Expenditure Response to Health Insurance Policies: Evidence from Kinks in Rural China

Yi Lu [*] Julie Shi [†] Wanyu Yang [‡]

Tsinghua and NUS Peking DUFE

This version: September 2018

Abstract

This paper utilizes administrative data to analyze expenditure responses to the health insurance policy in rural China, and clear visual evidence of bunching is observed at the kink point. A static response model with optimization frictions estimates that a complete elimination of the reimbursement would cause the total expenditure per visit to decrease by 34.5%, and approximately one third of the studied population make decisions with errors. Heterogeneous expenditure responses and optimization frictions are observed across demographic groups. Cost–benefit and counterfactual analyses indicate that the current policy generates the greatest welfare gains.

Keywords: bunching, health care, health insurance, optimization friction

JEL Classification: D12, G22, I13

[*]School of Economics and Management, Tsinghua University, Beijing, 100084, China; School of Economics, National University of Singapore, 117570, Singapore (luyi@sem.tsinghua.edu.cn)

[†]School of Economics, Peking University, Beijing, 100084, China (jshi@pku.edu.cn)

[‡]Department of Economics, Dongbei University of Finance and Economics, 116000, China (wanyu19@gmail.com)

1 Introduction

Universal health coverage (UHC) plays an important role in improving the well–being of a country's population, but is not widely provided. In 2006–2008, 58 out of 194 countries attained UHC (Stuckler et al., 2010). Although nearly all developed countries provide UHC, only a few developing countries have nationally implemented this plan (e.g., Cuba, Mongolia, Thailand). Many developing countries have instead managed to develop health insurance programs for formal sector workers and civil servants (e.g., Nigeria, China, Columbia, Indonesia, Mexico). As most rural residents work in informal sectors, and the low economic level in rural areas would repel nurses and doctors, the insurance coverage tends to be generally low in the rural areas of developing countries, for example, China, Georgia, and Rwanda. To manage such problems, some developing countries have started to develop rural insurance systems. Notably, how much rural residents value such insurance and the size of the welfare effect remain unclear.

This paper uses a bunching method [for a review, see Kleven (2016)] to estimate the elasticity of health expenditure in China, a large developing country that has only begun to provide social insurance to its massive rural population. Based on this estimate, we next conduct a cost–benefit analysis of the current policies. Finally, we examine several policy counterfactuals and provide novel information regarding the optimal design of health insurance policies.

Our analysis builds on a unique dataset drawn from China, which contains the medical claims information for each medical visit by all rural residents in a southwestern county. The collapse of China's old health insurance program for rural residents after the economic reform in 1978 resulted in a majority of rural residents remaining uninsured. To increase access to medical services and reduce the burden of out–of–pocket spending, the Chinese government launched a new health insurance program for its rural population in 2003, the New Rural Cooperative Medical Scheme (NRCMS). In our studied county, the NRCMS was established in 2005, and a kinked reimbursement rule for outpatient care was established in May 2010 and applied to the total expenditure net of a fixed payment on each visit since 2011. The structure of social insurance programs in rural China and the evolvement of the NRCMS over time in our studied county are discussed in Section 2. We also compare our research county to other counties in China, and its NRCMS policies to those of other countries to establish the external validity of our study.

Section 3 presents our theoretical model and empirical methodology to estimate the elasticity of the expenditure response to the coinsurance rate. Specifically, we assume that individuals are heterogeneous in their willingness to pay for the illness (i.e., the expenditure level without any insurance reimbursement). Additionally, based on our concerned insurance policies and data pattern, we also assume that individuals make static and frictionless decisions on health expenditure. Given that the insurance contract

reimburses patients for each outpatient visit in our studied county, we employ a utility function that depends on the per visit medical expenditure on outpatient services. This utility function is a modified version of that used by Einav et al. (2017). Next, we can derive the optimal choices of medical care expenditure under a linear insurance scheme and a nonlinear scheme, respectively.

The theoretical analysis shows that when a kink point is introduced into a linear scheme, individuals whose initial expenditure choice is greater than and close to the kink choose to bunch at the kink point exactly. The more sensitive an individual is to the coinsurance rate, the more likely he/she is to bunch at the kink. Hence, the bunching size is related to the elasticity of medical expenditure to health insurance. With a specified utility function, we can write the elasticity as a function of policy parameters (i.e., the kink value and coinsurance rates) and bunching size.

In the empirical part, with the administrative data, we observe the density distribution under a kinked contract. Assuming medical expenditures are smoothly distributed under a linear contract, we use the observed distribution of individuals who do not bunch to estimate the counterfactual one (i.e., when the kink point is not introduced) for individuals who do bunch around the kink. The extent of the excessive bunching under a nonlinear contract can hence be calculated by comparing the observed and counterfactual distributions. Combined with the policy parameters, we estimate the elasticity of medical expenditure to the coinsurance rate.

Notably, individuals usually face frictions when making decisions in reality; hence, by following the framework of Chetty et al. (2011), we extend our model in Section 3 to incorporate the optimization frictions. We assume that a fixed fraction of the population face frictions and cannot make optimal decisions. On the basis of this assumption, we can rewrite the excess bunching size as a function of two unknown parameters: the friction fraction and real elasticity. By exploiting the changes in the NRCMS policies over time, we construct two moment equations to simultaneously determine the two parameters. Finally, we discuss the reasons why we adopt a static response model instead of the dynamic model used by Einav et al. (2015).

We apply the model to outpatient service records in village clinics from March 9, 2012, to December 31, 2014, when a kinked insurance scheme was implemented. The graph of the results presented in Section 4 clearly demonstrates a spike in the distribution of the total medical expenditure around the kink point. To examine whether the bunching reflects the behavioral responses of individuals facing nonlinear insurance contracts or spurious relations due to unobserved factors, we examine the distribution patterns by using data from 2006–2009, when constant coinsurance rates were applied. We do not observe spikes around the kink points in the distributions for 2006–2009, and this supports our research design.

Three sets of empirical estimates are obtained in Section 5. First, we estimate the

elasticity without optimization frictions to be 0.4427 for the period from March 9, 2012, to October 30, 2012, during which the coinsurance rate increased from 40% to 100% at 10 RMB[1] of the total expenditure net of a fixed payment. This result implies that when the reimbursement rate was reduced from 60% to 0% at the kink, the total expenditure (net of a fixed payment) per visit would decrease by approximately 19%. We observe a similar elasticity (0.4478) for our second sample period (i.e., October 31, 2012, to December 31, 2014), during which the coinsurance rate increased from 30% to 100% at 11.43 RMB. The estimates are clearly insensitive to the choices of parameters in the baseline estimation specification, that is, the polynomial order, the bandwidth of the excluded region, and the reference points.

Our estimates are greater than those in the literature (Scitovsky and Snyder, 1972; Phelps and Newhouse, 1974; Eichner, 1998; Einav et al., 2017), and one possible reason is that the coinsurance rate in our research setting jumps from 30%/40% to 100%, whereas the literature have considered a copayment rate from 0%–25%. Another possible reason is that our focal sample is a population of poor residents from a developing country (i.e., rural residents in China) whose response to health insurance could be higher than that in the literature, which has exclusively focused on the United States.

Second, by using a model with optimization frictions, we estimate the degree of frictions to be 0.3427, indicating that the decisions made by about one third of our population are inconsistent with optimization. After correcting for frictions, we obtain an elasticity of 0.6415, approximately 40% larger than the elasticity estimated without friction correction. This result suggests that if the reimbursement rate were reduced from 70% to 60%, the total expenditure (net of a fixed payment) per visit would decrease by approximately 5%. A complete elimination of the reimbursement would cause the total expenditure (net of a fixed payment) per visit to decrease by 34.5%.

Third, we estimate heterogeneous responses across different subpopulations. Males are more responsive to insurance policies and display larger decision errors than females. This observation is consistent with the literature on sex differences in risk preferences regarding decision–making [for reviews, see Eckel and Grossman (2008), Croson and Gneezy (2009), and Bertrand (2011)]. Additionally, less educated individuals are more elastic to changes in coinsurance rates (reflecting their lower income levels and greater financial constraints) and face higher optimization frictions (in line with their lower cognitive abilities) than more educated people. Finally, older people display much larger elasticities than younger people (which is consistent with the insufficient financial resources for aging) and are more likely to make decision errors than young people (reflecting the deterioration of cognitive skills when aging).

[1]During our studied period (March 2012 to December 2014), the exchange rate between the Chinese currency (RMB) and US dollar was stable at approximately 6.80 before June 2010 and decreased steadily to approximately 6.10 after that.

With the baseline estimates, we conduct a cost–benefit analysis and identify several policy counterfactuals in Section 6. First, we calculate that per person per visit, the average benefit and government cost of the NRCMS program in 2014 were 14.87 RMB and 10.09 RMB, respectively. In total, the program generated a net welfare gain of 1.66 million RMB in 2014. Second, we conduct counterfactuals by changing the key policy parameters (the coinsurance rates and kink point) without increasing the government budget. Our analysis suggests that the current insurance policy (with the coinsurance rate jumping from 30% to 100% at 11.43 RMB) delivers the largest total welfare. Finally, we consider a counterfactual analysis applying the U.S. insurance schedule to our studied population. Specifically, we examine a standard plan in the Health Insurance Marketplaces created by the Affordable Care Act, which contains a deductible, coinsurance rate, and out–of–pocket maximum. We again maintain the government budget as unchanged and manipulate the deductible and out–of–pocket maximum thresholds to obtain the welfare–maximizing schedule. The results show that the insurance of the U.S. pattern delivers less welfare than that of the NRCMS. One potential explanation for this difference is that a substantial amount of the budget in the counterfactual plan is distributed to a small number of patients with high expenditures, resulting in the remaining low–spending patients receiving smaller reimbursements than they would obtain in the carried–out plan.

According to our review of the literature, the studies on expenditure responses to health insurance have been limited and focused exclusively on the United States. Three papers have used the regression approach to compute the expenditure elasticity (Scitovsky and Snyder, 1972; Phelps and Newhouse, 1974; Eichner, 1998), whereas more recent studies have used the kink design to calculate the response elasticity to Medicare Part D (Abaluck et al., 2018; Einav et al., 2015, 2017). Although our research design is similar to the static response framework presented in Einav et al. (2017), we examine a poor population in a developing country (i.e., rural residents in China), in which health care provisions are much more necessary.

Our work is also similar to the recent literature that has used the kink design to estimate behavioral responses. This methodology has been developed by Saez (2010) and Chetty et al. (2011) and applied to various settings with nonlinear budget sets, including a personal income tax (Saez, 2010; Chetty et al., 2011; Chetty et al., 2013; Bastani and Selin, 2014), value–added tax (Onji, 2009; Harju et al., 2016; Liu et al., 2017), pensions (Brown, 2013; Manoli and Weber, 2016), electricity prices (Ito, 2014), cellular service prices (Grubb and Osborne, 2015), health care procurement (Bajari et al., 2017), and housing transaction tax (Best et al., 2018; Best and Kleven, 2018). The contributions of our work focus on general health care and policy counterfactuals.

2 Research Background

This section introduces the institutional background of health insurance policies in rural China and discusses the external validity of our study by first showing the representativeness of our research region and then comparing insurance policies in China against those in other countries.

2.1 Health Insurance Policies in Rural China

Although China has experienced rapid economic development since the market reform in 1978, the degree of urbanization in its rural areas remains low. According to the 2011 China Health and Retirement Longitudinal Study (CHARLS), approximately 91.7% of the sampled villages had roads, but only 59.5% had paved roads; 30.6% of households on average had tap water to drink; 28.4% of households had natural gas or liquefied petroleum gas as a fuel source; 99.5% of households on average used electricity for daily life; the main type of toilet was open–air; 28.9% of villages had a public restroom; and 37.4% of agricultural products used machines for cultivation and harvest.

As for the medical system, two–tiered facilities provide basic services to residents: township health centers and village clinics. Each town has an intermediate–sized township health center. In 2011, a health center had on average 27.5 beds and 31.3 physicians (China Health Statistics Summary, 2012). Village clinics are small in size. In 2011, a village clinic had on average 2.3 physicians. Village clinics could have different types of ownership and 64.7% were publicly owned in 2011. According to the 2011 CHARLS, 68.8% of the sampled villages had village clinics, 3.3% had general hospitals, and 23.3% had township health clinics. For the villages without village clinics, township health clinics, or general hospitals, the residents visited the facility located on average 2.13, 5.09, or 24.32 km away, respectively, to receive care, and the travel time was on average 13.43, 24.69, or 65 minutes, respectively.

Since the economic reform in 1978, the old health insurance program in rural China, called the Cooperative Medical Scheme, had collapsed due to insufficient public financing. The majority of the rural residents remained uninsured for many years, for example, 87.3% of the rural residents were not covered by health insurance in 1998 (China Health and Family Planning Statistical Yearbook, 2013). In 2003, the Chinese government launched a new voluntary health insurance program, namely, the NRCMS, to mitigate the public concern regarding the inadequate access to medical services and high out–of–pocket spending. Because the program was heavily subsidized by the government, the enrollment rate was high: approximately 60% in 2003 (Bai and Wu, 2014). In 2011, the NRCMS covered 89.9% of the rural population, and 6.9% were covered by other public insurance. (China Health Statistics Summary, 2012). Private insurance was uncommon in rural areas, with only 0.1% of the population covered by private insurance, and the

remaining 3.1% were uninsured.

The NRCMS programs are administrated at the county level, the policies vary by county, and all the programs cover general medical services and prescription medication. The prices of services are regulated by the governments and set by the National Development Reform Commission on the basis of costs and patients' ability to pay. The prices are not frequently adjusted, and were fixed during our research period in the sampled county. Drugs and medical materials are procured through provincial platforms, and their prices are determined through a bidding process or negotiated between pharmaceutical firms and the governments. In general, medical facilities and physicians are price–takers and have little power to adjust medical prices. With the government's regulation, the labor–related services are generally priced below their market prices, and cheap drugs are available in rural areas. For example, the injection fee was only 3 RMB at a time during our research period, and a 24–count box of Aspirin tablets was 6.3 RMB. Hence, 10 RMB could allow the patients to receive basic medical care such as an injection. When spending increased from 10 RMB to 11.43 RMB, patients could obtain more services such as a few more pills.

Our study is based in a county in the southwestern part of China for which we could access medical claims data for all the enrollees in the NRCMS program. The program had been implemented in our research county since 2005, and the enrollment rate was almost inclusive, for example, 99% in 2014. Part of the premiums were contributed by enrollees, with a large portion subsidized by the government. For example, the total premium per enrollee in 2014 was 290 RMB, and the out–of–pocket premium was only 50 RMB. The insurance covered outpatient and inpatient services.

This paper focuses on outpatient care because it is the most common basic service. The NRCMS policies on outpatient reimbursement varied by year. In 2005–2009, enrollees received an annual payment (8 RMB in 2005–2008 and 14 RMB in 2009 per person) deposited in their family medical account. Family members could share the deposits, and the deposit amounts that remained at the end of the year could be carried forward to the next year. In May 2010, the insurance bureau canceled the family account and established a fixed reimbursement rule for outpatient care.[2] The insurance would pay 50% of the bill until the received per visit reimbursement reached 6 RMB if the visit was in village clinics, or 10 RMB if the visit was in township centers or hospitals. The annual upper limit for the total reimbursement was 30 RMB.

In 2011, the insurance introduced the General Diagnostic Payment (GDP), which provided a fixed reimbursement for the medical diagnostic fee for each outpatient visit. The medical diagnostic fee was set by the government. In our studied county, the per visit medical diagnosis fee was fixed at 6 RMB for hospitals, 10 RMB for township centers,

[2]The remaining deposits in the family account could still be used but no additional money would be credited into the account beyond 2010.

and 5 RMB for village clinics. For township centers and hospitals, the diagnosis fees were fully reimbursed by the GDP. For village clinics, the GDP only covered 4.5 RMB per visit if the medical diagnosis fee was charged. Since 2011, the reimbursement rule has applied to the total expenditure net of the GDP for each visit.

The reimbursement policy was adjusted frequently in the subsequent years to meet the changes in medical demand. The detailed adjustments were not well documented until early 2012. After that, the policies for village clinic visits became quite stable, but for visits to township health centers and hospitals, the reimbursement rate changed substantially and was also not well documented. Therefore, this paper focuses on the reimbursement policies in village clinics since early 2012.

Figure 1 shows the reimbursement schedules investigated in our study. The X axis is the total expenditure (net of the GDP) per visit (in RMB), and the Y axis is the marginal coinsurance rate.[3] The solid curve presents the schedule between March 9 and October 30, 2012, and the dashed curve presents the schedule between October 31, 2012, and December 31, 2014.

[Insert Figure 1 Here]

Between March 9 and October 30, 2012, patients could enjoy a 60% reimbursement rate until the amount of the received reimbursement reached 6 RMB per visit. Hence, the marginal coinsurance rate was 40% when the total expenditure (net of the GDP) per visit was less than 10 RMB (= 6/60%). Once the total expenditure (net of the GDP) per visit was greater than 10 RMB, any additional expenditure had to be paid fully out of pocket, that is, the marginal coinsurance rate increased to 100% at 10 RMB. Therefore, for example, when a patient visited a village clinic on March 10, 2012, and spent 16 RMB on outpatient services including diagnostics, he/she would receive a 4.5 RMB reimbursement as the GDP, an additional 6 RMB reimbursement, and pay 5.5 RMB out of pocket. Between October 31, 2012, and December 31, 2014, the reimbursement rate changed to 70%, and the reimbursement for each visit was capped at 8 RMB. Hence, the threshold of the jump changed to 11.43 RMB ($\simeq$ 8 RMB /70%): below this threshold, the marginal coinsurance rate was 30%, and above the threshold, it became 100%.

2.2 External Validity

2.2.1 Research County

Our analysis draws data from a county located in the southwestern part of China. To gain a sense of the external validity of our study, we compare our research county with all the counties in China in terms of the population structure, health and education levels, employment, income, and living standards by using the China 2010 Population Census.

[3]The annual upper limit of the total received reimbursement was 150 RMB in 2012 and 2013, and decreased to 100 RMB in 2014.

The results are presented in Appendix Table A1–A3. As shown in the tables, our research county is comparable to the median county. As a result, the empirical findings from our study could be generalizable regarding the responses of patients to health insurance in rural China.

2.2.2 Policy Comparison with Other Countries

We compare the policies of the health insurance programs in the sampled county with other countries. The comparison focuses on the two features of the NRCMS policy: the ceiling and the separated reimbursement for inpatient and outpatient care.

First, in the NRCMS program, patients have reduced/compensated prices when the total cost is less than the ceiling threshold, and incur full prices above the threshold. The design of the ceiling is not unique for programs in China. One example of the nonlinear health insurance contract is Medicare Part D in the United States. In a standard plan, the coinsurance rate is set as 100%, 25%, 100%, or 7%, if the total expenditure is below \$275, between \$275 and \$2,510, between \$2,510 and \$5,726, or greater than \$5,726, respectively. Hence, the consumer must pay the full expense out of pocket when the spending is between \$2,510 and \$5,726, which is called the donut hole. Einav et al. (2015) observe similar bunching evidence at the kink when the drug price sharply increases.

A ceiling is more popularly applied to insurance programs in developing countries. In 2003, the Voluntary Insurance Scheme in Vietnam had a ceiling policy similar to the NRCMS (Nguyen and Akal, 2003). Enrollees incurred a \$2 threshold per service for primary care and essential drugs, below which no copayments were required. Caps were also applied to high–cost surgical and medical interventions. For example, the cap was approximately \$940 per person for heart surgery, and approximately \$1,128 for kidney transplants. Since 2009, because of Vietnam's health insurance reform, several ceilings were eliminated, but high–technology treatments were remained covered up to a certain limit. Specifically, the ceiling was defined as 40 months of the minimum monthly salary, which was equivalent to \$35 (as of 2010) (Van Tien et al., 2011). In Thailand, patients are subject to copayments if the expenditure is greater than the ceiling for maternity care in its Social Security scheme. In Philippines, for the PhilHealth program, services are also subject to ceilings.

Second, the reimbursement policies of the NRCMS differ for outpatient and inpatient services. This distinction is also commonly found in programs in other countries (Giedion et al., 2013). In Colombia, the insurance program for formal sector employees, the Contributory Health Insurance Regime, categorizes services into nonlife–saving and inpatient care and defines different coinsurance rates. In Georgia, the Universal Benefit Package program covers primary care, preventive and emergency care, and treatment for selected diseases, but not inpatient care. In Mexico, the Seguro Popular program

categorizes services into three packages: community services, essential services such as prevention and rehabilitation, and high–cost tertiary care. In the United States, the Medicare program also establishes different policies for inpatient care, outpatient care, and prescription drugs.

Programs in other countries also have features that differ from the NRCMS. For example, Medicare mainly covers the elderly population in the United States, whereas the NRCMS covers the rural population. Because this study focuses on nonlinear insurance contracts, consumers respond with a similar pattern to the policies once they experience price changes. From this perspective, our results could be extended to other countries, especially developing countries with economic conditions and insurance policies similar to China.

3 Theory and Estimation

This section first presents the static, frictionless model developed by Einav et al. (2017) and estimation method, which follows Chetty et al. (2011). We next extend the model to incorporate potential optimization frictions by following the framework in Chetty et al. (2011). Finally, we discuss our adoption of a static model instead of the dynamic response model used by Einav et al. (2015).

3.1 Frictionless Response to Nonlinear Health Insurance Policies

To fit into our institutional setting, we consider a utility function that depends on the medical expenditure for each clinic visit. Specifically, the utility function of individual i is assumed to be

$$u_i(m_{i1}, m_{i2}, \ldots, m_{iJ}, y_i) = \sum_j \left\{ \left[2m_{ij} - \frac{A_{ij}}{1+\frac{1}{\alpha}} (\frac{m_{ij}}{A_{ij}})^{1+\frac{1}{\alpha}} + E * I_i \right] \right\} + \left[Y_i - \sum_j c_{ij} m_{ij} - T \right], \tag{1}$$

where m_{ij} ($j \in \{1, \ldots, J\}$) denotes the individual i's total expenditure (net of the GDP) on the jth clinic visit; A_{ij} is a fundamental variable, to be defined later; E is the GDP received; $I_{ij} = 1$ if individual i receives the GDP on the jth clinic visit and 0 otherwise; $y_i \equiv Y_i - \sum_j c_{ij} m_{ij} - T$ denotes the residual income; Y_i is the individual i's gross income; $c_{ij} \in [0, 1]$ denotes the fraction of the total expenditure to be paid out of pocket in the jth visit by the individual i, or the coinsurance rate; and T is the NRCMS premium. The first part of equation (1) presents the utility gained from the medical expenditure, whereas the second part presents that from other consumptions. α is our parameter of interest and denotes the elasticity of the medical expenditure to the insurance coverage.

This utility function is a modified version of that employed in Einav et al.'s (2017) framework. Einav et al. (2017) adapt Saez's (2010) static and frictionless model to the context of health insurance. Specifically, due to their specific research setting (i.e., Medicare Part D of the United States), they assume that the utility function depends on the total expenditure of a year instead of the expenditure per visit.

The optimization of the utility function (1) can determine that the optimal choice of the medical care expenditure of the individual i in jth visit is

$$m_{ij} = A_{ij}(2 - c_{ij})^{\alpha}. \tag{2}$$

When $c_{ij} = 1$, we have $m_{ij} = A_{ij}$. Hence, A_{ij} represents the individual i's total expenditure without health insurance in visit j; in other words, his/her willingness to pay for the illness. We assume that A_{ij} is distributed with density function $f(.)$ and cumulative function $F(.)$. And with full insurance coverage (i.e., $c_{ij} = 0$), $m_{ij} = 2^{\alpha}A_{ij}$. By taking the logarithm of both sides in equation (2), we have $\alpha = \frac{\Delta \ln m_{ij}}{\Delta(2-c_{ij})}$. Therefore, α represents the constant elasticity of the total expenditure on each visit with respect to two minus the coinsurance rate.

By following Saez (2010), we compare the optimal choices under a linear contract and a nonlinear contract to determine which group of individuals choose to bunch around the kink. With the optimal choices of the bunching individuals, we obtain the relationship between the elasticity α, the counterfactual density distribution, and the excess bunching size.

First, we consider a constant reimbursement schedule with $c_{ij} = c^0$. Let $H_0(m_{ij})$ be the cumulative distribution of m_{ij} and $h_0(m_{ij})$ be the corresponding density distribution. According to equation (2), we have $m_{ij} = A_{ij}(2-c^0)^{\alpha}$. Therefore, $H_0(m_{ij}) = Pr(A_{ij}(2-c^0)^{\alpha} \leqslant m_{ij}) = F(m_{ij}/(2-c^0)^{\alpha})$, and $h_0(m_{ij}) = f(m_{ij}/(2-c^0)^{\alpha})/(2-c^0)^{\alpha}$.

Next, we introduce a nonlinear health insurance plan kinked at the medical expenditure point m^*, as described in the previous section. As shown in Figure 2, when the total medical expenditure per visit m_{ij} is not greater than a threshold m^*, the marginal coinsurance rate is c^0, and when $m_{ij} > m^*$, the marginal coinsurance rate is c^1, with $c^1 > c^0$. Let $H(m_{ij})$ denote the cumulative distribution of m_{ij} under this nonlinear reimbursement schedule and $h(m_{ij})$ the corresponding density distribution.

[Insert Figure 2 Here]

With the kinked policy, the optimal response of the medical expenditure can be derived as follows. First, for the optimal per visit total expenditure below m^*, we have $m_{ij} = A_{ij}(2-c^0)^{\alpha}$. Hence, for $A_{ij} < m^*/(2-c^0)^{\alpha}$, we have $h(m_{ij}) = h_0(m_{ij})$. Second, for $A_{ij} \geq m^*/(2-c^0)^{\alpha}$, individuals have two options: an expenditure of m^*, and an expenditure

according to the new reimbursement rate c^1; specifically, for the latter, the optimal per visit total expenditure is $m_{ij} = A_{ij}(2-c^1)^\alpha$. By comparing the utilities from these two options, we have that for $A_{ij} > m^*/(2-c^1)^\alpha$, individuals choose $m_{ij} = A_{ij}(2-c^1)^\alpha$. Therefore, $H(m_{ij}) = Pr(A_{ij}(2-c^1)^\alpha \leqslant m_{ij}) = F(m_{ij}/(2-c^1)^\alpha)$ and $h(m_{ij}) = f(m_{ij}/(2-c^1)^\alpha)/(2-c^1)^\alpha = h_0(m_{ij}((2-c^0)/(2-c^1))^\alpha) * ((2-c^0)/(2-c^1))^\alpha$. As for individuals with $A_{ij} \in [m^*/(2-c^0)^\alpha, m^*/(2-c^1)^\alpha]$, they optimally choose $m_{ij} = m^*$ and hence bunch at the kink point. In summary, the optimal spending decision under the nonlinear health insurance plan is

$$m_{ij}(A_{ij}) = \begin{cases} A_{ij}(2-c^0)^\alpha & if \;\; A_{ij} < m^*/(2-c^0)^\alpha \\ m^* \quad if \;\; A_{ij} \in [m^*/(2-c^0)^\alpha, m^*/(2-c^1)^\alpha] \\ A_{ij}(2-c^1)^\alpha & if \;\; A_{ij} > m^*/(2-c^1)^\alpha \end{cases} . \tag{3}$$

An individual who bunches with the smallest value $A_{ij} = m^*/(2-c^0)^\alpha$ is the individual who chooses the expenditure level $m_{ij} = m^*$ under the constant reimbursement schedule $c_{ij} = c^0$. We denote this individual as L in Figure 2. An individual who bunches with the highest value $A_{ij} = m^*/(2-c^1)^\alpha$ would have chosen the expenditure level $m_{ij} = m^*(2-c^0)^\alpha/(2-c^1)^\alpha$ if the reimbursement schedule remained at constant $c_{ij} = c^0$. We denote this individual as H in Figure 2. Hence, any individual whose total medical expenditure ranges from m^* to $m^* + \Delta m^*$ under the constant reimbursement rate $c_{ij} = c^0$ bunches at m^* under the new nonlinear reimbursement schedule (c^0, c^1), where

$$\frac{\Delta m^*}{m^*} = \left(\frac{2-c^0}{2-c^1}\right)^\alpha - 1. \tag{4}$$

Therefore, the excess fraction of the population bunching is

$$B = \int_{m^*}^{m^*+\Delta m^*} h_0(m)dm \simeq h_0(m^*)\Delta m^*, \tag{5}$$

where the second approximation is based on the assumption that the density $h_0(m)$ is uniform around the kink point m^*, as in Saez (2010) and Chetty et al. (2011).

By combining equations (4) and (5), we can then solve elasticity α as a function of observable (i.e., c^0, c^1 and m^*) and empirically estimable variables (i.e., B and $h_0(m^*)$), that is,

$$\alpha = \ln\left(\frac{B}{h_0(m^*) \times m^*} + 1\right) / \ln\left(\frac{2-c^0}{2-c^1}\right). \tag{6}$$

3.2 Estimation Framework

As illustrated in Subsection 3.1, to estimate the elasticity α, we must explore the excess bunching B and counterfactual density $h_0(m^*)$ (i.e., the one that would appear

under a constant reimbursement schedule $c_{ij} = c^0$). However, from the administrative data, we observe the distribution with a kinked schedule only. To obtain the elasticity, α, we follow the empirical model of Chetty et al. (2011) to estimate the counterfactual density distribution $h_0(m_{ij})$ from the observed one. Specifically, we first group individuals into 0.5 RMB bins of the total outpatient service expenditure (net of the GDP); next, we exclude the observations around the kink and fit a polynomial to the observed counts in each bin. The estimation equation is

$$C_n = \sum_{p=0}^{q} \beta_p^0 (m_n)^p + \sum_{d=-R}^{R} \gamma_d^0 * 1[m_n = d] + \sum_{k \in K} \alpha_k^0 * 1[\frac{m_n}{k} \in \mathbb{N}] + \epsilon_n^0, \tag{7}$$

where C_n is the number of individuals in the expenditure bin n, m_n is the expenditure relative to the kink, q is the order of the polynomial, and R represents the width of the excluded region around the kink in 0.5 RMB (i.e., the fraction of individuals who choose to bunch at the kink point exactly under a nonlinear contract). To address the problem of reference point effects, we also add a control for multiples of 5 in equation (7). Specifically, $K = \{5\}$, and $\mathbb{N}$ is the set of natural numbers. This method of adding dummies to control for reference points is from Kleven and Waseem (2013).

The estimated counterfactual density distribution is defined as $\hat{C}_n^0 = \sum_{p=0}^{q} \hat{\beta}_p^0 (m_n)^p + \sum_{k \in K} \hat{\alpha}_k^0 * 1[\frac{m_n}{k} \in \mathbb{N}]$, and the excess mass of bunching is estimated as $\hat{B}^0 = \sum_{d=-R}^{R} \hat{\gamma}_d^0$. However, as individuals bunching around the kink are from the distribution to the right of the kink, the cumulated distribution under the estimated counterfactual density is not equal to that under the observed density, and this causes $\hat{B}^0$ to be overestimated. To address this problem and obtain a more accurate counterfactual density, we impose an integration constraint on our estimates that the total population of the counterfactual density should be equal to that of the observed one. This constraint is proposed by Chetty et al. (2011). With this constraint, the new counterfactual density can be estimated from the following equation

$$C_n \left(1 + 1[n > R] \frac{\hat{B}}{\sum_{n=R+1}^{\infty} C_n} \right) = \sum_{p=0}^{q} \beta_p (m_n)^p + \sum_{d=-R}^{R} \gamma_d * 1[m_n = d] + \sum_{k \in K} \alpha_k * 1[\frac{m_n}{k} \in \mathbb{N}] + \epsilon_n, \tag{8}$$

where $\hat{B} = \sum_{d=-R}^{R} \hat{\gamma}_d$ represents the excess number of individuals located around the kink with the counterfactual density $\hat{C}_n = \sum_{p=0}^{q} \hat{\beta}_p (m_n)^p + \sum_{k \in K} \hat{\alpha}_k * 1[\frac{m_n}{k} \in \mathbb{N}]$.

With this estimated counterfactual density, we can calculate Δm^* as $\Delta \hat{m}^* = \frac{\hat{B}}{\hat{h}_0(m^*)} = \frac{\hat{B}}{\sum_{d=-R}^{R} \hat{C}_d/(2R+1)}$. Hence, according to equation (6), elasticity α is computed as $\hat{\alpha} = \ln\left(\frac{\Delta \hat{m}^*}{m^*} + 1\right) / \ln\left(\frac{2-c^0}{2-c^1}\right)$. Standard errors are estimated by using a parametric bootstrap procedure. Specifically, by following Chetty et al. (2011), we redraw the estimated vector of errors ϵ_n in equation (8) with replacement to generate a new sample and calculate a

new estimate $\hat{\alpha}^n$. We repeat this procedure 200 times and estimate the standard error of $\hat{\alpha}$ as the standard deviation of the 200 estimates.

3.3 Discussion

3.3.1 Extension with Optimization Frictions

In the benchmark model, we consider a frictionless setting, that is, individuals optimally choose their medical expenditures given the health conditions and insurance policies without any errors. However, in practice, people often face optimization frictions when making decisions. The frictions may be from high adjustment costs, inertia, incomplete information, or the inability to calculate the optimal decision. These may be relevant in our research setting (i.e., the rural population in China), where the majority are not well educated (i.e., 81.36% with a middle school education or below). Optimization frictions act as a diffusion devise in our setting. With the frictions, people who should bunch at the kink point end up at other locations on the expenditure distribution. As a result, the bunching size, and hence the elasticity, might be under–estimated in our benchmark model.

To uncover the true elasticity with optimization frictions, we assume that a fraction, δ, of the population face frictions and cannot choose their optimal expenditures, whereas the remaining population have no frictions when considering their optimization decisions. This assumption is initially proposed in Chetty et al. (2011). With this assumption, a fraction $(1-\delta)$ of the population choose their optimal expenditures m_{ij} on the basis of their health conditions, and the remainder, δ, of the population choose an arbitrary expenditure by following the underlying distribution $H_0(m_{ij})$, as if they are irrational.

In the case of a constant reimbursement schedule (i.e., $c_{ij} = c^0$), the observed density at m^* contains two groups of individuals: the $(1-\delta)h_0(m^*)$ population facing no frictions and the $\delta h_0(m^*)$ population facing frictions.

With a nonlinear reimbursement schedule, the observed density at m^* contains three types of individuals: the $(1-\delta)h_0(m^*)$ population facing no frictions, the $(1-\delta)\int_{m^*}^{m^*+\Delta m^*} h_0(m^*)dm$ population with no frictions bunching at m^*, and the $\delta h_0(m^*)$ population with frictions. Hence, the observed excess bunching is $\hat{B} = (1-\delta)B$.

By combining equations (4) and (5), we can obtain

$$\left(\frac{2-c^0}{2-c^1}\right)^{\hat{\alpha}} = (1-\delta)\left(\frac{2-c^0}{2-c^1}\right)^{\alpha^t} + \delta, \tag{9}$$

where $\hat{\alpha} = \ln\left(\frac{\hat{B}}{\tilde{h}_0(m^*)\times m^*} + 1\right) / \ln\left(\frac{2-c^0}{2-c^1}\right)$ is the empirically estimated elasticity; α^t is the true elasticity to be estimated.

In equation (9), we have two unknown parameters α^t and δ; hence, we require two

empirical moments to estimate them simultaneously. By exploring the changes in the insurance policies in October 2012, we can obtain two $\hat{\alpha}$s (i.e., one for the period from March 9 to October 30, 2012, and the other for the period from October 31, 2012, to December 31, 2014), and therefore, two empirical moments. To further address the problem of seasonality, for the second sample period, we estimate $\hat{\alpha}$ by using the same days as the first sample period; that is, the March 9 to October 30 periods in 2013 and 2014.

3.3.2 Static Versus Dynamic Response Model

This subsection presents the reasons why we do not employ a dynamic response model.

In their study of the drug response to Medicare Part D, Einav et al. (2015) consider a dynamic response model to calculate the elasticity from the excess bunching at the kink point. Specifically, they consider risk–neutral and forward–looking individuals possibly facing a series of stochastic health shocks within one year. Because the timing of health shocks is uncertain, individuals make drug purchase decisions and update their beliefs regarding the state sequentially.

The adoption of the dynamic model is motivated by the institutional setting (i.e., the reimbursement of Medicare Part D is based on the annual total drug expenditure) and supported by the empirical regularities from the data (i.e., individuals purchase less in response to the kink and much of the response is concentrated in the late months within a calendar year). Einav et al. (2017) compare the dynamic response model with the static model developed by Saez (2010) (applied in Section 3.1). One key difference between the two models is whether the decisions are sequential within the policy coverage period.

Our research setting is similar to Einav et al. (2015), namely, an investigation of the expenditure responses to health insurance by employing excess bunching at the kink point and using individuals' insurance claim data with the exact day of the expenditure. Notably, there is one crucial difference between the policies we present and those in Einav et al. (2015): Medicare Part D provides reimbursement on an annual basis, and the NRCMS in China provides reimbursement per visit. The NRCMS has a cap on annual reimbursement, but this amount is generous compared with the cap per visit, that is, the former is approximately 25 times the latter. Additionally, individuals can use family members' unused annual reimbursement amount. When combined, the annual reimbursement cap does not have a strong influence on per visit decision. Hence, the sequential consideration of the medical expenditure is less relevant in our research setting than in Einav et al.'s (2015) and justifies our choice of a static response model.

To further support our argument, we present similar graphs regarding the empirical regularities that motivate the dynamic model in Einav et al. (2015). Specifically, in Figure A1, we plot the share of individuals with at least one visit in village clinics in each given

month against the total reimbursement received until the given month for the last four months of 2012. We observe that the propensity for clinic visits increases with the total amount of reimbursement received, and this result reflects the inherent health preference of individuals in a manner similar to the results in Einav et al. (2015). Notably, we do not observe large declines at any point, except for the annual upper limit, and especially for the large value of reimbursement received. These findings sharply contrast with those in Einav et al. (2015), which found lower–than–predicted frequencies when the total annual expenditure is close to the kink point. The results for 2013 and 2014, reported in Figures A2 and A3, show similar patterns.

These results suggest that individuals' medical decisions are largely based on each visit, instead of on an annual basis, given the nature of the NRCMS policy. Because of the absence of dynamic responses, we adopt the static model in our analysis.

4 Data and Descriptive Results

Our analysis is based on the health care administrative data in a southwestern county in China. The data are a record of every outpatient service visit at every counter of the local health institutions, namely, 313 village clinics, 21 township health centers, and 858 hospitals, in 2006–2014. Detailed information concerning each visit is contained in the data, including the date of visit, the diagnosis [i.e., the International Classification of Diseases (ICD) 10 code], medical organization visited, total outpatient expenditure, amount of the GDP received, amount of insurance reimbursement received, amount of deposits used from the family account, and amount of out–of–pocket payment. The data also contain the demographics of the patient, such as sex, birth date, marital status, education level, and occupation.

Table 1A presents the summary statistics for the total sample of the data. As shown in the table, the data contains 4 million outpatient service visits. Among these visits, approximately 46.6% occurred in village clinics, and the remainder were in township health centers and hospitals. Over the years, the number of visits increased and became stable after 2012, reflecting the growing enrollment in the NRCMS program. For each visit, patients spent on average approximately 30.98 RMB, a rate comparable to the reported average outpatient service expenditure per visit—56.9 RMB in the China Health and Family Planning Statistical Yearbook (2015). The reimbursement schedules for visits for chronic and special diseases differed from those for normal diseases in our focal county; thus, we exclude them from our analysis sample. The majority of visits were nonchronic and nonspecial diseases, and approximately 1.3% were chronic and special diseases. Of the total expenditure, on average, 3.05 RMB was paid by the deposits from the family account, 5.34 RMB was subsidized by the GDP [4], and 12.18 RMB was covered by the

[4] According to the insurance contract, the GDP provided a fixed reimbursement for medical diagnosis

reimbursement.

[Insert Table 1A Here]

Based on the availability of the health insurance policies discussed in Section 2.1, we restrict our sample to the outpatient service records in village clinics from March 9, 2012, to December 31, 2014. To study the responses of patients to the NRCMS, we exclude records with zero reimbursement subsidies. After the exclusion, approximately 984 thousand observations remain: 173,698 observations from March 9, 2012, to October 30, 2012, and 810,483 from October 31, 2012, to December 31, 2014. Much of our analysis focuses on the subset of observations around the kink point. Specifically, for the first sample period, we focus on 169,926 observations from −10 RMB to 20 RMB of the kink point at 10 RMB. For the second sample period, we use 806,260 observations from −12 RMB to 20 RMB of the kink point at 12 RMB. Combined, our focal analysis sample accounts for 99.19% of the full sample.

Table 1B presents the summary statistics for our analysis sample. The demographics of patients are reported in panel A. Of our sample population, 47.27% were male, 12.92% were single, 75.95% were married, 9.45% were widowed, and 1.67% were divorced. Individuals had 6 years of schooling on average, and 42.73% had attended middle school or higher education. The mean individual age was approximately 48 years, and 36.98% were old (aged older than 50 for females and 60 for males). Panel B shows the monetary information for each visit. Patients spent on average approximately 13.90 RMB on outpatient services per visit, among which 0.27 RMB was paid by the deposits from the family account, 2.60 RMB was covered by the GDP, and 6.50 RMB was reimbursed by the insurance.[5] The patient paid 11.30 RMB out of pocket. The per visit average expenditure on outpatient services is much less than the statistics in the total sample in Table 1A and the reported figure in China Health and Family Planning Statistical Yearbook (2015). One possible reason for this result is that our estimation sample excludes the costly diagnosis of special and chronic diseases. In addition, our estimation focuses on village clinics, where people generally visit for minor illness.

[Insert Table 1B Here]

The density distribution of the total expenditure (net of the GDP) per visit from March 9, 2012, to October 30, 2012, is presented in Figure 3A. The solid curve is the observed density; the dotted curve is the estimated counterfactual distribution from equa-

fees if applicable. Hence, for outpatient visits without medical diagnosis services, the GDP would be zero.

[5]In our studied county, the GDP only covered partial medical diagnosis fees for village clinics if applicable. Therefore, regardless of whether patients received medical diagnosis services and payments from the GDP, the total medical expenditure (net of the GDP) was always positive, as shown in Table 1B, Figure 3A, Figure 3B, and Figure 4.

tion (8); and the dashed lines represent the marginal coinsurance rates. There is a noticeable spike in the observed distribution of the total expenditure (net of the GDP) around the kink point that suggests a strong response to the kink in health insurance. Figure 3B shows the situation from October 31, 2012, to December 31, 2014, where the kink point changed from 10 RMB to 12 RMB. Accordingly, we observe that the sharp jump in the observed distribution of the total expenditure (net of the GDP) moves to the new kink point region, suggesting that the bunching reflects individuals' behavioral responses to the substantial increase in the coinsurance rate at the kink point.

[Insert Figures 3A and 3B Here]

The assumption underlying estimation equation (8) is that the expenditure distribution would be smooth if there were no increase in the coinsurance rate at the kink point. By using the data in previous years when there were no changes in the coinsurance rate throughout the distribution, we can examine the relevance of our identifying assumption.

Specifically, in 2006–2008, individuals received an annual lump sum transfer of 8 RMB per person into their family account, which increased to 14 RMB per person in 2009. After the family account had been spent, the coinsurance rate remained at 100% throughout the entire distribution in these years. We plot the distributions of the total expenditure per visit for 2006–2008 and 2009 in village clinics, respectively, in Figure 4, along with the distributions of the total expenditure (net of the GDP) per visit for the period from March 9, 2012, to December 31, 2014. Although the distribution from March 9, 2012, to October 30, 2012, and that from October 31, 2012, to December 31, 2014, have clear excess bunching at their corresponding kinks, no spikes are observed at these kinks in the distribution for 2006–2008 or 2009.

In addition, the distribution for 2006–2008 shows a bunching at 8 RMB that implies individuals' incentives to first pay the medical bills from their family accounts. Notably, no similar pattern is observed around 14 RMB for the distribution in 2009; one potential explanation is that individuals' behavior was affected by the change in the insurance policies. Specifically, the insurance program stopped making direct deposits into family accounts in 2010. Instead, enrollees were reimbursed through a fixed payment (i.e., the GDP) and payments proportional to total spending. Individuals could still use the balance from the family account to pay the cost–sharing part. These policy changes were announced in 2009. With this information, the individuals in our studied county might have had fewer incentives to use up the money in their family account, and hence, a smoothing in expenditure across the lump sum transfer of 14 RMB.

[Insert Figure 4 Here]

5 Empirical Estimates

5.1 Elasticities from the Frictionless Response Model

Table 2 presents the estimates of the response elasticity, α, from the frictionless framework presented in Section 3.1. We start with the period from March 9, 2012, to October 30, 2012, in panel A, in which the kink point was 10 RMB and the coinsurance rate increased from 40% to 100% at the kink. Our baseline uses the fifth–order polynomials to capture the counterfactual density distribution, sets the excluded region with a bandwidth of 3 RMB centered around the kink point of 10 RMB, and controls for multiplies of 5. The estimated counterfactual density distribution is plotted in Figure 3A: the red, dotted curve. With this counterfactual density, we calculate the normalized excess bunching as $\Delta\hat{m}^* = 2.3130$, with a statistical significance at the 1% level. According to equation (6), elasticity α is computed as $\hat{\alpha} = 0.4427$, which is statistically significant at the 1% level. With these estimates, we can calculate that when the reimbursement rate was reduced from 60% to 0% at the kink, the total expenditure (net of the GDP) per visit would decrease by approximately19%. This substantial response is mainly from the demand side, because prices are regulated by the government and the supply side has no market power.

[Insert Table 2 Here]

Panel B reports the estimates from October 31, 2012, to December 31, 2014, in which the kink point moved to 12 RMB and the coinsurance rate increased from 30% to 100% at the kink. The baseline parameters are set as follows: the fifth–order polynomials, the excluded region with a bandwidth of 2 RMB, and the controls for multiplies of 5. We observe that $\Delta\hat{m}^* = 3.2186$ and $\hat{\alpha} = 0.4478$, and both are statistically significant at the 1% level. With these estimates, we can calculate that when the reimbursement rate was reduced from 70% to 0% at the kink, the total expenditure (net of the GDP) per visit would decrease by approximately 24%. Additionally, if we consider an alternative experiment with the reimbursement rate reduced by 10% from 70%, the total expenditure (net of the GDP) per visit would decrease by approximately 3%.

Comparison with the existing literature. According to our review of the literature, much of the literature on expenditure response to health insurance exclusively has focused on developed countries, namely, the United States. Scitovsky and Snyder (1972) investigate the Group Health Plan offered by Stanford University to its staffs and observe that the per capita expenditure on physician services declines by 23.8% due to an increase in the copayment rate from 0% to 25%. Phelps and Newhouse (1974) use different sources of data from the United States to compute the total medical expenditure elasticity with respect to the coinsurance rate and arrive at an estimate of 0.043. Eichner (1998) studies three plans with different copayments offered by 500 firms in the United States and shows

that the overall elasticities of medical care with respect to the out–of–pocket price range from -0.22 to -0.32 for employees aged between 25 and 55. These three papers use the regression approach; however, Einav et al. (2017) calculate the response elasticity to Medicare Part D by using the kink design. Specifically, they use two bunching models: a static model similar to Saez (2010) and our paper with an elasticity of approximately 0.034–0.049, and a dynamic response model developed in Einav et al. (2015) with an elasticity between 0.22 and 0.26.

Our research design is similar to the static response framework in Einav et al. (2017), and the estimated elasticities are a bit larger than those in these studies. One difference between our work and the literature is that although these papers have investigated the coinsurance rate from 25% to 0%, the coinsurance rate in our research setting jumps from 30% to 100%. Another departure from the literature is that we investigate the setting in the poor population in a developing country (i.e., rural residents in China).[6] Hence, the response to health insurance could be higher than their counterparts in the rich countries.

Sensitivity to the choices of parameters. We now examine whether our findings are sensitive to the choices of parameters in the baseline specification: the polynomial order, the bandwidth of the excluded region, and the reference points. In Appendix Table A4, we experiment with two more flexible polynomial functions, namely, sixth and seventh orders, respectively. We consider four sizes of excluded regions (i.e., ± 2 bins of the baseline value) in Table A5. Finally, we compare the results without controlling for the reference points and the results that control for the integers in Table A6. Different specifications generate quite stable estimates of normalized excess bunching and elasticities. In addition, by comparing the estimated elasticities in Table 2 and Table A6, we conclude that the excess bunching barely changes when we exclude the controls for multiples of 5. Hence, the significant bunching size is largely from the kinked policy instead of the reference point effect.

5.2 Joint Estimates of Elasticities and Frictions

People often face optimization frictions when making decisions, for example, adjustment costs, inertia, and incomplete information. This phenomenon is particularly relevant

[6] In 2014, the Gross Domestic Product per capita was $54,597$ USD in the United States and $7,589$ USD in China, respectively (International Monetary Fund, World Economic Outlook Database, 2015). However, the average annual gross income per capita in our research county was only 12,700 RMB (approximately $2,082$ USD) (China County Statistical Yearbook, 2014). Our analysis focuses on outpatient expenditure on nonspecial and nonchronic diseases in village clinics. For individuals in this sample, the total outpatient expenditure in 2014 was approximately 1.37% of the monthly gross income and approximately 0.11% of the annual gross income. However, rural residents also visited township health centers and general hospitals. According to Shi et al. (2018), the average expenditure on outpatient services in clinics and hospitals in our research county in 2014 was approximately 187 RMB, approximately 1.47% of the annual gross income.

in the case of rural China, where the residents have low education levels. We develop a framework in Section 3.3.1 by following Chetty et al. (2011) to jointly estimate the elasticity of response and degree of friction. Two estimates require two moment conditions. To this end, we explore the two sample periods with different coinsurance rates and kinks, that is, from March 9, 2012, to October 30, 2012, and from October 31, 2012, to December 31, 2014. With the two $\hat{\alpha}_n$ $(n = 1, 2)$ values estimated from these two sample periods in panels A and B of Table 2 and the two coinsurance rates c_n^0 (and $c_n^1 = 1$), we can then calculate the true elasticity α^t and friction δ from equation (9).

The estimates of α^t and δ are reported in Table 3, panel A. We observe that $\delta = 0.5127$, and this result indicates that greater than 50% of the population were facing optimization errors. After correcting for such frictions, we observe that $\alpha^t = 0.8265$, which almost doubles the estimated elasticity without frictions.

[Insert Table 3 Here]

One concern with these estimates is that the second period covers more months within a calendar year than the first period; hence, the seasonality of health shocks may lead to biases in the estimates. To address this concern, we reselect the same days within the calendar year in the second sample period: March 9, 2013, to October 30, 2013, and March 9, 2014, to October 30, 2014. The refined estimates of α^t and δ are reported in panel B of Table 3. We observe that $\delta = 0.3427$, and this result indicates that approximately one third of the population were facing optimization errors; and $\alpha^t = 0.6415$, which is approximately 40% larger than the estimated elasticity without frictions. This result suggests that if the reimbursement rate were reduced by 10% from 70%, the total expenditure (net of the GDP) per visit would decrease by approximately 5%. Additionally, a complete elimination of the reimbursement would cause the total expenditure (net of the GDP) per visit to decrease by 34.5%.

5.3 Heterogeneity

These estimates represent the average response of the studied population to the nonlinear health insurance policies. However, different individuals may respond differently. To this end, we present the heterogeneous responses across different groups, that is, males versus females, low–educated versus high–educated residents, and young versus old residents. The results are reported in Table 4. Columns 1 and 2 present the corresponding elasticities without frictions in our two sample periods, as in Table 2, whereas columns 3 and 4 show frictions, δ, and elasticities, α^t, after being corrected for optimization frictions, as observed in panel B of Table 3.

[Insert Table 4 Here]

Heterogeneity across genders. In panel A, we observe that males are more likely to have errors in making decisions than females (i.e., 0.6272 for males vs. 0.3821 for females) and are more responsive to the insurance policies than females (i.e., 1.0204 for males vs. 0.6803 for females). These results are in line with the literature that demonstrate that females are more averse to risks than males [for reviews, see Eckel and Grossman (2008), Croson and Gneezy (2009), and Bertrand (2011)]: more cautious in decision–making (resulting in lower errors) and less sensitive to negative health shocks (explaining the smaller elasticities observed).

Heterogeneity across education levels. We divide the sample into two groups: one with education levels below middle school (*low education group*), and the other with education levels above middle school (*high education group*). The results are reported in panel B. We observe that approximately one third of highly educated individuals face frictions in decision–making, but approximately two thirds of the low education group make errors in their decisions. This difference may reflect the cognitive ability in calculating the optimal medical expenditure. Additionally, less educated people are more elastic to changes in the coinsurance rate than more educated people (i.e., 1.0058 for the former and 0.6273 for the latter); these results reflect that the former make less income and, hence, are more financially constrained, increasing their sensitivity to the kinks than the latter.

Heterogeneity across age cohorts. In panel C, we study the differences between the young and old populations (the division is based on the national classification; that is, the old population is defined as those aged older than 50 for females or older than 60 for males). We observe that older residents are more likely to make decision errors than young people, reflecting the deteriorating cognitive skills when aging. We also observe that old people have much larger elasticities than young people. Generally, old people suffer more serious diseases and, hence, tend to be less responsive to price changes. However, in rural China, due to the insufficient pension system, old people mostly rely on their own working or children for support. (Peng, 2011) The insufficient financial resources make old people more sensitive to the kinks in the health insurance policies than their young counterparts.

6 Cost–Benefit Analysis and Policy Counterfactuals

With the estimates of the model parameters, we now proceed with a cost–benefit analysis of the health insurance plan in 2014; we calculate the optimal policy design with the same government expenditure and finally conduct a counterfactual experiment with the U.S. health insurance policy.

6.1 Cost–Benefit Analysis

The average benefit of the NRCMS program per person per visit is

$$U = \int \left(2m_{ij} - \frac{A_{ij}}{1+\frac{1}{\alpha}} (\frac{m_{ij}}{A_{ij}})^{1+\frac{1}{\alpha}} + E * I_i - c_{ij} m_{ij} \right) \, dF(A_{ij}), \tag{10}$$

where a $(1-\delta)$ fraction of population follow the optimal expenditure decision $m_{ij}(A_{ij})$ in equation (3), and the remaining population face optimization frictions and are randomly distributed across the expenditure distribution.

The government cost of the program per person per visit is

$$G = \int \left[(1 - c_{ij}) \, m_{ij} + E \times I_i \right] dF(A_{ij}). \tag{11}$$

By combining equations (10) and (11), we have the total net welfare gain W of the NRCMS as

$$\begin{aligned} \frac{W(\alpha, \delta, F(A_{ij}), c^0, c^1, m^*, N)}{N} &= U - G \\ &= \int \left[m_{ij} - \frac{A_{ij}}{1+\frac{1}{\alpha}} (\frac{m_{ij}}{A_{ij}})^{1+\frac{1}{\alpha}} \right] dF(A_{ij}), \end{aligned} \tag{12}$$

where N is the number of total visits.

$\{c^0, c^1, m^*\}$ are known from the policy details. N is calculated from the data directly. $\{\alpha, \delta\}$ and $dF(A_{ij})$ are estimated from the data. Hence, we can calculate the total net welfare gain W from equation (12). We conduct this cost–benefit analysis by using the most recent data, that is, 2014.

The results are reported in Table 5. We calculate that per person per visit, the average benefit and government cost of the NRCMS program in 2014 were 14.87 RMB and 10.09 RMB, respectively. In total, the program generated a 1.66 million RMB net welfare gain. These results suggest that the benefits of the current NRCMS policies significantly outweigh their costs. In our estimation framework, there is no risk–sharing and only a possible moral hazard concern. The channel generating the positive benefits of the insurance program is the discrepancy between medical expenditure and willingness to pay for the illness. Specifically, unlike developed countries, the willingness to pay of rural residents in China is quite low due to the low income level.[7] According to Wagstaff et al. (2009), only 7.5% and 2.6% of households in the non–NRCMS counties had visited a doctor in the last 2 weeks and received inpatient services in the last 12 months in 2003, respectively. The situations in our studied county are similar, with approximately 25.9% of NRCMS enrollees not having any medical spending in 2014 (Shi et al., 2018).

[7] In 2014, the average annual gross income per capita in our research county was only 3.8% of the Gross Domestic Product per capita in the United States.

This low willingness to pay leads to low medical expenditure in rural China. With the introduction of the NRCMS, patients can receive partial reimbursement, and then increase their medical expenditure, which lends to an improvement in well–being.

[Insert Table 5 Here]

6.2 Optimal Policy Design

To provide additional details on the design of health insurance contacts, we consider several policy counterfactuals; that is, different combinations of $\{c^0, c^1, m^*\}$. Given the many potential choices, we constrain our counterfactual policies by setting $c^1 = 1$ and manipulating $\{c^0, m^*\}$, but without increasing the government expenditure. Specifically, we vary c^0 from 5% to 40% with an interval of 5%, and then solve m^* (rounded to the nearest 0.5) so that the new combination $\{c^0, m^*\}$ costs the government the same as the benchmark in 2014 (i.e., $\{30\%, 12\}$).

The results are reported in Table 6. When c^0 increases from 5% to 40%, m^* also increases from 8.0 to 20.0, reflecting the tradeoff between the generosity of low coinsurance rates and that of high reimbursement ceilings, given that the government budget is fixed. The total net welfare gain W first increases with c^0 until $c^0 = 30\%$, and then decreases. The optimal $\{c^0, c^1, m^*\}$ is $c^0 = 30\%$, $c^1 = 100\%$, and $m^* = 12$, corresponding to the parameters of the policy implemented in 2014. These results suggest that within our model considerations, the current NRCMS policies are optimal for welfare.

[Insert Table 6 Here]

6.3 Counterfactual Using U.S. Policy

We now consider a further counterfactual by introducing an insurance schedule of a U.S. pattern into our studied population. The insurance system in the United States is fragmented; that is, different markets could offer different insurance contracts. We now consider plans in the Health Insurance Marketplaces created by the Affordable Care Act as an example. Specifically, a typical plan is set with a deductible, a coinsurance rate, and an out–of–pocket maximum. Insurers start to reimburse patients with the set coinsurance rate when the medical bills surpass the deductible threshold and cover all additional bills once the out–of–pocket maximum threshold is met. In our analysis, we restrict the coinsurance rate to 30% and vary the two thresholds to obtain a schedule that maximizes the total welfare at the same government budget as our focal plan carried out in 2014 in China.

Table 7 presents the results. The deductible threshold and out–of–pocket maximum threshold are 5.5 RMB and 8.5 RMB, respectively. The total welfare is calculated to be

approximately 1.28 million RMB, which is less than the 1.66 million RMB induced by the NRCMS policy in 2014.

[Insert Table 7 Here]

The intuition behind the reduced welfare in the counterfactual scenario is as follows. In our NRCMS plan, the insurance covers a large fraction of the population, with each individual receiving a moderate reimbursement. In the counterfactual plan, a large amount of the budget is distributed to a small number of patients with high spending, resulting in the remaining low–spending patients receiving a smaller reimbursement than they obtain in the carried–out plan. Additionally, patients pay the general diagnosis fee out of pocket under the counterfactual plan, which is instead covered largely by the carried–out policy. In 2014, the general diagnosis fee was 5 RMB, with 4.5 RMB covered by the carried–out plan. As a result, patients would pay more under the counterfactual plan if the medical cost is not too high. For example, for the patient with the 20 RMB total medical expenditure including diagnosis service fees, the total out–of–pocket payment under the plan actually carried out is $12 \times 0.3 + (20 - 4.5 - 12) = 7.1$ RMB, whereas that under the counterfactual plan is 8.5 RMB.

7 Conclusion

This article has used a bunching method to explore the expenditure responses to the kink in health insurance policies. According to our review of the literature, the studies on expenditure responses have focused exclusively on policies in the developed countries. By contrast, this paper investigate a population in a developing country, where public insurance coverage is of greater need.

The basis for our work is a sample of rural residents in a southwestern county in China. We first provide graphical evidence to illustrate that the distribution of medical expenditures bunches at the kink point of the insurance. The spike is only observed for the periods when the kinks exist, but not for periods when coinsurance rates are flat. We next estimate that the elasticity of the expenditure with respect to the coinsurance rate without optimization frictions is approximately 0.44. These findings are insensitive to the choice of model parameters. By using a model with optimization frictions, we estimate the degree of frictions to be 0.34, indicating that approximately one third of the population are inconsistent with optimization when making decisions. This model with frictions leads to a larger estimation of the elasticity, suggesting that an elimination of reimbursement would decrease medical expenditures per visit by 34.5%. We finally conduct a cost–benefit analysis and simulate several policy counterfactuals. The insurance is estimated to generate a net welfare gain of 1.66 million RMB in 2014, which equals 47.41% of the program cost. The counterfactual analysis indicates that with the govern-

ment budget unchanged, the current policy delivers the largest welfare gain among all counterfactual plans.

Our heterogeneity analysis has shown that less educated individuals and older people are more likely to make decision errors than their counterparts largely because of their limited cognitive skills. It is common for people to have difficulties understanding insurance. People with low cognitive ability are usually less healthy and have greater needs for medical care; thus, providing them with information and explanations regarding insurance policies should improve their decision–making skills regarding their use of medical services.

Bibliography

Abaluck, J., Gruber, J., and Swanson, A. (2018). Prescription Drug Use under Medicare Part D: A Linear Model of Nonlinear Budget Sets. *Journal of Public Economics*, 164:106–138.

Bai, C.-e. and Wu, B. (2014). Health Insurance and Consumption: Evidence from China's New Cooperative Medical Scheme. *Journal of Comparative Economics*, 42(2):450–469.

Bajari, P., Hong, H., Park, M., and Town, R. (2017). Estimating Price Sensitivity of Economic Agents Using Discontinuity in Nonlinear Contracts. *Quantitative Economics*, 8(2):397–433.

Bastani, S. and Selin, H. (2014). Bunching and Non-Bunching at Kink Points of the Swedish Tax Schedule. *Journal of Public Economics*, 109:36–49.

Bertrand, M. (2011). New Perspectives on Gender. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4b, chapter 17, pages 1543–1590. Elsevier B.V., Amsterdam: North-Holland.

Best, M. C., Cloyne, J., Ilzetzki, E., and Kleven, H. (2018). Estimating the Elasticity of Intertemporal Substitution Using Mortgage Notches.

Best, M. C. and Kleven, H. J. (2018). Housing Market Responses to Transaction Taxes: Evidence From Notches and Stimulus in the U.K. *Review of Economic Studies*, 85(1):157–193.

Brown, K. M. (2013). The link between pensions and retirement timing: Lessons from California teachers. *Journal of Public Economics*, 98:1–14.

Chetty, R., Friedman, J. N., Olsen, T., and Pistaferri, L. (2011). Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records. *The Quarterly Journal of Economics*, 126(2):749–804.

Chetty, R., Friedman, J. N., and Saez, E. (2013). Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings. *American Economic Review*, 103(7):2683–2721.

Croson, R. and Gneezy, U. (2009). Gender Differences in Preferences. *Journal of Economic Literature*, 47(2):448–474.

Eckel, C. C. and Grossman, P. J. (2008). Men, Women and Risk Aversion: Experimental Evidence. In Plott, C. and Smith, V., editors, *Handbook Of Experimental Economics Results*, number 1, chapter 113, pages 1061–1073. Elsevier B.V., New York.

Eichner, M. J. (1998). The Demand for Medical Care: What People Pay Does Matter. *The American Economic Review*, 88(2):117–121.

Einav, L., Finkelstein, A., and Schrimpf, P. (2015). The Response of Drug Expenditure to Nonlinear Contract Design: Evidence from Medicare Part D. *The Quarterly Journal of Economics*, 130(2):841–899.

Einav, L., Finkelstein, A., and Schrimpf, P. (2017). Bunching at the Kink: Implications for Spending Responses to Health Insurance Contracts. *Journal of Public Economics*, 146:27–40.

Giedion, U., Alfonso, E. A., and Diza, Y. (2013). The Impact of Universal Coverage Schemes in the Developing World: A Review of the Existing Evidence. Technical report, The World Bank., Washington DC.

Grubb, M. D. and Osborne, M. (2015). Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock. *American Economic Review*, 105(1):234–271.

Harju, J., Matikka, T., and Rauhanen, T. (2016). The Effects of Size-Based Regulation on Small Firms: Evidence from VAT Threshold.

Ito, K. (2014). Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing. *American Economic Review*, 104(2):537–563.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8:435–464.

Kleven, H. J. and Waseem, M. (2013). Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan. *The Quarterly Journal of Economics*, 128(2):669–723.

Liu, L., Lockwood, B., and Almunia, M. (2017). VAT Notches, Voluntary Registration , and Bunching: Theory and UK Evidence.

Manoli, D. and Weber, A. (2016). Nonparametric Evidence on the Effects of Financial Incentives on Retirement Decisions. *American Economic Journal: Economic Policy*, 8(4):160–182.

Nguyen, T. K. P. and Akal, A. (2003). Recent Advances in Social Health Insurance in Vietnam: A Comprehensive Review of Recent Health Insurance Regulations.

Onji, K. (2009). The Response of Firms to Eligibility Thresholds: Evidence from the Japanese Value-Added Tax. *Journal of Public Economics*, 93:766–775.

Peng, X. (2011). China's Demographic History and Future Challenges. *Science*, 333(6042):581–587.

Phelps, C. E. and Newhouse, J. P. (1974). Coinsurance, the Price of Time, and the Demand for Medical Services. *The Review of Economics and Statistics*, 56(3):334–342.

Saez, E. (2010). Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy.*, 2(3):180–212.

Scitovsky, A. A. and Snyder, N. M. (1972). Effect of Coinsurance on Use of Physician Services. *Social Security Bulletin*, 35(6):3–19.

Shi, J., Yao, Y., and Liu, G. (2018). Modeling Individual Health Care Expenditures in China : Evidence to Assist Payment Reform in Public Insurance. *Health Economics*, (May):1–18.

Stuckler, D., Feigl, A. B., Basu, S., and McKee, M. (2010). The Political Economy of Universal Health Coverage.

Van Tien, T., Phuong, H. T., Mathauer, I., and Nguyen, T. K. P. (2011). A Health Financing Review of Vietnam with a Focus on Social Health Insurance.

Wagstaff, A., Lindelow, M., Jun, G., Ling, X., and Juncheng, Q. (2009). Extending Health Insurance to the Rural Population : An Impact Evaluation of China's New Cooperative Medical Scheme. *Journal of Health Economics*, 28:1–19.

Figures and Tables

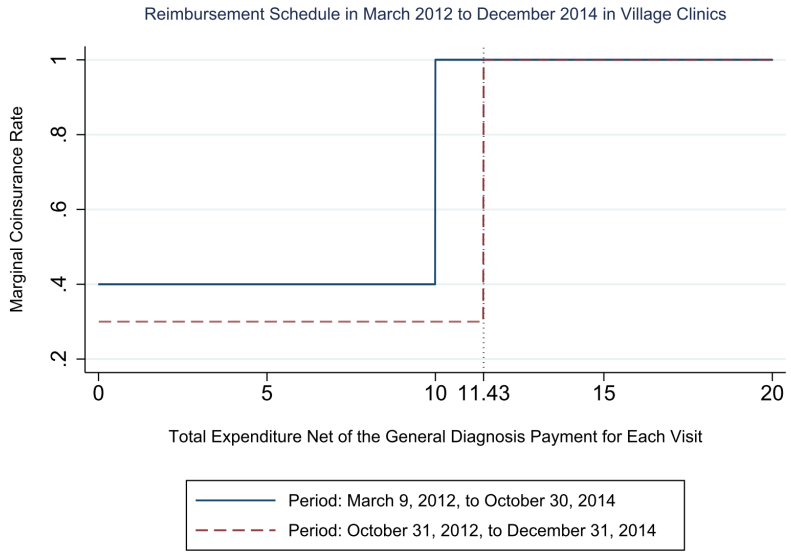


Figure 1

Reimbursement Schedule in 2012–2014

Notes: This figure shows the reimbursement schedule in the periods from March 9, 2012, to October 30, 2012, and from October 31, 2012, to December 31, 2014. The solid line plots the schedule in the first period and the dashed line plots the schedule in the second one. Patients could purchase outpatient services at 40% (30%) coinsurance rate in the period from March 9, 2012, to October 30, 2012, (from October 31, 2012, to December 31, 2014,) when the total expenditure net of the GDP was below 10 RMB (11.5 RMB) in each visit. The coinsurance rate changed to 100% when the total expenditure net of the GDP was above the corresponding threshold.

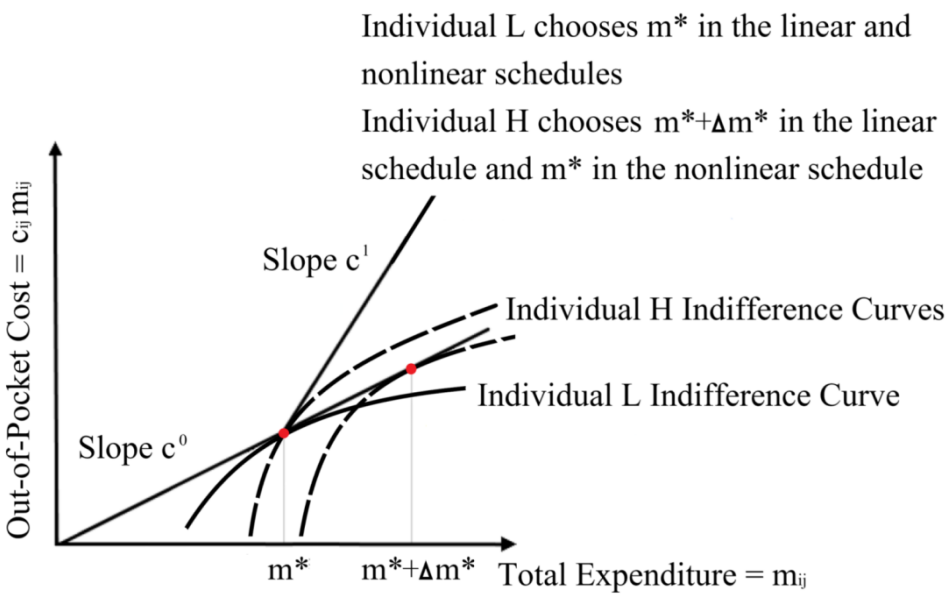


Figure 2

Bunching in Frictionless Setting

Notes: This figure illustrates how people respond to nonlinear budget sets. The solid line demonstrates the nonlinear budget set with a kink at m^* created by the nonlinear reimbursement schedule. The dashed line shows the budget set when the reimbursement rate is constant at c^0. The solid curve represents the indifference curve of the individual L who bunches at the kink under the nonlinear reimbursement schedule but with the lowest level of A_i; his/her indifference curve is tangent to both the linear and nonlinear budget set and hence he/she is not affected by the introduction of the kink. The dashed curves represent the indifference curves of the individual H who bunches at the kink under the nonlinear schedule but with the highest level of A_i; he/she consumes $m^* + \Delta m^*$ in the linear scenario and m^* in the nonlinear one with his/her indifference curve tangent to the new budget set. Any individual whose indifference curve is tangent to the linear budget set at an expenditure level between m^* and $m^* + \Delta m^*$ decreases his/her spending to m^* under the nonlinear schedule as well. Therefore, the density distribution shows a bunching at the kink in the nonlinear scenario.

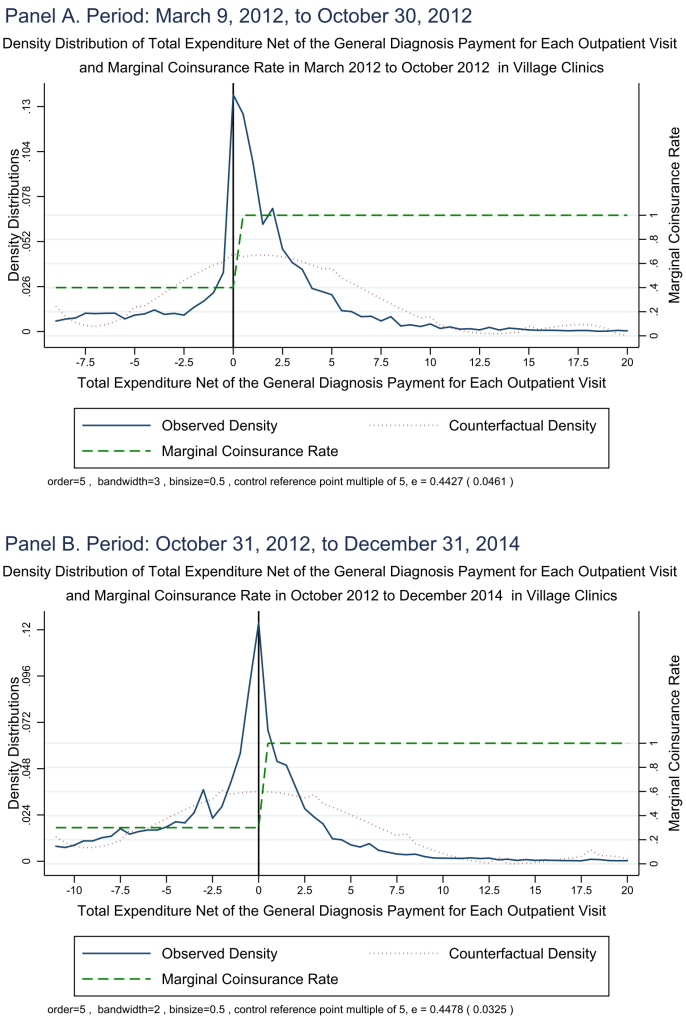


Figure 3

Density Distribution of Total Expenditure Net of the GDP in 2012–2014

Notes: These figures show the density distributions of the per visit total expenditure net of the GDP around the kink point (demarcated by the vertical line at 0). Panel A depicts the densities in the period from March 9, 2012, to October 30, 2012, and panel B shows those in the period from October 31, 2012, to December 31, 2014. The solid curves display the observed densities in 0.5 RMB bins, and the dotted curves display the counterfactual densities by excluding a window of 3 RMB (2 RMB) centered around the kink point, controlling for multiples of 5, and fitting a fifth–order polynomial to the observed distributions in panel A (B).

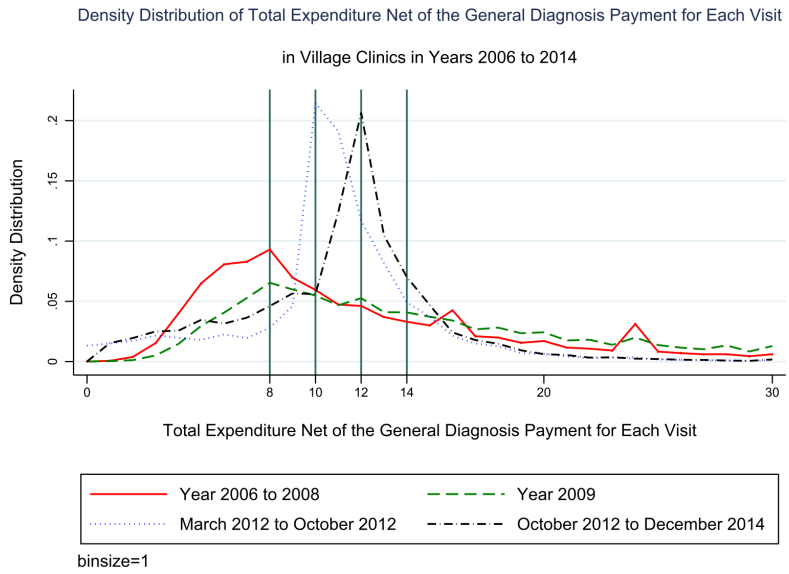


Figure 4

Density Distribution of Total Expenditure Net of the GDP in 2006–2014

Notes: This figure plots the density distributions of the per visit total expenditure net of the GDP in 2006–2014 in 1 RMB bins.

Table 1A. Summary Statistics for the Whole Sample

Panel A. Sample Sizes

Variables	(1) N	(2) Percentage
Whole Sample	4,130,744	100%
Village Clinics Subsample	1,923,537	46.57%
2006 Subsample	76,309	1.85%
2007 Subsample	75,918	1.84%
2008 Subsample	59,772	1.45%
2009 Subsample	92,544	2.24%
2010 Subsample	373,109	9.03%
2011 Subsample	580,381	14.05%
2012 Subsample	866,261	20.97%
2013 Subsample	1,016,099	24.60%
2014 Subsample	990,351	23.98%
Special and Chronic Diseases Subsample	54,612	1.32%
NonSpecial and NonChronic Diseases Subsample	4,076,132	98.68%

Panel B. Costs and Payments Variables

Variables	(1) N	(2) mean	(3) sd	(4) min	(5) max
Total Expenditure	4,130,744	30.9766	36.5743	0	3010
Deposits Used from the Family Account	4,130,744	3.0464	10.6989	0	350
Reimbursement Received	4,130,744	12.1755	15.1492	0	150
The General Diagnosis Payment (GDP) Received	4,130,744	5.3432	4.6056	0	10
Total Expenditure net of the GDP	4,130,744	25.6335	35.1790	0	3000

Notes: This table displays the summary statistics for the whole sample of the administrative data on outpatient service visits. Panel A lists the number of observations for the whole sample and the subsamples of village clinics, year 2006, ..., year 2014, special and chronic diseases, and nonspecial and nonchronic diseases. Panel B presents the summary statistics for all relevant costs and payments variables in the whole sample. Variables Total Expenditure, Deposits Used from the Family Account, Reimbursement Received, The General Diagnosis Payment (GDP) Received, and Total Expenditure net of the GDP denote the corresponding monetary amounts in each visit.

Table 1B. Summary Statistics for the Estimation Subsamples: March 9, 2012, to December 31, 2014

Variables	(1) N	(2) mean	(3) sd	(4) min	(5) max
Panel A. Demographics					
Male	143,176	0.4727	0.4993	0	1
Single	132,949	0.1292	0.3355	0	1
Married	132,949	0.7595	0.4274	0	1
Widowed	132,949	0.0945	0.2925	0	1
Divorced	132,949	0.0167	0.1283	0	1
Years of Schooling	133,513	6.3494	3.5901	0	18
High Education Level	133,513	0.4273	0.4947	0	1
Age	143,148	48.2754	18.7878	1	124
Old Individual	143,148	0.3698	0.4827	0	1
Panel B. Costs and Payments					
Total Expenditure	984,181	13.9037	4.8778	0.20	1,017
Deposits Used from the Family Account	984,181	0.2731	2.0446	0	139.60
Reimbursement Received	984,181	6.5033	1.9989	0.10	8
The General Diagnosis Payment (GDP) Received	984,181	2.6036	2.1424	0	4.50
Total Expenditure net of the GDP	984,181	11.3002	5.1356	0.10	1,012.50

Notes: This table presents the summary statistics for the estimation subsample from March 9, 2012, to December 31, 2014. Panel A displays the statistics for demographics variables. Variables Male, Single, Married, Widowed, Divorced, High Education Level, and Old Individual are dummies. High Education Level variable indicates whether an individual has attended middle school or higher education. Old Individual variable indicates whether an individual is older than 50 for female or 60 for male. Variable Years of Schooling is generated based on the variable Highest Education Level Attended. Variable Age is calculated as the calendar year age, i.e. the difference between the year getting treated and the year of birth. Panel B documents the statistics for costs and payments variables, which denote the corresponding monetary amounts in each visit.

Table 2. Baseline Results

	(1) α	(2) Δm^*
Panel A. Period: March 9, 2012, to October 30, 2012		
	0.4427	2.3130
	(0.0474)	(0.2436)
Panel B. Period: October 31, 2012, to December 31, 2014		
	0.4478	3.2186
	(0.0293)	(0.2127)

Notes: This table shows the elasticity and normalized excess bunching estimates in the studied periods from March 9, 2012, to October 30, 2012, in panel A and from October 31, 2012, to December 31, 2014, in panel B, with standard errors in parenthesis. The results are computed by employing the empirical model of Chetty et al. (2011) in the subsamples with the total health care expenditure net of the GDP ranging from 0 RMB to 30 RMB in panel A and from 0 RMB to 32 RMB in panel B, respectively. The corresponding counterfactual density distribution in the panel A (panel B) is estimated by excluding a window of 3 RMB (2 RMB) centered around the kink point 10 RMB (12 RMB), controlling for multiples of 5 reference points, and fitting a fifth–degree polynomial to the observed density.

Table 3. Estimates for True Elasticity and Friction Fraction

			(1) α	(2) δ
Panel A.	Period 1:	March 9, 2012, to October 30, 2012;		
	Period 2:	October 31, 2012, to December 31, 2014.		
			0.8265	0.5127
			(0.3981)	(0.1963)
Panel B.	Period 1:	March 9, 2012, to October 30, 2012;		
	Period 2:	March 9, 2013, to October 30, 2013, and March 9, 2014, to October 30, 2014.		
			0.6415	0.3427
			(0.3002)	(0.1299)

Notes: This table shows the true elasticity and friction fraction estimates with standard errors in parenthesis when patients are assumed to face frictions in making decisions. The elasticity in column (1) and the friction fraction in column (2) are solved from equation (9) by using two periods with different reimbursement schedules. Panel A employs the periods from March 9, 2012, to October 30, 2012, and from October 31, 2012, to December 31, 2014; Panel B uses the period from March 9, 2012, to October 30, 2012, as the first one, and the periods from March 9, 2013, to October 30, 2013, and from March 9, 2014, to October 30, 2014, as the second one.

Table 4. Heterogeneity of True Elasticity and Friction Fraction

	(1) $\hat{\alpha}_1$	(2) $\hat{\alpha}_2$	(3) α	(4) δ
Panel A. Sex Subgroups				
Female	0.4453	0.4484	0.6803	0.3821
	(0.0481)	(0.0278)	(0.3297)	(0.1379)
Male	0.4415	0.4473	1.0204	0.6272
	(0.0469)	(0.0266)	(0.4730)	(0.2639)
Panel B. Education Level Subgroups				
Low	0.4419	0.4495	1.0058	0.6181
	(0.0463)	(0.0307)	(0.4813)	(0.2580)
High	0.4441	0.4465	0.6273	0.3231
	(0.0500)	(0.0202)	(0.2688)	(0.1003)
Panel C. Age Subgroups				
Young	0.4432	0.4473	0.7534	0.4550
	(0.0497)	(0.0270)	(0.3641)	(0.1923)
Old	0.4402	0.4498	1.1464	0.6781
	(0.0455)	(0.0314)	(0.5237)	(0.3172)

Notes: This table shows the heterogeneity of the bunching response across sex (in panel A), education level (in panel B), and age (in Panel C) subgroups. Column (1) presents the estimated elasticities in the period from March 9, 2012, to October 30, 2012; column (2) presents the elasticities in the periods from March 9, 2013, to October 30, 2013, and from March 9, 2014, to October 30, 2014. The elasticities in column (1) [(2)] are estimated with the empirical model of Chetty et al. (2011). We exclude a window of 3 RMB (2 RMB) centered around the kink point 10 RMB (12 RMB), control for multiples of 5 reference points, and fit a fifth–degree polynomial to the observed density. Columns (3) and (4) show true elasticities and friction fractions estimated from equation (9) by using the elasticities in columns (1) and (2). Standard errors are presented in parenthesis. People are defined as single if they are unmarried, widowed or divorced, with high education level if they have attended middle school or higher education, and old if they are older than 50 for females or 60 for males.

Table 5. Cost–Benefit Analysis in Year 2014

(1) α	(2) δ	(3) c_0	(4) c_1	(5) m^*	(6) N	(7) U	(8) G	(9) W
0.6415	0.3427	0.3	1	12	347,522	14.8691	10.0870	1,661,878
(0.3002)	(0.1299)							

Notes: This table shows the cost–benefit analysis of the reimbursement schedule in Year 2014. Columns (1)–(9) present the true elasticity, friction fraction, marginal coinsurance rate below the kink point, marginal coinsurance rate above the kink point, kink point, total number of visits, benefit per visit, government cost per visit, and total net welfare gain. Standard errors are presented in parenthesis. The benefit per visit and government cost per visit are calculated from equations (10) and (11). The total net welfare gain is estimated from equation (12).

Table 6. Counterfactual Reimbursement Schedules in 2014

(1) c_0	(2) c_1	(3) m^*	(4) W
0.05	1	8.0	1,614,019
0.10	1	8.5	1,631,773
0.15	1	9.5	1,636,227
0.20	1	10.0	1,652,607
0.25	1	10.5	1,656,643
0.30	1	12.0	1,661,878
0.35	1	14.0	1,649,133
0.40	1	20.0	1,619,050

Notes: This table evaluates different counterfactual reimbursement schedules costing the government the same amount of money as the carried–out policy in 2014. The true elasticity and friction fraction are assumed to be of the estimated values 0.6415 and 0.3427 as in Table 3, respectively. Columns (1)–(4) present the counterfactual marginal coinsurance rate below the kink point, the marginal coinsurance rate above the kink point, the counterfactual kink point, and the total net welfare gain. The total net welfare gain is estimated from equation (12).

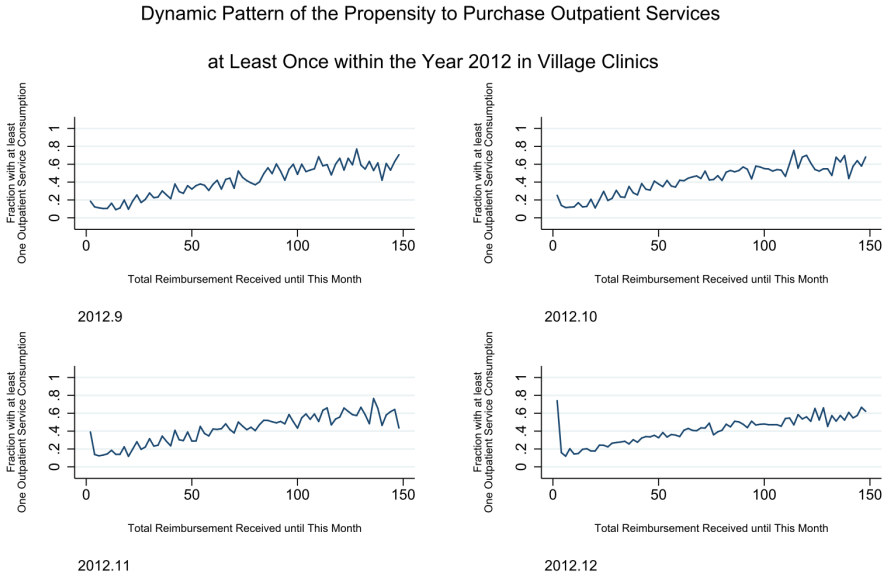
Table 7. Counterfactual Insurance Schedule of U.S. Pattern

(1) Deductible	(2) Out–of–Pocket Maximum	(3) c	(4) W
5.5	8.5	0.3	1,277,264

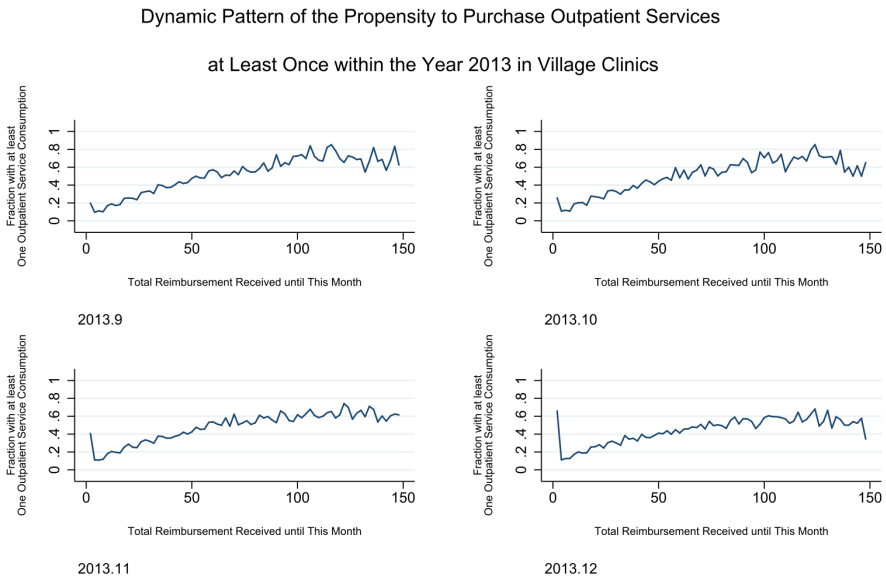
Notes: This table presents the insurance schedule of U.S. pattern which costs the government the same amount of money as the policy carried out in 2014 and maximizes the total net welfare gain. The true elasticity and friction fraction are assumed to be of the estimated values 0.6415 and 0.3427 as in Table 3, respectively. The coinsurance rate is assumed to be 0.3. Columns (1), (2) and (4) present the deductible threshold, the out–of–pocket maximum threshold, and the total net welfare gain. The total net welfare gain is estimated from equation (12).

Appendix

Panel A. Year 2012



Panel B. Year 2013



Panel C. Year 2014

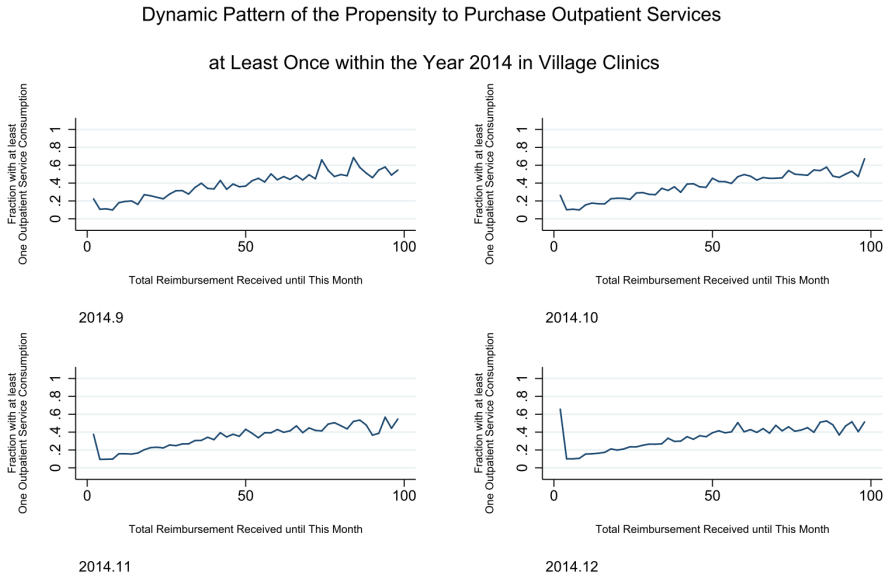


Figure A1

Density Distribution of Total Expenditure Net of the GDP in 2012–2014

Notes: These figures show the dynamic buying patterns of patients for outpatient services on nonspecial and nonchronic diseases in village clinics in 2012–2014. Panels A, B, and C plot the patterns for the last 4 months of years 2012, 2013 and 2014, respectively. The x–axis reports the total reimbursement received until the corresponding month in 2 RMB bins. The y–axis reports the propensity to purchase outpatient services at least once in the corresponding month, which is estimated by the share of individuals with at least one visit to village clinics during that month.

Table A1. Population Compositions for the Studied County and All Counties in China

	Resident Population (1)	Percentage of Male Population (2)	Percentage of Population of Han Race (3)	Percentage of Population Aged below 20 (4)	Percentage of Population Aged between 20 and 59 (5)	Percentage of Population Aged Older than 59 (6)	Population with Local Hukou (7)	Percentage of Population with Rural Local Hukou (8)
Panel A. The Studied County								
	412,758	0.4974	0.9983	0.1728	0.6471	0.1802	422,740	0.7924
Panel B. Percentiles for All Counties in China								
1%	20,319	0.4852	0.0167	0.1229	0.4852	0.0593	20,237	0.0631
5%	52,741	0.4947	0.0834	0.1525	0.5224	0.0781	53,070	0.1726
10%	104,609	0.4992	0.2629	0.1695	0.5450	0.0917	105,102	0.3025
25%	221,505	0.5052	0.8748	0.1997	0.5856	0.1119	222,229	0.6008
50%	379,941	0.5118	0.9849	0.2413	0.6239	0.1299	386,731	0.7996
75%	625,119	0.5197	0.9971	0.2878	0.6638	0.1515	625,216	0.8761
90%	900,581	0.5285	0.9992	0.3296	0.6954	0.1707	939,590	0.9155
95%	1,101,077	0.5349	0.9996	0.3572	0.7100	0.1818	1,132,528	0.9300
99%	1,558,663	0.5569	0.9999	0.4144	0.7465	0.2063	1,598,895	0.9583

Notes: Figures for all counties in China and the studied county are collected from the China 2010 Population Census. The percentages in columns (2) to (6) are calculated by dividing the corresponding subpopulation by the total resident population. The percentages in column (8) are calculated by dividing the corresponding subpopulation by the total population with local hukou.

Table A2. Health Levels and Education Levels for the Studied County and All Counties in China

	Percentage of Very Healthy Population (1)	Percentage of Basically Healthy Population (2)	Percentage of Unhealthy but Capable of Caring for Oneself Population (3)	Percentage of Very Unhealthy Population (4)	Percentage of Illiterate Population (5)	Percentage of Population with Primary School or Below Education Level (6)	Percentage of Population with Middle School Education Level (7)	Percentage of Population with High School Education Level (8)	Percentage of Population with College or Above Education Level (9)	Average Schooling Years (10)
Panel A. The Studied County										
	0.4160	0.4580	0.1107	0.0153	0.0595	0.4095	0.3989	0.1347	0.0569	8.31
Panel B. Percentiles for All Counties in China										
1%	0.0000	0.1794	0.0000	0.0000	0.0067	0.1050	0.0881	0.0202	0.0183	3.96
5%	0.2527	0.3307	0.0591	0.0000	0.0115	0.1504	0.1992	0.0530	0.0257	6.38
10%	0.3121	0.3445	0.0791	0.0000	0.0155	0.1847	0.2720	0.0713	0.0298	7.27
25%	0.3519	0.3522	0.1040	0.0164	0.0248	0.2715	0.3481	0.0991	0.0380	8.06
50%	0.3919	0.3711	0.1319	0.0259	0.0420	0.3553	0.4130	0.1303	0.0528	8.62
75%	0.4190	0.4545	0.1556	0.0342	0.0729	0.4451	0.4717	0.1757	0.0971	9.38
90%	0.4429	0.4614	0.1775	0.0430	0.1249	0.5728	0.5193	0.2371	0.1921	10.69
95%	0.4605	0.5119	0.1883	0.0490	0.1874	0.7094	0.5438	0.2693	0.2600	11.29
99%	0.5007	0.5130	0.2183	0.0683	0.3956	0.8513	0.5839	0.3117	0.3834	12.28

Notes: Figures for all counties in China and the studied county are collected from the China 2010 Population Census. The percentages in columns (1) to (4) refer to the corresponding percentages in population aged 60 or above. The percentages in column (5) refer to the corresponding percentages in population aged 15 or above. The percentages in columns (6) to (10) refer to the corresponding percentages in population aged 6 or above.

Table A3. Employment, Income, and Living Standard Levels for the Studied County and All Counties in China

	Percentage of Population Working Last Week (1)	Percentage of Population Living on Labor Income (2)	Percentage of Population Living on Pension (3)	Percentage of Population Living on Unemployment Insurance (4)	Percentage of Population Living on Minimum Living Allowance (5)	Percentage of Population Living on Property Income (6)	Percentage of Population Living on Family Support (7)	Percentage of Population Living in Bungalows (8)	Percentage of Population Renting Houses (9)
Panel A. The Studied County									
	0.7010	0.7010	0.0461	0.0022	0.0112	0.0067	0.2059	0.6739	0.0943
Panel B. Percentiles for All Counties in China									
1%	0.5264	0.5330	0.0000	0.0000	0.0000	0.0000	0.1143	0.0314	0.0000
5%	0.5807	0.5798	0.0044	0.0000	0.0015	0.0000	0.1560	0.0694	0.0093
10%	0.6078	0.6080	0.0083	0.0000	0.0034	0.0000	0.1723	0.1184	0.0161
25%	0.6557	0.6536	0.0147	0.0000	0.0062	0.0006	0.1978	0.2642	0.0317
50%	0.6942	0.6904	0.0261	0.0000	0.0101	0.0024	0.2246	0.5443	0.0639
75%	0.7187	0.7146	0.0605	0.0007	0.0156	0.0054	0.2596	0.7905	0.1324
90%	0.7376	0.7320	0.1380	0.0018	0.0246	0.0095	0.2968	0.8864	0.2593
95%	0.7471	0.7436	0.1898	0.0030	0.0332	0.0131	0.3214	0.9211	0.3508
99%	0.7722	0.7656	0.2765	0.0064	0.0600	0.0198	0.3858	0.9825	0.5215

Notes: Figures for all counties in China and the studied county are collected from the China 2010 Population Census. The percentages in columns (1) to (7) refer to the corresponding percentages in population aged 15 or above.

Table A4. Sensitivity Analysis with Respect to the Polynomial Order

	(1)	(2)	(3)	(4)
	order = 6		order = 7	
	Δm^*	α	Δm^*	α
Panel A. Period:	March 9, 2012, to October 30, 2012			
	2.3655	0.4518	2.1211	0.4093
	(0.3867)	(0.0739)	(0.2049)	(0.0403)
Panel B. Period:	October 31, 2012, to December 31, 2014			
	3.2562	0.4524	2.6799	0.3799
	(0.2385)	(0.0324)	(0.1765)	(0.0249)

Notes: This table shows the sensitivity analyses with respect to the choices of the polynomial order in the studied periods from March 9, 2012, to October 30, 2012, in panel A and from October 31, 2012, to December 31, 2014, in panel B, respectively. Columns (1) and (3) present the normalized excess bunching estimates, and columns (2) and (4) display the elasticity estimates. The results are estimated with the empirical model of Chetty et al. (2011) by excluding a window of 3 RMB (2 RMB) centered around the kink point 10 RMB (12 RMB), controlling for multiples of 5 reference points, and fitting the corresponding degree polynomial to the observed density.

Table A5. Sensitivity Analysis with Respect to the Bandwidth of the Excluded Region

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Period: March 9, 2012, to October 30, 2012							
bandwidth=2		bandwidth=2.5		bandwidth=3.5		bandwidth=4	
Δm^*	α	Δm^*	α	Δm^*	α	Δm^*	α
2.2648	0.4344	2.2992	0.4403	2.4339	0.4635	2.3882	0.4557
(0.2563)	(0.0491)	(0.2561)	(0.0494)	(0.1922)	(0.0377)	(0.2337)	(0.0461)
Panel B. Period: October 31, 2012, to December 31, 2014							
bandwidth=1		bandwidth=1.5		bandwidth=2.5		bandwidth=3	
Δm^*	α	Δm^*	α	Δm^*	α	Δm^*	α
2.4584	0.3512	2.9735	0.4172	3.2014	0.4457	3.4950	0.4817
(0.2804)	(0.0387)	(0.2649)	(0.0363)	(0.1883)	(0.0262)	(0.2250)	(0.0315)

Notes: This table shows the sensitivity analyses with respect to the choices of the bandwidth of the excluded region in the studied periods from March 9, 2012, to October 30, 2012, in panel A and from October 31, 2012, to December 31, 2014, in panel B, respectively. Columns (1), (3), (5), and (7) present the normalized excess bunching estimates and columns (2), (4), (6), and (8) display the elasticity estimates. The results are estimated with the empirical model of Chetty et al. (2011) by excluding a window of the corresponding bandwidth centered around the kink point 10 RMB (12 RMB), controlling for multiples of 5 reference points, and fitting a fifth–degree polynomial to the observed density.

Table A6. Sensitivity Analysis with Respect to the Reference Point Controls

	(1)	(2)	(3)	(4)
	No Controls		Controlling Integers	
	Δm^*	α	Δm^*	α
Panel A. Period:	March 9, 2012, to October 30, 2012			
	2.3144	0.4429	2.3015	0.4407
	(0.2420)	(0.0470)	(0.2434)	(0.0474)
Panel B. Period:	October 31, 2012, to December 31, 2014			
	3.2106	0.4468	3.2174	0.4476
	(0.2138)	(0.0294)	(0.2136)	(0.0294)

Notes: This table shows the sensitivity analyses with respect to the choices of the reference points controls in the studied periods from March 9, 2012, to October 30, 2012, in panel A and from October 31, 2012, to December 31, 2014, in panel B, respectively. Columns (1) and (3) present the normalized excess bunching estimates and columns (2) and (4) display the elasticity estimates. The results are estimated with the empirical model of Chetty et al. (2011) by excluding a window of 3 RMB (2 RMB) centered around the kink point 10 RMB (12 RMB), controlling for the corresponding reference points, and fitting a fifth–degree polynomial to the observed density.