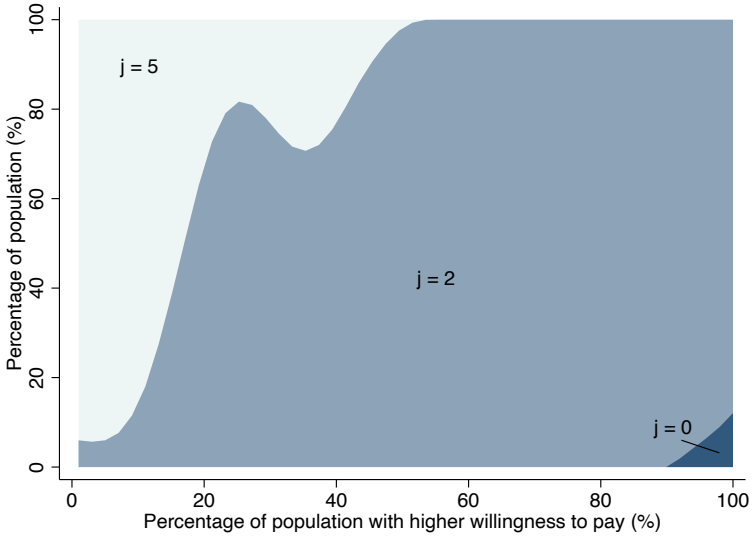


Figure A6: Willingness to Pay and Spending

Notes: This graph illustrates the relationship between total spending and marginal willingness to pay for transferring from a non-contracted county hospital ($j = 0$) to a contracted THC with low generosity ($j = 3$) among AECOPD patients. It consists of a fractional polynomial fit based on the scatters.

(a) Privately Efficient Hospitals



(b) Socially Efficient Hospitals

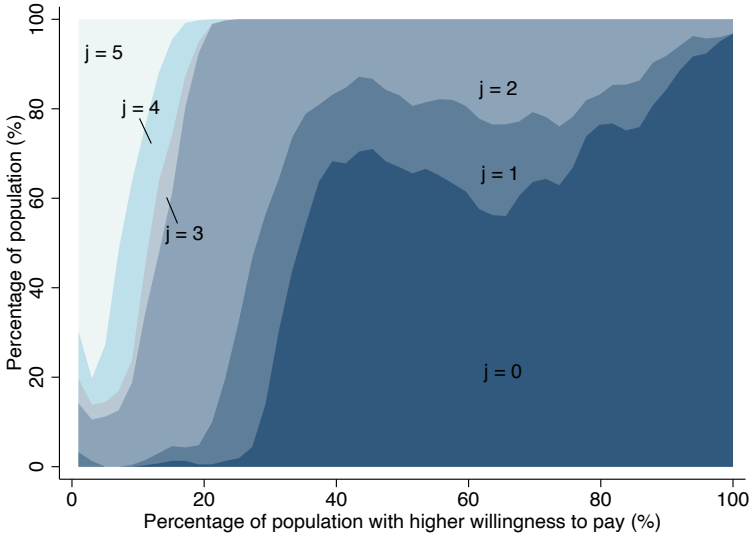


Figure A7: Efficient Hospital by Willingness to Pay

Notes: The graph shows the percentage of patients at each percentile of willingness to pay for whom each hospital is (a) privately optimal and (b) socially optimal, assuming that there are zero additional (e.g., transportation) costs. Each panel includes several local polynomial smoothed area plots based on 100 binned scatters.