# Essays on China's Health Care and Skill Upgrading Decisions

Chen Zhuang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Yuya Takahashi, Chair

Xu Tan

Jing Tao

Vanessa Oddo

Program Authorized to Offer Degree:
Economics

University of Washington

## Abstract

Essays on China's Health Care and Skill Upgrading Decisions

Chen Zhuang

Chair of the Supervisory Committee:
Dr. Yuya Takahashi
Department of Economics

In this dissertation, I explore three seemingly disparate but inter-connected questions related to China's strategies for human capital maintenance and accumulation. Hereinafter, a brief outline of each of the three chapters is given.

In the first chapter, I investigate how efficiently the Chinese government designs a public health insurance program, namely the New Rural Cooperative Medical Scheme (NRCMS), in a hierarchical medical system. To study the welfare effects of adjusting the plan's policies, I construct and structurally estimate a two-stage choice model based on a set of medical claims data of all inpatients enrolled in the NRCMS between 2012 and 2014 from a representative county of China.[1] In the first stage, a patient chooses one of the hospitals available to treat his or her disease with imperfect information of health status; in the second stage, the patient decides a spending level according to his or her realized health risk, moral hazard type, and the cost-sharing structure in the chosen hospital. According to my model, patients can exhibit diverse risk attitudes that affect their hospital choices, and a higher-deductible plan can improve social welfare if generosity increases mistrust among patients with minor diseases. Indeed, empirical results confirm these concerns, and eventually lead me to suggest the Chinese government to increase both the deductible and reimbursement maximum for social welfare gains as well as public acceptance. In this first chapter, I focus

[1]The data are provided by Dr. Julie Shi and Dr. Xi Wang in the School of Economics at Peking University. They also provide helpful comments on my manuscript. Emails: jshi@pku.edu.cn and wang.x@pku.edu.cn, respectively.

on the demand side of China's healthcare system to understand how China maintains the health capital of its rural residents. In the meanwhile, I identify efficiency issues caused by how consumers respond to the way in which China manages health capital, and propose policy recommendations to improve economic efficiency.

I realize that the efficiency issues in China's healthcare system come from not only the demand side but also the supply side. Thus, I present my second chapter, which is a joint work with Qifan Huang[2] and Zhentong Lu[3]. In this chapter, we consider the incentives of physicians and the price bargaining between the Chinese government and pharmaceutical companies, in a prescription drug market, in addition to considering consumer incentives. We take advantage of a policy shock, namely the Zero Markup Drug Policy (ZMDP), to measure physician incentives and the bargaining power of the Chinese government relative to pharmaceutical companies regarding wholesale pricing and to identify the impacts of the ZMDP on market structure and patient welfare. We find that: prescription choices of physicians are more sensitive to out-of-pocket costs of patients than drug markups when the coinsurance rate is below 35 percent; wholesale pricing is mostly dominated by provincial governments; branded drugs are more preferable and less price elastic than generic ones, but the ZMDP improves the favorability and thus the profitability of generic drugs; while total sales can be negatively affected by the ZMDP, patient welfare can be improved by a sizable amount because of the lowered prices, holding other medical activities constant.

The third chapter describes a study which is a joint work with two of my classmates, Chujian Shao[4] and Qiliang Chen[5]. It emphasizes a different aspect of China's human capital accumulation—job skills. In this project, we study the impact of a regional trade agreement,

---

[2]Qifan Huang is a Ph.D. student in the Department of Economics at the University of Washington at the time of our collaboration. He and I split the (manual) data collection work equally. He initiates the Python coding, and I initiate the programming in Stata. Email: qifan@uw.edu.

[3]Zhentong Lu is a Senior Economist in the Financial Stability Department at the Bank of Canada when he first joins our project. He provides comments for manuscript revision. Email: zlu@bankofcanada.ca.

[4]Chujian Shao is my Ph.D. classmate. Email: shaox103@uw.edu.

[5]Qiliang Chen is my Ph.D. classmate. Email: qlchen@uw.edu. For this joint work, the co-authors equally contribute to the direction of this project, data collection, theoretical analysis, and empirical analysis.

namely the Asia-Pacific Trade Agreement (APTA), on skill upgrading by manufacturers in China. We first develop a general equilibrium model of trade with heterogeneous firms and endogenous export and employee training decisions to explain firm performance following trade liberalization. Then, we test our theory based on a general difference-in-differences strategy, showing that manufacturing firms in some sectors that face higher reductions in India's tariffs increase their investment in on-the-job training faster. The effects of trade openness on export participation and training spending of firms are the largest in the middle range of productivity, which corresponds to our model prediction. This third chapter shows that one of the effective human capital accumulation strategies can be to encourage voluntary skill upgrading of firms via bilateral trade liberalization, and it works for China, especially during its transition (i.e., privatization) era.

Understanding how China can manage to maintain and improve one of its drivers of economic growth and development, human capital, is meaningful. It not only provides policy implications for other developing countries, but also points out the areas of improvement for China and other countries. I hope that this dissertation can contribute to the discussion of human capital maintenance and accumulation.

## TABLE OF CONTENTS

Page

## LIST OF FIGURES

## LIST OF TABLES

## ACKNOWLEDGMENTS

I would like to express my deepest gratitude to the chair of my supervisory committee, Yuya Takahashi, as well as my committee members, Rachel Heath, Xu Tan, Jing Tao, and Vanessa Oddo, for their unwavering support and guidance throughout my doctoral study. Each of them plays an indispensable role in leading me to finish writing this dissertation. Yuya Takahashi provides many insightful suggestions for several research projects of mine during our one-to-one and group meetings; Rachel Heath offers constructive criticism that guides me to pursue meaningful research topics; Xu Tan inspires me to think more critically, deeply, and rigorously about my research questions with her teaching; Jing Tao's genuine compassion helps me through a difficult time during the pandemic; Vanessa Oddo kindly engages me in additional research and funding opportunities. Admittedly, the completion of this dissertation would not have been possible without their involvements.

I would also like to send my sincere thanks to my like-minded peers Qifan Huang, Ying-chin Chen, Chujian Shao, Qiliang Chen, and Minyan Shen, for their encouragements in my research journey and daily life. In addition, I really appreciate the financial support from the Grover & Creta Ensley Fellowship and Rachel M. Storer Award in Labor Economics offered by the Department of Economics at the University of Washington, and I am forever grateful to the faculty and staff members at the University of Washington for creating such a friendly environment for me to grow. Special thanks go to Yu-chin Chen, Quan Wen, Eric Zivot, Shi Chen, Shawna Reimers, Heidi Hannah, and Kim Lee.

Finally, I am particularly thankful for my parents. It is their unconditional love that allows me to chase my dreams with passion. They are the role models of my life. Without them, I would not be the person I am.

## DEDICATION

to my parents and my soul mate,

for their love,

and trust.

Chapter 1

# UNDERSTANDING DEDUCTIBLE AND REIMBURSEMENT MAXIMUM IN A TIERED MEDICAL SYSTEM

## *1.1 Introduction*

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

It is well documented by randomized controlled experiments and quasi-experiments that consumers respond to demand-side incentives. While the literature focuses mainly on cost containment (Newhouse and the Insurance Experiment Group, 1993; Finkelstein et al., 2012) and how spending is reduced (Brot-Goldberg et al., 2017), a more important but currently less discussed aspect is how the efficiency of a medical system is influenced. Particularly, what remains as a question is whether the expenditure reduction incentivized by high cost-sharing is achieved by consuming more high-value care, which would improve welfare, or

through reduction of necessary services, which may in contrast deteriorate population health. However, as medical systems are complex, and it is difficult to clearly distinguish between high-value care and low-value care, the evidence on this aspect is limited. As the costs of medical waste due to overtreatment or low-value care are estimated to add up to 75.7–101.2 billion US$ in the United States alone,[1] this issue is of great policy relevance. Recent studies thus have started to conduct welfare analyses of healthcare programs by imposing assumptions about the structures of those programs, such as Finkelstein et al. (2019).

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

To identify how patients respond to financial incentives, we adopt a structural approach by leveraging the exogenous variation in hospital options and isolated variation along the dimension of coinsurance level. Our model entails two decision stages. First, patients make a discrete choice over hospitals under uncertainty about health status; then, they make a continuous spending choice upon realizing their health risks. As a usual rational decision maker, when choosing a hospital, a patient forms an expectation under uncertainty and

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

takes into consideration his/her own opinion about financial risk (e.g., he/she would decide if a higher coinsurance rate is riskier than a potentially lower service quality associated with a lower price), the expected service utility from the second stage, the hospital level fixed effects, and the taste shock. In the second stage, the patient incorporates health status, moral hazard type, and the cost structure to choose expenditure optimally. Note that, this model allows for heterogeneity in risk preferences, moral hazard types, and health states, so that we can obtain the richest possible understanding of how patients select hospitals and make utilization decisions in a consecutive manner. When building the model, we consult the well-established insurance choice literature including Cardon and Hendel (2001), Carlin and Town (2009), Bundorf et al. (2012), Einav et al. (2013), Handel (2013), Azevedo and Gottlieb (2017), and Marone and Sabety (2022). Their settings allow for variation in coinsurance level among insurance plans, and our setting allows for variation in cost-sharing across hospitals under the same plan.

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

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

There could be efficiency loss due to the current policy. Owing to low deductibles and the potential mistrust of quality in more generous hospitals, patients with low willingness to pay, who also are more likely to be at low spending risk, tend to choose less generous hospitals. These hospitals tend to be higher-tiered hospitals, which are supposed to deal with more complicated diseases. Mistrust not only distorts the allocation of resources, but can also reduce willingness to pay (and thus consumer and social welfare) directly. Therefore, we believe that delaying patients' exposure to reimbursement by having higher deductibles can alleviate the negative impact of mistrust[3] and promote a more efficient allocation of medical resources at the lower end of the willingness-to-pay distribution.[4] This is in line with the purpose of HDHPs—encouraging patients to make higher-value choices. Furthermore, at the higher end of the willingness-to-pay distribution, it is possible that the reimbursement cap is causing the distortion of resource allocation. Raising the cap, for example, might increase the willingness to pay and thus promote the welfare of very risk-averse patients who value financial risk protection more, while pushing some to higher-tiered hospitals that become relatively more attractive than before, leading to a lower increase in the overall willingness to pay; at the same time, part of the patients who are mistrustful of quality associated with generosity may experience a decrease in willingness to pay, and these patients might be pushed to higher-tiered hospitals that become relatively less unattractive than before, leading to a smaller decrease in the average willingness to pay and consumer welfare. Since the current cap is not binding (i.e., none of our patients exhausted the reimbursement limit), we do not expect insurer/government costs to change much in response to a slight change of the cap.

Our model focuses on the financial dimension of the policy design associated with the multi-tiered hospital system and permits a rich space of potential (counterfactual) contracts.

---

[3]There remains a question for us to empirically investigate if lower prices are indeed associated with lower quality in different tiers of hospitals. Anecdotal evidence suggests that some patients with common diseases or minor illnesses do not require treatment in a hospital but still get hospitalized by their physicians as the number of inpatients is one of their key performance indicators. For these patients who do not look for reimbursement initially, discounted prices may indeed lead to mistrust of quality.

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

We utilize our structural model estimates to investigate three types of alternative policies. First, we experiment with a few higher-deductible policies, and find that higher deductibles within a certain degree can lead to increased social welfare. However, larger gaps between high- and low-tiered hospitals could lead to unexpected efficiency loss, potentially hurting more risk-averse patients. As a result, the scenario in which social welfare is increased the most is to increase deductibles moderately for all hospitals without increasing the gaps between them. Second, we experiment with policies with alternative caps, and it turns out that the current reimbursement cap is close to the optimal level—largely raising or lowering the reimbursement limit all lead to efficiency loss. Raising the maximum can potentially improve efficiency but the effect is quite modest. Nevertheless, we find that raising or even removing the caps has limited impact on welfare and insurer/government cost, making it a candidate way to compensate for high deductibles to improve policy acceptance. Third, we test the combination of the two policies. It turns out that, accompanying a higher deductible with a slight increase in the reimbursement limit can further improve social welfare, and the welfare gain is higher than the sum of those obtained separately, indicating synergy.

This chapter is closely related to a few literature branches. First, it builds on the prior studies about determinants of hospital choice including Burns and Wholey (1992), Roh et al. (2008), Brown and Theoharides (2009), Escarce and Kapur (2009), Ho and Pakes (2011, 2014b), Sanders et al. (2015), Baker et al. (2016), Mak (2018), Avdic et al. (2019), and Zhu et al. (2019). They emphasize hospital features (such as distance and quality) and patient characteristics (such as health) as the determinants. Particularly, Gaynor and Vogt (2003) and Ho and Pakes (2011, 2014b) estimate a discrete choice model of hospital demand.

Second, it is linked to the literature on the value-based insurance design (VBID) that aims to encourage the use of higher-value services by aligning cost with value (Perez et al., 2019). As suggested by Agarwal et al. (2018) and Ma et al. (2019), the VBID can promote primary care by modifying cost-sharing without an increase in total health spending. There has been a rising branch of literature on the design of an optimal health insurance menu (Einav et al., 2010; Bundorf et al., 2012; Geruso, 2017; Ho and Lee, 2019; Marone and Sabety, 2022). However, very few discussions are given to the design of an appropriate hospital menu within a single insurance contract.

Third, this work connects the literature on (ex post) moral hazard in healthcare. The moral hazard issues induced by cost-sharing are well documented both theoretically and empirically, by Pauly (1968), Manning and Marquis (1996), Cutler and Zeckhauser (2000), Aron-Dine et al. (2015), Keane and Stavrunova (2016), Hudson et al. (2017), Brot-Goldberg et al. (2017), and many others. However, while offering compelling evidence, most of these reduced form studies provide little guidance for forecasting health-care spending especially under situations not directly observed in the data. As suggested by Einav and Finkelstein (2018), to complement the limitations of prospective policy analysis in guiding the optimal design of healthcare system to address moral hazard, we need economic models that rely on deeper economic primitives. Following Einav et al. (2013), Bajari et al. (2014), Kowalski (2015), Einav et al. (2015, 2017), and Lu et al. (2019), we thus rely on a more sophisticated economic model of individual behavior and investigate an optimal plan design.

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

Our contributions to the literature are two-fold. First, we emphasize the cost-sharing

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

[6]While higher cost-sharing is reported to be effective, some studies have reported that it is associated with adverse outcomes, especially among vulnerable populations such as elderly and poor patients (Zeber et al., 2007; Hartung et al., 2008; Trivedi et al., 2010).

structure as another important feature affecting hospital choices, while incorporating other determinants (such as distance, moral hazard, and subjective risk attitude) in our flexible model. When investigating hospital choices, we take the consecutive utilization decisions into account. Second, we add to the VBID literature by considering a design based on different tiers of hospitals. For example, if hospitals of lower tiers provide services with higher social values, patients should be incentivized to choose lower-tiered hospitals. Our welfare measure considers both insurer costs and patient benefits and suggests a potential tool for cost control—high deductibles. If healthcare facilities are also sorted by the efficiency of services, applying different cost-sharing structures to different facilities may reduce medical waste and improve welfare.

The rest of the chapter proceeds as follows. In Section 1.2, we introduce the hierarchical medical system in China as well as the insurance plan that covers all the patients in our study. Section 1.3 first illustrates our theoretical framework, and then presents the empirical implementation of our model. Section 1.4 describes our data and the variation it provides. Section 1.5 shows the model estimates and calculates willingness to pay and social surplus. Section 1.6 evaluates welfare and distributional outcomes under alternative pricing policies. Finally, we note concluding remarks in Section 1.7.

## *1.2 Institutional Background*

### *1.2.1 The Hierarchical Medical System in Rural China*

In rural areas of China, residents face three main tiers of medical facilities: village clinics, township health centers (THCs), and county-level hospitals (Wang et al., 2014), as shown in Figure 1.1. Village clinics are usually with a single physician, and only responsible for treating very common illnesses. THCs have a group of physicians and several departments, such as internal medicine, gynecology, and traditional Chinese medicine. County hospitals are comprehensive facilities and are able to treat almost all common illnesses. Patients with rare diseases or severe conditions can transfer to more advanced hospitals in metropolitans. Village clinics only provide outpatient care, while THCs and county hospitals can admit patients for inpatient care. Patients can freely choose the facilities they want, and no referral

is needed.

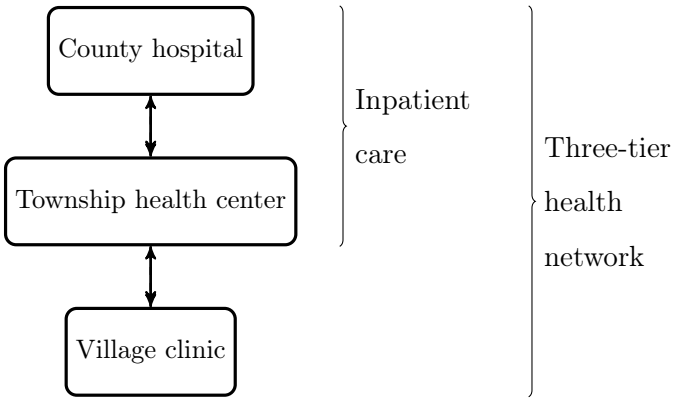



Figure 1.1: The Hierarchical Medical System in the Study Area

### *1.2.2 Research County and Healthcare System*

We draw data from a county in the southwestern part of China. According to China's 2010 Population Census, this county is comparable to the median county of China in terms of the population structure, health and education levels, employment, income, and living standards;[7] based on China City Statistical Yearbook and Statistical Yearbooks of different provinces, the research county's fiscal revenue per capita and healthcare features are close to the national average in 2015.[8] These facts provide us with the external validity to generalize empirical results regarding inpatient responses to healthcare policies in rural China.

The area of the study county is around 330 square miles ($mi^2$), which is comparable to New York City (around 303 $mi^2$); its resident population is around 440 thousand, which is

[7]For example, the percentage of population with rural local hukou in the research county is 79.2%, while the national median is 80.0% with an interquartile range of 60.1%–87.6%. For more details, please refer to Lu et al. (2019).

[8]For example, the fiscal revenue per capita is 5.8 thousand Chinese Yuan (CNY) in the research county, while the national average is 5.2 thousand CNY with a standard deviation (SD) of 7.0; the number of medical institutions per thousand people is 6.5 versus 6.7 with an SD of 15.5; the number of physicians per ten thousand people is 24.7 versus 23.9 with an SD of 16.3; the number of beds in medical institutions per ten thousand people is 57.6 versus 51.3 with an SD of 44.2.

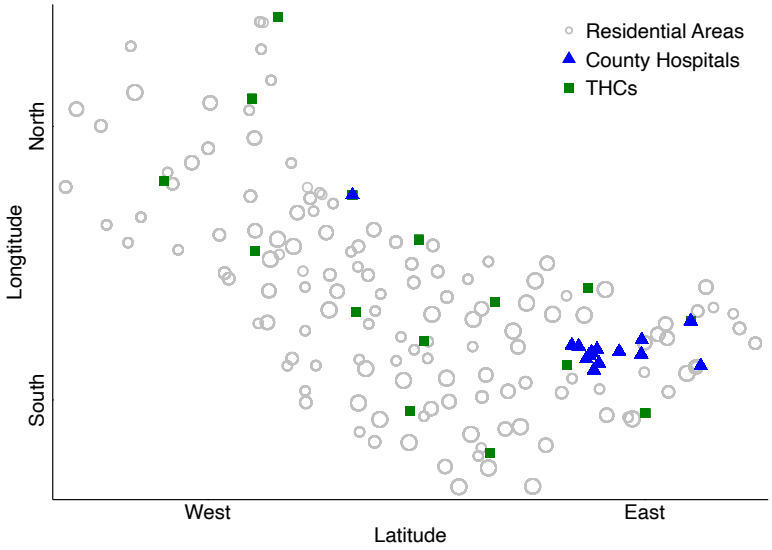



Figure 1.2: Geography of the Study County

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

comparable to Oakland in California (around 400 thousand), as of the study period. To better visualize the geographic distribution of the local health institutions and residential areas in our sample, we provide an illustrative map in Figure 1.2. There are fifteen THCs in the county, dispersed evenly across the residence areas. All of the fourteen county hospitals locate in the southeast of the county, which is the urban area, except for one locating in the middle north. Health care is quite accessible across the county, but some neighborhoods (e.g., those located in the southeast) have slightly better access to healthcare than others (e.g., those located in the northwest). Characteristics of facilities are described in Table 1.1. As shown, compared to THCs, county hospitals have much more licensed physicians and less occupational physician assistants, leading to a much higher physician-assistant ratio, potentially reflecting higher service quality; county hospitals have more beds, but while there are more authorized beds than those actually needed in THCs, the number of authorized beds do not meet the number of actual ones in county hospitals, showing a potential overcrowding

Table 1.1: Basic Characteristics of the Study County's Healthcare System

| | THCs | | | County hospitals | | |
|---|---|---|---|---|---|---|
| | 2012 | 2013 | 2014 | 2012 | 2013 | 2014 |
| Licensed physicians | 12.133 | 11.067 | 11.533 | 41.750 | 46.583 | 47.250 |
| | (6.174) | (4.935) | (6.174) | (61.458) | (67.641) | (67.293) |
| Physician assistants | 7.667 | 8.231 | 7.385 | 2.667 | 2.750 | 2.250 |
| | (6.055) | (4.986) | (4.788) | (2.015) | (1.422) | (1.138) |
| Authorized beds | 50.000 | 52.133 | 52.133 | 121.833 | 112.833 | 119.583 |
| | (25.506) | (23.348) | (23.348) | (145.284) | (143.848) | (146.421) |
| Actual beds | 47.533 | 42.600 | 42.867 | 152.000 | 165.083 | 171.833 |
| | (25.562) | (17.521) | (18.228) | (193.478) | (205.794) | (217.895) |
| Total revenue (CNY | 7.318 | 7.407 | 7.371 | 34.418 | 39.356 | 42.509 |
| million/year) | (3.772) | (4.193) | (4.295) | (60.746) | (69.742) | (74.072) |
| Medical revenue (CNY | 4.504 | 4.608 | 4.923 | 32.539 | 36.247 | 40.466 |
| million/year) | (2.747) | (3.054) | (3.188) | (58.901) | (63.133) | (71.032) |
| N | 15 | 15 | 15 | 12 | 12 | 12 |

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

issue; the total revenue of a typical county hospital is five times higher than that of a typical THC, and the medical revenue also accounts for a greater proportion in a typical county hospital, showing that THCs rely on the governments more to survive.[9]

In our target population, patients are covered by a single health insurance program, the New Rural Cooperative Medical Scheme (NRCMS), aiming to achieve universal access to healthcare and reduce financial burden for all rural residents in China. The NRCMS was initiated in July 2003 in some pilot counties[10] and then was rapidly expanded to all regions nationwide (Bai and Wu, 2014; Chen et al., 2019). By the start of our study period (2012), the program covered about 805 million individuals, or 98.3% of rural residents in

[9]Total revenue consists of medical revenue, government subsidy, science and education project revenue, and other revenue.

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

China (China Health and Family Planning Statistical Yearbook, 2015). Our focal county was one of the NRCMS pilot counties in 2005, and the participation rate reached 74.8% at the end of the year; by 2012, all counties in the province had initiated the NRCMS, and the participation rate of our focal province was 98.7%. The program is heavily subsidized by the government. For example, the out-of-pocket premium was 60 CNY in 2012, which only accounts for less than one fourth of the total premium (290 CNY). The insurance covers both outpatient and inpatient care.

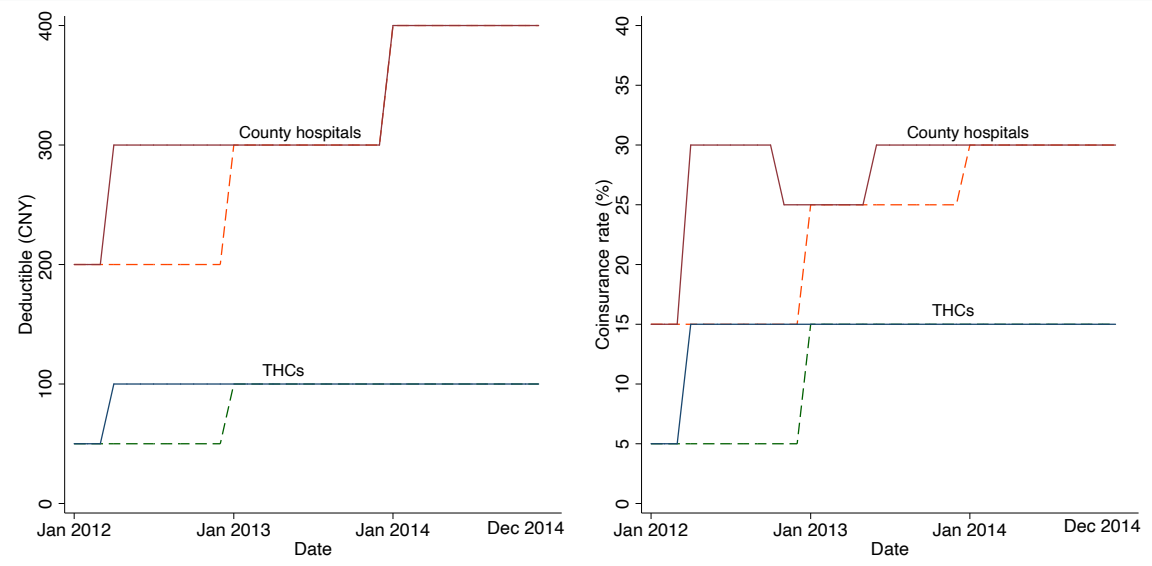



Figure 1.3: Stated and Revealed Policies in the Study County

Notes: The dashed lines are based on available NRCMS policy documents, while solid lines are based on our data.

According to the NRCMS policy documents, in most cases, the deductible per visit in THCs was 50 CNY in 2012 and 100 CNY in 2013 and 2014, and that in county hospitals was 200 CNY in 2012, 300 CNY in 2013, and 400 CNY in 2014. Reimbursement maximum was 100 thousand CNY in all hospitals in all three years. After paying the deductible and before reaching the reimbursement maximum, the coinsurance rate (one minus the reimbursement rate) in THCs was 5% in 2012 and 15% in 2013 and 2014, and that in county hospitals was 15% in 2012, 25% in 2013, and 30% in 2014. Our data find some discrepancy from the policy

documents: in our data, the deductible and coinsurance rate in THCs were increased to the 2013 level in April 2012; in county hospitals, the deductible was also increased to the 2013 level in April 2012; the coinsurance rate in county hospitals actually jumped to 30% in April 2012 and fell slightly to 25% in October 2012, which lasted only until the middle of 2013. As we can see, the government kept reducing the generosity of the NRCMS, potentially trying to control costs.

We further summarize the effective reimbursement rates in each year in each hospital tier, and the effective rates are lower (for example, in 2013, the specified reimbursement rate in THCs should be 85%, but the average effective rate is <75%) due to some non-reimbursable services and items used during the treatment. Although the majority of treatments and services are reimbursable, there are still some services (such as cosmetic surgery) and drugs (such as newly-patent drugs) are not reimbursable. As illustrated in Figure 1.4, the average reimbursement rate for inpatient care is 71.3% in THCs and 55.6% in county hospitals, between 2012 and 2014. The average reimbursement rates received in these two tiers of healthcare facilities in each year declined from 2012 to 2014.

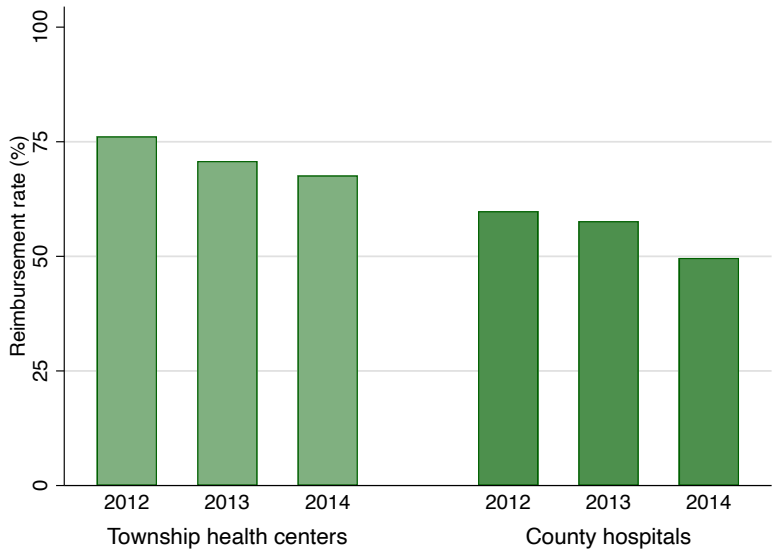



Figure 1.4: Average Effective Reimbursement Rates in the Study County

## 1.3 Model

### 1.3.1 Theoretical Framework

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

*Demand and Patient Incentives*

We consider a two-stage demand model resembling that of Einav et al. (2013) and Marone and Sabety (2022). In the first stage, a forward-looking utility-maximizing patient chooses a health-care provider without knowing her exact health status (i.e., amount needed to treat her disease).[11] The patient forms an expectation regarding her health realization based on all available information. In the second stage, the patient learns about her health state (after being admitted to a hospital) and decides how much to spend on health care.

Patients are characterized by type $\theta$: $\{F, \omega, \psi\}$, where $F$ is a patient's belief about her subsequent health status $\lambda \in \mathbb{R}_+^*$; $\omega \in \mathbb{R}_+$ is a "moral hazard" parameter capturing the extra spending that would occur by moving from no insurance to full coverage; $\psi \in \mathbb{R}$ is the "risk attitude" parameter—it measures how patients evaluate uncertainty in OOP costs and quality. Population is then defined by the distribution $G(\theta)$.

A patient chooses a hopsital from a set of health-care providers denoted by $J = \{1, 2, ..., j, ..., N_J\}$ in the first stage. Specifically, we denote $j = 0$ as the hospital that charges the full cost, which is excluded from the empirical choice set.[12] After choosing a hospital, patients realize their health status $\lambda$ and decide the dollar amount $m \in \mathbb{R}_+$ of health care utilization. Health care utilization provides patients with benefits—we denote the money-metric valuation of benefits as $b(m; \lambda, \omega)$. It also costs patients money—we denote the OOP

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

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

cost as $c(m, j)$. A utility-maximizing patient should trade off the benefits and OOP costs to find the optimal spending $m^*(\lambda, \omega, j) = \arg\max_m \{b(m; \lambda, \omega) - c(m, j)\}$. We define the indirect benefit by substituting $m^*$, i.e., $b_j^*(\lambda, \omega) = b(m^*(\lambda, \omega, j); \lambda, \omega)$; similarly, the indirect OOP cost is $c_j^*(\lambda, \omega, j) = c(m^*(\lambda, \omega, j), j)$, while the indirect payoff from hospital service is $x_j^*(\lambda, \omega, j) = b_j^*(\lambda, \omega) - c_j^*(\lambda, \omega, j)$.

The patient's utility function is given by $v(m, y, j) = v_\psi(y + l_j + b(m; \lambda, \omega))$, where $v_\psi$ is strictly increasing and its shape depends on the value of $\psi$; $y = \hat{y} - c(m, j) - p$ is the "residual income" defined by subtracting the OOP cost of health care $c$, and other costs (such as transportation costs) $p$, from the initial income $\hat{y}$; $l_j$ is the average "goodwill" of hospital $j$.[13] Due to the uncertainty in health, the patient forms an expected utility, given by $U(j, p, \theta) = \mathbb{E}[v_\psi(\hat{y} - p - c_j^*(\lambda, \omega, j) + l_j + b_j^*(\lambda, \omega))|\lambda \sim F]$, when choosing a hospital. We can also write the expected utility as

$$U(j, p, \theta) = \mathbb{E}[v_\psi(\hat{y} - p + l_j + x_j^*(\lambda, \omega, j))|\lambda \sim F]. \tag{1.1}$$

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

$$\begin{aligned} \Delta x_j^*(\lambda, \omega, j) =& x_j^*(\lambda, \omega, j) - x_0^*(\lambda, \omega, j) \\ =& \underbrace{[b_j^*(\lambda, \omega) - b_0^*(\lambda, \omega)]}_{\text{induced benefit from extra spending}} - \underbrace{[c_j^*(\lambda, \omega, j) - c_0^*(\lambda, \omega, j)]}_{\text{OOP cost of extra spending}} \end{aligned} \tag{1.2}$$

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

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

[13] Note that, $l_j$ does not depend on the individual consumption level, and we can regard $p$ as a "price" or opportunity cost patients have to pay for each hospital's admission before utilize health care.

define $WTP = p - p_0$ such that

$$U(j, p, \theta) = U(0, p_0, \theta). \tag{1.3}$$

It can be shown[14] that

$$\begin{aligned} & WTP(j, \theta, c_j) \\ = & \underbrace{\mathbb{E}_\lambda[c_0^*(\lambda, \omega, 0) - c_0^*(\lambda, \omega, j)]}_{\text{mean reduced OOP cost at uninsured spending}} + \underbrace{\mathbb{E}_\lambda[\Delta x_j^*(\lambda, \omega, j)]}_{\text{mean payoff from moral hazard spending}} \\ & + \underbrace{RP(0, \theta) - RP(j, \theta)}_{\text{value of risk change}} + \underbrace{l_j - l_0}_{\text{average quality premium}} \end{aligned} \tag{1.4}$$

The above willingness to pay is comprised of four terms. The first term captures the transfer of health-care cost liability from the patient to the insurer associated with hospital $j$, which occurs even without moral hazard. The next two terms are relevant to social welfare, and they depend on patient preferences: the second term suggests that patients value the ability to consume more health care when they are insured; the third term tells how patients value the ability to smooth consumption across health states and how they rate hospital $j$'s ability to help them do so, and $RP(j, \theta) = \mathbb{E}_\lambda[\hat{y} - p + l_j + x_j^*(\lambda, \omega, j)] - v_\psi^{-1}(U(j, p, \theta))$ is the lottery-like risk premium that does not depend on $\hat{y} - p - l_j$. These first three terms are mentioned in a similar fashion by Azevedo and Gottlieb (2017). Their risk-sharing value term does not consider a possible increase in (subjective) risk as consumers are not tied to any specific service provider in their setting. In addition, we have a fourth term that measures the average quality premium of hospital $j$ (a summary of all characteristics), and it is independent of individual preferences.

*Supply Regulation*

We denote the insurer cost as $k(m, j) = m - c(m, j) + l_j - l_0$ (supposing that it costs the insurer $l_j - l_0$ to improve average quality). Thus, the reduced OOP cost in Equation (1.4) is the increased insurer cost. Define $k_j^*(\lambda, \omega, j) = k(m^*(\lambda, \omega, j), j)$ and $k_0^*(\lambda, \omega, j) = k(m^*(\lambda, \omega, 0), j)$.[15] Then, the social surplus (SS) of choosing hospital $j$ against 0 can be

[14] See Appendix A.5.1 for derivation.

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

written as:

$$\begin{aligned} SS(j,\theta) &= WTP(j,\theta,c_j) - \mathbb{E}_\lambda[k_j^*(\lambda,\omega,j)] \\ &= \underbrace{RP(0,\theta) - RP(j,\theta)}_{\text{value of risk change}} - \underbrace{\mathbb{E}_\lambda\left[k_j^*(\lambda,\omega,j) - k_0^*(\lambda,\omega,j) - \Delta x_j^*(\lambda,\omega,j)\right]}_{\text{social cost of moral hazard}}. \end{aligned} \tag{1.5}$$

In Equation (1.5), the social cost is independent of uncertain payoffs as we assume that the insurer is risk neutral. Hospital assignment faces the trade-off between the value of subjective risk change and social cost of moral hazard, and the efficient choice would maximize social surplus, i.e., $j^{\text{eff}}(\theta) = \arg\max_{j\in J} SS(j,\theta)$. Given prices $\mathbf{p} = \{p_j\}_{j\in J}$, cost-sharing structures $\mathbf{c} = \{c_j\}_{j\in J}$ associated with all potential hospitals, and fixed quality premiums $\mathbf{l} = \{l_j\}_{j\in J}$, patients choose the privately optimal hospital by trading off their private utility and prices (opportunity costs): $j^*(\theta,\mathbf{p},\mathbf{c}) = \arg\max_{j\in J}\{WTP(j,\theta,c_j) - p_j\}$.

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

$$W(\mathbf{p},\mathbf{c}) = \int SS(j^*(\theta,\mathbf{p},\mathbf{c}),\theta)dG(\theta). \tag{1.6}$$

In our counterfactual analyses, we explore the welfare effect of how the regulator provides a vertical menu of hospitals with different reimbursement policies. For policymakers, there is a trade-off between risk-smoothing and moral hazard. While more risk-smoothing can be welfare-improving, higher coverage also promotes more unnecessary expenditures, and therefore higher social costs. The following two propositions (proven in Appendix A.5.2) guide our counterfactual analyses.

**Proposition 1.1**. A higher deductible can improve social welfare if mistrustfulness increases with generosity among patients with minor diseases.

**Proposition 1.2**. A higher reimbursement maximum does not necessarily lead to a large increase in insurer costs when moral hazard is modest.

### *1.3.2 Empirical Model*

To empirically investigate the theoretical predictions and explore other patterns in our data, we first parameterize patient utility and the distribution of health states. Then, we discuss

the sources of variation in our data that can help with identification. Last, estimation methods are explained.

*Parameterization*

As explained in section 1.3.1, our model entails two stages. Here, we extend our theoretical model by accounting for the fact that in our empirical setting, the opportunity costs of visiting a hospital can be valued by certain amounts of CNY in OOP spending, and patients make repeated hospital choices over time.

*Second Stage: Utilization Decision.* Following Einav et al. (2013), Lu et al. (2019), as well as Marone and Sabety (2022), we assume that the benefit of health-care spending $m$ is quadratic in its difference from the health risk $\lambda$ (the amount of spending necessary to treat one's disease). That is,

$$b(m_{it}; \lambda_{it}, \omega_i) = (m_{it} - \lambda_{it}) - \frac{1}{2\omega_i}(m_{it} - \lambda_{it})^2 \tag{1.7}$$

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

$$m^*_{jt}(\lambda_{it}, \omega_i) = \arg\max_m \{b(m; \lambda_{it}, \omega_i) - c_{jt}(m)\}. \tag{1.8}$$

The first order condition is

$$m^*_{jt}(\lambda_{it}, \omega_i) = \omega_i(1 - c'_{jt}(m^*_{jt}(\lambda_{it}, \omega_i))) + \lambda_{it}. \tag{1.9}$$

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

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

*First Stage: Hospital Choice.* Before choosing a hospital, the patient receives a private signal about her latent health status $\lambda_{it} \sim F_{it}^{\lambda}$. She therefore chooses a health-care provider $j$ from set $J_{it}$ (all the hospitals available[16] to the patient) to maximize the objective function below:

$$v_{ijt}\left(F_{it}^{\lambda}(\cdot), \omega_i, \psi_i\right) = \int \frac{1-\exp\left(-\psi_i u_{ijt}^{*}(\lambda, \omega_i)\right)}{\psi_i} dF_{it}^{\lambda}(\lambda), \quad \psi_i \neq 0. \tag{1.10}$$

and $v_{ijt}\left(F_{it}^{\lambda}(\cdot), \omega_i, \psi_i\right) = \int u_{ijt}^{*}(\lambda, \omega_i) dF_{it}^{\lambda}(\lambda)$ if $\psi_i = 0$. Here, in line with our theoretical model, preferences are assumed to exhibit CARA and the coefficient of absolute risk aversion is $\psi_i$.[17] It is important to note that, this risk attitude parameter can reflect two effects that are countervailing: (i) attitudes toward financial uncertainty, and (ii) subjective perceptions about service quality and its relation with financial uncertainty. With moral hazard, the von Neumann Morgenstern (vNM) utility function is defined over the payoff from health spending (in the second stage) and some hospital and individual characteristics. By extending Equation (1.1), we define this payoff by

$$u_{ijt}^{*}(\lambda, \omega_i) = \beta_{0,j} + \beta_1\left(b_{jt}^{*}(\lambda, \omega_i) - c_{jt}^{*}(\lambda, \omega_i)\right) + \beta_2 D_{ijt} + Z_{jt}'\beta_3 + \sigma_\epsilon \epsilon_{ijt} \tag{1.11}$$

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

$$j^{*}\left(F_{it}^{\lambda}(\cdot), \omega_i, \psi_i\right) = \arg\max_{j \in J_{it}} v_{ijt}\left(F_{it}^{\lambda}(\cdot), \omega_i, \psi_i\right). \tag{1.12}$$

*Health Information.* Suppose patients believe that health risks are drawn from a right-truncated lognormal distribution of health states,[18]

$$\log \lambda_{it} \sim N(\mu_{\lambda,it}, \sigma_{\lambda,it}^2)\mathbf{1}\{0 < \lambda \leq \overline{\lambda}\}, \tag{1.13}$$

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

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

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

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

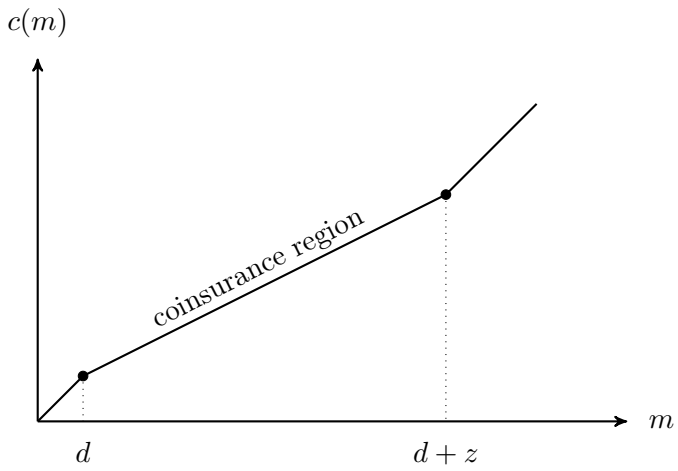



Figure 1.5: Out-of-Pocket Cost Function, $a \in (0, 1)$

*Reimbursement Scheme.* In our context, the OOP cost function is nonlinear—or more precisely, piecewise linear (see Figure 1.5 for an illustration). The marginal OOP cost function $c'(m)$ in Equation (1.9) is thus piecewise constant. In the region $m \leq d$ or $m \geq d + z$ where $d$ denotes deductible and $z$ denotes the maximum reimbursable spending, we have $c'(m) = 1$; in the region $d < m < d + z$, $c'(m) = 1 - a \in (0, 1)$, where $a$ is the reimbursement rate in the coinsurance region. As a result, the optimal $m^*$ would be a step function of $\lambda$, also depending on $\omega$. We next derive cutoff values on the health state that determine which OOP region a patient will find a specific solution optimal.

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

$$m^* = \begin{cases} \lambda & \text{if } \lambda \leq d - \omega a/2 \\ \lambda + \omega a & \text{if } d - \omega a/2 < \lambda \leq d + z - \omega a \\ d + z & \text{if } d + z - \omega a < \lambda \leq d + z \\ \lambda & \text{if } \lambda > d + z \end{cases} . \tag{1.14}$$

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

$$m^* = \begin{cases} \lambda & \text{if } \lambda \leq d - \omega a/2 \\ \lambda + \omega a & \text{if } d - \omega a/2 < \lambda \leq d + z - \sqrt{2\omega a z} \\ d + z & \text{if } d + z - \sqrt{2\omega a z} < \lambda \leq d + z \\ \lambda & \text{if } \lambda > d + z \end{cases} . \tag{1.15}$$

If $z = 0$ (no coinsurance region), then $m^* = \lambda$. As we can see, different coinsurance regions create different financial incentives—the larger the region, the more likely patients incur extra spending. All hospitals in our empirical setting have large coinsurance regions (for example, $z$ is greater than 100 thousand CNY, while $\omega$ is smaller than 1 thousand CNY), and thus only Equation (1.14) is used. Derivations can be found in Appendix A.5.3.[19]

*Identification*

Our goal is to recover the joint distribution across patients of willingness to pay, risk attitude, and the social cost of moral hazard associated with different hospitals. Variation in these objects comes from variation in either patient preferences (the risk attitude and moral-hazard parameters) or in the distribution of health states. Major concerns include (1) distinguishing preferences ($\omega_i$) from private information about health ($\mu_{\lambda,it}$), (2) distinguishing taste for reimbursed spending ($\beta_1$) from risk attitude ($\psi_i$), and (3) identifying heterogeneity in the risk attitude ($\psi_i$) and moral hazard ($\omega_i$) parameters.

First, when observing a positive correlation between reimbursement generosity (caused by both the treatment styles of health-care providers and patient choices) and total health-care spending (conditional on observable characteristics) in the data, we can explain it as either the effect of private health information affecting hospital choice (selection) or physician styles driving utilization (moral hazard). To distinguish one explanation to the other, we take the advantage of the variation in hospital menus $J_{it}$ (sets of available or accessible hospitals). When hospital choices vary with menus, the degree of moral hazard can be identified by the

[19] We can see from Equation (1.14) that there is bunching at the convex kink point $d + z$, as discussed by Einav et al. (2017). However, unlike the "donut hole" in the context of prescription drug insurance for the elderly in Medicare Part D, our kink point is quite high in the budget set, and empirically almost none of our patients are near there. Thus, our identification will not rely on bunching.

extent to which patients facing more generous hospital menus also have higher health-care spending. When observing patients who face similar menus making different hospital choices, we can identify the amount of private information about health and the magnitude of the idiosyncratic shock $\sigma_\epsilon$: conditional on observables and the predicted effects of moral hazard, if patients who choose more generous hospital inexplicably spend more on health care, this variation in hospital choice is attributed to private information about health; otherwise, we attribute any unexplained residual variation in hospital choice to the idiosyncratic shock.

Second, in our model, both risk parameter $\psi_i$ and taste for reimbursed spending $\beta_1$ affect hospital choice but not spending. To distinguish between them, we can utilize cases in which observably different patients face similar hospital menus. Risk attitude is then identified by analyzing how a patient associates uncertainty in reimbursed spending with reimbursement rate, holding the expected OOP cost fixed. The taste for reimbursed spending is identified by the rate at which patients trade off other hospital characteristics (e.g., distance) with expected OOP cost, holding uncertainty in OOP cost fixed.

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

*Estimation*

*Structural Settings.* First, we make additional parametric assumptions about the distribution of individual health status $F^{\lambda}_{it}(\cdot)$: let's assume that $\mu_{\lambda,it}$ has a fixed-effect structure and $\sigma_{\lambda,it}$ can be projected on time-varying patient characteristics. That is,

$$\mu_{\lambda,it} = \overline{\mu_{\lambda,i}} + (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\beta_\mu \tag{1.16}$$

$$\sigma_{\lambda,it} = \mathbf{x}^{\sigma}_{it}\beta_\sigma \tag{1.17}$$

where $\overline{\mathbf{x}}_i$ is a vector of within-individual averages of $\mathbf{x}_{it}$, including a constant and an individual health risk predictor[20] calculated based on age, gender, education, marital status, and the International Classification of Diseases (ICD) 10 codes; $\overline{\mu_{\lambda,i}}$ is the average (over time) of $\mu_{\lambda,it}$ drawn from the jointly right-truncated normal distribution described below.

$$\begin{pmatrix} \overline{\mu_{\lambda,i}} \\ \log\omega_i \\ \psi_i \end{pmatrix} \sim N\left( \begin{pmatrix} \overline{\mathbf{x}}_i\beta_\mu \\ \overline{\mathbf{x}}_i^{\omega}\beta_\omega \\ \overline{\mathbf{x}}_i^{\psi}\beta_\psi \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma_\mu^2 & \sigma_{\mu,\omega} & \sigma_{\mu,\psi} \\ \sigma_{\mu,\omega} & \sigma_\omega^2 & \sigma_{\omega,\psi} \\ \sigma_{\mu,\psi} & \sigma_{\omega,\psi} & \sigma_\psi^2 \end{pmatrix}}_{\Sigma_1} \right) \mathbf{1}\{0 < \omega \leq \overline{\lambda}\} \tag{1.18}$$

There are both observed (via mean) and unobserved (via covariance) heterogeneity in each parameter. Covariates $\mathbf{x}_{it}^{\sigma}$, $\mathbf{x}_{it}^{\omega}$ and $\mathbf{x}_{it}^{\psi}$ include a standardized risk predictor[21] and a constant.

The parameters to be estimated include the four vectors of mean shifters $(\beta_\mu, \beta_\sigma, \beta_\omega, \beta_\psi)$, six variance and covariance parameters $(\sigma_\mu, \sigma_\omega, \sigma_\psi, \sigma_{\mu,\omega}, \sigma_{\mu,\psi}, \sigma_{\omega,\psi})$, and five (vectors of) taste or magnitude parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma_\epsilon)$.

*Algorithm.* We resort to a maximum likelihood approach.

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

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

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

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

is given by

$$f_m(m_{it}|c_{jt},\alpha_{its},\theta,\mathbf{x}_{it}) = \frac{\Phi'\left(\frac{\log\lambda_{ijts}-\mu_{\lambda,its}}{\sigma_{\lambda,it}}\right)}{\Phi\left(\frac{\log\overline{\lambda}-\mu_{\lambda,its}}{\sigma_{\lambda,it}}\right)} \tag{1.19}$$

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

Second, we calculate the probability of each hospital choice. Given $\theta$ and $\alpha_{its}$, we simulate the distribution of health states by $\lambda_{ijtsd} = \exp(\mu_{\lambda,its}+\sigma_{\lambda,it}Z_d)$ where $Z_d$ is a vector of points approximating a standard normal distribution, and we denote $W_d$ as the associated Gaussian quadrature weights. Then, we calculate the optimal health-care spending $m_{ijtsd}$ associated with each potential health state realization based on formula (1.14):[22]

$$m^*_{ijtsd} = \begin{cases} \lambda_{ijtsd}+\omega_{is}a_{jt} & \text{if } d_{jt}-\frac{\omega_{is}a_{jt}}{2} < \lambda_{ijtsd} \leq d_{jt}+z_{jt}-\omega_{is}a_{jt} \\ d_{jt}+z_{jt} & \text{if } d_{jt}+z_{jt}-\omega_{is}a_{jt} < \lambda_{ijtsd} \leq d_{jt}+z_{jt} \\ \lambda_{ijtsd} & \text{otherwise} \end{cases}. \tag{1.20}$$

Now, we have the distributions of privately optimal spending $m^*_{ijtsd}$ for each admission and draw of $\alpha_{its}$ to calculate the patient's expected utility from choosing each potential hospital. Then, the numerical approximation to Equation (1.10) is constructed using the quadrature weights $W_d$ mentioned above:

$$v_{ijts} = \frac{1-\sum_{d=1}^{N_d} W_d \cdot \exp\left(-\psi_{is}u^*_{ijts}(\lambda_{ijtsd},\omega_{is})\right)}{\psi_{is}}, \quad \psi_{is} \neq 0, \tag{1.21}$$

and $v_{ijts} = \sum_{d=1}^{N_d} W_d \cdot u^*_{ijts}(\lambda_{ijtsd},\omega_{is})$ if $\psi_{is} = 0$, where $N_d$ is the number of support points and the payoff $u^*$ is calculated based on Equation (1.11). We estimate the model in certainty-equivalent (CE) units of $v_{ijts}$ to avoid numerical issues when dealing with double-exponentiation:

$$v^{CE}_{ijts} = \overline{u}_{ijts} - \frac{1}{\psi_{is}}\log\left(\sum_{d=1}^{N_d} W_d \cdot \exp\left(-\psi_{is}(u^*_{ijts}(\lambda_{ijtsd},\omega_{is})-\overline{u}_{ijts})\right)\right), \quad \psi_{is} \neq 0 \tag{1.22}$$

and $v^{CE}_{ijts} = \sum_{d=1}^{N_d} W_d \cdot u^*_{ijts}(\lambda_{ijtsd},\omega_{is})$ if $\psi_{is} = 0$, where $\overline{u}_{ijts} = \mathbb{E}_d[u^*_{ijts}(\lambda_{ijtsd},\omega_{is})]$.

The choice probabilities conditional on $\alpha_{its}$ are given by the standard logit formula

$$L_{ijts} = \frac{\exp\left(v^{CE}_{ijts}/\sigma_\epsilon\right)}{\sum_{j\in J_{it}}\exp\left(v^{CE}_{ijts}/\sigma_\epsilon\right)}. \tag{1.23}$$

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

Third, we write the numerical approximation to the likelihood of the sequence of hospital choices and medical expenditures for a given patient:

$$L_i = \sum_{s=1}^{N_s} W_s \prod_{t=1}^{T} \prod_{j \in J_{it}} \left( f_m(m_{it}|c_{jt}, \alpha_{its}, \theta, \mathbf{x}_{it}) L_{ijts} \right)^{d_{ijt}} \tag{1.24}$$

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

$$LL(\theta) = \sum_{i=1}^{N} \log(L_i). \tag{1.25}$$

*Recovering Individual Types.* We assume that individual types $\alpha_{it}(\theta) = \{\mu_{\lambda,it}, \omega_i, \psi_i\}$ are distributed multivariate normal with observable heterogeneity in the mean vector based on Equation (1.18). The maximum likelihood algorithm provides us with $\hat{\theta}$, an estimate of $\theta$, which helps us back out individual types, using a sequence of observed hospital choices and medical expenses denoted as $\mathbf{y}$. Denote the population distribution of types as $g(\alpha|\hat{\theta})$, the probability of observed outcomes as $p(\mathbf{y}|\hat{\theta})$, and the conditional probability of observed outcomes $p(\mathbf{y}|\alpha)$ (the "conditioning of individual tastes"). Then, according to Bayes' rule, the density of $\alpha$ conditional on parameters and observed outcomes $h(\alpha|\hat{\theta}, \mathbf{y})$ (the posterior distribution of $\alpha$) can be written as

$$h(\alpha|\hat{\theta}, \mathbf{y}) = \frac{p(\mathbf{y}|\alpha) \cdot g(\alpha|\hat{\theta})}{p(\mathbf{y}|\hat{\theta})}. \tag{1.26}$$

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

$$h_{is}(\alpha|\hat{\theta}, \mathbf{y}_i) = \frac{L_{is} \cdot W_s}{L_i}, \tag{1.27}$$

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

ity are hence

$$\mathbb{E}\overline{\lambda}_{it} = \sum_{s=1}^{N_s} h_{is}(\alpha|\hat{\theta}, \mathbf{y}_i) e^{\mu_{\lambda,its}+\frac{1}{2}\sigma^2_{\lambda,it}} \cdot \frac{\Phi\left(\frac{\log\overline{\lambda}-\mu_{\lambda,its}-\sigma^2_{\lambda,it}}{\sigma_{\lambda,it}}\right)}{\Phi\left(\frac{\log\overline{\lambda}-\mu_{\lambda,its}}{\sigma_{\lambda,it}}\right)}, \tag{1.28}$$

$$\mathbb{E}\omega_i = \sum_{s=1}^{N_s} h_{is}(\alpha|\hat{\theta}, \mathbf{y}_i)\omega_{is}, \tag{1.29}$$

$$\mathbb{E}\psi_i = \sum_{s=1}^{N_s} h_{is}(\alpha|\hat{\theta}, \mathbf{y}_i)\psi_{is}. \tag{1.30}$$

## *1.4 Data*

### *1.4.1 Summary Statistics*

We use a set of unique medical claims data for all the enrollees of the NRCMS program from a county located in the southwestern part of China. The data include records on all inpatient care in all health institutions within the county. Detailed information concerning each admission includes the date of visit, diagnosis (the ICD 10 code), medical organization visited, total medical expenditure, and the amount of insurance reimbursement received. The data also contain information on patient demographics, including age, gender, marital status, and education level.

We use data for visits to all health institutions that provide inpatient services, including THCs and county hospitals. Our data indicate that at least 9.3% of the rural residents (i.e., if we assume that all of them are covered by the NRCMS) in the study county get hospitalized at least once. We select a sample from the data with complete information of our interest (e.g., home address), which leaves us with 79,531 admissions and 46,577 unique patients who visited 15 THCs and 14 county hospitals between 2012 and 2014. The summary statistics of our main variables of interest are reported in Table 1.2.

About 42.9% of patients in our sample are male, and an average person visits a hospital almost twice during our study period (less than once a year). Among these visits, about 56.4% of them occur in county hospitals, and the remainder are in township health centers. Approximately half of the visits are paid by patients aged 18 to 60 years, and the average age of patients at the time of their visits is 56 years. The average patient has completed

Table 1.2: Descriptive Statistics for the Estimation Sample

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

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

only five years of schooling (i.e., mostly finishes elementary school)—this is not surprising given the fact that our respondents are predominantly elderly rural residents—only 29.0% of them have been to middle school or above. Interestingly, around 73.1% of the visits are paid by married patients, while the remainder are paid by patients who are single, divorced, or widowed. Given the fact that half of the sample are under the age of 18 or over the age of 60, the married proportion seems high. The rescaled risk score is obtained from the Johns Hopkins ACG system (v. 12.1), and more details can be found in Appendix A.2.[23] For each inpatient visit, patients spend on average 3.2 thousand Chinese Yuan (CNY) in our study sample, and the deductible is about 0.2 thousand CNY. On average, after paying for the deductible, patients can reimburse 62.4% of the rest of the total spending.[24]

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

### *1.4.2 Variation in Reimbursement Rates and Hospital Menus*

It's important for this research to check two key features of our setting—the plausibly exogenous variation in hospital menus and isolated variation along the dimension of coinsurance level. To better illustrate the variation, we focus on the most popular disease among our patients—acute exacerbations of chronic obstructive pulmonary disease (AECOPD, ICD 10: J40–J44).[25] This disease can be treated in either a THC or a county hospital. Since most (around 75%) of the AECOPD patients are treated in THCs, we focus on patients who choose a THC in this discussion.

---

[23] 9.5% of the diagnoses suggest a serious disease in the department of general surgery [e.g., severe acute pancreatitis (ICD 10: K85)], neurosurgery [e.g., acoustic neuroma (ICD 10: D33.3), urinary surgery [e.g., muscle-invasive bladder cancer (ICD 10: C67)], orthopedics [e.g., spinal tuberculosis (ICD 10: A18.0 and M49.0)], or hematology [e.g., myelodysplastic syndrome (ICD 10: D46)].

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

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

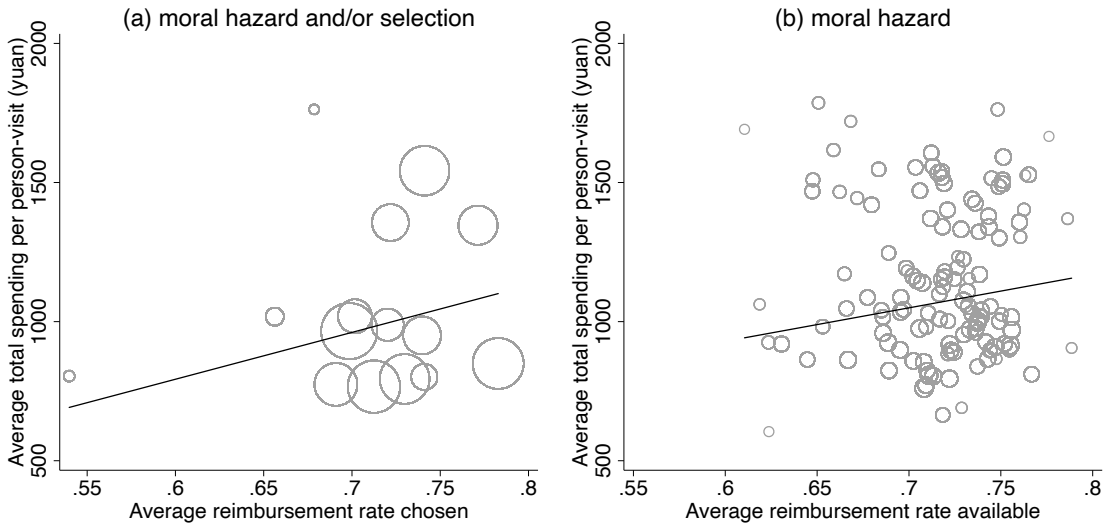



Figure 1.6: Average Spending by Generosity Chosen and Available

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

As shown in Figure 1.6, conditional on AECOPD and THC, health-care spending is positively correlated with reimbursement generosity. In the left panel, individuals are grouped by their chosen hospital, and the plot shows the average spending per person-visit in each of the THCs, weighting each hospital by admission. Consistent with our expectation, individuals who chose more generous THCs have higher spending, indicating moral hazard, adverse selection, or both. To distinguish between moral hazard and selection, the right panel, grouping individuals by their community (and thus their corresponding hospital menu), plots the average rate of reimbursement in the hospitals available/accessible to that community against the average spending of individuals living in the community. We notice that individuals with access to more generous THCs, thus arguably more likely to choose a THC

with a higher reimbursement rate, have larger spending. This suggests the presence of moral hazard as well as the coexistence of adverse selection on unobservables.

Our structural model is identified in a similar way. A key identifying assumption is that, conditional on observables, hospital menus are not related to unobservables of individuals that could affect health-care spending. This can be threatened by township governments or community leaders trying to set hospital generosity in response to unobservable information about residents that would drive spending. For example, if more generous hospitals are open for communities with unobservably healthier residents, the extent of moral hazard can be underestimated. To see if this is the case, we seek to explain hospital menu generosity by individual health risk predictor and other observables. We argue that, if hospital menus are not responding to our risk predictor, it is unlikely that they are responding to unobservables, because we should have better information on these patients (collected when they are admitted to the hospitals) than township governments and community leaders do before their admission. To hold hospital level and disease effects fixed, we again focus on AECOPD patients treated in THCs. As shown by Table A.2, conditional on community features and hospital features chosen (such as number of beds and number of doctors), we do not find any correlation between hospital menu generosity and the individual health indicator.

## *1.5 Results*

### *1.5.1 Parameter Estimates*

The parameter estimates of our structural model are shown in Table 1.3. Based on the results, we estimate an average moral hazard parameter $\omega$ of 0.4 thousand CNY (less than 0.1 thousand US$ in 2014). This is smaller than Einav et al. (2013)'s estimate of the average $\omega$, 1.3 thousand US$. This difference is due to several reasons. First, our $\omega$ represents the extra total spending per visit,[26] while Einav et al. (2013)'s $\omega$ is the extra total spending per year. Second, we focus on inpatient care in rural China, where the average price level is much lower and resources are more limited; our average per-visit spending is around 3.2 thousand CNY (about 0.5 thousand US$ in 2014), while Einav et al. (2013)'s average

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

Table 1.3: Parameter Estimates

| Variable | Parameter | Robust Std. Err. |
|---|---|---|
| County hospital fixed effect, $\beta_0$ | 4.9225 | 0.0204 |
| Net benefit from utilization stage, $\beta_1$ | 4.4367 | 0.0062 |
| Distance (km), $\beta_2$ | -0.1000* | – |
| Number of doctors, $\beta_3$ | -0.0151 | 0.0003 |
| Number of beds, $\beta_3$ | -0.0413 | 0.0001 |
| Taste shock's scale, $\sigma_\epsilon$ | -5.6434 | 0.0013 |
| Health state mean × risk predictor, $\beta_\mu$ | 1.0129 | 0.0002 |
| Health state mean intercept, $\beta_\mu$ | -0.0092 | 0.0001 |
| Health state mean's std. dev., $\sigma_\mu$ | 0.0330 | 0.0000 |
| Health state std. dev. × standardized risk predictor§, $\beta_\sigma$ | 0.1104 | 0.0003 |
| Health state std. dev. intercept, $\beta_\sigma$ | 0.2258 | 0.0003 |
| Risk attitude × standardized risk predictor§, $\beta_\psi$ | 2.5258 | 0.0231 |
| Risk attitude intercept, $\beta_\psi$ | 2.3387 | 0.0238 |
| Risk attitude std. dev., $\sigma_\psi$ | 4.5979 | 0.0164 |
| Log moral hazard × standardized risk predictor, $\beta_\omega$ | 10.3156 | 0.0049 |
| Log moral hazard intercept, $\beta_\omega$ | -8.0730 | 0.0048 |
| Log moral hazard std. dev., $\sigma_\omega$ | 4.1135 | 0.0002 |
| Corr. b/w health and log moral hazard, $\rho_{\mu,\omega}$ | 0.5519 | 0.0003 |
| Corr. b/w health and risk attitude, $\rho_{\mu,\psi}$ | 0.6168 | 0.0063 |
| Corr. b/w log moral hazard and risk attitude, $\rho_{\omega,\psi}$ | 0.9753 | 0.0010 |

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

annual spending is more than ten times our level. Therefore, the estimated $\omega$ is still quite significant in our case. Note that, $\omega$ is the additional total spending induced by moving a patient from no coverage to full reimbursement. Thus, this estimate implies that moving from a hospital with half the prices (after a deductible and before reaching a cap) to one

with zero costs is expected to increase inpatient-care spending by six percent of the mean spending. Interestingly, our model suggests that moral hazard is idiosyncratically more serious among people who privately expect that they are less healthy, as $\rho_{\mu,\omega} > 0$.

We find that patients in rural China have a wide range of risk attitudes, with the mean (median) coefficient of absolute risk aversion being -0.2 (0.4). We may translate it to an amount of money, say $\$X$, such that individuals are indifferent between (i) a payoff of zero and (ii) an equal-odds gamble between gaining \$100.0 and losing $\$X$. Based on our calculations, the mean (median) value of such indifferent value ($\$X$) is \$101.6 (\$96.6). The fact that some patients are willing to lose more money than their potential gain does not necessarily suggests that they love gambling in our context. Rather, they may perceive high cost-sharing as an indicator of high service quality (and are willing to pay \$3.4 for a chance to enjoy a higher quality). In our data, we observe that some patients would prefer a less generous hospitals given similar hospital characteristics, and these behaviors or preferences could not be explained by a hospital tier fixed effect. We reflect these preferences by allowing for negative coefficients, but we may not use the traditional term "risk taking" even though it is shown widespread among rural Chinese patients in various forms (Carlsson et al., 2012; Jin et al., 2017).[27] Our estimation suggests that risk aversion also increases idiosyncratically with private information about higher spending expectation, as $\rho_{\mu,\psi} > 0$, which is intuitive. Finally, we do find that more risk averse people (who may care less about service quality) are idiosyncratically more prone to moral hazard. For the unconditional joint distribution of the three dimensions of patient type, please refer to Figure A.2.

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

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

### *1.5.2 Model Fit*

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

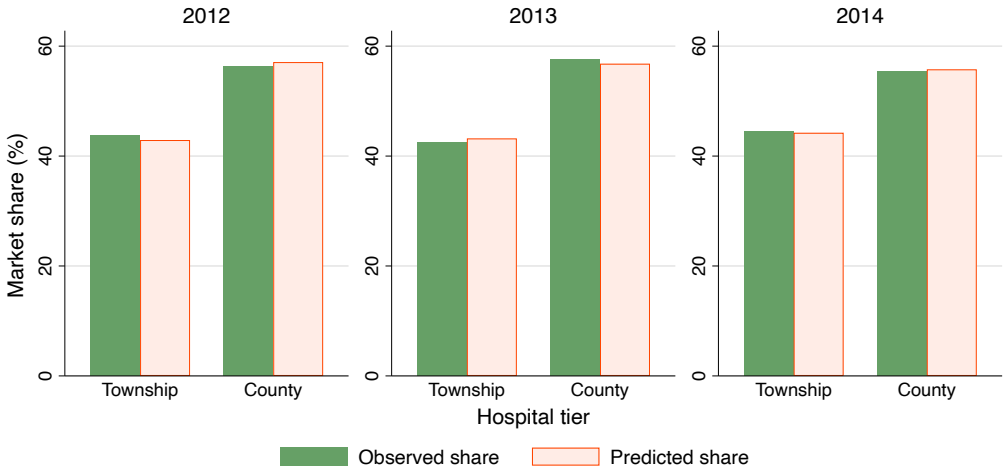



Figure 1.7: Model Fit: Hospital Tier Choices

Notes: The figure shows the observed and predicted market shares at the hospital tier level calculated based on Table 1.3. An observation is a person-visit in each year.

First, we can compare the observed and predicted market shares for each hospital. According to Figure 1.7, the model prediction is quite good at the hospital tier level. To inspect the flexibility of the model with respect to the choice of a specific hospital within a tier, we also show the market shares at the hospital level in Figure A.3. It turns out to be reasonably good as well.

Second, we can compare the observed and predicted distributions of patients' total inpatient-care spending per visit each year. The expected spending of each patient is used to construct the predicted spending distribution in the population of patients. We show the kernel density plots of spending on a log scale in Figure 1.8. If we pool THCs and county hospitals together, our model tends to overestimate the spending slightly in 2012. The predicted mean matches the observed mean well in 2013 and 2014, however. This

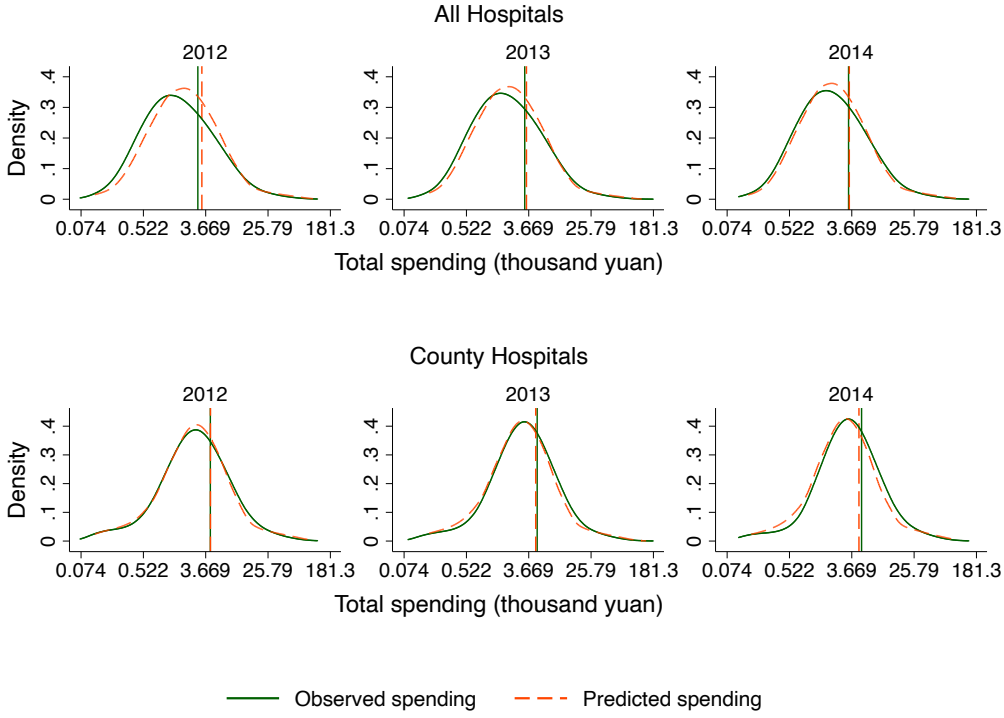



Figure 1.8: Model Fit: Inpatient-Care Spending

Notes: The kernel density plots of the observed and predicted distributions of total inpatient-care spending are on a log scale. An observation is a person-visit in each year. Predicted distributions are calculated based on Table 1.3.

could be partly because our estimation procedure pools all the three years of data together, while there might be some heterogeneity in 2012. If we focus on only county hospitals, nevertheless, the fit is much better, although there might be a slight underestimation of spending in 2014. On average, the inpatient-care spending is predicted to be 3,186 CNY across patient-visit observations in our data, which is close to the observed average (3,150 CNY). By implementing a Kolmogorov-Smirnov test, we cannot reject the equality of the observed and predicted distributions of spending.[28]

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

### *1.5.3 Willingness to Pay*

In this subsection, we construct each patient's willingness to pay for different levels of hospitals and reimbursement rates according to our previous parameter estimation. To map our empirical model to the theoretical framework, a few simplifications are needed. First, we limit our focus to the AECOPD patients who only have one inpatient visit between 2012 and 2014 (N = 4,612). This allows us to assign a single type $\alpha_i = \{F_i^\lambda, \omega_i, \psi_i\}$ to each patient, where $F_i^\lambda$ is a right-truncated lognormal distribution described by $\{\mu_{\lambda,i}, \sigma_{\lambda,i}, \overline{\lambda}\}$.[29] Second, we assume that the idiosyncratic shock is utility-irrelevant.[30] Next, we hold all non-financial features fixed to limit our attention to the cost-sharing dimension. Last, we assume that all patients have the same per unit opportunity cost of travel, which is 22.5 CNY per km.[31]

The reference hospital $j = 0$ is a county hospital that does not reimburse any cost, which is not observed in our data. Based on Equation (1.21), we calculate the utility of choosing hospital $j > 0$ in CE units, denoted as $v_{ij}^{CE}$ and then calculate willingness to pay as $WTP_{ij} = v_{ij}^{CE} - v_{i0}^{CE}$. We decompose WTP into four terms according to Equation (1.4)—a "transfer" term that represents the mean reduced OOP cost holding patient behavior constant, a "moral hazard" term that describes the mean net payoff from moral hazard spending, a "risk attitude" term that shows how much a patient values the reduction of financial uncertainty (financial risk protection) over the mistrust of quality associated with low prices, and finally a fixed "tier change value" term that reflects the monetary value of a hospital level change to an average patient.

For tractability, we summarize our tiered hospital system under NRCMS into a list of four focal hospitals ($j = 1, ..., 4$) with the same reimbursement maximum (100 thousand CNY per year). In addition, we consider a free hospital. Their deductibles are 0.4, 0.3, 0.2, 0.1, and 0.0 thousand CNY, while reimbursement rates are 60%, 70%, 80%, 90%, and 100%

---

[29] This is done by integrating over everyone's posterior distribution of types described by Equations (1.28)–(1.30).

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

[31] Since all patients live in the same county, they are limited to very few means of transportation. Thus, for simplicity, we assume all of them to have the same per unit opportunity cost, backed out from the coefficient $\beta_1$ in Table 1.3 (i.e., $0.1 \times 1{,}000/4.4367$).

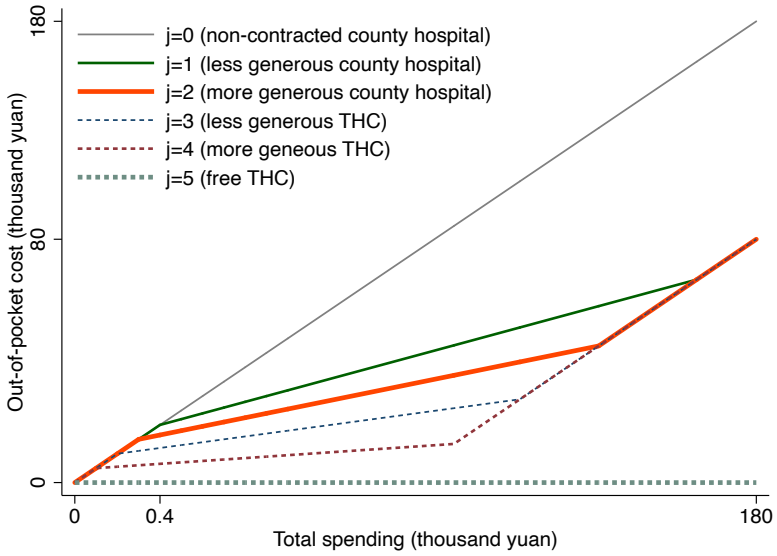



Figure 1.9: A Simplified Tiered Medical System

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

(full coverage), respectively. Moreover, $j = 0, 1, 2$ are county hospitals, while $j = 3, 4, 5$ are THCs. Figure 1.9 shows the OOP cost functions of these four focal hospitals, the null county hospital, and the counterfactual free THC.

We present the distributions of willingness to pay among these AECOPD patients in Figure 1.10. We sort patients according to their values of willingness to pay on the horizontal axis, and those with lower willingness to pay are on the right as in a demand curve. Some patients, especially those at the lower end of the willingness-to-pay distribution, seem to perceive generosity as an indicator of lower quality more than a financial risk protection, and thus are not willing to visit a more generous hospital (unless being compensated).[32] In

[32]We notice a positive correlation between the reimbursement rate and total number of visits during 2012–

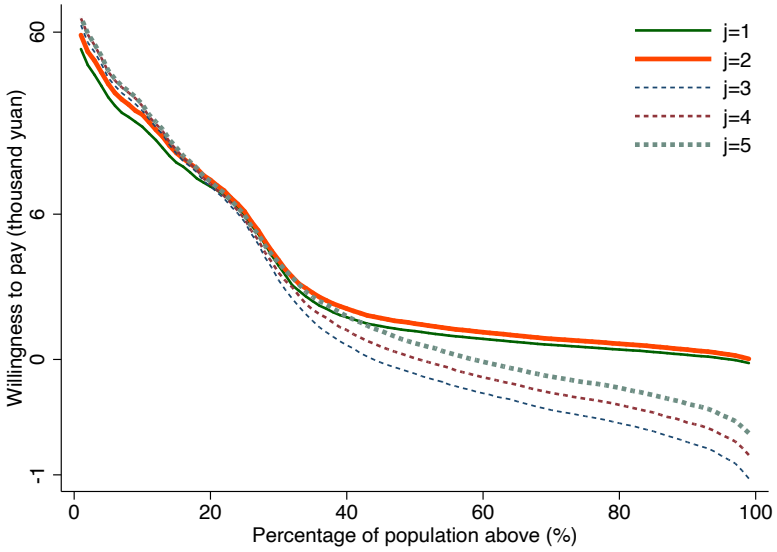



Figure 1.10: Marginal Willingness to Pay

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

order to encourage 99 percent of the population to go to THC $j = 3$ with a deductible of 200 CNY and a coinsurance rate of 20 percent instead of a non-contracted county hospital ($j = 0$) holding other characteristics fixed, a travel subsidy that is worth 1 thousand CNY should be given (or the THC needs to be 45 km closer). On the other hand, the patients with the top 1 percent willingness to pay are willing to pay 60 thousand CNY for the full coverage in a THC (or to travel about 2.7 thousand km). Clearly, the range of willingness to pay is wide among these AECOPD patients. Slightly more than half of them prefer county hospitals ($j = 1, 2$) to THCs ($j = 3, 4, 5$) holding other characteristics fixed; interestingly, these are

2014 in our data. It seems to suggest that patients who visit more generous hospitals also tend to get hospitalized more frequently (due to less efficient treatments). This makes the association between low quality and high reimbursement one of the plausible explanations for "risk taking".

also the people with lower willingness to pay for any coverage, suggesting a tendency to bypass primary care. Some of them spent quite little (see Figure A.6), suggesting common diseases or minor illnesses.

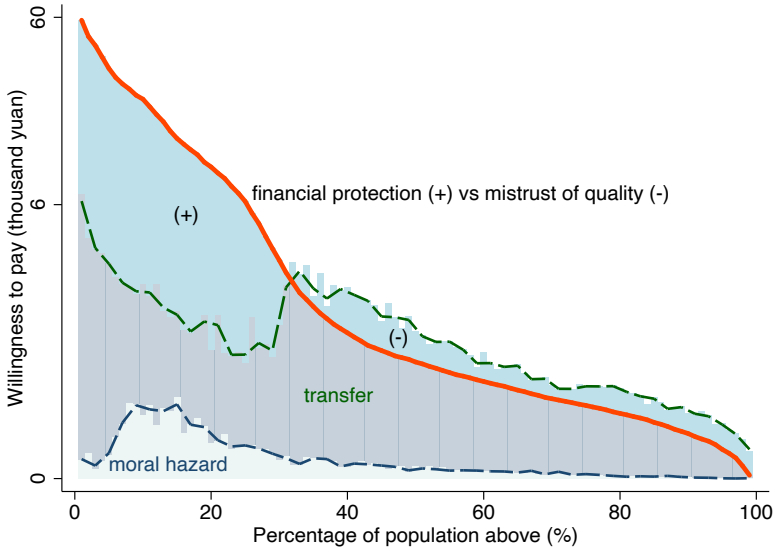



Figure 1.11: Decompose Marginal Willingness to Pay

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

We further decompose the marginal WTP for $j = 2$ as Figure 1.11 shows. Note that, since both $j = 0$ and $j = 2$ are county hospitals, the "tier change value" is zero (assuming that all non-financial characteristics are the same) and thus we are left with three components. As we can see, for most (more than 60%) of the AECOPD patients, willingness to pay mainly comes from the "transfer" term, and the net payoff from moral hazard spending only represents a very small portion of the willingness to pay, while mistrust of quality can lead to a lower willingness to pay (by -0.1 to -1 thousand CNY), perhaps due to subjective perceptions about how more generous hospitals may not handle their health risks as efficient.

For those with high (top 35%) willingness to pay, the value of financial protection finally outweighs the mistrust (of quality associated with generosity) and explains most of the willingness to pay, although "transfer" and the net payoff from moral hazard spending are also relatively high compared to those with low willingness to pay. At the top 1% percentile of the willingness-to-pay distribution, in addition to paying 6 thousand CNY (for transportation) to avoid paying nearly 6 thousand RMB in expected OOP costs, patients are also willing to pay an additional 50–55 thousand CNY to reduce financial uncertainty by 70% (i.e., the reimbursement rate in $j = 2$). This suggests that social surplus could be improved by allocating more of these patients with high valuation of financial risk protection to hospitals with higher reimbursement rates.

It's important to recall that, we determine patients' privately optimal choices given transportation subsidies/costs here, while these choices may not be socially optimal. To discuss socially optimal choices, we shall calculate the social surplus generated by allocating a patient to a given hospital based on Section 1.3.1.

### *1.5.4 Social Surplus*

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

According to Equation (1.5), we may decompose SS into two parts, as illustrated by Figure A.4. From the figure, we can see that the social cost of moral hazard is relatively small compared to willingness to pay especially for those with very high willingness to pay. As a result, social welfare gains from more generous hospitals are mainly driven by patients with the highest willingness to pay. This is driven by the shape of risk attitude (see Figure A.5) as well as the shape of risk itself.[33]

Eventually, we show the marginal social surplus generated by allocating patients to each hospital relative to the non-contracted (null) county hospital in Figure 1.12, by subtracting

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

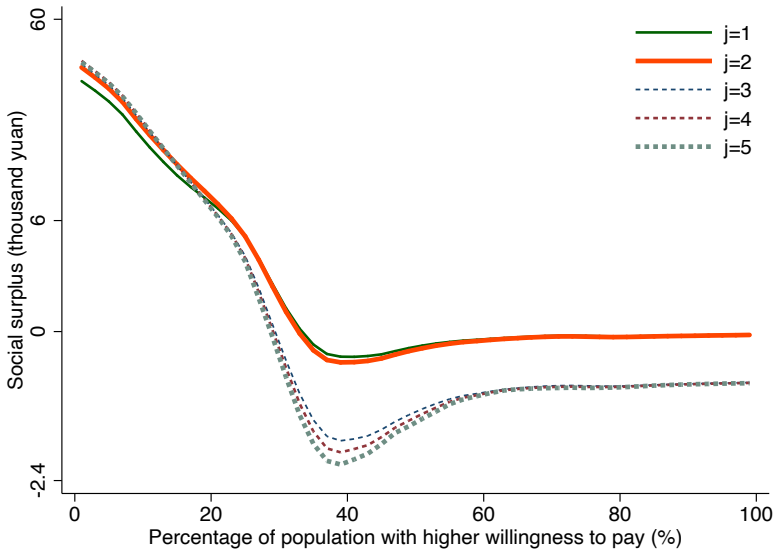



Figure 1.12: Social Surplus

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

Figure A.4b from Figure A.4a. Since patients can be screened by their willingness to pay, this is relevant for the optimal design of a health insurance program.

As we can see, for 65–70% of the population, social surplus curves for all contracted hospitals lay below zero, indicating that a non-contracted county hospital is the best hospital (from a social welfare perspective) when willingness to pay is low. This is because cost transfer at the lower end does not generate enough willingness to pay due to mistrust of quality. We can also find that, none of these hospitals are strictly the best. That is, the upper envelope of these social surplus curves is composed of multiple hospitals. At low levels of willingness to pay, the county hospital with the least generosity ($j = 0$) is the best (from a social welfare perspective); as willingness to pay increases, the more generous county hospitals ($j = 1$ and then $j = 2$) become the best; at very high levels of willingness

to pay, the more generous THCs ($j = 4$ and $j = 5$) become the best. Clearly, cost transfer is beneficial to the society only when consumers value it enough, and vertical differentiation is necessary for maximizing social welfare.

Nevertheless, the socially efficient hospital is an average or overall concept and is not necessarily the best for every patient. From Figure 1.10, we can notice that it is not even the best for an average patient sometimes.[34] To further investigate the heterogeneity in privately versus socially optimal hospitals across patients, Figure A.7 shows the distribution of efficient hospitals at every percentile of the average willingness-to-pay distribution. On average, when we assume zero additional (e.g., transportation) subsidies or costs, the non-contracted county hospital ($j = 0$) is only privately efficient for 0.8% of AECOPD patients, but is socially efficient for 49.0% of them; the most generous county hospital ($j = 2$) is privately efficient for 78.3% of them, but is only socially efficient for 25.7% of them; the THC with full insurance ($j = 5$) is privately efficient for 20.9% of them, but is only socially efficient for 6.7% of them; the less generous county hospital and THCs ($j = 1, 3, 4$) are never privately efficient, but are socially efficient for 13.2%, 1.4%, and 4.0% of these patients, respectively. Therefore, new policies can be designed to guide patients to choose the socially efficient hospitals over the privately efficient ones to achieve a larger social welfare target, which will be explored in the next section.

### *1.6 Counterfactual Policies*

The main objective of designing alternative (counterfactual) policies is to see how deductible and reimbursement maximum work and if there is a more socially efficient way to allocate patients to resources.[35] We consider three types of policies: (1) higher deductibles, (2) reimbursement cap adjustments, and (3) the combinations of the previous two.

---

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

[35] We do not consider the capacity constraint of each hospital, and thus do not check any potential over-crowding issue mentioned by Liu et al. (2018).

### *1.6.1 High Deductibles*

Based on the discussions in Section 1.5.3 and especially the evidence shown by Figure 1.11, we can see that, patients with low willingness to pay (who tend to be healthier) do not value the financial protection aspects in more generous hospitals as much as the disutility from their subjective perceptions about the low quality associated with price discounts. Then, based on Section 1.5.4, the more socially optimal allocation of resources would be to have these patients choose less generous hospitals (even the non-contracted ones). Under low deductibles, however, they start to receive price discounts too early, which does not generate more social welfare but leads them to have more moral hazard spending in more generous hospitals. This leads to a further social welfare loss. We may thus increase the deductibles and reserve the benefit of cost transfer to less "mistrustful" and more risk-averse patients with higher willingness to pay. By doing so, we improve the social welfare by the amount of disutility from mistrust plus the social cost of moral hazard. The question is, how high should deductibles be set at and should different hospitals with different reimbursement generosity also increase deductibles differently?

The NRCMS policy features low deductibles for all hospitals ($\leq$ 0.4 thousand CNY), and lower deductibles for lower-tiered hospitals, but the differences between them are small in absolute value compared to the differences in average spending. We hence experiment with six alternative high-deductible policies: (i) increase the deductibles of all hospitals slightly (by 0.5 thousand CNY); (ii) increase the deductibles of all hospitals moderately (by 1 thousand CNY); (iii) increase the deductibles of all hospitals greatly (by 2 thousand CNY); (iv) increase the deductibles and the gaps between hospitals slightly (multiply deductibles by 3); (v) increase the deductibles and the gaps moderately (multiply deductibles by 5); and (vi) increase the deductibles and the gaps greatly (multiply deductibles by 10).

Table 1.4 provides outcomes under the current (original) deductibles, as well as those under each of the six alternative policies with higher deductibles. Due to mistrust of quality associated with price generosity at the lower end of the spending distribution, increasing the gaps may encourage more patients to switch to higher-tiered (county) hospitals as they become relatively more attractive. If we increase deductibles without increasing the gaps,

Table 1.4: Outcomes of Alternative High-Deductible Policies

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

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

those mistrustful patients may maintain their hospital choice while more risk-averse patients can switch to lower-tired hospitals, leading to slightly more patients (especially those with small spending) to switch from county hospitals to THCs. Among more risk-averse patients, higher deductibles reduce the willingness to pay as they take away the value of "transfer" and moral hazard spending, but this is partially compensated by higher reimbursement rates in THCs when they are encouraged to switch from county hospitals; on the other hand, "mistrust" associated with price discounts can be mitigated. Since patients tend to be more mistrustful than risk-averse at the lower end of the spending distribution, and the "transfer" term does not contribute to social surplus, the overall social welfare can be increased under higher deductibles by the delayed exposure to mistrust and moral hazard. Of course, as we continue to increase deductibles, the loss of patient welfare will eventually outweigh the gain from insurer cost saving.

In our population, it seems that policy (ii), increasing the deductibles of all hospitals moderately (by 1 thousand CNY) without increasing their gaps, is a more efficient and logical choice. It encourages more patients with low health risks/needs to choose lower-tiered hospitals and save the medical resources in higher-tiered hospitals to those with higher risks/needs, and at the same time increases social welfare and reduces insurer/government

costs. Consumer surplus (the sum of willingness to pay) is slightly lower, but the allocation of resources becomes more efficient from the societal perspective.

### *1.6.2 Reimbursement Maximum Adjustments*

From Sections 1.5.3 and 1.5.4, we learn that patients with very high willingness to pay tend to be quite risk-averse and thus value the financial protection aspects very much, while their degrees of moral hazard are modest. Adjusting the shape of risk/uncertainty for them could lead to efficiency gains. However, since there are both mistrustful and risk-averse patients at the higher end of spending distribution (implied by the non-monotonic trend of the willingness to pay explained by "transfer" in Figure 1.11, as well as Figure A.6), and due to the fact that reimbursement maximum is working only on the higher end of the spending distribution, we may not have a definite answer to how we should adjust the cap. We experiment with fifteen alternative policies in Table 1.5.

The current reimbursement cap is set at 100 thousand CNY per year for every hospital, which seems to be close to the optimal level. First, changing the caps within a certain range (e.g., -10 to 10 thousand CNY) does not affect the average insurer cost significantly. This is partly because none of the patients in our data use up the reimbursement limit and most of them are quite far away from it. Second, changing the cap too much in either direction (e.g., $\pm$50 thousand CNY) seems to affect both consumer welfare and social welfare negatively. Changing the maximum can alter the shape of risks facing patients and lead to redistribution of hospital choices and reevaluation of financial protection and mistrust. Since the relationship between spending and how patients value financial protection is neither linear nor monotonic, there could be a certain level of reimbursement that is socially optimal. Third, we find that increasing the maximum of all hospitals or just county hospitals slightly (by 5 thousand CNY) can slightly improve both consumer welfare and social welfare—how much patients appreciate this financial protection aspect outweighs how much they are mistrustful of it. Fourth, it is interesting to note that, having the maximum in THCs higher than that in county hospitals can further reduce welfare. This is probably because it worsens mistrust of quality in THCs and reduces willingness to pay, and at the same time reallocates

Table 1.5: Outcomes of Alternative Reimbursement Caps

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

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

more patients to less generous county hospitals in which the value of financial protection is lower and further reduces willingness to pay. Fifth, when we lower the maximum in county hospitals greatly (by 50 thousand CNY), it starts to become binding for some patients, and insurer costs can be reduced. However, due to the large reduction in risk protection value and considerable increase in moral hazard spending (and social cost associated with it) from more risk-averse patients who switch to THCs,[36] both patient welfare and social welfare drop significantly, and the latter drops more.

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

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

### *1.6.3 Combination Policies*

We have discussed high deductibles and reimbursement cap adjustments separately. There is a concern that when we implement two sets of policies together, unexpected effects could arise. In this section, we are particularly interested in compensating higher deductibles (+0.5 to +1 thousand CNY) by higher reimbursement caps (+5 thousand to unlimited). Table 1.6 lists the outcomes under these combination policies.

Table 1.6: Outcomes of Combination Policies

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

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

First, there seems to be a "synergy" effect. For example, increasing deductibles in all hospitals by 1 thousand CNY alone can lead to a 2.22% increase in social welfare as shown by Table 1.4, and raising the reimbursement maximum in all hospitals by 5 thousand CNY alone can lead to a 0.09% increase in social welfare as shown by Table 1.5; however, if we

combine these two policies, the social welfare increase is 2.37%, which is larger than 2.22% + 0.09% = 2.31%. Similar agglomeration effects can be found in other combination policies in Table 1.6.

Second, this table shows that, the positive effects of high deductibles can outweigh the negative impacts of completely removing the reimbursement caps (i.e., allowing unlimited reimbursement). Thus, there is plenty of wiggle room for reimbursement maximum adjustments if policymakers intend to reduce resistance of high-deductible policies by a higher reimbursement limit.

## *1.7 Concluding Remarks*

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

There is rich variability in consumer preferences, and vertical differentiation is needed to improve allocative efficiency of medical resources in our context. Patients with lower willingness to pay tend to be mistrustful of quality associated with generosity and thus a lower coverage should be offered; on the other hand, high willingness-to-pay patients value financial protection enough to make a higher coverage efficient. The current policy, nevertheless, assigns lower coverage in higher-tiered hospitals, which further encourages patients with common diseases and minor illnesses to bypass primary care, as they tend to have lower willingness to pay. We propose to delay exposure to cost sharing by introducing higher deductibles, to mitigate the negative impact of mistrust, encourage primary care, and save insurer cost. Our counterfactual analysis suggests that a moderate increase of the deductibles in all hospitals (by 1 thousand CNY) can achieve a 2-percentage point increase

in social welfare, and significantly lower insurer cost by almost 30 percentage points (from 1.8 to 1.3 thousand CNY). Patient welfare is lower initially due to higher out-of-pocket costs, and thus policymakers may need to consider compensating tools to not only improve policy acceptance among patients but also make up for their welfare loss. The first compensating tool we consider is an increase of reimbursement limit. We find that there is plenty of leeway. Since moral hazard is modest compared to how much patients value financial protection at the higher end of the spending distribution and the reimbursement cap is not binding for most patients, removing the maximum (i.e., allowing unlimited reimbursement) would not completely take away the efficiency improvement from moderately higher deductibles. Other compensating tools focusing on the lower-income patients using the budget saving should be considered. Eventually, reallocation of the budget saving should make up for the initial welfare loss of patients, and improve the social welfare by a sizable amount.

It is important to be mindful that there are a few limitations that need be taken into consideration when interpreting the above conclusions. First, since we only observe patients who make a visit to a hospital (either a THC or a county hospital) within our study area, we do not model how patients decide whether to go to a hospital to treat their diseases when needed.[37] Thus, our counterfactual policies do not measure the welfare loss of patients when they are discouraged from seeking health care. In this sense, the welfare gains due to cost saving by high-deductible policies mainly reflect higher-value choices made by patients, rather than reduced needed care,[38] by assuming that they would continue to seek health care. Second, we do not consider protection by limited liability such as bankruptcy protection (Gross and Notowidigdo, 2011) and liquidity constraints (Ericson and Sydnor, 2018), which could potentially affect the shape of risks facing our patients. It would be interesting to explore how these distortions can affect consumer behaviors and our conclusions in future work. Third, we do not consider externalities of health care utilization, such as crowding out because of limited capability, by assuming that the socially optimal level is the one chosen

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

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

by patients without insurance. If there are positive externalities, the socially desirable level could include some additional health utilization induced by insurance. It could be challenging to evaluate externalities and determine the truly socially optimal level of health care utilization, but it should be considered a direction of future research. Fourth, to simplify our estimation of moral hazard, we assume health care to be a homogenous good conditional on the hospital chosen. However, the reality can be multidimensional and complex, and it could be important to extend our parsimonious model to capture more behavioral characteristics as a next step. Last, future research should try to separate mistrust of quality from risk aversion when studying consumers' health-care provider decisions.

Chapter 2

# EXAMINING THE ZERO-MARKUP DRUG POLICY IN CHINA: A STRUCTURAL APPROACH

## *2.1 Introduction*

To support the maintenance of health stock, policymakers around the world have been regulating their pharmaceutical industries for decades to lower drug prices and make health care more affordable, and China is not an exception. Since 2010, there have been a series of important regulatory changes in China's pharmaceutical industry, which affect the pricing decisions of firms and the drug choices of hospitals, physicians, and patients. In this chapter, we consider an influential reform, namely the Zero-Markup Drug Policy (ZMDP), which requires that hospitals cannot profit from dispensing drugs.

Although previous work has documented the aggregate effects of the ZMDP on some equilibrium outcomes such as drug prices, little is known about the underlying mechanism of how it works: e.g., how it changes physician choices, drug prices, firm profitability, and consumer welfare. We try to fill this void by first estimating a structural model of China's pharmaceutical industry and then quantifying the impacts on different parties in the market using counterfactual simulations.

One important feature of the demand for drugs in China is that there is an "expert-client" relationship such that a physician acts as a patient's agent. This relationship naturally generates agency problems because physicians concern both hospitals' profit from selling drugs and patients' welfare.[1] Since the 1950s, due to the lack of funding, the Chinese government explicitly allowed public hospitals to add a 15% markup to the wholesale prices

[1]The integration between drug prescription and dispensation has a long history in China, dating back to the Eastern Han Dynasty. Inspired by Zhang Zhongjing (A.D. 150–219), Chinese physicians started to "sit" in the pharmacies to provide services and name themselves after *zuotangyi* (on-site physicians). It cultivated the partnership of physicians and drug sellers. Sometimes, physicians may even open pharmacies themselves, known as *langzhong*. With the rapid transformation of the pharmaceutical and healthcare systems in Mao Era, on-site physicians flooded in public hospitals and became their employees.

of drugs when selling them to patients. Part of hospitals' profit became physicians' income. Consequently, Chinese physicians had been taking hospitals' drug markups into account when making drug choices for or with patients by deliberately prescribing more and expensive drugs. Although this agency problem was somewhat restricted by the 15% markup itself and a direct price upper limit regulation, these pricing constraints are not stringent enough to eliminate "distortions" in physicians' prescription decisions. To mitigate this incentive problem, in 2015, China started the ZMDP in prefecture-level public hospitals. This policy was then implemented nationwide in 2017 [2]

Eliminating hospitals' drug markups affects drug retail prices through three mechanisms:

1. *Direct effect:* if wholesale prices are fixed, it would directly lower retail prices.

2. *Dethronement effect:* it makes the expensive branded drugs less attractive than before (because of changes in physicians' incentive), which might decrease the market power of branded drugs and lower their prices (either wholesale or retail); on the contrary, the market power of generic drugs will likely increase, which leads to higher prices.

3. *Push-out effect:* removing markups also makes physicians less likely to prescribe in general (such as encouraging patients to go on a healthy diet instead), so the overall market power of prescription drugs decreases, which might lead to lower prices. Through these mechanisms, the ZMDP is very likely to reduce the retail prices of branded drugs. However, the overall impact on prices of generic drugs is ambiguous. The policy effects on retail prices translate into those on manufactures' profitability, patient welfare, etc., and quantifying these different aspects is the main goal of this chapter.

One difficulty is how to single out the effect of ZMDP from those of other policy changes that happen around the same time and aim for the same objectives. To tackle this problem, we develop a structural model of demand and supply of China's pharmaceutical market

[2]The pilot reform was launched earlier for county-level or township hospitals. But due to the lack of detailed data, in this research, we focus on prefecture-level public hospitals.

that exploits some institutional features and help us tease out the impact of ZMDP. This structural model will be more sensitive to the ZMDP than to other policy changes. Using data on drug sales and observed binding constraints on the prices of lipid-lowering drugs between 2012 and 2018, we first estimate the structural parameters in the model and then simulate the counterfactual unregulated equilibrium outcome in the absence of ZMDP. The comparison between the actual and counterfactual market outcomes gives us a quantitative account of the impact of ZMDP.

Our first step is to estimate the demand for differentiated lipid-lowering drugs. We follow the standard approach in empirical IO (Berry, 1994; Berry et al., 1995; Iizuka, 2007; Berry and Jia, 2010), and set up a two-type mixed nested logit model of the joint preference of a physician-patient pair based on observed drug characteristics, where the mixture captures the unobserved preference heterogeneity due to our partial observation on whether a hospital is subject to the ZMDP. We find that the estimated demand is strongly affected by hospital drug markups, implying that physicians do not fully represent the patients' interests. Also, physicians put more weight on patient welfare than hospitals' profits from drugs, as long as the coinsurance rate for drugs is low enough.

Once we have estimated demand and the implied substitution patterns, we explore the impact of the ZMDP on retail prices in a setting where competing drug manufacturers simultaneously negotiate with the provincial government about wholesale prices in a Nash bargaining game (Horn and Wolinsky, 1988; Crawford and Yurukoglu, 2012; Grennan, 2013; Gowrisankaran et al., 2015; Ho and Lee, 2017; Dubois et al., 2019a) given the observed constraints imposed by the regulator. The model allows us to separately identify costs and bargaining parameters, the latter of which captures how the provincial policymakers in China trade off between firm profits and patients' welfare.

Finally, given the estimated parameters of preference, production cost, and bargaining power, we can quantify how much the observed decline in China's prescription drug prices can be explained by ZMDP using counterfactual simulations. In particular, we calculate the new equilibrium prices in a hypothetical scenario in which ZMDP did not happen. Then, conditional on all other policies implemented in 2018, we can compare the counterfactual retail prices without the ZMDP and the observed actual retail prices in 2018. The mar-

ket shares, revenues, profits and social welfare under our counterfactual scenario are also compared to the estimates under the actual situation.

Our results have a few implications. First, though the prescription choices of physicians can be influenced by drug markups, physicians are more sensitive to patient's medication costs than hospitals' profits from drug markups, as long as the coinsurance rate is lower than 35 percent. Second, if we assume that the provincial government represents the consumer welfare, and simplify the centralized drug procurement into a 1-to-1 bargaining process between each firm and the government, then the pricing equilibrium is mostly dominated by provincial governments due to their strong bargaining power. Third, branded drugs are more preferred than generic drugs in China, and the demand elasticity for generic drugs is about 23 percent more elastic than branded drugs on average, suggesting a higher market power of the latter. Fourth, ZMDP makes generic drugs relatively more favorable, and thus can increase their profitability. Lastly, overall drug demand are weakened by the ZMDP, but due to the reduced prices, overall patient welfare is improved by more than 12 percent.

Our work is related to several strands of literature. First, it builds upon the broad research on estimating demand for pharmaceuticals using various methods to estimate preferences for drugs and substitution patterns, from the log-log models (Berndt et al., 1995) to the discrete choice models such as logit (Berndt et al., 2003b), nested logit (Iizuka, 2007; Donohue and Berndt, 2013; Song et al., 2017), and random coefficient logit (Björnerstedt and Verboven, 2016; Dubois and Lasio, 2018; Dubois et al., 2019a). Also, it relates to the research on physicians' agency problem. For example, Ho and Pakes (2014a) investigate the agency problem in physicians' referral decisions, and Lu (2014) studies physicians' overprescription behaviours. Furthermore, the demand model in this chapter is similar to Iizuka (2007), which shows Japanese physicians' prescription decisions respond to drug markups when diagnoses and drug sales are vertically integrated. Moreover, it relates to the empirical studies of double marginalization (Berto Villas-Boas, 2007; Bonnet and Dubois, 2010; Gayle, 2013).

This chapter belongs to the literature on the program evaluation of China's healthcare reforms such as the ZMDP (Zhou et al., 2015; Yi et al., 2015; Fu et al., 2018), Samming model and "two invoices" system (Meng et al., 2019), and Shenzhen's experiment with group

purchasing organizations (Yang et al., 2020). These existing studies are either case studies using data from only a sample city or are based on county-level hospitals. Case studies may fail to distinguish the effects of different components of a systemic reform, while studies that focus on county-level hospitals leave the effects in the cities unanswered. Our work fills the gap by evaluating the nationwide implementation of ZMDP among public hospitals in China using a nationally representative sample.

Finally, we contribute to the discussion on how a drug procurement system affects market outcomes. Baldi and Vannoni (2017) use the data on tender prices of selected drugs for hospital usage provided by 52 Italian local health service providers during 2009–2012 and find that centralized procurers pay lower prices than decentralized units, conditional on measures of institutional quality, corruption, and some other covariates. Duggan and Scott Morton (2006) study the effects of government drug procurement using the data on Medicaid prescription drug purchasing, and find that a set fraction of the average price paid by non-Medicaid consumers can not only increase the overall prices but also lead to the introduction of new products that are free of the Medicaid price regulation and thus more expensive. In this chapter, we investigate the role that centralized procurement plays in lowering the prescription drug prices. Different from Duggan and Scott Morton (2006), we focus on China's basic medical insurance that covers more than 95% of its population. Similar to Baldi and Vannoni (2017), we are also able to compare a centralized system with a decentralized system, through replacing the Nash bargaining between the government and firms by the Bertrand competition among firms.

The remainder of the chapter is organized as follows. Section 2.2 describes the empirical settings, including China's prescription drug market, the incentive problem, the regulatory efforts to solve the problem, the recent policy changes, the data used, and reduced-form evidence of price and quantity decline. In Section 2.3, we present the structural model of the demand and supply for each market, as well as the identification and estimation strategy. Section 2.4 presents the estimation results of the structural model. In Section 2.5, we then provide the counterfactual price equilibrium and profitability calculations in the absence ZMDP in 2018, and then calculate the welfare change for patients. Finally, we conclude in Section 2.6.

## 2.2 Background and Data

### 2.2.1 Drug Procurement Reform in China

In 2009, the Chinese government formally initiated a nationwide centralized drug procurement (henceforth CDP) scheme after 9 years of development and experiment in 4 provinces since 2000. The scheme is outlined in two documents released in 2010, namely *Notice on the Issuance of the Centralized Drug Procurement in Health Facilities* (Ministry of Health, 2010) and *State Council Office's Notice on Establishing and Standardizing Essential Drug Procurement in Government-sponsored Primary Health Facilities* (State Council's General Office, 2010). The new policy required that all public healthcare institutions could procure drugs only via their provincial governments' CDP platforms.

The procurement procedure can be described as the following. First, each hospital takes physicians' advice into account and submits a proposal of drug demand. Then, the provincial government evaluates those proposals and approves a list of drugs to participate in the procurement process. Finally, drug suppliers (e.g., manufacturers, domestic agencies of foreign pharmaceutical companies) compete on the drugs they would like to provide via a rather complicated bidding process.[3] The bidding process is not a standard scoring auction and the specific rules are different across provinces. Without detailed information, it's hard to exactly model this process. So in our empirical analysis, we proceed with a parsimonious model of bargaining between the governments and drug suppliers on drug prices a la Dubois et al. (2019a,b). After December 2018, the procurement process is changed/enhanced,[4] and so the data in 2019 are only used to generate summary statistics but not for estimating the structural model.

[3]For example, one popular bidding framework is the so-called "two envelope" bidding, in which drug suppliers are required to submit prices in one envelope (termed a price envelope) and the information of their drugs (such as indications) and suppliers (such as reputation) in another envelope (termed a quality envelope). The government then groups suppliers according to their proposals. Next, for each group, the government does a quality screening and chooses qualified candidates based on the quality envelope. Within each group, if there are only a few candidates (e.g., two), the government would directly negotiate with them, otherwise the government may simply choose several low bids (not necessarily the lowest one) as the winning suppliers.

[4]Joint procurement was carried out by "4+7" large cities in December 2018 and then by 27 provinces in September 2019.

The major players in the CDP scheme, who are the subjects of our research, are included in Figure 2.1. As mentioned in Ministry of Health (2010), the bargaining should be between pharmaceutical companies and provincial governments. Renegotiation between pharmaceutical companies and hospitals is prohibited.

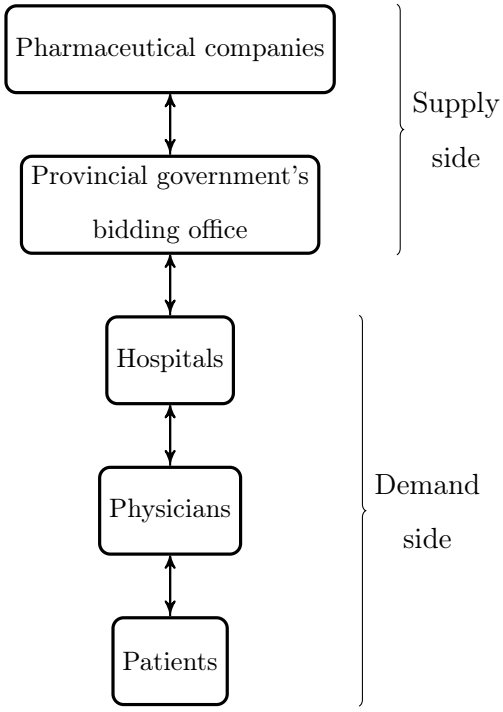



Figure 2.1: Basic Structure of China's Drug Market at the Provincial Level

### *2.2.2 The ZMDP and Price Regulations*

We briefly summarize the major regulatory policy changes that may affect drug prices during 2014-2018 in Table 2.1. The key policy change during this period is the ZMDP, which was implemented only among prefecture-level public hospitals in May 2015 and later extended to all public hospitals nationwide in September 2017.

To explain the implications of the price regulations on the retail price of a drug, let us denote $p^W$ as a wholesale price, which is the same for all hospitals in the same province and

Table 2.1: Major Policy Changes from 2014 to 2018

| Time | Description |
|---|---|
| Apr 2014$^{\diamond}$ | Remove the retail price caps for Lovastatin, Fenofibrate, Gemfibrozil, Xuezhiang, and Zhibituo. |
| May 2015* | Initiate the ZMDP among prefecture-level public hospitals (the start of the phase-in period). |
| Jun 2015$^{\diamond}$ | Remove retail price caps for all other lipid-lowering drugs. |
| 2015–2017 | Based on State Council's General Office (2015), the revenue from drugs should be no more than 30% of the total medical revenues in the urban public hospitals by 2017. |
| 2016–2018 | Encourage local governments to experiment with joint procurement. For example, Shanghai and Shenzhen experimented with some Group Purchasing Organizations (GPOs) in 2016; Beijing, Tianjin, and Hebei united in the procurement of medical supplies in 2017. |
| Mar 2016 | Launch the Generic Consistency Evaluation (GCE) program to test the quality and efficacy of generic drugs. The deadline for chemical drugs that entered before October 2007 was set to December 2018 but then it was canceled/extended. |
| 2017–2018 | According to State Council's Healthcare Reform Committee (2016), a "two invoices" system should be phased in among publicly owned medical institutions and implemented nationwide by 2018. |
| Sep 2017* | Zero hospital drug markup for all public hospitals. |
| Dec 2018 | "4+7" large cities joint procurement of Atorvastatin and Rosuvastatin. Winners take 60%–70% public hospital market shares in those cities. |

is decided by the bargain between the firm and the provincial government. Let $p^R$ denote the retail price of the drug at the hospital. Before April 2014, the regulations require that:

$$\frac{p^R - p^W}{p^W} \leq 15\% \text{ and } p^R \leq p^{Highest}, \tag{2.1}$$

where $p^{Highest}$ is the price cap imposed by the provincial government (may be different across provinces). We can rewrite (2.1) as

$$p^R \leq \min\{p^{Highest}, 1.15p^W\}. \tag{2.2}$$

After June 2015, the price cap is removed, so we simply have $p^R \leq 1.15p^W$. Finally, it is replaced by $p^R = p^W$ since 2017Q4.

### *2.2.3 Data and Descriptive Statistics*

We obtain quarterly data between 2012Q1 and 2019Q3 from the Pharmaceutical DataBase (PDB) on revenues and quantities of the prescription drugs in the "national drug catalog"[5] treating hyperlipidemia in the sample hospitals. The sample covers around 700 hospitals in 24 provinces of China. Among these hospitals, about 79% are tertiary and about 20% are secondary.[6]

In the raw data, the same drug can come with different forms (e.g., tablets and capsules) and sizes (e.g., 5mg and 10mg). We aggregate drug products (defined by a molecule-firm pair) with the same name but with multiple forms and sizes by "standard unit", the recommended daily dose of a given molecule produced by a given firm.[7] We obtain aggregate sales of different drug products at the province-quarter (defined as a "market" later) level, and then compute quarterly wholesale prices as the ratio of total revenue to total quantity in standard units. Retail prices are not directly observed from the data. We calculate them by assuming that the price constraint (2.2) is binding.[8]

Drug characteristics (including standard units, indications, and contraindications) are manually collected from the package inserts provided by YAOZH.COM and various sources (most of which are publicly available). Information on price caps are from YAOZH.COM as well. Firm characteristics (such as the time a firm was first allowed to produce each drug in China, and the time each firm was certified by GSP for distribution) are obtained from MENET.com.cn. We also manually collect the county-level minimum wages facing each manufacturer each quarter from the policy documents posted by local governments. Table 2.2 lists all of these variables.

We present the average wholesale prices in Figure 2.2 for some best-selling drugs, i.e.,

---

[5]The "national drug catalog" is designed for the basic medical insurance, work-related injury insurance, and maternity insurance.

[6]Very few hospitals are either lower-level or not classed and thus are negligible. For more details, visit http://pdb.pharmadl.com.

[7]We treat firms that share the same parent company as one firm.

[8]That is, we assume that the hospitals set the highest possible retail prices, as hospitals typically add a 15% drug markup when they can. Anecdotal evidence suggests that this assumption almost certainly hold in reality.

Table 2.2: Definitions of Main Variables

| Variable | Definition |
|---|---|
| *Drug characteristics:* | |
| Dose | Amount (mg) of drug taken at one time |
| Frequency | How often each drug is taken every day |
| Standard unit | Daily dose = dose × frequency |
| # of indications | Number of indications |
| # of contraindications | Number of situations in which the drug should not be used with another drug (drug contraindication) or by a patient (patient contraindication) |
| Chinese | Dummy = 1 if the drug contains Chinese herbal medicine ingredients |
| Old Statins | Dummy = 1 for the first / second generation of Statins |
| New Statins | Dummy = 1 for the third generation of Statins |
| Fibrates | Dummy = 1 if the drug belongs to Fibrates |
| Niacin | Dummy = 1 if the drug belongs to Niacin |
| # of forms | Number of drug forms by each firm |
| # of sizes | Number of drug sizes by each firm |
| *Firm characteristics:* | |
| First generic drug | Dummy = 1 if the drug is the first generic drug available in China |
| Branded | Dummy = 1 if the drug is branded |
| Time from entry | Number of quarters from entry in Chinese market |
| Foreign | Dummy = 1 if the firm is foreign-invested |
| *Cost shifters:* | |
| Min wage | Minimum hourly wage of the county in which the manufacturer is located |
| Imported | Dummy = 1 if the drug is imported |
| GSP | Dummy = 1 if the firm has the GSP certification for distribution |
| *Policy shocks:* | |
| Pilot rate | The proportion of cities in a province that pilot the systemic public hospital reform |
| Start GCE | Dummy = 1 if the firm has started the generic consistency evaluation |
| *Market performance:* | |
| Retail price | Price charged by each hospital per standard unit |
| Wholesale price | Procurement price per standard unit |
| Hospital markup | Difference between retail price and wholesale price |
| Market share | The ratio of the sales volume of a firm/drug to the total market sales volume |

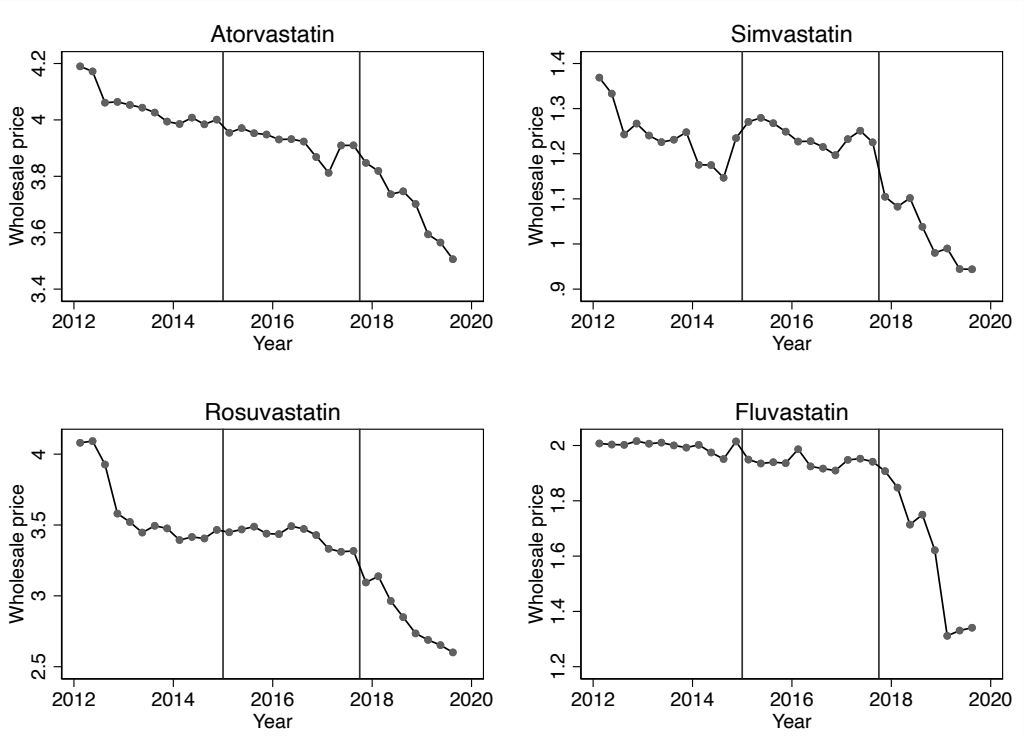



Figure 2.2: Average Prices of Top-Selling Lipid-Lowering Drugs in China

Atorvastatin, followed by Simvastatin, Rosuvastatin, and Fluvastatin since 2012. As shown, each of the best-selling prescription drugs treating hyperlipidemia experienced a modest decline in price before 2017Q3 and a drastic decline in price after the nationwide implementation of the ZMDP.

Finally, the summary statistics of drug characteristics and other variables used in this chapter are shown in Table 2.3. Patient-day unit retail prices vary across molecules, but it is 3.67–3.70 CNY on average before ZMDP pilot kicked in (roughly speaking 1 CNY equaled \$0.16 during 2012–2014). Physicians/hospitals in turn earned 0.43–0.46 CNY on average per patient-day (or 157–168 CNY per patient-year) by prescribing a lipid-lowering drug. Starting from 2017Q4, physicians could not earn such profits directly from dispensing drugs anymore. According to the number of quarters from first entry, we learn that some drugs are relatively new while some are quite old. And most of the drugs entered the Chinese market

Table 2.3: Summary Statistics

| | Obs. | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| *Price and markup (CNY):* | | | | | |
| *(2012Q1–2014Q4)* | | | | | |
| Retail price | 10,178 | 3.70 | 2.74 | 0.06 | 14.94 |
| Hospital markup | 10,178 | 0.46 | 0.36 | -5.07 | 1.95 |
| *(2015Q1–2017Q3)* | | | | | |
| Non-pilot retail price | 8,944 | 3.79 | 2.81 | 0.04 | 19.47 |
| Non-pilot hospital markup | 8,944 | 0.49 | 0.36 | -0.19 | 2.54 |
| Pilot retail price | 8,944 | 3.30 | 2.45 | 0.03 | 16.93 |
| Pilot hospital markup | 8,944 | 0 | 0 | 0 | 0 |
| *(2017Q4–2018Q4)* | | | | | |
| Retail price | 4,025 | 3.28 | 2.49 | 0.12 | 14.4 |
| Hospital markup | 4,025 | 0 | 0 | 0 | 0 |
| *Product and firm features:* | | | | | |
| # of indications | 23,147 | 3.05 | 0.97 | 1 | 4 |
| # of contraindications | 23,147 | 5.31 | 1.71 | 2 | 7 |
| First generic drug | 23,147 | 0.22 | 0.42 | 0 | 1 |
| Branded | 23,147 | 0.25 | 0.43 | 0 | 1 |
| Time from entry | 23,147 | 48.97 | 20.16 | 4 | 138 |
| Foreign | 23,147 | 0.28 | 0.45 | 0 | 1 |
| Chinese | 23,147 | 0.06 | 0.24 | 0 | 1 |
| Old Statins | 23,147 | 0.43 | 0.49 | 0 | 1 |
| New Statins | 23,147 | 0.28 | 0.45 | 0 | 1 |
| Fibrates | 23,147 | 0.19 | 0.39 | 0 | 1 |
| Niacin | 23,147 | 0.04 | 0.19 | 0 | 1 |
| *Cost shifters:* | | | | | |
| Min wage | 23,147 | 18.71 | 13.61 | 6 | 80.39 |
| Imported | 23,147 | 0.22 | 0.42 | 0 | 1 |
| GSP | 23,147 | 0.72 | 0.45 | 0 | 1 |
| *Policy shocks:* | | | | | |
| Pilot rate | 23,147 | 0.41 | 0.40 | 0 | 1 |
| Start GCE | 23,147 | 0.02 | 0.15 | 0 | 1 |

Note: Please refer to Table 2.2 for variable definitions.

before our sample period.

### *2.2.4 Reduced-Form Analysis of Price and Quantity*

As listed in Table 2.1, there are several regulatory changes that might affect demand or supply of lipid-lowering drugs in China during the period of study. Before showing how we structurally identify and estimate the impact of ZMDP on prices, demand, and profitability, we run reduced-form regressions of wholesale price and quantity over the three implementation phases of the ZMDP: pre-reform (2012–2014), partial reform (2015–2017Q3), and post-reform (2017Q4–2018) periods. Admittedly, we do not intend to explore any causal relationship with the reduced-form regressions as we may not completely rule out the effects of the "two invoices" system, GCE program, local joint procurement attempts, restricted revenue composition (e.g., revenues from drugs should account for less than 30 percent of total hospital revenues), and other regulations that were phased in during the same period of time. Nevertheless, the reduce-form regressions portray the trends of price and quantity over the three policy periods.

To make the price and quantity comparisons across different phases of the ZMDP meaningful, we control for the firm and drug fixed effects, and other characteristics of the drug products. We also try our best to control for measures of policy shocks that are uneven to different markets and firms in each quarter. Table 2.4 reports the results of the fixed-effect regression of the log wholesale price and quantity of drugs on the reform phase dummies and drug characteristics. When we control for the fixed effects, we see an evident price drop following the implementation of the ZMDP. In the first column, we can also see that branded drugs and those with more indications are more expensive; drugs with more contraindications are cheaper; older generic drug tends to have a lower price. Other policy shocks are also associated with lower wholesale prices. In the second column, we notice that the zero-markup drug policy was associated with lower quantity as well, seemingly indicating the "push-out" effect.[9]

This reduced-form evidence seems to confirm that vertical separation was associated

[9]The results are similar when we jointly estimate both equations.

Table 2.4: Fixed-Effect Regressions of Log Wholesale Price and Quantity

| | (1) log price | (2) log quantity |
|---|---|---|
| 2012–2014 (pre-reform) | (reference group) | |
| 2015–2017Q3 (partial reform) | -0.042*** | -0.197*** |
| | (0.005) | (0.027) |
| 2017Q4–2018 (post-reform) | -0.074*** | -0.246*** |
| | (0.007) | (0.037) |
| # of indications | 0.461*** | 1.848*** |
| | (0.090) | (0.466) |
| # of patient contraindications | -0.239*** | 0.685*** |
| | (0.035) | (0.178) |
| # of drug contraindications | -0.684*** | -0.460* |
| | (0.054) | (0.278) |
| First generic drug | -0.083*** | 0.681*** |
| | (0.020) | (0.103) |
| Branded | 0.075** | 3.209*** |
| | (0.035) | (0.182) |
| Pilot rate | -0.033*** | 0.936*** |
| | (0.006) | (0.033) |
| Start GCE | -0.220*** | 0.512*** |
| | (0.016) | (0.083) |
| Firm fixed effect | Yes | Yes |
| Molecule fixed effect | Yes | Yes |
| Observations | 23,147 | 23,147 |
| $R^2$ | 0.551 | 0.394 |

Notes: (1) Standard errors in parentheses under each coefficient. (2) Dependent variables are the natural log of wholesale price in CNY, and the natural log of quantity in "standard unit". (3) Data for China in 2012–2018. (4) ***, **, * denote significance level at 1%, 5%, and 10%, respectively.

with lower prices, conditional on other environmental changes being modest. Our structural estimation will allow us to interpret those results, test whether observed price changes are due to demand or supply conditions, and separate the effect of the ZMDP from other regulations

and environmental changes. As a result, we can tell how much of the price drop can be explained by vertical separation. Moreover, our structural approach will allow us to perform simulations of counterfactual policies.

## *2.3 Model*

In this section, we set up an empirical model of demand, supply and market equilibrium in the Chinese lipid-lowering drug market (focusing on hospital pharmacies).

### *2.3.1 Demand Side*

Given our data, we define a market as a province-quarter pair and label it by $t = 1, ..., T$. Each market $t$ consists a set of competing products, labeled by $j = 1, ..., J_t$, which are defined as molecule-firm pairs.

A patient and her physician jointly decide on which drug product to buy and use. So we model the joint preference of a patient-physician pair, labeled by $i$, using a standard nested-logit random utility model, i.e., the utility that $i$ obtains from choosing product $j$ is

$$U_{ijt} = \underbrace{X_{jt}\theta_1 + \xi_{jt}}_{\equiv \delta_{jt}} - \alpha P_{ijt} + \gamma M_{ijt} + \zeta_{igt}(\lambda) + (1-\lambda)\varepsilon_{ijt}, \tag{2.3}$$

where

- $\delta_{jt}$ represents the mean utility, in which $X_{jt}$ is a vector of observed product or market characteristics, including molecule dummies, the proportion of public hospital reform pilot cities in market $t$ (to control for reform intensity), the indicator for the start of GCE process, and a constant term, and $\xi_{jt}$ is an unobserved product-market level demand shock;

- $P_{ijt}$ and $M_{ijt}$ are the retail price and (hospital) markup that patient-physician $i$ faces for drug product $j$ in market $t$, and $\alpha$ is the dis-utility of price and $\gamma$ measures the severity of the expert agency problem (Iizuka, 2007);

- Depending on whether the hospital associated with $i$ is subject to the ZMDP, $P_{ijt}$ and $M_{ijt}$ differ across $i$'s: $P_{ijt} = p^W_{jt}$ and $M_{ijt} = 0$ if $i$ is subject to the ZMDP and

$P_{ijt} = p^R_{jt}$ and $M_{ijt} = m_{jt}$ otherwise;[10] also, for the partial ZMDP periods, we do not observe whether $i$ is subject to the ZMDP, so $P_{ijt}$ and $M_{ijt}$ become unobserved heterogeneity and we shall estimate the fraction of $i$'s that are not subject to the ZMDP as a parameter $\phi \in [0, 1]$;

- $\zeta_{igt}$ is a random variable that is common to all products in nest $g$, whose distribution depends on $\lambda$. $\lambda \in [0, 1)$ is the "nesting parameter" capturing the within-group correlation between choices. Larger $\lambda$ means nests matter more. $\varepsilon_{ijt}$ is an i.i.d. idiosyncratic preference shock following the standard Type I extreme value distribution. In our empirical analysis, a group $g$ is defined by a molecule and there are 17 of them (0 for outside goods, and 1–16 for the 16 molecules in Table B.1 except Jiaogulan).

Each decision maker $i$ in market $t$ maximizes her utility by choosing the best option in $\mathcal{J}_t$. Given nested-specification, the choice probability that $i$ chooses $j$ in $t$ can be written as

$$\sigma_j\left(\delta_t, P_{it}, M_{it}\right) = \underbrace{\frac{\exp\left(\frac{\delta_{jt}-\alpha P_{ijt}+\gamma M_{ijt}}{1-\lambda}\right)}{\sum_{j\in g}\exp\left(\frac{\delta_{jt}-\alpha P_{ijt}+\gamma M_{ijt}}{1-\lambda}\right)}}_{\text{within-group share}} \underbrace{\frac{\left(\sum_{j\in g}\exp\left(\frac{\delta_{jt}-\alpha P_{ijt}+\gamma M_{ijt}}{1-\lambda}\right)\right)^{1-\lambda}}{\sum_{g\in\mathcal{G}_t}\left(\sum_{j\in g}\exp\left(\frac{\delta_{jt}-\alpha P_{ijt}+\gamma M_{ijt}}{1-\lambda}\right)\right)^{1-\lambda}}}_{\text{group share}}. \quad (2.4)$$

Thus we can obtain the aggregate market share $E\left[\sigma_j\left(\delta_t, P_{it}, M_{it}\right)\right]$ by integrating out the heterogeneous $P_{it}$ and $M_{it}$. For the pre-2015 periods, all the $i$'s are not subject to ZMDP and thus

$$E\left[\sigma_j\left(\delta_t, P_{it}, M_{it}\right)\right] = \sigma_j\left(\delta_t, p^R_t, M_{ijt}\right). \quad (2.5)$$

Also, between 2015Q1 and 2017Q3 (partial implementation of ZMDP), whether each $i$ is subject to the ZMDP is an unobserved heterogeneity and thus

$$E\left[\sigma_j\left(\delta_t, P_{it}, M_{it}\right)\right] = \phi\sigma_j\left(\delta_t, p^R_t, M_{ijt}\right) + (1-\phi)\,\sigma_j\left(\delta_t, p^W_t, 0\right). \quad (2.6)$$

Finally, after 2017Q3 (full implementation of ZMDP), the market share equation is

$$E\left[\sigma_j\left(\delta_t, P_{it}, M_{it}\right)\right] = \sigma_j\left(\delta_t, p^W_t, 0\right). \quad (2.7)$$

[10] The variation in $p^{Highest}$ across markets will ensure non-colinearity between $p^R$ and the markup $m$, which will help us identify how the retail price and hospital drug markup affect the utility of consumers separately.

With the above specified market share function, we can write the demand system as

$$s_{jt} = \bar{\sigma}_{jt}\left(\delta_t; \theta_2\right), \ \forall j, t \tag{2.8}$$

where $s_{jt}$ is the observed market share of $j$ in $t$, $\bar{\sigma}_{jt}\left(\delta_t; \theta_2\right) \equiv E\left[\sigma_j\left(\delta_t, P_{it}, M_{it}\right)\right]$, and $\theta_2 = (\theta_1, \alpha, \gamma, \lambda, \phi)$.

To estimate the model, we invert the demand systems[11], (2.5), (2.6) and (2.7), to obtain

$$X_{jt}\theta_1 + \xi_{jt} = \bar{\sigma}_{jt}^{-1}\left(s_t; \theta_2\right) \tag{2.9}$$

and assume the following identification condition

$$\mathbb{E}\left[Z_{jt}^d \xi_{jt}\right] = 0, \tag{2.10}$$

where $Z_{jt}^d$ is a vector of exogenous variables, including arguably exogenous product characteristics, cost shifters ("Min wage" and "Imported" in Table 2.2) and BLP-type IVs: (1) the number of drugs and the sum of characteristics for other drugs sharing the same molecular class at market $t$ (the crowdedness of the product space), and (2) the number of drugs and the sum of characteristics for other drugs sold by the same firm at market $t$ (the ownership pattern).

Based on the moment condition (2.10), We estimate the demand model using GMM. Standard errors of the estimates are calculated according to the formulas provided in Section B.1.

### *2.3.2 Supply Side*

As discussed earlier, the wholesale price of a drug is determined jointly by its pharmaceutical company and the local government, which typically have distinct objective functions. In particular, we assume that pharmaceutical firms try to maximize their profits while governments concern the welfare of patients and physicians, following the literrature convention

[11] We solve the following contraction mapping and obtain $\xi_{jt}$, whose validity has been proven by Iizuka (2007) and Berry and Jia (2010):

$$\delta_{jt}^M = \delta_{jt}^{M-1} + (1-\lambda)\left\{\ln s_{jt} - \ln s_{jt}(\delta_{jt}^{M-1}, \theta_2)\right\}$$

where $M$ is the iteration number.

(Crawford and Yurukoglu, 2012; Grennan, 2013; Gowrisankaran et al., 2015; Ho and Lee, 2017; Dubois et al., 2019a). This is a parsimonious characterization of the trade-offs facing policymakers, who should balance producer profits against consumer welfare.

To capture the clear conflict of interests between firms and governments, we model the determination of wholesale prices of drugs using a simultaneous "Nash-in-Nash" bargaining model (Dubois et al., 2019a), in which each drug's wholesale price is negotiated bilaterally between its firm and a local government given the equilibrium prices of other bargain pairs. Following Dubois et al. (2019a), we assume that bargaining takes place at product-by-product level.

In each market $t$, the profit function of a firm supplying a set of products $\mathcal{F}_t$ is

$$\Pi_{\mathcal{F}_t,t}\left(\mathbf{p}_t^W\right) = N_t \sum_{j\in\mathcal{F}_t}\left(p_{jt}^W - c_{jt}\right)\bar{\sigma}_{jt}\left(\delta_t;\theta_2\right) \tag{2.11}$$

where $N_t$ is the market size of $t$. Note that we can write the profit function as a function of wholesale price only (given everything else) because the retail price is a fixed function of wholesale price (recall the discussion in Section 2.2.2).

For a given market $t$, the welfare is defined as the sum of the expected patient-physician joint utility produced by each drug available in market (Small and Rosen, 1981),

$$\Lambda_t\left(\mathbf{p}_t^W\right) = N_t E\left[\ln\left(\sum_{g\in\mathcal{G}_t}\left(\sum_{j\in g}\exp\left\{\frac{\delta_{jt}-\alpha P_{ijt}+\gamma M_{ijt}}{1-\lambda}\right\}\right)^{1-\lambda}\right)\right] \tag{2.12}$$

where the expectation is taken with respect to the heterogeneity in $P_{it}$ and $M_{it}$.

In each market $t$, the equilibrium prices solve the Nash-in-Nash bargaining problem

$$\max_{p_{jt}^W}\left\{\left[\Pi_{\mathcal{F}_t,t}\left(\mathbf{p}_t^W\right) - \Pi_{\mathcal{F}_t\setminus\{j\},t}\left(\mathbf{p}_t^W\right)\right]^{\rho_j}\left[\Lambda_{\mathcal{J}_t,t}\left(\mathbf{p}_t^W\right) - \Lambda_{\mathcal{J}_t\setminus\{j\},t}\left(\mathbf{p}_t^W\right)\right]^{1-\rho_j}\right\},\ \forall j \tag{2.13}$$

where $\rho_j \in [0,1]$ represents the relative bargaining power of the firm in the bargaining of product $j$'s price. The firm's objective is the change in profit generated by offering drug $j$ in market $t$. The government's objective is the change in consumer welfare generated by the presence of drug $j$ in market $t$. Note that we have assumed the bargaining power parameter of a product does not vary across markets. It's different from Dubois et al. (2019a), who studied seven countries while we only focus on China.

The first order condition of product $j$ in market $t$ is

$$c_{jt} = p_{jt}^W + \frac{1}{\underbrace{\frac{\partial \ln \bar{\sigma}_{jt}\left(\delta_t;\theta_2\right)}{\partial p_{jt}^W}}_{\text{Demand semi-elasticity}} + \frac{1-\rho_j}{\rho_j} \underbrace{\frac{\partial \ln \Lambda_{\mathcal{J},t}\left(\mathbf{p}_t^W\right)}{\partial p_{jt}^W}}_{\text{Welfare semi-elasticity}}}. \tag{2.14}$$

Note that (2.14) collapses to the first order condition of standard Bertrand-Nash equilibrium when $\rho_j$ equals to 1, i.e., when government's preference is not taken into account.

Next, we parameterize the marginal cost as follows:

$$c_{jt} = \left(Z_{jt}^s\right)' \beta + \omega_{jt}, \tag{2.15}$$

where $Z_{jt}^s$ includes a constant, the three cost shifters from Table 2.2, duration since entry, molecule and province-year dummies. Combining (2.14) and (2.15), we estimate the $\beta$ and $(\rho_1, ..., \rho_J)$ based on the least square criteria, i.e.,

$$\min_{\beta \in R^{k_\beta}, (\rho_1,...,\rho_J) \in [0,1]^J} \sum_{j,t} \omega_{jt}^2. \tag{2.16}$$

Given that $\beta$ enters the first order condition linearly, We simplify the optimization problem by concentrating out $\beta$ in close-form

$$\tilde{\omega}_{jt}\left(\rho_j\right) = \left[1 - \left(Z_{jt}^s\right)' \left[Z_{jt}^s \left(Z_{jt}^s\right)'\right]^{-1} Z_{jt}^s\right] \tilde{c}_{jt}\left(\rho_j\right), \tag{2.17}$$

where

$$\tilde{c}_{jt}\left(\rho_j\right) \equiv p_{jt}^W + \frac{1}{\frac{\partial \ln \bar{\sigma}_{jt}(\delta_t;\theta_2)}{\partial p_{jt}^W} + \frac{1-\rho_j}{\rho_j} \frac{\partial \ln \Lambda_{\mathcal{J},t}\left(\mathbf{p}_t^W\right)}{\partial p_{jt}^W}}. \tag{2.18}$$

Then we solve the simplified optimization problem

$$\min_{(\rho_1,...,\rho_J) \in [0,1]^J} \sum_{j,t} \left[\tilde{\omega}_{jt}\left(\rho_j\right)\right]^2. \tag{2.19}$$

## *2.4 Estimation Results*

### *2.4.1 Demand Estimation Results*

Demand estimation results are reported in Table 2.5. We can see that the physicians care both patients' and hospitals' interests, since the coefficients on retail price and hospital

markup are significant. To make sense of the estimated coefficients, we illustrate how physicians trade off the markup and patients' out-of-pocket cost via a simple example. Suppose that patients on average pay 20% of the cost of medication. Since the coefficient of hospital drug markup is approximately 2.89 times of that (absolute value) of retail price, a patient-physician pair is willing to give up 1 dollar of markup for a reduction of drug price (to a patient) by 58 cents ($\approx 2.89 \times 0.2$). That is, a patient-physician puts a greater weight on patient welfare than hospital profit (derived from drug) unless the coinsurance rate is higher than 35%. This finding resembles Iizuka (2007)'s results on Japanese market, where Japanese physicians are willing to give up 1 dollar if patient's cost is reduced by 28 cents, suggesting that the agency problem of physicians in Japan might be lighter than China.

Table 2.5: Demand Estimation Results

| | Coef. | St. Err. |
|---|---|---|
| # of indications | 5.085*** | 0.519 |
| # of patient contraindications | -0.188* | 0.102 |
| # of drug contraindications | -2.266*** | 0.260 |
| First generic drug | 0.251*** | 0.057 |
| Branded | 1.148*** | 0.063 |
| Time from entry | 0.034*** | 0.007 |
| (Time from entry)$^2$ | -0.000*** | 0.000 |
| Pilot rate | -0.042*** | 0.012 |
| Start GCE | 0.176 | 0.240 |
| $\alpha$ | 0.439*** | 0.080 |
| $\gamma$ | 1.268*** | 0.264 |
| $\lambda$ | 0.668*** | 0.005 |
| $\phi$ | 0.796* | 0.440 |
| Constant | -14.63*** | 1.100 |
| *Molecule dummies:* | | |
| *(Reference: Acipimox, Rosuvastatin, Simvastatin, Xuezhikang)* | | |
| Atorvastatin | 0.146 | 0.151 |
| Bezafibrate | -5.784*** | 0.347 |
| Ezetimibe | -0.161 | 0.236 |
| Fenofibrate | -7.855*** | 0.618 |

| | | |
|---|---|---|
| Fluvastatin | 0.617*** | 0.210 |
| Gemfibrozil | -11.991*** | 0.987 |
| Inositol Nicotinate | -2.270 | 1.601 |
| Lovastatin | -2.384*** | 0.321 |
| Pitavastatin | 1.904*** | 0.361 |
| Pravastatin | 2.082*** | 0.282 |
| Probucol | 6.436*** | 0.572 |
| Zhibituo | -0.854*** | 0.258 |
| Observations | 23,147 | |
| Objective function value | 0.140 | |

Notes: (1) ***, **, * denote significance level at 1%, 5%, and 10%, respectively. (2) See Table 2.2 for variable definitions. (3) $\alpha$ measures the disutility of price, $\gamma$ measures physician's marginal utility from drug markup (or the severity of expert agency problem) if there is any, $\lambda$ is the nesting parameter, and $\phi$ measures the average proportion of type 1 consumers between 2015Q1 and 2017Q3.

Other parameters in Table 2.5 also provide some interesting insights. For example, the number of indications significantly increases the demand, and branded drugs are also favored over generic ones. First mover advantage appears to exist in China's prescription drug market as the first generic drug marketed in China in its molecular class has a significantly higher demand. There is an upward trend in the demand for lipid-lowering drugs after entry, but the growth rate drops a little over time. Molecule dummies suggest that the demand for Statins is usually larger than drugs of other therapeutic class, except for Probucol. Public hospital reform seems to negatively impact the market share, while the generic consistency evaluation program may increase the market share (although not significant, probably due to a small sample issue because it's relatively new).

From the estimated demand model, we calculate the price elasticities and summarize them in Table 2.6 and Table 2.7. First, the mean own-price elasticity across products and markets in China in 2018 is -2.84 and ranges from -3.44 to -1.68 across markets. As expected, generics are more elastic than branded drugs (-2.98 versus -2.42), suggesting that even in 2018, after ZMDP is fully implemented, the branded drugs generally have higher market power in China. To see how the price elasticities change over time, Table 2.6 and Table 2.7 also report own- and cross-price elasticities for the main lipid-lowering drugs in China from 2012 to 2018. Own-price elasticity tends to be lower for both branded and generic drugs.

Table 2.6: Own-Price Elasticities for Main Lipid-Lowering Drugs, 2012–2018 (China)

| | Branded | | | | | | | Generic | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Subclass | Fibrate | Statin | Statin | Statin | Statin | CAI | Statin | Statin | Niacin | Statin |
| Company | Fournier | Luye | Novartis | Pfizer | AstraZ | SGP | MSD | Jialin | Lunan | Lunan |
| Molecule | Feno. | Xuezhikang | Fluva. | Atorva. | Rosuva. | Ezetimibe | Simva. | Atorva. | Acipimox | Rosuva. |
| Drug name | Tcicor | Xuezhikang | Lescol | Lipitor | Crestor | Zetia | Zocor | – | – | – |
| Year | Estimate | | | | | | | | | |
| 2012 | -9.918 | -3.206 | -2.478 | -7.414 | -4.305 | -7.473 | -3.119 | -10.713 | -6.373 | -9.149 |
| 2013 | -9.772 | -3.419 | -2.396 | -7.118 | -6.506 | -10.413 | -3.040 | -9.850 | -4.956 | -10.870 |
| 2014 | -8.450 | -3.119 | -2.111 | -7.114 | -5.538 | -9.878 | -2.932 | -9.100 | -5.187 | -9.038 |
| 2015 | -6.442 | -3.155 | -1.790 | -5.307 | -5.310 | -7.290 | -2.261 | -6.838 | -4.598 | -8.351 |
| 2016 | -5.834 | -2.855 | -1.817 | -4.690 | -4.740 | -6.319 | -1.997 | -5.315 | -4.190 | -7.485 |
| 2017 | -4.612 | -2.926 | -1.517 | -4.372 | -4.536 | -5.479 | -1.638 | -5.152 | -3.735 | -6.447 |
| 2018 | -3.570 | -2.020 | -0.704 | -2.818 | -2.987 | -3.267 | -1.159 | -3.261 | -2.268 | -3.708 |

Notes: (1) Each number is the estimated own-price elasticity of demand for the drug defined in the first few rows. (2) Company names: Luye stands for Luye Pharma Group, AstraZ is AstraZeneca, SGP is Schering-Plough, and MSD is Merck Sharp & Dohme. (3) Molecules: Feno. is Fenofibrate, Fluva. is Fluvastatin, Atorva. is Atorvastatin, Rosuva. is Rosuvastatin, and Simva. is Simvastatin. (4) Subclass: CAI stands for Cholesterol absorption inhibitors.

Table 2.7: Average Cross-Price Elasticities among Main Lipid-Lowering Drugs, 2012–2018 (China)

| | Branded | | | | | | | Generic | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Subclass | Fibrate | Statin | Statin | Statin | Statin | CAI | Statin | Statin | Niacin | Statin |
| Company | Fournier | Luye | Novartis | Pfizer | AstraZ | SGP | MSD | Jialin | Lunan | Lunan |
| Molecule | Feno. | Xuezhikang | Fluva. | Atorva. | Rosuva. | Ezetimibe | Simva. | Atorva. | Acipimox | Rosuva. |
| Drug name | Tcicor | Xuezhikang | Lescol | Lipitor | Crestor | Zetia | Zocor | – | – | – |
| Year | Estimate | | | | | | | | | |
| 2012 | 0.027 | 0.021 | 0.048 | 1.363 | 0.718 | 0.009 | 0.080 | 0.342 | 0.013 | 0.080 |
| 2013 | 0.027 | 0.023 | 0.051 | 1.429 | 0.924 | 0.017 | 0.063 | 0.314 | 0.018 | 0.119 |
| 2014 | 0.022 | 0.018 | 0.042 | 1.300 | 0.823 | 0.011 | 0.051 | 0.316 | 0.010 | 0.112 |
| 2015 | 0.019 | 0.015 | 0.033 | 1.019 | 0.702 | 0.012 | 0.036 | 0.261 | 0.008 | 0.123 |
| 2016 | 0.015 | 0.014 | 0.022 | 0.836 | 0.597 | 0.015 | 0.026 | 0.244 | 0.007 | 0.136 |
| 2017 | 0.012 | 0.012 | 0.012 | 0.796 | 0.512 | 0.015 | 0.019 | 0.230 | 0.006 | 0.152 |
| 2018 | 0.007 | 0.007 | 0.007 | 0.494 | 0.279 | 0.011 | 0.009 | 0.154 | 0.004 | 0.113 |

Notes: (1) Each number is the average of the estimated cross-price elasticities of demand for the drug defined in the first few rows with respect to (the price changes) of the other drugs. (2) Company names: Luye stands for Luye Pharma Group, AstraZ is AstraZeneca, SGP is Schering-Plough, and MSD is Merck Sharp & Dohme. (3) Molecules: Feno. is Fenofibrate, Fluva. is Fluvastatin, Atorva. is Atorvastatin, Rosuva. is Rosuvastatin, and Simva. is Simvastatin. (4) Subclass: CAI stands for Cholesterol absorption inhibitors.

Table 2.8: Revenue Per Market in 2012-2018 (China)

| Year | | All firms | Bottom 90% | Top 10 % |
|---|---|---|---|---|
| | All drugs | 24.62 | 6.17 | 18.45 |
| 2012 | Branded | 18.11 | 2.48 | 15.63 |
| | Generic | 6.51 | 3.69 | 2.82 |
| | All drugs | 25.20 | 6.02 | 19.18 |
| 2013 | Branded | 18.44 | 2.43 | 16.01 |
| | Generic | 6.76 | 3.59 | 3.17 |
| | All drugs | 25.42 | 5.72 | 19.70 |
| 2014 | Branded | 18.55 | 2.24 | 16.31 |
| | Generic | 6.87 | 3.48 | 3.39 |
| | All drugs | 25.31 | 5.89 | 19.42 |
| 2015 | Branded | 17.60 | 2.17 | 15.43 |
| | Generic | 8.05 | 4.06 | 3.99 |
| | All drugs | 25.33 | 6.18 | 19.15 |
| 2016 | Branded | 17.28 | 2.28 | 15.00 |
| | Generic | 8.05 | 3.9 | 4.15 |
| | All drugs | 25.03 | 5.87 | 19.16 |
| 2017 | Branded | 17.09 | 2.28 | 14.81 |
| | Generic | 7.94 | 3.59 | 4.35 |
| | All drugs | 24.59 | 5.70 | 18.89 |
| 2018 | Branded | 16.67 | 2.25 | 14.41 |
| | Generic | 7.92 | 3.44 | 4.48 |

Notes: (1) Market is defined by a specific quarter of a year in a province in China. (2) Revenue is sample estimation, which is just 20-30% of the real-world values. (3) Revenue is in 100 million CNY.

Also, lipid-lowering drugs becomes less substitutable as indicated by lowering magnitudes of cross-price elasticities. These drugs are more substitutable within a molecular class (e.g., Atorvastatin produced by Pfizer versus Jialin, or Rosuvastatin produced by AstraZeneca versus Lunan) than between branded and generic groups.

The decreasing price sensitivity might seem a bit surprising given that retail prices are also decreasing, because standard oligopoly theory tells us that they should be inversely related. However, recall that the overall demand becomes much weaker (less prescriptions from physicians) after ZMDP so the market become more competitive, which explains the decreasing price.

Table 2.8 shows that since 2015 the total revenue of all drugs keeps decreasing, which could be attributed to the push-out effect. Also, the total revenue of top 10% generic drugs is increasing in the meantime, which may be due to the dethronement effect (i.e., the market power of top generic drugs is increasing).

### *2.4.2 Supply Side Estimation*

We first present our estimates of bargaining power parameters $\rho_j$ in Figure 2.3. It's not surprising to see that most firms/products have lower bargaining power than the provincial governments (indicated by $\rho_j < 0.5$), and only a small fraction of firms/products show higher bargaining power than the government.

To show the goodness of fit of the bargaining model, we predict the wholesale prices using our estimated marginal cost function $c_{jpt}(\rho_j)$, following Pakes (2017) and Wollmann (2018). The predicted prices and actual prices are largely centering around a 45-degree line. The linear regression of actual prices on predicted prices without a constant gives a coefficient of 0.998, which is almost 1 (see Figure 2.4 for illustration).

We also look at the predicted price index and compare it with the actual one (like the one in Table B.2). As shown by Figure 2.5, predicted price index is rather close to (although slightly lower than) the actual one and captures the general declining trend over time.

Using our estimated demand parameters, bargaining power parameters, and pricing equilibrium, we can then estimate total revenue and profit of each market. Before showing the

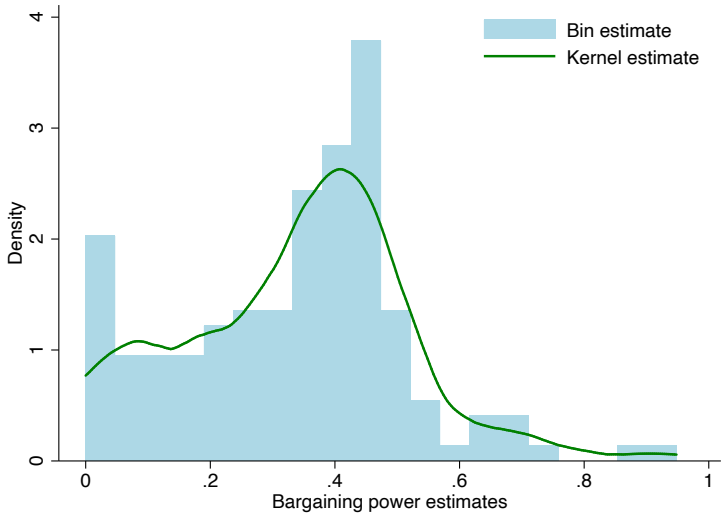



Figure 2.3: Distribution of Bargaining Parameters

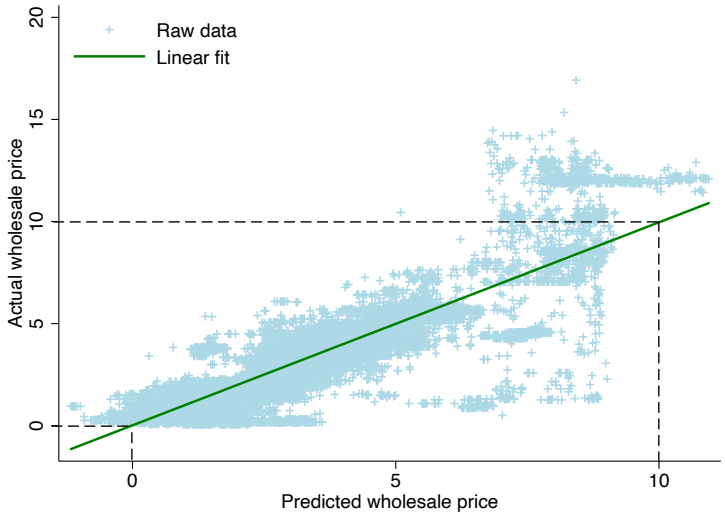



Figure 2.4: Predicted Prices and Actual Prices

total revenues and total profits, we provide the distribution of estimated margins of each product in 2018 in Figure 2.6 and Figure 2.7. The average profit per each standard unit across products and markets (i.e., each observation is weighted by the corresponding amount

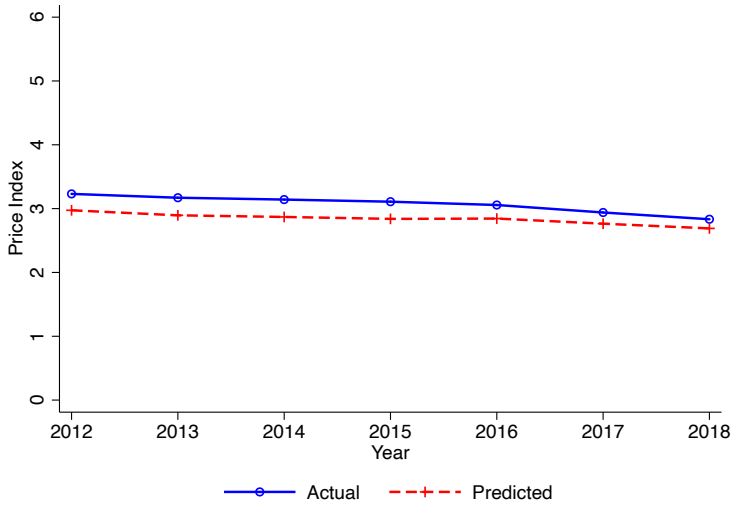



Figure 2.5: Predicted Price Index and Actual Price Index

of standard units sold) in 2018 is 0.63 CNY, ranging from nearly 0 to 2.15 CNY. Profit margin, or price-cost margin (also known as the Lerner index), is 0.27 on average, and most products exhibit a relatively low market power.

We noticed that branded drugs typically have a higher price-cost margin than generic drugs.[12] As pointed out by Dubois and Lasio (2018), it is known in the industry that generic firms have lower marginal costs. Our calculations suggest that, in 2018, the (weighted) average cost of a standard unit of generic drugs is 2.48 CNY, compared to 2.91 CNY for branded drugs. The prices of generic drugs, however, are much lower than branded drugs (2.79 versus 3.69), suggesting lower margins of generic drugs.

We summarize the average revenue and profit per market in Table 2.9. In an average market, branded drugs take up the majority (61 percent) of the market share, and the top 10 percent best selling branded products account for 83 percent of the branded market share, indicating high market concentration. The total manufacture revenue of an average market (defined by a season-province pair) in 2018 is 2.46 billion CNY, and the total manufacture

[12]Price-cost margin is defined as the difference between wholesale price and marginal cost as a fraction of wholesale price. Weighted averages are 0.28 versus 0.19.

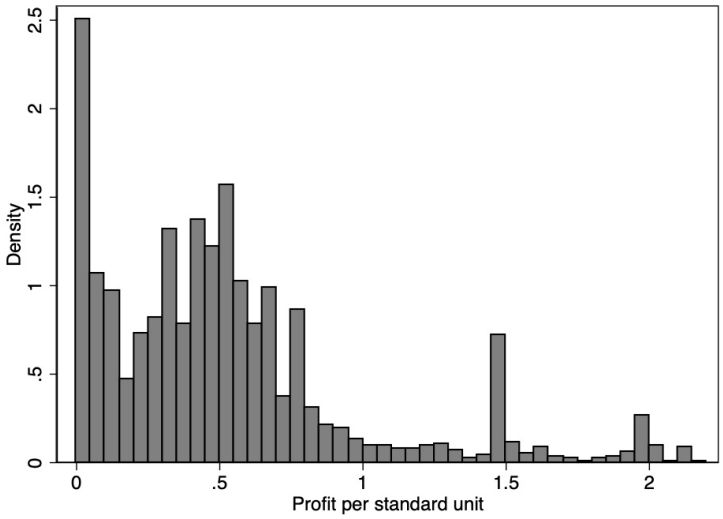



Figure 2.6: Estimated Profit Per Standard Unit in 2018

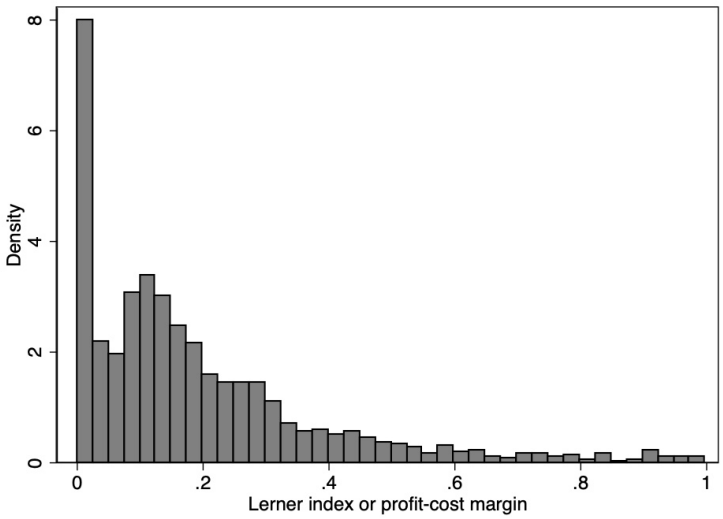



Figure 2.7: Lerner Index in 2018

profit of a market is 0.45 billion CNY. Due to higher market shares and higher prices, branded drugs are more lucrative. The average revenue of all branded drugs is 110 percent higher than that of all generic drugs, while total profit is 321 percent higher.

Table 2.9: Market Share, Revenue and Profit Per Market in 2018 (China)

| | | All firms | Bottom 90% | Top 10 % |
|---|---|---|---|---|
| Market share (%) | All drugs | 38.99 | 11.97 | 27.02 |
| | Branded | 23.73 | 3.93 | 19.80 |
| | Generic | 15.26 | 8.04 | 7.22 |
| Revenue | All drugs | 24.59 | 5.70 | 18.89 |
| | Branded | 16.67 | 2.25 | 14.41 |
| | Generic | 7.92 | 3.44 | 4.48 |
| Profit | All drugs | 4.47 | 0.87 | 3.60 |
| | Branded | 3.61 | 0.27 | 3.34 |
| | Generic | 0.86 | 0.60 | 0.26 |

Notes: (1) Market is defined by a specific quarter of a year in a province in China. (2) Revenue and profit are in 100 million CNY.

### *2.5 Counterfactual: Quantifying the Effects of ZMDP*

In this section, we examine how profit and consumer surplus were affected by ZMDP that breaks the integration between prescribing and dispensing drugs (Iizuka, 2007). To avoid the complication of price caps that were in place during the transition periods, we conduct the counterfactual simulation based on on the data of the post-reform era, i.e., 2018. Specifically, we assume the absence of ZMDP such that pre-reform hospital markup, i.e., 15 percent of wholesale price, is restored. Then, we calculate counterfactual equilibrium prices, market shares, profits, etc., using the estimates and data of 2018.[13]

Figure 2.8 compares the counterfactual retail prices to the actual prices, showing that the distribution of counterfactual prices shifts to the right, i.e., if the 15% hospital drug markup still existed, the average retail price would be higher. The profit from selling a standard unit of lipid lowering drug (defined by the difference between the retail price and the marginal

[13]We solve for new equilibrium prices using firms' first-order conditions. A fixed point algorithm was used to solve the system with a numerical tolerance level smaller than $10^{-6}$.

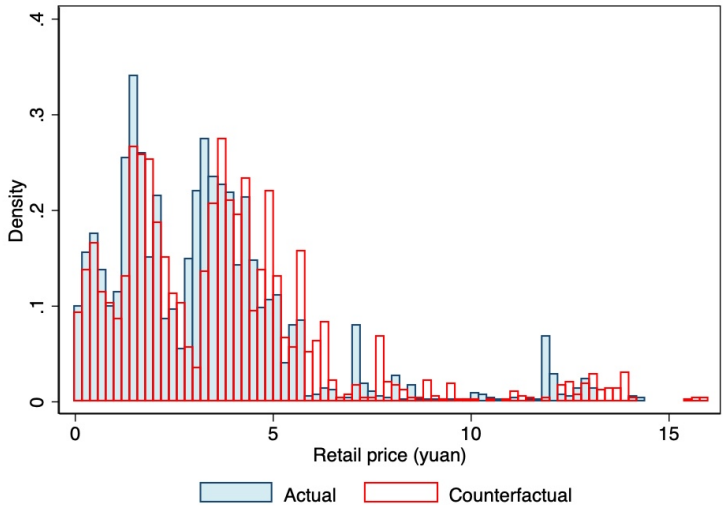



Figure 2.8: Counterfactual and Actual Retail Prices in 2018

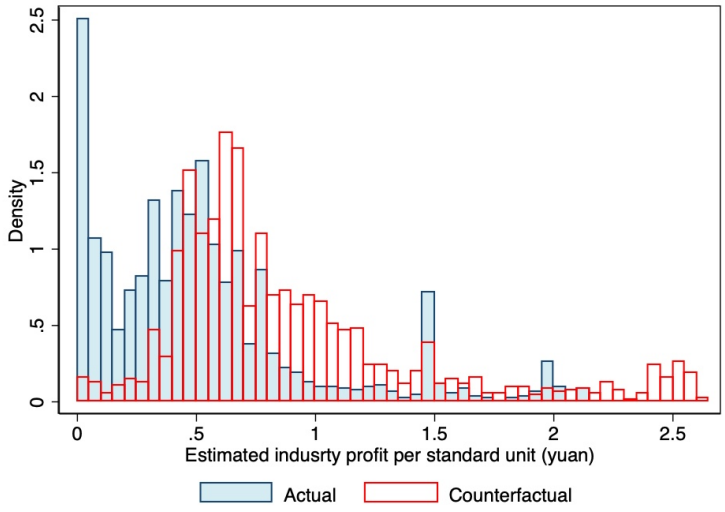



Figure 2.9: Counterfactual and Actual Industry Profits Per Unit in 2018

cost) is shown in Figure 2.9. Again, it is clear that hospital drug markup would generate higher profits.

Based on the counterfactual market equilibrium, we calculate the implied market shares,

revenues, and profits that are summarized in Table 2.10. The share of an average market could expand to nearly 51 percent of the hyperlipidemia population with the hypothetical 15 percent hospital drug markup in 2018. Compared to the estimates in Table 2.9, branded drugs could experience a much larger increase in market share, and become even more concentrated; this again suggests that, without ZMDP, top selling branded drugs are more preferred by physicians.

Table 2.10: Counterfactual Share, Profit, Revenue, and Surplus Per Market (2018)

| | | All firms | Bottom 90% | Top 10 % |
|---|---|---|---|---|
| Market share (%) | All drugs | 50.58 | 13.52 | 37.06 |
| | Branded | 34.57 | 4.69 | 29.88 |
| | Generic | 16.01 | 8.83 | 7.18 |
| Revenue | All drugs | 33.81 | 7.53 | 26.28 |
| | Branded | 24.29 | 2.93 | 21.36 |
| | Generic | 9.52 | 4.61 | 4.92 |
| Profit | All drugs | 4.83 | 0.82 | 4.01 |
| | Branded | 4.09 | 0.26 | 3.83 |
| | Generic | 0.74 | 0.56 | 0.18 |
| Patient surplus change | | | -12.24% | |

Notes: (1) We use 2018 product attributes for all counterfactual exercises. The counterfactuals assume that there is a 15% drug markup just like 2012-2014. (2) Other cautions are in Table 2.9. (3) Revenue and profit are in 100 million CNY.

Comparing to actual data, the total revenue under the counterfactual scenario goes up by 37.5 percent, and the total profit increases by 8 percent. Generic drugs would lose some profit if ZMDP were removed in 2018. This is consistent with our conjecture – without the drug markup, the dethronement effect should lead to a higher relative market power of generic drugs compared to the branded drugs. Note that this result is also consistent with the estimated revenues in Table 2.8.

Finally, we measure the changes in patients' welfare due to the counterfactual drug markup. This is done by assuming that the utility function fully represents patients' preference (so $\gamma$ is fixed at 0). Our calculation suggests that, patient's welfare would drop by 12.24 percent if there were a 15 percent hospital drug markup. This results suggest that the ZMDP is overall beneficial to patients.

### *2.6 Concluding Remarks*

In this chapter, we develop a structural model of China's prescription drug market to investigate the impact of the ZMDP. Using the data from PDB and various sources (e.g., *menet*, *yaozh*, package inserts, policy documents, etc.) on wholesale transactions of lipid-lowering drugs in a sample of hospitals during 2012–2018, we first estimate a mixed nested-logit demand model that accommodates the consumer heterogeneity due to zero-markup drug policy pilot programs. The demand estimation suggests that lipid-lowering drugs are highly differentiated. Brand-name drugs are preferred to their generic versions, which is in line with the literature. Moreover, physicians' prescription decisions are affected by the hospital markups, although they care more about patient welfare and choose drugs that have less out-of-pocket costs, unless coinsurance rate is high.

Under the assumption that prices are set according to Nash bargaining between each firm and the corresponding provincial government in China, we separately identify costs and bargaining parameters, the latter of which can be interpreted as the degree to which the government leaders choose to trade off between firm profits and immediate consumer welfare. Results suggest that most policymakers value consumer welfare more, and thus firms typically have a bargaining power parameter that is less than 0.5.

We then perform a counterfactual analysis by removing the ZMDP in 2018 and then quantify its impact on firms' profitability and patients' welfare. Our calculations indicate that, ZMDP leads to an increase in patients' welfare by about 12 percent. Moreover, ZMDP benefits generic drugs more than branded ones and make the branded drugs less concentrated. Overall, this counterfactual exercise confirms that ZMDP largely achieves its policy goal of reducing drug prices and increasing patients' welfare.

Finally, we close this chapter by mentioning some caveats and limitations of our work

left for future research. First, our results are based on a static model and does not include dynamic considerations, such as investment, entry and exit, etc., so they only reflect a short-term evaluation of the policy effects. Also, we only focus on lipid-lowering drugs in a selected sample of hospitals in this chapter, a more comprehensive investigation that covers more hospitals and types of drugs would be helpful to understand ZMDP to a greater extent.

Chapter 3

# TRADE LIBERALIZATION AND SKILL UPGRADING: EVIDENCE ON THE IMPACT OF APTA ON CHINESE MANUFACTURERS

## *3.1 Introduction*

China's performance in economic growth is remarkable, and one significant contributor is its human capital accumulation. Human capital accumulation has improved sharply in China in the past twenty years, beyond maintaining health stock. For instance, its tertiary school enrollment rate has increased from 7.7% in 2000 to 53.8% in 2018 according to the World Bank. Workers upgrade their skill levels through education or on-the-job training. In the meanwhile, we observe a contemporaneous increase in its trade openness. After joining the World Trade Organization (WTO) in 2001, China engages further in international trade and attracted more foreign direct investments (FDI). It benefits greatly from trade integration through acting as a leading country of exports, inducing capital inflows and promoting economic growth. Besides the WTO accession, regional trade liberalization such as the Asian Pacific trade agreement (APTA) and Regional Comprehensive Economic Partnership (RCEP) also foster economic development in China. Our firm level data show a pattern that Chinese new exporters have greater incentives in providing labor training than non-exporters. This indicates that there can be a relationship between exports and human capital investment, but it is not enough to explain whether expanded export opportunities encourages firms to invest more in human capital for innovation or vice versa. In this chapter, we intend to study the effect of trade liberalization (APTA) on export and skill upgrading decisions of Chinese manufacturers theoretically and empirically.

The trade model with heterogeneous firms in Melitz (2003) and Bernard et al. (2003) emphasizes that trade integration reallocates market shares towards exporters who are larger, more productive and more skill- and capital-intensive than non-exporters. In our benchmark model, following the literature, more productive firms find it profitable to pay for the fixed

costs of entering the export market, and those with even higher productivity choose to invest in human capital with fixed skill-upgrading costs because firms receiving larger sales are able to provide labor training, which is also consistent with our data pattern. A reduction in trade cost increases export sales and encourages more new entrants in the export market, and then it also induces more firms to invest in human capital and produce skill-intensive products.

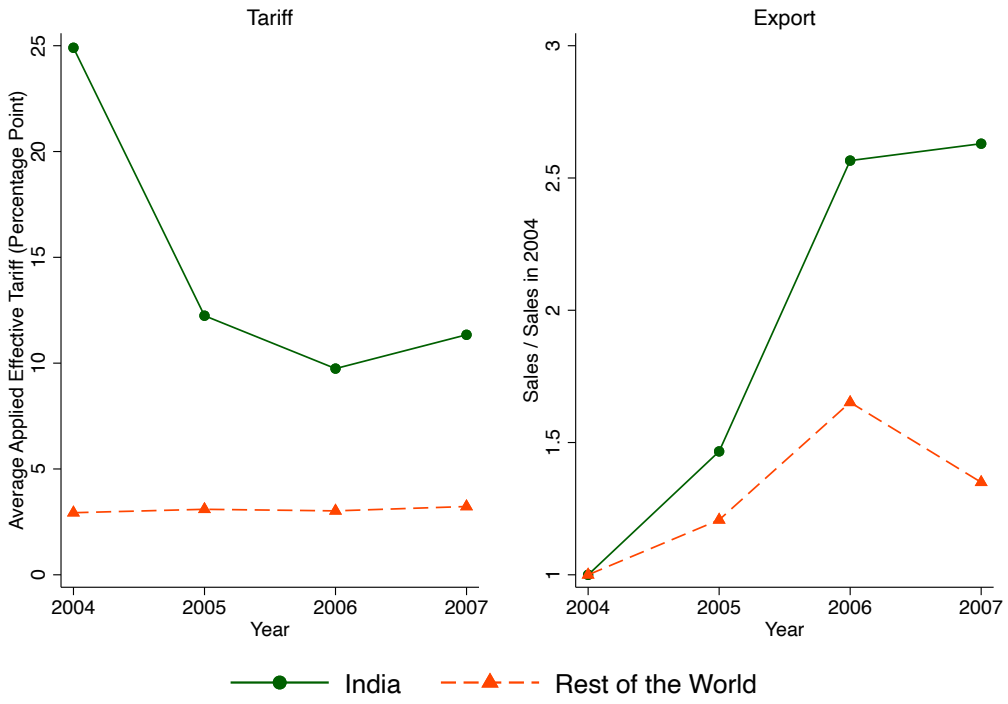



Figure 3.1: Trends of Tariffs and China's Export Sales (2004–2007)

Notes: The left panel shows the average applied effective tariff China faces when exporting goods to India and the rest of the world; the average for the rest of the world is weighted by the export sales from China to each country. The right panel depicts China's total export sales in each year relative to 2004.

Figure 3.1 (green solid line) reflects that a reduction in India's tariff leads to a boost in exports to the Indian market by Chinese firms. India's average applied effective tariff declines

by about 15 percentage points from 2004 to 2007, but the change in the average tariff in the rest of the world (Figure 3.1's red dashed line) is nearly zero during the same period. Based on this fact, we extend our model to distinguish export destinations—the "main" trading countries with lower and stable trade barriers (the rest of the world) and the less preferential trading partners such as India who impose relatively higher (but declining) tariffs—and then analyze the impacts of trade liberalization on firms' export participation and investment in on-the-job training in the home country (China). Regional trade liberalization policies are expected to boost regional trade and investment, facilitate technology and skill upgrading, and stimulate economic growth. We test the model in the context of a regional trade liberalization episode, APTA, through estimating the impact of the reduction in India's tariffs on firm entry in the export market and labor training provided by firms between 2004 and 2007.

We start our empirical analysis by exploring data patterns. In the first check, we investigate whether the sorting pattern predicted by the benchmark model is consistent with the observed differences between exporters and non-exporters. The model implies that productivity differences produce a sorting of firms in four groups: the least productive firms exit the market (not in data), the low productivity group employs low-skilled workers and serve the domestic market only, the middle group exports but still produces unskilled goods with demand for low-skilled workers, and the most productive firms both export and produce skilled goods through upgrading labor skill levels. Indeed, the data confirm that exporters provide more labor training than non-exporters in 2004. Moreover, new exporters increase investment in human capital faster than continuing exporters during the 2004–2007 liberalization period. It is plausible that new exporters produce more skill-intensive products in order to become more competitive in the foreign market.

In the second data check, we investigate whether the sorting pattern predicted by the extended model is consistent with the observed differences between exporters to main trading partners and exporters to less preferential trading partners. In the extended model, the least productive firms exit the market; the lower-middle group exports unskilled goods to main trading countries; the upper-middle group exports unskilled goods to less preferential trading partners; the high productivity group upgrades labor skills, and only exports skilled goods

to main trading countries; the most productive firms are able to both provide labor training and export skilled goods to less preferential trading countries. The extended model assumes that some firms find it more profitable to export skilled goods to main countries than to export unskilled goods to other countries who impose higher trade barriers. This assumption comes directly from the data pattern, as we notice that switching exporters from India to the rest of the world increase training spending per worker faster during 2004–2007. The data pattern also shows that exporters to less preferential trading countries such as India invest more in on-the-job training than those exporting to main countries (the rest of the world) in 2004, except for those who switch destination countries in 2007. In particular, both new and switching exporters to India increase labor training slightly more than continuing, exiting, and never exporters during the regional trade liberalization period.

The data patterns describe above show that there is a coincidence between export participation and skill upgrading but do not address the question of whether trade liberalization induces firms to invest in human capital and produce skill-intensive goods instead of unskilled goods. Thus, we attempt to establish causality by linking exporting and skill upgrading directly to the reduction in India's tariffs for imports from China. This is a direct test of the model where firm decisions to enter the export markets and provide labor training are endogenous. In the meantime, we compare two export destinations imposing different trade costs, and analyze whether exporters switch to another export market following trade liberalization.

First, the benchmark model predicts that the productivity cutoffs to enter the export market and to upgrade labor skills fall more when tariffs fall more. Firms find it easier and more profitable to participate in foreign markets and provide labor training following trade integration. Thus, we estimate the change in the probability that a firm enters the export market as a function of the change in India's tariffs. The average reduction in tariffs (15 percentage pints) increases the probability of entering the export market by 1.55 to 1.88 percentage points from 2004 to 2007. Then, we estimate the change in spending on labor training as a function of the change in tariffs. The average reduction in tariffs increases spending on labor training by 0.11 to 0.13 log points during the same period. The above

empirical results are from a sample of selected sectors.[1] In the other sectors, we find that the effect of trade liberalization on export participation is positive but insignificant, and the lower trade costs even decrease spending on training. The different results between selected sectors and other sectors are due to sector heterogeneity in productivity and policy in China. Therefore, some of our empirical findings can go beyond the benchmark model predictions, which indicates a need for model extension.

Next, the extended model predicts that the reduction in tariffs of less preferential trading partners (country $o$) increases the probability of entering the export markets of both main trading partners (country $m$) and less preferential trading partners (country $o$). Meanwhile, lower trade costs induce more spending on labor training provided by exporters to country $o$, but discourages skill upgrading of exporters to country $m$ as predicted by the extended model. The extended estimation of selected sectors shows that the reduction in India's tariffs induce more export entry in the Indian market and increase spending on labor training provided by new exporters to India. Consistent with model prediction, new exporters have a larger likelihood to enter the Indian market and new exporters to India induce more investment in on-the-job training following the reduction in India's tariffs.

This chapter contributes to the theoretical literature that studies the mechanism of how trade openness affects firms' investment training. The theoretical model in this chapter builds on Melitz (2003) and Bustos (2011b).[2] The heterogeneous-firm model offers new insights into the causes and consequences of international trade.[3] There are model spec-

[1]They are as follows: "wood processing and wood, bamboo, rattan, palm and grass products" (No. 32), "coatings, inks, pigments and similar products" (No. 42), "daily chemical products" (No. 45), "rubber products" (No. 48), "plastic products" (No. 49), "brick, stone and other building materials" (No. 52), "boilers and prime movers" (No. 64), "metal processing machines" (No. 65), "mining, metallurgy, and special equipment for construction" (No. 69), "special machines for agriculture, forestry, animal husbandry and fishery" (No. 71), "ship and floating devices" (No. 75), "household electric and non-electric appliances" (No. 80), "instrumentation" (No. 88), and "cultural and office machines" (No. 89).

[2]Bustos (2011a) points out that firms upgrading skills also upgrade technology, and analyze skill upgrading in the context of the employment share of skilled workers. In this chapter, firms make skill upgrading decisions through increasing spending in labor training.

[3]Recent literature also incorporates firm dynamics in models of international trade. Burstein and Melitz (2013) generate substantial aggregate-transition dynamics from endogenous shifts in firm-size distribution in response to trade liberalization and find that the responses of trade volumes, innovation, and aggregate output depends on the assumption for firm dynamics, endogenous innovation, and the expected time path of trade liberalization.

ifications that study trade-induced economic outcomes.[4] In the context of human capital adjustment, however, Falvey et al. (2010) build a traditional two-sector Heckscher-Ohlin trade model with skilled and unskilled labor to address when and whether unskilled workers opt for skill upgrading in response to trade liberalization in a skill-abundant country; Van Long et al. (2007) develop a model of firm-specific human capital accumulation, and focus on the decision of workers to accumulate firm-specific skills following trade liberalization. The major differences between our chapter and their papers are that 1) we apply the "new" trade theory (heterogeneous-firm model) and 2) we focus on the decision of firms. In terms of studies on China's economy, some papers associate China's economic growth with its human capital accumulation. China has been sustaining the fastest growth for a long period of time after it started economic reform and engaged in global economy. Li et al. (2017) point out that human capital is also an important source and prospect for the future economic growth in China as higher per capita income is positively associated with higher levels of human capital. This chapter focuses on the impact of a regional trade liberalization policy, APTA, on export participation, as well as labor training decisions by Chinese manufacturing firms. Firms' investment in on-the-job training is an important way for skill enhancement, human capital adjustment and product quality improvement in China.

The empirical work presented herein is related to the fields of trade liberalization and manufacturing firms' performance. A wide range of studies have investigated the impacts of trade integration on export market entry, technology adaption, skill upgrading, productivity, wage inequality and other economic outcomes. For instance, Bustos (2011b) empirically analyzes the impact of free trade on export participation and technology upgrading of Argentinian firms. Bas (2012) extends the previous work by also considering skill upgrading with plant-level data from Chile's manufacturing sector. There, however, can be heterogeneity in the trade effect on skill upgrading by export destinations. For example, Yamashita (2008) finds that fragmentation trade with high income countries has a skill downgrading effect, in contrast to skill upgrading among firms with developing East Asian countries, based on a panel dataset covering 52 Japanese manufacturing industries. There are empirical studies

[4]For instance, Helpman et al. (2004) highlight the important role of within-sector firm productivity differences in explaining the structure of international trade and FDI with heterogeneous firms.

about trade adjustment and human capital development of less developed countries, such as India (Edmonds et al., 2010),[5] and Indonesia (Bazzi et al., 2016).[6] Wang (2007) uses data from manufacturing industries in 25 developing countries to study the role of human capital in trade-related technology spillovers. Regarding trade liberalization in China, Brandt et al. (2017) focus on how the WTO accession influences markups and productivity of Chinese manufacturing firms. Our research is closely related to Bustos (2011b). The departure of this chapter from the literature is that we introduce two different export markets to discuss firms' export decisions in response to the reduction in one destination's tariffs, and examine how a regional trade liberalization (APTA) affects investment in on-the-job training of Chinese manufacturers.[7]

The rest of the chapter is organized as follows. Section 3.2 presents the benchmark model. Section 3.3 derives the extended model in which we distinguish two distinct export markets. Section 3.4 describes trade policies and data sets. Section 3.5 provides an empirical framework to exam the effects of trade liberalization on export participation and skill upgrading and test the predictions of the baseline and extended models; in particular, section 3.5.5 makes a discussion about selected and other sectors. Section 3.6 concludes the whole chapter.

### *3.2 Benchmark Model*

The model is built on Melitz (2003) and Bustos (2011b) to study the impact of trade liberalization on firms' human capital investment decisions. There are two identical countries, and each country has two sectors, the skill-intensive sector $s$ and the unskilled sector $u$. We consider a monopolistically competitive setup with heterogeneous firm productivity, endoge-

[5]They examine the impact of India's 1991 trade reform on schooling and child labor. They find that rural India experiences an increase in schooling and decline in child labor, but the rural districts with employment subject to larger changes in final product protection have a relative rise in poverty and smaller improvements in schooling.

[6]They study the role of location-specific human capital and skill transferability in shaping productivity in Indonesia.

[7]In the context of labor training, Liu and Lu (2016) and Huang and Zhuang (2021) apply a large panel data of manufacturing firms in China to investigate the effects of on-the-job training on firm productivity and wages.

nous skill upgrading decision and endogenous export participation. The least productive firms have to exit the market due to negative profits. Some firms in the middle range of productivity can export to the foreign country even though they are not productive enough to invest in labor training. The most productive firms are able to export and invest in labor training. Precisely, they employ high-skilled labor and produce skill-intensive products.

### *3.2.1 Preferences*

In each country, there are two sectors, indexed by $i \in (s, u)$, the skill-intensive sector $s$ and unskilled sector $u$. The preferences of a representative consumer in the home country is give by the following CES function combing skilled and unskilled goods:

$$\max_{y_{s,t}(\omega), y_{u,t}(\omega)} \left[ \left( \int_{\omega \in \Omega_u} y_{u,t}(\omega)^{\frac{\theta-1}{\theta}} d\omega \right)^{\frac{\theta}{\theta-1}\frac{\rho-1}{\rho}} + \left( \int_{\omega \in \Omega_s} y_{s,t}(\omega)^{\frac{\theta-1}{\theta}} d\omega \right)^{\frac{\theta}{\theta-1}\frac{\rho-1}{\rho}} \right]^{\frac{\rho}{\rho-1}}$$

subject to

$$\int_{\omega \in \Omega_u} p_{u,t}(\omega) y_{u,t}(\omega) d\omega + \int_{\omega \in \Omega_s} p_{s,t}(\omega) y_{s,t}(\omega) d\omega = E$$

where $\Omega_i$ is the mass of varieties available in sector $i$ coming from home and foreign countries, $E$ is the aggregate level of spending, $y_i(\omega)$ and $p_i(\omega)$ are the consumption of good $\omega$ and the price of this good respectively, $\theta$ is the elasticity of substitution within sector varieties and $\rho$ is the elasticity of substitution between sector varieties.

These preferences generate demand functions in sector $u$ and $s$, and they are

$$y_u(\omega) = \left( \frac{p_u(\omega)}{P_u} \right)^{-\theta} \frac{P}{P_u} Y = \rho_u(\omega)^{-\theta} \rho_1^{\theta-\rho} Y$$
$$y_s(\omega) = \left( \frac{p_s(\omega)}{P_s} \right)^{-\theta} \frac{P}{P_s} Y = \rho_s(\omega)^{-\theta} \rho_2^{\theta-\rho} Y$$

where the relative prices are defined as $\frac{p_u(\omega)}{P} = \rho_u(\omega)$, $\frac{p_s(\omega)}{P} = \rho_s(\omega)$, $\frac{P_u}{P} = \rho_1$, $\frac{P_s}{P} = \rho_2$ and aggregate consumption good defined as $Y \equiv U$ (utility) and $PY = E$.

The aggregate price index is denoted as

$$P = \left[ P_u^{1-\rho} + P_s^{1-\rho} \right]^{\frac{1}{1-\rho}}$$

where $P_u = \left( \int_{\omega \in \Omega} p_u(\omega)^{1-\theta} d\omega \right)^{\frac{1}{1-\theta}}$ and $P_s = \left( \int_{\omega \in \Omega} p_s(\omega)^{1-\theta} d\omega \right)^{\frac{1}{1-\theta}}$ are the prices of unskilled and skilled goods respectively.

### *3.2.2 Firm Entry*

Firms under monopolistic competition are heterogeneous in their productivity $z$, and pay a sunk entry cost $f_e$ in units of aggregate consumption good. Following Ghironi and Melitz (2005), the firm entrant draws its productivity $z$ with a Pareto distribution $G(z) = 1 - z^{-\kappa}$ after entering the market. Then, firms can make decisions for exporting and human capital investment. Human capital investment in this chapter refers to how much training spending firms are able to provide workers with for skill upgrading.

### *3.2.3 Production*

There is a continuum of firms with heterogeneous productivity $z$. Let $z \in \Omega$ be a particular variety. Firms endogenously choose to produce unskilled or skilled goods. Firm technology is represented by a total cost function, and the total cost under the unskilled sector is

$$TC_u(z) = f_u + \frac{w_l}{z} y_u(z)$$

where $f_u$ is fixed production costs of the unskilled sector measured in units of aggregate consumption goods, and $w_l$ is the real wage of low-skilled workers. More productive firms can hire high-skilled workers to produce skill-intensive goods with paying higher fixed costs $f_s > f_u$, and deliver lower marginal production costs with $\gamma > 1$ and $\beta \in (0, 1)$. $w_h$ is the real wage of high-skilled workers. The total cost of skill-intensive goods is

$$TC_s(z) = f_s + \frac{w_h^{\beta} w_l^{1-\beta}}{\gamma z} y_s(z)$$

The profit maximization of these two sectors yields the following pricing rules of domestic sales:

$$\rho_u^d(z) = \frac{\theta}{\theta - 1} \frac{w_l}{z}$$
$$\rho_s^d(z) = \frac{\theta}{\theta - 1} \frac{w_h^{\beta} w_l^{1-\beta}}{\gamma z}$$

The two pricing rules of exporting are $\rho_s^x(z) = \tau \rho_s^d(z)$, $\rho_u^x(z) = \tau \rho_u^d(z)$. Hence, $\rho_s^d(z) = \rho_u^d(z)/\lambda$ where $\lambda \equiv \gamma \left(\frac{w_l}{w_h}\right)^{\beta}$.

Profits if producing unskilled goods and only serving the domestic market:

$$\begin{aligned}\pi_u^d(z) &= \frac{1}{\theta}\left(\frac{\theta}{\theta-1}\frac{w_l}{z}\right)^{1-\theta}\rho_1^{\theta-\rho}Y - f_u \\ &= \frac{r_u^d(z)}{\theta} - f_u\end{aligned}$$

where firm revenue $r_u^d(z) = \left(\frac{\theta}{\theta-1}\frac{w_l}{z}\right)^{1-\theta}\rho_1^{\theta-\rho}Y$.

Profits if producing unskilled goods and exporting:

$$\begin{aligned}\pi_u^x(z) &= (1+\tau^{1-\theta})\frac{1}{\theta}\left(\frac{\theta}{\theta-1}\frac{w_l}{z}\right)^{1-\theta}\rho_1^{\theta-\rho}Y - f_u - f_x \\ &= (1+\tau^{1-\theta})\frac{r_u^d(z)}{\theta} - f_u - f_x\end{aligned}$$

Exporting is costly, incurring iceberg trade costs $\tau$ and fixed exporting costs, $f_x$, measured in units of aggregate consumption goods.

Profits if producing skill-intensive goods and only serving the domestic market:

$$\begin{aligned}\pi_s^d(z) &= \frac{1}{\theta}\left(\frac{\theta}{\theta-1}\frac{w_h^\beta w_l^{1-\beta}}{\gamma z}\right)^{1-\theta}\rho_2^{\theta-\rho}Y - f_s \\ &= \lambda^{\theta-1}\frac{r_u^d(z)}{\theta}\left(\frac{\rho_2}{\rho_1}\right)^{\theta-\rho} - \phi f_u\end{aligned}$$

Profits if producing skill-intensive goods and exporting:

$$\begin{aligned}\pi_s^x(z) &= (1+\tau^{1-\theta})\frac{1}{\theta}\left(\frac{\theta}{\theta-1}\frac{w_h^\beta w_l^{1-\beta}}{\gamma z}\right)^{1-\theta}\rho_2^{\theta-\rho}Y - f_s - f_x \\ &= \lambda^{\theta-1}(1+\tau^{1-\theta})\frac{r_u^d(z)}{\theta}\left(\frac{\rho_2}{\rho_1}\right)^{\theta-\rho} - \phi f_u - f_x\end{aligned}$$

where $\phi = \frac{f_s}{f_u} > 1$. High productivity firms find it profitable to export skilled goods with incurring higher fixed production costs, $\phi f_u$.

After learning the idiosyncratic productivity $z$, firms endogenously choose to produce unskilled or skilled goods. The least productive firms must exit the market if the domestic sales profit is negative, so the exit cutoff $z_e$ is defined as:

$$z_e = \{z|\pi_u^d(z) = 0\}$$

$z_x$ denotes the productivity level above which firms producing unskilled goods and finding export profitable, so

$$z_x = \{z|\pi_u^d(z) = \pi_u^x(z)\}$$

Thus, $z_x$ can be expressed as a function of $z_e$ with the zero profit condition for marginal exporters:

$$z_x = \tau z_e \left(\frac{f_x}{f_u}\right)^{\frac{1}{\theta-1}} \tag{3.1}$$

This condition shows that $z_x > z_e$ as long as $\tau \left(\frac{f_x}{f_u}\right)^{\frac{1}{\theta-1}} > 1$.

More productive firms are able to provide training to upgrade workers' skill levels, so they can enter the skill-intensive sector. The productivity cutoff $z_s$ is the cutoff level where firms obtain equal profits from producing unskilled and skilled goods:

$$z_s = \{z|\pi_s^x(z) = \pi_u^x(z)\}$$

The zero profit condition for the marginal firm to produce skill-intensive goods gives the following expression of $z_s$ as a function of $z_e$:

$$z_s = z_e \left[\frac{\phi - 1}{(1+\tau^{1-\theta})(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - 1)}\right]^{\frac{1}{\theta-1}} \tag{3.2}$$

The restriction required for $z_s > z_e$ is $\phi - 1 > (1+\tau^{1-\theta})(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - 1)$. Given $\tau \left(\frac{f_x}{f_u}\right)^{\frac{1}{\theta-1}} > 1$, the ratio of $z_s$ and $z_x$ is larger than one:

$$\frac{z_s}{z_x} = \left[\frac{\tau^{1-\theta}(\phi-1)f_u}{(1+\tau^{1-\theta})(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - 1)f_x}\right]^{\frac{1}{\theta-1}} > 1$$

In equilibrium, there are four groups of firms. The least productive firms $z < z_e$ exit the market, the low productivity firms $z_e < z < z_x$ are not able to investment in labor training and they only serve the domestic market, the moderate productivity firms $z_x < z < z_s$ also cannot invest in human capital but can export to the foreign market, and the most productive firms $z > z_s$ are able to both export and upgrade skill levels of their workers. The productivity cutoffs in the model $z_s > z_x$ are consistent with the data since some firms find it profitable to export, but not profitable to provide labor training and produce skilled goods.

### *3.2.4 Equilibrium*

*Labor Market*

The aggregate demand for low-skilled workers under both the unskilled and skill-intensive sectors is:

$$\begin{aligned} L &= L_u + L_s \\ &= \int_{z_e}^{z_x} l_u^d(z)dz + \int_{z_x}^{z_s} l_u^x(z)dz + \int_{z_s}^{\infty} l_s^x(z)dz \end{aligned}$$

The aggregate demand for high-skilled workers under the skill-intensive sector is

$$H = \int_{z_s}^{\infty} h_s^x(z)dz$$

*Free Entry*

The present value of the average profit flows $\tilde{v} = \sum_{t=0}^{\infty}(1-\delta)^t\tilde{\pi} = \frac{\tilde{\pi}}{\delta}$ and the net value of entry is $v_e = \frac{1}{1-G(z_e)}\tilde{v} - f_e$, so the free entry condition is

$$f_e = (1 - G(z_e))\frac{\tilde{\pi}}{\delta} \tag{3.3}$$

The average profit is $\tilde{\pi} = \tilde{\pi}_u^d + n_x\tilde{\pi}_u^x + n_s\tilde{\pi}_s^x$, where $\tilde{\pi}_u^d$ is the average profit for firms that produce unskilled goods and serve the domestic market only, $n_x \equiv \frac{1-G(z_x)}{1-G(z_e)} = \left(\frac{z_x}{z_e}\right)^{-\kappa}$ is the fraction of firms that export but employ low-skilled labor and produce unskilled goods, $\tilde{\pi}_u^x$ is the average profits for exporters producing unskilled goods, and $n_s \equiv \frac{1-G(z_s)}{1-G(z_e)} = \left(\frac{z_s}{z_e}\right)^{-\kappa}$ is the fraction of exporters providing labor training and produce skilled goods, and $\tilde{\pi}_s^x$ is their average profits.

In Appendix C.2, we derive the average revenues of surviving firms is

$$\tilde{r} = \theta f_u \left(\frac{\tilde{z}_e}{z_e}\right)^{\theta-1} + n_x\theta f_x \left(\frac{\tilde{z}_x}{z_x}\right)^{\theta-1} + n_s\theta f_u(\phi-1)\left(\frac{\tilde{z}_s}{z_s}\right)^{\theta-1}.$$

After substituting $\tilde{r}$ into the free entry condition, we obtain

$$z_e = \left(\frac{1}{f_e\delta}\frac{\theta-1}{\kappa-(\theta-1)}\right)^{\frac{1}{\kappa}} \left[f_u + n_x f_x + n_s f_u(\phi-1)\right]^{\frac{1}{\kappa}} \tag{3.4}$$

where $n_x = \left(\frac{z_x}{z_e}\right)^{-\kappa} = \tau\left(\frac{f_x}{f_u}\right)^{\frac{-\kappa}{\theta-1}}$ and $n_s = \left(\frac{z_s}{z_e}\right)^{-\kappa} = \left[\frac{\phi-1}{(1+\tau^{1-\theta})(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right]^{\frac{-\kappa}{\theta-1}}$

Substituting $n_x$ and $n_s$ into equation (3.4), we get

$$z_e = \Lambda\Phi \tag{3.5}$$

where $\Lambda \equiv \left(\frac{f_u}{f_e\delta}\frac{\theta-1}{\kappa-(\theta-1)}\right)^{\frac{1}{\kappa}}$ and

$\Phi \equiv \left[1+\frac{f_x}{f_u}\left(\frac{f_x}{f_u}\right)^{\frac{-\kappa}{\theta-1}}\tau^{-\kappa}+(\phi-1)\left(\frac{\phi-1}{(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right)^{\frac{-\kappa}{\theta-1}}(1+\tau^{1-\theta})^{\frac{\kappa}{\theta-1}}\right]^{\frac{1}{\kappa}}$.

By substituting the solution for the exit cutoff, we can get a solution for the export and skill upgrading cutoffs below.

$$z_x = \tau\Lambda\Phi \tag{3.6}$$

$$z_s = \Lambda\Phi\left[\frac{\phi-1}{(1+\tau^{1-\theta})(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right]^{\frac{1}{\theta-1}} \tag{3.7}$$

### *3.2.5 Trade Liberalization*

In this section, we analyze the impact of trade liberalization on export participation and skill upgrading. We find that the reduction in iceberg trade costs increases export profits, inducing more firm to enter the export market and encourage exporters to provide more labor training and produce skilled goods.

**Proposition 3.1**. A reduction in iceberg trade costs ($\tau$):

a. increases the equilibrium skill premium, $\frac{\partial\frac{w_h}{w_l}}{\tau} < 0$

b. increases the average profit, $\frac{\partial\tilde{\pi}}{\tau} < 0$

c. increases the exit productivity cutoff, $\frac{\partial z_e}{\tau} < 0$

d. reduces the export productivity cutoff, $\frac{\partial z_x}{\tau} > 0$

e. reduces the skill upgrading cutoff, $\frac{\partial z_s}{\tau} > 0$

Proof: see Appendix C.2.1.

There is an asymmetric effect of trade liberalization since firms are heterogeneous. Market shares are reallocated from the firms producing unskilled goods to the firms providing

skilled goods with a reduction in trade costs, which increases the relative demand for skilled labor. This leads to an increase in the skill premium. We also can conclude that trade integration increases firms' revenues, encourages more firms in the middle range of productivity levels to enter the export market, and makes labor training more profitable for productive exporters.

## *3.3 Extended Model*

### *3.3.1 Production*

Different from the benchmark model, we assume that the foreign country can be either a main trading partner of the home country, denoted as country $m$, or a less preferential trading partner of the home country, denoted as country $o$. The home country and its main trading partner are assumed to impose the most-favored-nation (MFN) tariff, which is the lowest possible tariff a country can assess from another country. The less preferential trading partner imposes larger tariffs. In this section, two export productivity cutoffs are considered to distinguish if the firm can export to only the main country $m$ or both countries $o$ and $m$. In section 3.5.5, we regard India as a representative country $o$ because empirically India impose higher tariffs on Chinese products compared to the rest of the world during our study period 2004–2007.

There are two types of iceberg trade costs, $\tau_m < \tau_o$, and fixed export costs, $f_{mx} < f_{ox}$, since trade barriers are lower if the home firms export to country $m$. Four pricing rules of export are $\rho_s^{mx}(z) = \tau_m \rho_s^d(z)$, $\rho_s^{ox}(z) = \tau_o \rho_s^d(z)$, $\rho_u^{mx}(z) = \tau_m \rho_u^d(z)$ and $\rho_u^{ox}(z) = \tau_o \rho_u^d(z)$.

Profits if producing unskilled goods and only serving the domestic market:

$$\begin{aligned}\pi_u^d(z) &= \frac{1}{\theta}\left(\frac{\theta}{\theta-1}\frac{w_l}{z}\right)^{1-\theta}\rho_1^{\theta-\rho}Y - f_u \\ &= \frac{r_u^d(z)}{\theta} - f_u\end{aligned}$$

Profits if producing unskilled goods and exporting to country $m$:

$$\begin{aligned}\pi_u^{mx}(z) &= (1+\tau_m^{1-\theta})\frac{1}{\theta}\left(\frac{\theta}{\theta-1}\frac{w_l}{z}\right)^{1-\theta}\rho_1^{\theta-\rho}Y - f_u - f_{mx} \\ &= (1+\tau_m^{1-\theta})\frac{r_u^d(z)}{\theta} - f_u - f_{mx}\end{aligned}$$

Profits if producing unskilled goods and exporting to country $o$:

$$\pi_u^{ox}(z) = (1+\tau_o^{1-\theta})\frac{1}{\theta}\left(\frac{\theta}{\theta-1}\frac{w_l}{\gamma_u z}\right)^{1-\theta}\rho_1^{\theta-\rho}Y - f_u - f_{ox}$$
$$= (1+\tau_o^{1-\theta})\frac{r_u^d(z)}{\theta}\gamma_u^{\theta-1} - f_u - f_{ox}$$

Profits if producing skill-intensive goods and exporting to country $m$:

$$\pi_s^{mx}(z) = (1+\tau_m^{1-\theta})\frac{1}{\theta}\left(\frac{\theta}{\theta-1}\frac{w_h^{\beta}w_l^{1-\beta}}{\gamma_m z}\right)^{1-\theta}\rho_2^{\theta-\rho}Y - f_{ms} - f_{mx}$$
$$= \lambda_m^{\theta-1}(1+\tau_m^{1-\theta})\frac{r_u^d(z)}{\theta}\left(\frac{\rho_2}{\rho_1}\right)^{\theta-\rho} - \phi_m f_u - f_{mx}$$

Profits if producing skill-intensive goods and exporting to country $o$:

$$\pi_s^{ox}(z) = (1+\tau_o^{1-\theta})\frac{1}{\theta}\left(\frac{\theta}{\theta-1}\frac{w_h^{\alpha}w_l^{1-\alpha}}{\gamma_o z}\right)^{1-\theta}\rho_2^{\theta-\rho}Y - f_{os} - f_{ox}$$
$$= \lambda_o^{\theta-1}(1+\tau_o^{1-\theta})\frac{r_u^d(z)}{\theta}\left(\frac{\rho_2}{\rho_1}\right)^{\theta-\rho} - \phi_o f_u - f_{ox}$$

where $f_{ox} > f_{mx}$, $f_{os} > f_{ms} > f_u$, $\alpha > \beta$, $\gamma_o > \gamma_m > \gamma_u > 1$ and $\phi_o > \phi_m > 1$.

Firms with productivity above $z_{mx}$ export to country $m$ (the main trading partner) while they can export to country $o$ if their productivity is above $z_{ox}$. Thus, these two export productivity cutoffs are

$$z_{mx} = \{z|\pi_u^d(z) = \pi_u^{mx}(z)\} \qquad z_{ox} = \{z|\pi_u^{mx}(z) = \pi_u^{ox}(z)\}$$

The productivity cutoff of producing skill-intensive goods (skill upgrading) is also different from the one in the benchmark model. The highly productive firms find it profitable to provide labor training when trading with country $m$, and the most productive firms are able to export to country $o$ and invest in human capital; thus the two productivity cutoffs of skill upgrading are

$$z_{ms} = \{z|\pi_u^{ox}(z) = \pi_s^{mx}(z)\} \qquad z_{os} = \{z|\pi_s^{mx}(z) = \pi_s^{ox}(z)\}$$

In equilibrium, firms sort into six different groups: the least productive firms ($z < z_e^{'}$) exit the market, the low productivity firms ($z_e^{'} < z < z_{mx}$) employ low-skilled labor and serve

only the home country, the lower-middle productivity firms ($z_{mx} < z < z_{ox}$) employ low-skilled labor and export to country $m$; the upper-middle productivity firms ($z_{ox} < z < z_{ms}$) export unskilled goods to country $o$; the high productivity firms ($z_{ms} < z < z_{os}$) export to country $m$ but is able to employ skilled labor, the most productive firms ($z > z_{os}$) can export to country $o$ and provide labor training.

$$z_{mx} = z_e^{'} \tau_m \left( \frac{f_{mx}}{f_u} \right)^{\frac{1}{\theta-1}} \tag{3.8}$$

$$z_{ox} = z_e^{'} \left( \frac{f_{ox} - f_{mx}}{((1+\tau_o^{1-\theta})\gamma_u^{\theta-1} - \tau_m^{1-\theta} - 1) f_u} \right)^{\frac{1}{\theta-1}} \tag{3.9}$$

$$z_{ms} = z_e^{'} \left[ \frac{(\phi_m - 1) f_u + f_{mx} - f_{ox}}{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u) f_u} \right]^{\frac{1}{\theta-1}} \tag{3.10}$$

$$z_{os} = z_e^{'} \left[ \frac{f_{ox} - f_{mx} + (\phi_o - \phi_m) f_u}{(\lambda_o^{\theta-1}(1+\tau_o^{1-\theta}) - \lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho} f_u} \right]^{\frac{1}{\theta-1}} \tag{3.11}$$

### *3.3.2 Equilibrium*

*Labor Market*

The aggregate demand for low-skilled workers in both the unskilled and skill intensive sectors is

$$\begin{aligned} L^{'} &= L_u^{'} + L_s^{'} \\ &= \int_{z_e^{'}}^{z_{mx}} l_u^d(z)dz + \int_{z_{mx}}^{z_{ox}} l_u^{mx}(z)dz + \int_{z_{ox}}^{z_{ms}} l_u^{ox}(z)dz + \int_{z_{ms}}^{z_{os}} l_s^{mx}(z)dz + \int_{z_{os}}^{\infty} l_s^{ox}(z)dz \end{aligned}$$

while the aggregate demand for high-skilled workers in the skill-intensive sector is

$$H^{'} = \int_{z_{ox}}^{z_{ms}} h_s^{mx}(z)dz + \int_{z_{os}}^{\infty} h_s^{ox}(z)dz$$

*Free Entry*

The numbers of firms exporting unskilled or skilled goods to country $m$ and country $o$ can be derived as: $n_{mx} = \left(\frac{z_{mx}}{z_e^{'}}\right)^{-\kappa}$, $n_{ox} = \left(\frac{z_{ox}}{z_e^{'}}\right)^{-\kappa}$, $n_{ms} = \left(\frac{z_{ms}}{z_e^{'}}\right)^{-\kappa}$ and $n_{os} = \left(\frac{z_{os}}{z_e^{'}}\right)^{-\kappa}$. The average profit is $\tilde{\pi} = \tilde{\pi}_u^d + n_{mx}\tilde{\pi}_u^{mx} + n_{ox}\tilde{\pi}_u^{ox} + n_{ms}\tilde{\pi}_s^{mx} + n_{os}\tilde{\pi}_s^{ox}$, and it can be described in

this way:

$$\tilde{\pi}^{'} = \frac{\tilde{r}^{'}}{\theta} - f_u - n_{mx}f_{mx} - n_{ox}(f_{ox} - f_{mx}) - n_{ms}((\phi_m - 1)f_u + f_{mx} - f_{ox}) \\ - n_{os}((\phi_o - \phi_m)f_u + f_{ox} - f_{mx})$$

Similar to the benchmark model, we can derive the average revenues $\tilde{r}^{'}$ expressed as the productivity cutoffs:

$$\tilde{r} = \theta f_u \left(\frac{\tilde{z}_e^{'}}{z_e^{'}}\right)^{\theta-1} + n_{mx}\theta f_{mx}\left(\frac{\tilde{z}_{mx}}{z_{mx}}\right)^{\theta-1} + n_{ox}\theta(f_{ox} - f_{mx})\left(\frac{\tilde{z}_{ox}}{z_{ox}}\right)^{\theta-1} \\ + n_{ms}\theta(f_u(\phi_m - 1) + f_{mx} - f_{ox})\left(\frac{\tilde{z}_{ms}}{z_{ms}}\right)^{\theta-1} + n_{os}\theta(f_{ox} - f_{mx} + (\phi_o - \phi_m)f_u)\left(\frac{\tilde{z}_{os}}{z_{os}}\right)^{\theta-1}$$

After substituting $\tilde{r}^{'}$ into the free entry condition, we obtain

$$z_e^{'} = \left(\frac{\theta - 1}{\kappa - (\theta - 1)}\frac{1}{\delta f_e}\Psi\right)^{1/\kappa} \tag{3.12}$$

where $\Psi^{\kappa} = f_u + n_{mx}f_{mx} + n_{ox}(f_{ox} - f_{mx}) + n_{ms}(f_u(\phi_m - 1) + f_{mx} - f_{ox}) + n_{os}(f_{ox} - f_{mx} + (\phi_o - \phi_m)f_u)$.

### *3.3.3 Trade Liberalization*

We have two additional propositions below.

**Proposition 3.2**. A reduction in iceberg trade costs of country $m$ ($\tau_m$):

a. increases the average profit, $\frac{\partial \tilde{\pi}^{'}}{\tau_m} < 0$

b. increases the exit productivity cutoff, $\frac{\partial z_e^{'}}{\tau_m} < 0$

c. reduces the productivity cutoff of exporting to country $m$, $\frac{\partial z_{mx}}{\tau_m} > 0$

d. reduces the skill-upgrading cutoff of exporters to country $m$, $\frac{\partial z_{ms}}{\tau_m} > 0$

Proof: Please see Appendix C.2.2. Part d is established only when certain conditions are met.

**Proposition 3.3**. A reduction in iceberg trade costs of country $o$ ($\tau_o$):

a. increases the average profit, $\frac{\partial \tilde{\pi}^{'}}{\tau_o} < 0$

b. increases the exit productivity cutoff, $\frac{\partial z_e^{'}}{\tau_o} < 0$

c. reduces the productivity cutoff of exporting to country $m$, $\frac{\partial z_{mx}}{\tau_o} > 0$

d. reduces the productivity cutoff of exporting to country $o$, $\frac{\partial z_{ox}}{\tau_o} > 0$

e. increases the skill-upgrading productivity cutoff of exporters to country $m$, $\frac{\partial z_{ms}}{\tau_o} < 0$

f. reduces the skill-upgrading productivity cutoff of exporters to country $o$, $\frac{\partial z_{os}}{\tau_o} > 0$

Proof: Please see Appendix C.2.2. Parts c and d are established only when certain conditions are met.

## *3.4 Trade Policies and Data*

### *3.4.1 Post-WTO Accession Trade Liberalization in China*

Trade liberalization policies undertaken in China after WTO accession are described in this session. First, China joined WTO in 2001 and continued trade liberalization from 2001-2005. For instance, China bounded all tariff lines and the average applied most-favored nation (MFN) rate dropped from 15.6% in 2001 to 9.7% in 2005, with the manufactured goods rate declining from 14.3% to 8.9%, and agricultural products rate decreasing from 23.2% to 14.6% during the same period (Bin, 2015). This indicates that China made a great achievement of import liberalization. Meanwhile, China's industrial goods conquered the global markets after it joined the WTO in 2001. China doubled its share of trade in manufactured goods from 7.9% in 2000 to 17.7% in 2012 (Hilpert, 2014). According to UNCTADStat, its share of global export goods was 3.9% in 2000, and went up to 14.7% in 2020. FDI in China was less restricted after WTO accession, so China became the most important global investment destination. According to the World Bank, the net inflows FDI in China started with 42.1 billion dollars in 2000 and achieved a peak level at 290.9 billion

dollars in 2013. Furthermore, China overtook US as the world's leading destination for FDI in 2020.

Next, we describe an important regional trade liberalization—Asian Pacific Trade Agreement (APTA). APTA, previously the Bangkok Agreement, was signed in 1975 and renamed in 2005. Bangladesh, China, India, Lao PDR, Mongolia, Republic of Korea, and Sri Lanka are the current parties. APTA promotes intra-regional trade and contributes to economic development of the seven developing countries through trade and investment liberalization. China acceded APTA in 2001 and endorsed a preferential trade arrangement among developing Asian countries. In 2005, the first Ministerial Council was held in Beijing, China and the third round of negotiation results was implemented in 2006. Trade and tariff data from the World Integrated Trade Solution (WITS) between 2004 and 2007 (see Figure 3.1) show that India became the most preferential trading partner of China. This chapter emphasizes the impact of APTA on exports and human capital investment of Chinese manufactured firms theoretically and empirically. India is selected as a representative country, which imposed much higher tariffs than the rest of the world in 2004, but also reduced trade barriers with China more significantly from 2004 to 2007. More trade agreements among APTA members were made after 2007,[8] which will not be our focus.

### *3.4.2 Firm-Level Data*

We resort to two data sources to construct a balanced panel of manufacturing firms in China. First of all, we obtain data from the Chinese Industrial Enterprises Database (CIED). The CIED is constructed by China's National Bureau of Statistics (CNBS) mainly based on the annual or quarterly reports submitted to local bureau of statistics. The database contains all industrial enterprises that are non-state-owned and "large-scale"[9] or state-owned. In the

[8] A new Asia-Pacific trade deal was created in November 2020. It is a new overarching Regional Comprehensive Economic Partnership (RCEP) Free Trade Agreement (FTA) between 15 Asia-Pacific countries. Its signatories are the 10 members of the Association of Southeast Asian Nations (ASEAN) countries and Japan, Korea, China, Australia and New Zealand. The signing of the RCEP aims to facilitate regional or even world trade and investment further. RCEP connects about 30% of the world's population and output, makes the Asian economies more efficient, and improves technology and solidifies global value chains. China continues to benefit from trade openness.

[9] That is, the main business income of an enterprise was larger than 5 million RMB, and this standard was revised to 20 million RMB in 2011.

database, approximately 90% of the enterprises are manufacturing firms, which will be what we focus on in this research. Although the database spans from 1998 to 2013, information about on-the-job training spending (TS), which is our main measure of skill upgrading in the empirical analysis, is only available in 2004–2007, which is exactly the regional trade liberalization period we emphasize. This database has been exploited by economists such as Hsieh and Klenow (2009), Song et al. (2011), Brandt et al. (2017), as well as Huang and Zhuang (2021), who provide more details about this database. From CIED, we also learn about each firm's total sales (which include both domestic and export sales) and number of employees, other than TS.

Second, we collect dis-aggregated information from the China Customs Database (CCD). It provides the data of exports by firm, 8-digit HS product, and destination country. We merge the information from CCD to the CIED according to the firms' names, postal codes, or telephone numbers, following Ruiqin et al. (2019).[10] We then aggregate export data at the 4-digit CIC industry level for each firm-destination country pair (some firms operate in multiple industries domestically and/or globally).

Next, we select the firms in the sectors that are covered by India's consolidated list of concessions of the first 3 rounds of negotiations to APTA member countries and in the 4-digit CIC industries with information on India's tariffs. We ended up with a balanced panel of 110,632 manufacturing firms (operating in 131,460 4-digit CIC industries) in each year from 2004 to 2007. The sample is representative of firms owning establishments with more than 10 employees that can potentially be affected by APTA.

By merging the CIED and CCD data, we can calculate the total sales per employee for each firm, which will be one of our firm-level controls. Moreover, domestic sales can be calculated by subtracting the export sales from total sales. One special feature of the data is that we know each firm's export sales to each destination country, including India—we can therefore learn about whether a firm exports to India or not.

Table C.1 in Appendix C.3 contains summary statistics by export status for the main variables of interest for the initial year, 2004.

[10]The numbers of merged manufacturing enterprises are 52,046 in 2004, 51,026 in 2005, 60,345 in 2006, and 61,749 in 2007. These correspond to 38.1%, 37.7%, 27.8%, and 28.9% of the enterprises in CCD.

### *3.4.3 Industry-Level Data*

In the empirical section we use controls for 4-digit CIC industry characteristics that might be correlated with changes in tariffs. We first obtain average capital and skill intensity in the industry in the United States in the 1980s from the National Bureau of Economic Research (NBER) productivity database (see Appendix C.3 for details). We also use the import demand elasticity and export supply elasticity as estimated by Broda and Weinstein (2006) and Broda et al. (2008).

## *3.5 Empirics*

In this section we test the theoretical predictions developed in Sections 3.2 and 3.3. First, we check whether the sorting pattern of firms into exporting and training predicted by the model is consistent with the observed characteristics of exporters to different countries and non-exporters operating in the same four-digit CIC industry. Second, we test the main predictions of our model: a reduction in export tariffs encourages firm entry in the export market and induces skill upgrading. We focus on the regional trade liberalization (APTA) effect, as we select the sectors covered by India's consolidated list of concessions of the first 3 rounds of negotiations to APTA member countries.

### *3.5.1 Within-Industry Patterns in the Data*

In the benchmark model, underlying productivity differences produce a sorting of firms into four groups: the least productive firms exit the market (not in data), the low productivity firms produce unskilled goods and serve the domestic market only, the middle ones still produce unskilled goods but also export, and the most productive firms both export and provide labor training for skilled goods production. In this setting, a reduction in trade costs $\tau$ increases export profits, inducing more firms in the middle range of the productivity distribution to enter the export market and upgrade workers' skill levels. Figure 3.2 indicates the effects of trade liberalization for firms in each range of the productivity distribution through reflecting the changes in productivity cutoffs from 2004 to 2007. Precisely speaking, firms with intermediate productivity find it easier to start exporting and the more productive

firms also have greater incentives to provide labor training.

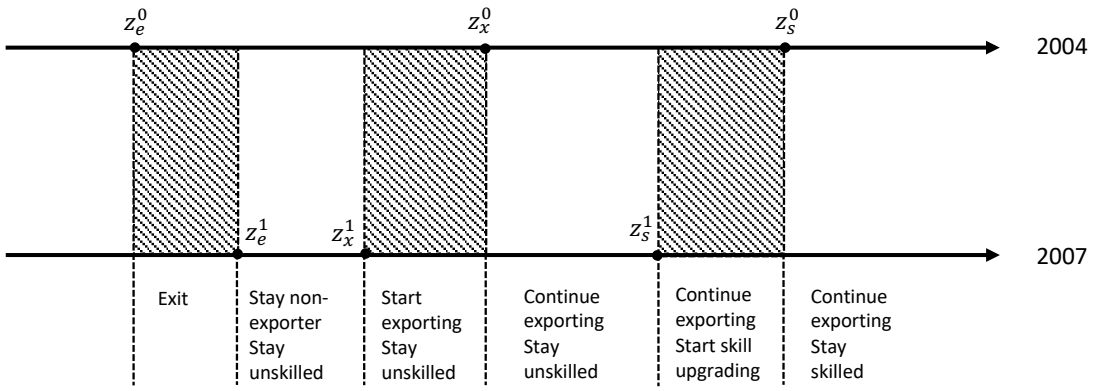



Figure 3.2: Effect of Lowering Variable Trade Costs: Benchmark Model

In the extended model, we distinguish the trade effects of two export destinations, countries $m$ and $o$. Country $m$ refers to a main trading partner of the home country, while country $o$ is a less preferential trading partner since it imposes a higher tariff ($\tau_o > \tau_m$). This chapter refers to India as a representative country $o$. In this case, firms sort into six groups: the least productive firms exit the market, the low productivity firms produce unskilled goods and serve the domestic market only, the lower-middle ones still produce unskilled goods but also export to country $m$, the upper-middle ones export unskilled goods to country $o$, the high productivity firms are able to provide labor training but export skilled goods to country $m$, and the most productive firms can export skilled goods to country $o$. Figure 3.3 also displays the effect of trade liberalization for firms in each part of the productivity distribution. In particular, as shown by the shaded areas, firms switching export markets from country $m$ to $o$ could continue producing unskilled goods, start skill downgrading, or begin skill upgrading.[11]

To check whether the sorting patterns depicted in Figures 3.2 and 3.3, and the parameter restrictions required to obtain the model are consistent with the data, we follow Bustos (2011b) to divide firms into four groups: continuing exporters, new exporters, exiting ex-

[11]The extended model does not reflect trade-induced switching from country $o$ to $m$. There should be an exogenous shock that leads to this switching.

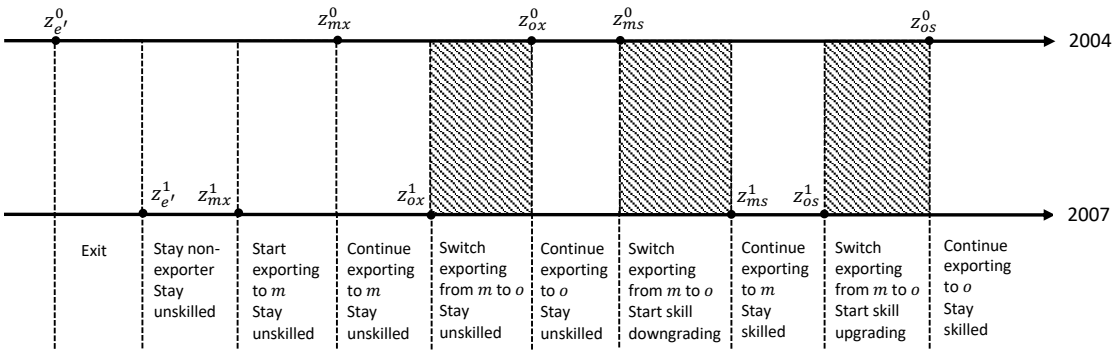



Figure 3.3: Effect of Lowering Variable Trade Costs: Extended Model

porters and firms serving the domestic market only (never exporters), and compute the differences in characteristics, including sales, employment and training spending per worker, for firms operating within the same four-digit CIC industry.

First, Table 3.1 indicates that, based on the basic model, all types of exporters have higher sales, employment, and training spending per worker than never exporters in 2004 on average, which mirrors the fact that (potential) exporters are larger, more productive and more skill-intensive. Second, although sales and employment of new exporters are relatively lower than incumbent exporters in 2004, their per capita training spending is higher than continuing exporters on average, preparing them to enter the skill-intensive sector and employ more high-skilled workers. Third, the increase in training spending per worker for continuing exporters are almost zero from 2004 to 2007 (trade liberalization period), but exiting exporters that later serve the domestic market only reduce labor training.[12] On the contrary, new exporters have the largest average increases in sales, employment, and training spending per worker compared to never exporters. These results indicate that new exporters benefit more from trade liberalization and has greater incentives in human capital investment. It is intuitive that new exporters might demand more high-skilled workers in

[12]Based on the number of industry-level observations and the number of distinct firms, we can see that only continuing exporters operate in multiple industries, while exiting and new exporters both operated in a single industry.

Table 3.1: Differences between Different Types of Exporters and Non-Exporters in APTA Sectors

| | Levels in 2004 | | | Changes 2004–2007 | | | Size | |
|---|---|---|---|---|---|---|---|---|
| Firm characteristic | Sales | Employment | Training per worker | Sales | Employment | Training per worker | Observations | Firms |
| *Basic model:* | | | | | | | | |
| New exporters | 0.596*** | 0.505*** | 0.412*** | 0.176*** | 0.146*** | 0.129*** | 3,957 | 3,957 |
| | (0.020) | (0.017) | (0.041) | (0.014) | (0.010) | (0.048) | | |
| Continuing exporters | 1.240*** | 1.027*** | 0.150*** | -0.095*** | 0.058*** | 0.011 | 40,873 | 20,045 |
| | (0.009) | (0.008) | (0.017) | (0.006) | (0.004) | (0.019) | | |
| Exiting exporters | 1.086*** | 0.852*** | 0.438*** | -0.130*** | -0.035*** | -0.087* | 3,919 | 3,919 |
| | (0.021) | (0.018) | (0.040) | (0.015) | (0.010) | (0.046) | | |
| *Extended model:* | | | | | | | | |
| New exporters to non-Indian countries ($m$) | 0.532*** | 0.462*** | 0.329*** | 0.152*** | 0.132*** | 0.129** | 3,203 | 3,203 |
| | (0.021) | (0.018) | (0.045) | (0.015) | (0.011) | (0.052) | | |
| New exporters to India ($o$) | 0.876*** | 0.691*** | 0.769*** | 0.278*** | 0.205*** | 0.128 | 754 | 754 |
| | (0.049) | (0.042) | (0.091) | (0.027) | (0.021) | (0.111) | | |
| Continuing exporters to $m$ | 1.160*** | 0.997*** | 0.047*** | -0.113*** | 0.046*** | -0.001 | 34,448 | 17,227 |
| | (0.010) | (0.008) | (0.018) | (0.006) | (0.004) | (0.020) | | |
| Continuing exporters to $o$ | 1.751*** | 1.234*** | 0.813*** | -0.024* | 0.102*** | 0.015 | 2,200 | 1,831 |
| | (0.032) | (0.026) | (0.058) | (0.014) | (0.011) | (0.063) | | |
| Switching exporters from $m$ to $o$ | 1.562*** | 1.131*** | 0.594*** | 0.061*** | 0.163*** | 0.080 | 2,971 | 2,379 |
| | (0.029) | (0.024) | (0.049) | (0.012) | (0.010) | (0.054) | | |
| Switching exporters from $o$ to $m$ | 1.545*** | 1.157*** | 0.435*** | -0.165*** | 0.014 | 0.140* | 1,254 | 1,082 |
| | (0.042) | (0.034) | (0.071) | (0.020) | (0.016) | (0.078) | | |
| Exiting exporters to $m$ | 0.997*** | 0.794*** | 0.338*** | -0.140*** | -0.044*** | -0.080 | 3,348 | 3,348 |
| | (0.022) | (0.019) | (0.043) | (0.016) | (0.011) | (0.049) | | |
| Exiting exporters to $o$ | 1.618*** | 1.194*** | 1.032*** | -0.072** | 0.015 | -0.129 | 571 | 571 |
| | (0.053) | (0.045) | (0.104) | (0.031) | (0.023) | (0.124) | | |
| Observations | 131,460 | 131,460 | 131,460 | 131,460 | 131,460 | 131,460 | | |
| Firms | 110,632 | 110,632 | 110,632 | 110,632 | 110,632 | 110,632 | | |

Notes: (1) Robust standard errors are in parentheses. (2) Exporter premia are estimated from a regression of the form $\ln Y_{ij} = \alpha_1 \text{Type } 1_{ij} + \alpha_2 \text{Type } 2_{ij} + ... + I_j + \varepsilon_{ij}$ where $i$ indexes firms, and $j$ indexes four-digit CIC industries; the reference category relative to which differences are estimated is non-exporters; $I$ are industry dummies, and $Y$ is the firm characteristic for which the differences are estimated. (3) ***, **, and * denote significance level at 1%, 5%, and 10%, respectively. (4) Some continuing and switching exporters operated in multiple industries.

order to become more competitive in the foreign market.

In terms of the extended model, we distinguish two export markets, country $m$ and $o$. Exporters to India have higher sales, employment, and training spending per worker in 2004 than those exporting to non-Indian countries (except for those who switch destinations) as shown in Table 3.1, which is consistent with the model setting that firms exporting to India are more productive, conditional on skill level. Moreover, firms switching from non-India to Indian markets in 2004 have slightly higher training spending per worker than those doing the converse, which shows that more of them started to produce high-skilled products earlier. From 2004 to 2007, there is a larger decline in India's tariffs and a greater increase in exports to India compared to those of other countries (Figure 3.1). Therefore, new exporters targeting Indian markets have even larger average increases in sales and employment than those targeting non-Indian markets. However, interestingly, new exporters targeting India have a slightly less significant average increase in training spending per worker than those targeting non-Indian countries. This is probably due to the fact that they have already spent more in training their workers in 2004 before changes in trade costs $\tau_o$, or there could be sector heterogeneity. Moreover, continuing exporters targeting non-Indian markets had even lower average increases in sales and employment than those targeting India, although their increases in average per capita training spending were almost the same. Additionally, there are firms switching export destinations during the same period. Table 3.1 shows that the trade liberalization period coincides with a slightly higher increase in training pending per worker among firms who switch export markets from country $o$ to $m$, which supports our model assumption that some firms could find it more profitable to export skilled goods to country $m$ than to export unskilled goods to country $o$ who imposes higher (but declining) trade barriers. For firms who switch destinations from $m$ to $o$, the increase in per capita training is not so significant. This result is actually also in line with the model in some way, as two shading areas in Figure 3.3 imply that these firms either start skill downgrading or continue producing unskilled goods. Thus, the average effect of labor training could be ambiguous.

The pattern in the Table 3.1 shows that there is coincidence between entry in the export market and skill upgrading, and the performance of exporters targeting the Indian market

differs from other firms, but we cannot establish whether it is better export opportunities that induce human capital investment or vice versa, or whether both are caused by a third factor. The next empirical exercise is to establish causality by linking exporting and skill upgrading directly to the reduction in India's tariffs for imports from China.

### *3.5.2 The Impact of the APTA: Identification Strategy*

After China joined APTA in 2001, a reduction in India's tariffs for imports from China across four-digit CIC industries leads to changes in Chinese firms' entry in the export market and skill upgrading. There are two features of the source of identification that make it exogenous with respect to these two outcomes of interest. First, the tariff reductions were constantly adjusted and negotiated among APTA members during 2004 and 2007. In our data, the average tariff facing manufacturing firms in the sectors covered by APTA decreases from about 28.59 percentage points in 2004 to 13.43 percentage points in 2007. These tariff reductions are not likely to be determined by any individual firm in a specific country. Second, the 2004 simple average effectively applied (AHS) import tariff of India for China was close to those of the rest of the world.[13] Import tariffs of India are unlikely to be targeted to industry characteristics particular to China. Share of India's imports from China was 6.1% in 2004, but rose to 11.2% in 2007, since India's average tariff declined by more than a half during the same period.

The reverse causality problem may not be a concern, but India's initial tariff structure is surely not random. India's trade policy is correlated with some industry characteristics, so omitting them could lead to a source of bias. Hence, we estimate all the equations in first differences to eliminate constant industry characteristics. Still, India's tariffs could capture some omitted industry-level time-varying variable if industries with different initial characteristics are on different trends. In order to further address this issue, we include in the first-differenced equations sector dummies to control for unobserved sector trends, and also include four-digit CIC level controls for industry characteristics such as import demand

[13] According to WITS, the simple average AHS tariffs of India for China and the world are 28.8 and 28.6 percentage points respectively in 2004. In terms of the simple average MFN tariff, the 2004 tariff rates of India are 28.8 percentage points for China and 29.5 percentage points for the world.

elasticity, export supply elasticity, and capital and skill intensity.

We use the India's tariffs to measure the effect of increased export opportunities on export participation and skill upgrading, but these influences might be correlated with changes in China's tariffs. Thus, we control for the changes in China's tariffs with respect to the world between 2004 and 2007, as well as the changes in China's tariffs with respect to India.[14]

Under the benchmark model, the reduction in India's tariffs induces entry in the export market and skill upgrading for firms in the middle range of the productivity distribution rather than affects the least and most productive groups. The extended model also predicts that firms in the middle or upper-middle range have a higher likelihood of exporting, switching export destinations and starting skill downgrading or upgrading following trade integration. To study these heterogeneous effects of firm productivity, we use firm size relative to the four-digit CIC industry mean in 2004 as a proxy for initial productivity and divide firms into quartiles. In the next section, we discuss the empirical results of how the reduction in India's tariffs affects each quartile of the firm size distribution through emphasizing export entry and skill upgrading decisions, and compare them with our theoretical findings.

### *3.5.3 Export Markets Entry Decision*

In this section, we intend to recover the signs of the partial derivatives of interest in the model of the export markets entry choices described by equations (3.6), as well as (3.8) and (3.9). To do so, we estimate a linearized version of the entry models and access the economic significance of the estimated coefficients. We first describe estimation of the average effect of a reduction in India's tariffs on entry in the export market for all firms. Next, we distinguish the export markets, comparing non-Indian countries with India.

Consistent with the benchmark model, we empirically analyze the export entry decision using an index model:

[14]Both final goods and intermediate inputs tariffs are controlled.

$$EX^k_{ijst} = \begin{cases} 1 & \text{if } \beta^k_{\tau^{ex}}\tau^{ex}_{jt} + \alpha^k_{st} + \mu^k_i + \epsilon^k_{ijst} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

where $i$ indexes firms; $j$ indexes four-digit CIC industries; $s$ indexes sectors; $t$ indexes years from 2004 to 2007; $\tau^{ex}_{jt}$ are India's tariffs that vary across four-digit CIC industries and time; $\alpha^k_{st}$ are sector dummies that capture time-varying sector features; $\mu^k_i$ are firm fixed effects capturing unobserved constant heterogeneity including firm heterogeneity $z$ defined in the model and other characteristics affecting productivity cutoffs; $EX^k_{ijst}$ is a dummy that captures firms' export decisions to any partner or India. When $k = 0$, firms can either be exporters or non-exporters; $EX^0_{ijst}$ takes the value of 1 if a firm exports to any country in the world in year $t$ and 0 otherwise. If $k = 1$, the firm is an exporter; $EX^1_{ijst} = 1$ when the firm exports to India (and potentially other counrties at the same time), and $EX^1_{ijst} = 0$ if the firm exports only to non-Indian countries.

*First-Differenced Specification*

We take first differences to eliminate time-invariant plant and sector heterogeneity, and obtain

$$\Delta EX^0_{ijst} = \beta^0_{\tau^{ex}}\Delta\tau^{ex}_{jt} + \Delta\alpha^0_{st} + \Delta\epsilon^0_{ijst} \tag{3.14}$$

In the meantime, we control for changes in China's import tariffs for both outputs and inputs with respect to the world and India ($\Delta\tau^{im}_{jt}$), the firm characteristics in the initial year (2004) such as the number of workers and sales per worker ($z_{ij2004}$), and four-digit industry characteristics like the import demand and export supply elasticities, skill and capital intensity in the United States ($c_j$).[15] Hence, we have the following equation:

$$\Delta EX^0_{ijst} = \beta^0_{\tau^{ex}}\Delta\tau^{ex}_{jt} + \beta^0_{\tau^{im}}\Delta\tau^{im}_{jt} + \beta^0_z Z_{ij2004} + \beta^0_c c_j + \Delta\alpha^0_{st} + \Delta\epsilon^0_{ijst} \tag{3.15}$$

Estimation of equation (3.14) is reported in column 1 of Table 3.2, and the regression coefficients including other controls (equation 3.15) are shown in columns 2 to 8. From panel A of Table 3.2, we find that a reduction in India's tariffs increases the likelihood of entry in

[15] We calculate elasticities sand intensity following Broda and Weinstein (2006) and Broda et al. (2008).

Table 3.2: Entry in the Export Markets Stratified by Sector Group

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| *Panel A: Sample of selected sectors. Dependent variable: year-over-year change in export status* | | | | | | | | |
| ΔIndia's tariffs | -0.124*** | -0.125*** | -0.120*** | -0.104*** | -0.103** | -0.112*** | -0.104*** | -0.103*** |
| | (0.032) | (0.031) | (0.034) | (0.038) | (0.040) | (0.031) | (0.032) | (0.033) |
| ΔChina's tariffs w.r.t. world | | | | | | | | |
| Outputs | | | yes | yes | yes | | | |
| Inputs | | | | yes | yes | | | |
| ΔChina's tariffs w.r.t. India | | | | | | | | |
| Outputs | | | | | | yes | yes | yes |
| Inputs | | | | | | | yes | yes |
| Sector FE | yes | yes | yes | yes | yes | yes | yes | yes |
| Year FE | yes | yes | yes | yes | yes | yes | yes | yes |
| Firm-level controls | | yes | yes | yes | yes | yes | yes | yes |
| Industry controls | | | | | yes | | | yes |
| Observations | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 |
| $R^2$ | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 |
| *Panel B: Sample of other sectors. Dependent variable: year-over-year change in export status* | | | | | | | | |
| ΔIndia's tariffs | -0.014 | -0.014 | -0.011 | -0.016 | -0.016 | -0.012 | -0.015 | -0.015 |
| | (0.015) | (0.015) | (0.017) | (0.015) | (0.015) | (0.015) | (0.015) | (0.015) |
| Observations | 302,511 | 302,511 | 302,511 | 302,511 | 302,511 | 302,511 | 302,511 | 302,511 |
| $R^2$ | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 |
| *Panel C: Sample of selected sectors. Dependent variable: export status in the current year* | | | | | | | | |
| ΔIndia's tariffs | -0.147*** | -0.147*** | -0.143*** | -0.127*** | -0.116*** | -0.135*** | -0.128*** | -0.116*** |
| | (0.036) | (0.036) | (0.038) | (0.042) | (0.043) | (0.036) | (0.036) | (0.036) |
| Export status in the | 0.933*** | 0.932*** | 0.932*** | 0.932*** | 0.931*** | 0.932*** | 0.932*** | 0.931*** |
| previous year | (0.008) | (0.008) | (0.008) | (0.008) | (0.008) | (0.008) | (0.008) | (0.008) |
| $R^2$ | 0.884 | 0.884 | 0.884 | 0.884 | 0.884 | 0.884 | 0.884 | 0.884 |
| *Panel D: Sample of baseline non-exporters in selected sectors. Dependent variable: export status in the current year* | | | | | | | | |
| ΔIndia's tariffs | -0.128*** | -0.126*** | -0.122*** | -0.121*** | -0.090** | -0.127*** | -0.120*** | -0.092*** |
| | (0.038) | (0.039) | (0.039) | (0.043) | (0.035) | (0.039) | (0.038) | (0.031) |
| Export status in the | 0.729*** | 0.726*** | 0.726*** | 0.726*** | 0.725*** | 0.726*** | 0.726*** | 0.725*** |
| previous year | (0.021) | (0.021) | (0.021) | (0.021) | (0.020) | (0.021) | (0.021) | (0.020) |
| Observations | 60,273 | 60,273 | 60,273 | 60,273 | 60,273 | 60,273 | 60,273 | 60,273 |
| $R^2$ | 0.317 | 0.318 | 0.318 | 0.318 | 0.319 | 0.318 | 0.318 | 0.319 |

Notes: (1) Standard errors are clustered at the 4-digit CIC industry level. (2) Δ denotes the year-over-year change in a variable during the period 2004–2007. (3) Firm-level controls include number of employees and sales per worker, all measured in the initial year (2004). (4) Industry controls include import demand elasticity, export supply elasticity, skill intensity, and capital intensity in the United States. (5) In panels B and D, remaining controls are the same as in the corresponding column in panel A. (6) In panel C, controls and number of observations are the same as in the corresponding column in panel A. (7) ***, **, * denote significance level at 1%, 5%, and 10%, respectively. (8) Selected sectors include 14 manufacturing sectors in footnote 1.

the export market for the sample of selected sectors. For instance, columns 5 and 8 indicate that the probability of firm entry in the export market increases by 1.55 percentage points when the average reduction in India's tariffs is around 15 percentage points from 2004 to 2007. This empirical result is consistent with the model prediction that a reduction in trade costs increases firm profit and encourages export participation. Interestingly, only certain sectors show a significant relationship between trade openness and export participation by Chinese firms, and the coefficients are not statistically significant for other sectors as shown in panel B of Table 3.2. Due to sector heterogeneity, there are some limitations of the model, which may not be able to distinguish firm performance (in particular, export decisions) in different sectors after trade liberalization. The different results in panels A and B imply that diverse trade policies could be necessary to further encourage exports because of the heterogeneous features across sectors.

*Lagged-Dependent Variable*

To check for robustness, we implement two more exercises. First, current export decisions might be influenced by lagged export status because of sunk export costs. Therefore, we control the export status in the previous year and estimate the equation in levels with the following regression:

$$EX^0_{ijst} = \beta^0_{\tau^{ex}} \Delta\tau^{ex}_{jt} + \gamma^0 EX^0_{ijs,t-1} + \alpha^0_{st} + \epsilon^0_{ijst} \tag{3.16}$$

The second check is to create a sample of baseline non-exporters under selected sectors and estimate the equation (3.16) that is restricted to non-exporters in 2004. This estimation highlights the effects of changing tariffs on initial non-exporters as we notice that trade liberalization between 2004 and 2007 has a greater positive impact on sales, employment and labor training of new exporters in Table 3.1. The estimates in panels C and D of Table 3.2 are very close to the coefficients of changes in India's tariffs in panel A. This implies that our estimated results of the export entry decision are fairly robust.

*Export Decision by Quartile of the Firm Size Distribution*

The benchmark model predicts that lower trade costs induce entry in the export market for firms with intermediate productivity levels since a reduction in trade costs decreases the export productivity cutoff $z_x$. More precisely, as depicted by Figure 3.2, the export productivity cutoff in 2007 ($z_x^1$) is much lower than the initial cutoff ($z_x^0$). Firms with productivity in the range $z_x^1 < z < z_x^0$ become exporters following trade liberalization. The less productive firms still stay out of the market or serve domestic market only and the most productive firms continue exporting. Empirically, we estimate the impact of the change in India's tariffs on each quartile of the initial firm size distribution with the following equation:

$$\Delta EX_{ijst}^0 = \sum_{n=1}^{4} \beta_{\tau^{ex},n}^0 (\Delta\tau_{jt}^{ex} \times Q_{ij,n}) + \sum_{n=1}^{4} \delta_n^0 Q_{ij,n} + \beta_{\tau^{im}}^0 \Delta\tau_{jt}^{im} + \Delta\alpha_{st}^0 + \Delta\epsilon_{ijst}^0 \qquad (3.17)$$

where $n$ is each of the four quartiles of the firm size distribution and $Q_{ij,n}$ are dummy variables being 1 when firm $i$ belongs to quartile $n$. Columns 1 to 9 of Table 3.3 show that the effect of the reduction in India's tariffs on firm entry in the export market is significant in the last three quartiles of the firm size distribution, while firms in the fourth quartile ($\beta_{\tau^{ex},4} = -0.17$) actually receive larger influences from changing in tariffs than those in the second ($\beta_{\tau^{ex},2} = -0.11$) and third quartiles ($\beta_{\tau^{ex},3} = -0.13$) within the selected sectors. Columns 4 to 6 present estimation of the above equation in levels controlling for lagged export status. The point estimates of $\beta_{\tau^{ex},n}$ are a bit larger, but still have the same pattern. Additionally, the estimated results of the sample of non-exporters in 2004 are smaller than those of the full sample, and the impacts of trade costs on the third and fourth quartiles are more significant compared to the first and second quartiles. In particular, the point estimates of $\beta_{\tau^{ex},3}$ in columns 4 and 7 imply that the the average decline in India's tariffs (15 percentage points) increases the probability to participate in the export market by 2.13 percentage points for all firms in the selected sectors and 1.64 percentage points for the sample of non-exporters in 2004.

All coefficients ($\beta_{\tau^{ex},n}$) are negative within the selected sectors (columns 1 to 9) though some firms in the first quartile are not always statistically significant. This suggests that some firms in the first quartile are less likely to be induced to export with a reduction in

India's tariffs, which is consistent with the model prediction. Nevertheless, the firm size distribution may not be a good measure of firm productivity and the export productivity cutoffs could differ across industries, which explains the significance of the fourth quartile.

Overall, most of firms in the first quartile are still below the productivity threshold of exports after liberalization, while firms in the middle range (second and third quartiles) of the size distribution are more likely to be induced to enter in the export market. However, the empirical findings show that firms in the fourth quartile have the largest incentive to enter the export market as the absolute value of $\beta_{\tau^{ex},4}$ is the biggest in each column of selected sectors. This result is not in line with the model prediction that the most productive firms would always export regardless of tariffs. Besides the initial firm size not being a perfect measure, some relevant policies in China could help explain this finding. Former Chinese leader Hu Jintao encouraged large-scale enterprises to participate the export market and implemented the the Eleventh Five-Year Plan in 2006. In particular, the chapter 11 of the Eleventh Five-Year Plan includes goals to revitalize manufacturing of major technical equipment, strengthen the shipbuilding industry and improve the performance of the automotive industry. Therefore, this policy can explain why more firms in the fourth quartiles enter the export market during the trade-integration period.

Even though the impact of trade liberalization on other sectors is not significant on average as shown in panel B of Table 3.2, the export decision of the fourth quartile of the firm size distribution is significantly and negatively affected by trade costs with point estimates around -0.05 presented in columns 10–12 of Table 3.3. This suggests that the 15 percentage point decline in India's tariffs also increases the probability of firms in the fourth quartile to enter in the export market by 0.75 percentage points in other sectors from 2004 to 2007. Thus, trade liberalization induces more entry in the export market for moderately productive firms (consistent with the benchmark model) only within certain specific sectors.

### *3.5.4 Skill Upgrading Decision*

In this section, we focus on the skill upgrading decisions made by firms. The decision of providing labor training is described in equations (3.7), (3.10) and (3.11). Following the

Table 3.3: Entry in the Export Market by Quantile of the Firm Size Distribution and Sector Group

| Sample | Selected sectors | | | | | | | | | Other sectors | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full sample | | | | | | Baseline non-exporters | | | Full sample | | |
| Dependent variable | Change in status | | | Status in the current year | | | Status in the current year | | | Change in status | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| ΔIndia's tariffs | | | | | | | | | | | | |
| × first size quartile | -0.070* | -0.054 | -0.050 | -0.079** | -0.064 | -0.060 | -0.060* | -0.059 | -0.060* | 0.016 | 0.015 | 0.014 |
| | (0.036) | (0.042) | (0.036) | (0.039) | (0.046) | (0.040) | (0.036) | (0.040) | (0.036) | (0.016) | (0.015) | (0.014) |
| × second size quartile | -0.113*** | -0.095** | -0.095** | -0.124*** | -0.107** | -0.107** | -0.084** | -0.082** | -0.083** | 0.001 | -0.001 | -0.001 |
| | (0.035) | (0.041) | (0.037) | (0.039) | (0.047) | (0.042) | (0.038) | (0.041) | (0.038) | (0.015) | (0.016) | (0.015) |
| × third size quartile | -0.126*** | -0.107*** | -0.107*** | -0.142*** | -0.124*** | -0.124*** | -0.109*** | -0.107*** | -0.107*** | -0.016 | -0.018 | -0.016 |
| | (0.033) | (0.041) | (0.034) | (0.037) | (0.044) | (0.038) | (0.034) | (0.039) | (0.035) | (0.019) | (0.020) | (0.020) |
| × fourth size quartile | -0.169*** | -0.150*** | -0.149*** | -0.194*** | -0.176*** | -0.175*** | -0.162*** | -0.160*** | -0.160*** | -0.051** | -0.054** | -0.051** |
| | (0.040) | (0.048) | (0.038) | (0.041) | (0.048) | (0.039) | (0.041) | (0.047) | (0.041) | (0.021) | (0.022) | (0.022) |
| ΔChina's tariffs w.r.t. world | | yes | | | yes | | | yes | | | yes | |
| ΔChina's tariffs w.r.t. India | | | yes | | | yes | | | yes | | | yes |
| Sector FE | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | |
| Year FE | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | |
| Firm-level controls | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | |
| Industry controls | | yes | yes | | yes | yes | | yes | yes | | yes | yes |
| Observations | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 | 60,273 | 60,273 | 60,273 | 302,511 | 302,511 | 302,511 |
| $R^2$ | 0.004 | 0.004 | 0.005 | 0.884 | 0.884 | 0.884 | 0.320 | 0.320 | 0.320 | 0.005 | 0.005 | 0.005 |

Notes: (1) Standard errors are clustered at the 4-digit CIC industry level. (2) Δ denotes the year-over-year change in a variable during the period 2004–2007. (3) Firm-level controls include dummies for the second, third, and fourth quartile of the firm-size distribution in the initial year (2004). (4) Industry controls include import demand elasticity, export supply elasticity, skill intensity, and capital intensity in the United States. (5) Controls for changes in China's tariffs with respect to the world and India include both output and input tariffs. (6) ***, **, * denote significance level at 1%, 5%, and 10%, respectively. (7) Selected sectors include 14 manufacturing sectors mentioned in footnote 1.

estimation of equation (3.15), in addition to India's tariffs, we control for China's import tariffs regarding outputs and inputs, four-digit CIC industry characteristics, sector dummies and plant fixed effects; thus, the level of investment in on-the-job training can be expressed as:

$$\log TS_{ijst} = \beta_{\tau^{ex}} \tau_{jt}^{ex} + \beta_{\tau^{im}} \tau_{jt}^{im} + \alpha_{st} + \mu_i + \epsilon_{ijst} \tag{3.18}$$

where $TS$ denotes a firm's spending of labor training; $\tau_{jt}^{im}$ are China's import tariffs, which also affect firm revenues and skill upgrading decisions.

*First-Differenced Estimation*

Similarly, we estimate equation (3.18) in first differences to eliminate constant plant and sector heterogeneity:

$$\Delta \log TS_{ijst} = \beta_{\tau^{ex}} \Delta\tau_{jt}^{ex} + \beta_{\tau^{im}} \Delta\tau_{jt}^{im} + \Delta\alpha_{st} + \Delta\epsilon_{ijst} \tag{3.19}$$

Panel A of Table 3.4 indicates that trade liberalization between 2004 and 2007 induces more investment in on-the-job training by manufacturing firms in selected sectors. In particular, columns 5 and 8 of panel A with inclusion of additional controls shows that the 15 percentage point decline in India's tariffs increases labor training provided by firms by about 0.11 to 0.13 log points. When trade costs become lower, the productive firms earn greater revenues, so they have higher incentives to increase human capital investment and produce skill-intensive products more.

Nevertheless, there are reverse findings in other sectors from panel B of Table 3.4 with average point estimates around 0.15 (a 0.02 log point reduction in training spending occurs when India's tariffs drop by 15 percentage points). In these sectors, firms reduce labor training even though trade barriers decline. There could be several reasons. First, the productivity (not directly observed) of more firms in these sectors may fall between $z_{ms}^0$ and $z_{ms}^1$ in Figure 3.3, so the reduction in India's tariffs may not affect their export participation (not specific to any country) as shown by Table 3.2, but can cause skill downgrading. Second, the local government may finance labor training when trade barriers are high, but reduce fiscal support when trade barriers become lower. Findings from panels A and B imply that

Table 3.4: Investment in On-the-Job Training Stratified by Sector Group and Initial Export Status

| Dependent variable: year-over-year change in log(training spending) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| *Panel A: Sample of selected sectors.* | | | | | | | | |
| ΔIndia's tariffs | -0.854*** | -0.854*** | -0.903*** | -0.796** | -0.753* | -0.859*** | -0.883*** | -0.842*** |
| | (0.304) | (0.304) | (0.325) | (0.397) | (0.409) | (0.316) | (0.311) | (0.320) |
| ΔChina's tariffs w.r.t. world | | | | | | | | |
| Outputs | | | yes | yes | yes | | | |
| Inputs | | | | yes | yes | | | |
| ΔChina's tariffs w.r.t. India | | | | | | | | |
| Outputs | | | | | | yes | yes | yes |
| Inputs | | | | | | | yes | yes |
| Sector FE | yes | yes | yes | yes | yes | yes | yes | yes |
| Year FE | yes | yes | yes | yes | yes | yes | yes | yes |
| Firm-level controls | | yes | yes | yes | yes | yes | yes | yes |
| Industry controls | | | | | yes | | | yes |
| Observations | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 | 91,869 |
| $R^2$ | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| *Panel B: Sample of other sectors.* | | | | | | | | |
| ΔIndia's tariffs | 0.140* | 0.139* | 0.144* | 0.145* | 0.141 | 0.142* | 0.152* | 0.148* |
| | (0.081) | (0.081) | (0.083) | (0.086) | (0.086) | (0.082) | (0.085) | (0.085) |
| Observations | 302,511 | 302,511 | 302,511 | 302,511 | 302,511 | 302,511 | 302,511 | 302,511 |
| $R^2$ | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| *Panel C: Sample of baseline non-exporters in selected sectors.* | | | | | | | | |
| ΔIndia's tariffs | -0.810** | -0.804** | -0.818** | -0.778* | -0.683 | -0.820** | -0.908*** | -0.824** |
| | (0.324) | (0.324) | (0.344) | (0.443) | (0.441) | (0.339) | (0.325) | (0.320) |
| Observations | 60,273 | 60,273 | 60,273 | 60,273 | 60,273 | 60,273 | 60,273 | 60,273 |
| $R^2$ | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| *Panel D: Sample of baseline exporters in selected sectors.* | | | | | | | | |
| ΔIndia's tariffs | -0.917 | -0.915 | -1.018 | -0.931 | -1.096 | -0.878 | -0.889 | -1.050 |
| | (0.788) | (0.785) | (0.851) | (0.869) | (0.900) | (0.805) | (0.805) | (0.833) |
| Observations | 31,596 | 31,596 | 31,596 | 31,596 | 31,596 | 31,596 | 31,596 | 31,596 |
| $R^2$ | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |

Notes: (1) Standard errors are clustered at the 4-digit CIC industry level. (2) Δ denotes the year-over-year change in a variable during the period 2004–2007. (3) Firm-level controls include number of employees and sales per worker, all measured in the initial year (2004). (4) Industry controls include import demand elasticity, export supply elasticity, skill intensity, and capital intensity in the United States. (5) In panels B, C, and D, remaining controls are the same as in the corresponding column in panel A. (6) ***, **, * denote significance level at 1%, 5%, and 10%, respectively. (7) Selected sectors include 14 manufacturing sectors mentioned in footnote 1.

sector heterogeneity in productivity and policy plays a role in affecting skill upgrading decisions following trade integration. The theoretical model explains trade-induced investment in on-the-job training within some sectors, while providing a possibility of skill downgrading in other sectors where firms are around the cutoff of switching destination countries due to an increase in skill upgrading productivity cutoff in country $m$.

In terms of the sample of non-exporters in 2004 in selected sectors, the estimation in panel C of Table 3.4 implies that skill upgrading in response to a reduction in India's tariff is still positive but less precise. The estimated coefficients for the sample of exporters in 2004 in selected sectors are also insignificant as presented in panel D. One possible explanation is that some continuing exporters switch to the Indian market due to an increase in skill upgrading productivity cutoff in non-Indian countries and no longer produce high-skill products, and some exiting exporters no longer demand high-skill workers when serving the domestic market only.[16]

*Skill Upgrading Decision by Quartile of the Firm Size Distribution*

The benchmark model predicts that lower trade costs encourage firms operating in the range $z_s^1 < z < z_s^0$ to provide more labor training, since a reduction in trade costs decreases the skill upgrading productivity cutoff $z_s$. As shown in Figure 3.2, these firms are in the middle range of the productivity distribution, and they invest more in human capital following trade liberalization. The least and the most productive firms wouldn't change their decisions of labor skill upgrading in response to trade openness. Empirically, we estimate the impact of the change in India's tariffs on each quartile of the initial firm size distribution with the following equation:

$$\Delta \log TS_{ijst} = \sum_{n=1}^{4} \beta_{\tau^{ex},n}(\Delta\tau_{jt}^{ex} \times Q_{ij,n}) + \sum_{n=1}^{4} \delta_n Q_{ij,n} + \beta_{\tau^{im}}\Delta\tau_{jt}^{im} + \Delta\alpha_{st} + \Delta\epsilon_{ijst} \quad (3.20)$$

where $n$ is each of the four quartiles of the firm size distribution and $Q_{ij,n}$ are dummy variables being 1 when firm $i$ belongs to quartile $n$. Columns 1–3 of Table 3.5 show that the effect of the reduction in India's tariffs on investment in on-the-job training is significant in

[16] Although not shown in Table 3.2, the effect of trade liberalization on entry in the export market is less statistically significant (although larger) in the sample of baseline exporters.

the first three quartiles of the firm size distribution for the full sample of selected sectors. Trade liberalization has a relatively larger impact on the skill upgrading decision of firms in the second quartile. For instance, the 15 percentage point reduction in India's tariffs from 2004 to 2007 increases spending on training of firms in the second quartile by 0.18 log points, while firms in the first and third quartiles increase labor training by only about 0.14 log points. As the model predicts that firms with productivity in the middle range are sensitive to changes in trade costs, so the reduction in the tariffs positively affects skill upgrading decisions of firms in the second and third quartiles.

One question is: why firms in the low or lower-middle range of the size distribution choose to increase human capital investment after liberalization? The benchmark model cannot match this empirical result. One possible reason is that Chinese local governments protect some smaller domestic firms and state-owned enterprises. The less productive firms could receive subsidies from governments or benefit from new regulations, so that they can follow their high productive competitors to provide more labor training and employ more high-skilled workers during the liberalization period, especially among those who are encouraged to "go out" (a famous slogan from the Chinese government during that period).

In terms of the sample of initial non-exporters and the sample of initial exporters in 2004, the trade-integration effect on labor training by firms in the first quartile is very similar to the findings in panels C and D of Table 3.4. Regardless of the initial export status, the effect of the reduction in tariffs on the fourth quartile is less precisely estimated. It is consistent with the model prediction that the most productive firms ($z > z_s$) still find it profitable to provide labor training even when the India's tariffs are lower.

The point estimates in other sectors are positive but not significant for the first and third quartiles as shown in columns 10–12 of Table 3.5. The reduction in tariffs induces a statistically significant reduction in spending on labor training in the fourth quartile of the firm size distribution. Precisely, the 15 percentage point decline in India's tariffs from 2004 to 2007 reduces spending on training by about 0.07 log points for firms in the fourth quartile. Some of these large firms may actually be in the range $z^1_{mx} < z < z^1_{ms}$ in Figure 3.3, as initial firm size may not be a perfect measure of productivity. As a result, trade liberalization may both encourage the large firms to export and downgrade skill level on average. Interestingly,

Table 3.5: Investment in On-the-Job Training by Quantile of the Firm Size Distribution, Sector Group, and Initial Export Status

| Dependent variable: year-over-year change in log(training spending) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | Selected sectors | | | | | | | | | Other sectors | | |
| | Full sample | | | Baseline non-exporters | | | Baseline exporters | | | Full sample | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| ΔIndia's tariffs | | | | | | | | | | | | |
| × first size quartile | -0.977*** | -0.913** | -0.992*** | -1.065*** | -0.999** | -1.117*** | -0.860 | -0.895 | -0.845 | 0.015 | 0.018 | 0.022 |
| | (0.338) | (0.416) | (0.344) | (0.330) | (0.419) | (0.330) | (0.904) | (0.981) | (0.919) | (0.126) | (0.129) | (0.129) |
| × second size quartile | -1.222*** | -1.153** | -1.246*** | -1.045** | -0.972* | -1.117*** | -1.838** | -1.857* | -1.824** | -0.100 | -0.092 | -0.086 |
| | (0.380) | (0.478) | (0.386) | (0.403) | (0.523) | (0.407) | (0.877) | (0.959) | (0.897) | (0.097) | (0.101) | (0.101) |
| × third size quartile | -0.924*** | -0.854** | -0.953*** | -0.650* | -0.573 | -0.740** | -1.537* | -1.560* | -1.511* | 0.088 | 0.097 | 0.104 |
| | (0.332) | (0.427) | (0.335) | (0.352) | (0.469) | (0.351) | (0.846) | (0.938) | (0.875) | (0.116) | (0.122) | (0.120) |
| × fourth size quartile | -0.168 | -0.099 | -0.195 | 0.048 | 0.124 | -0.051 | -0.420 | -0.432 | -0.388 | 0.476*** | 0.485*** | 0.493*** |
| | (0.343) | (0.420) | (0.354) | (0.338) | (0.423) | (0.353) | (0.866) | (0.952) | (0.884) | (0.173) | (0.179) | (0.176) |
| ΔChina's tariffs w.r.t. world | | yes | | | yes | | | yes | | | yes | |
| ΔChina's tariffs w.r.t. India | | | yes | | | yes | | | yes | | | yes |
| Sector FE | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | |
| Year FE | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | |
| Firm-level controls | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | |
| Industry controls | | yes | yes | | yes | yes | | yes | yes | | yes | yes |
| Observations | 91,869 | 91,869 | 91,869 | 60,273 | 60,273 | 60,273 | 31,596 | 31,596 | 31,596 | 302,511 | 302,511 | 302,511 |
| $R^2$ | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |

Notes: (1) Standard errors are clustered at the 4-digit CIC industry level. (2) Δ denotes the year-over-year change in a variable during the period 2004–2007. (3) Firm-level controls include dummies for the second, third, and fourth quartile of the firm-size distribution in the initial year (2004). (4) Industry controls include import demand elasticity, export supply elasticity, skill intensity, and capital intensity in the United States. (5) Controls for changes in China's tariffs with respect to the world and India include both output and input tariffs. (6) ***, **, * denote significance level at 1%, 5%, and 10%, respectively. (7) Selected sectors include 14 manufacturing sectors mentioned in footnote 1.

trade liberalization induces more labor training of firms in the second quartile in these other sectors with $\beta_{\tau^{ex},2}$ equal to about -0.09.

From the findings in Table 3.3 and Table 3.5, trade liberalization induces a significant increase in both the probability of export participation and spending on labor training by firms in the second and third quartiles in selected sectors, which is consistent with the benchmark model prediction. Results based on other sectors and for other quartiles are partially explained by our extended model. In the next section, we attempt to extend our empirical analysis to focus more on the extended model, and make some further discussions including sector characteristics and specific policies or regulations in China to understand those empirical results.

### *3.5.5 Extension*

Recall that the data pattern shows that India's tariffs are much higher than those of other countries, but the reduction in India's tariffs is much more drastic between 2004 and 2007. Thus, we build the extended model that distinguishes two different export destinations, countries $m$ and $o$ in order to highlight the effect of trade liberalization on exporters to the Indian market or other foreign markets. Next, we estimate two similar first-differenced models as in subsections 3.5.3 and 3.5.4, but analyze the export destination decisions of new exporters and the skill upgrading decisions of new exporters to the Indian market.

Rewriting equation (3.13) when $k = 1$ yields:

$$EX^1_{ijst} = \begin{cases} 1 & \text{if exporting to India} \\ 0 & \text{if exporting to non-Indian countries} \end{cases}$$

Figure 3.3 shows that more continuing exporters switch to India and new exporters are more likely to enter the Indian market after India's tariffs are reduced. Meanwhile, within-industry patterns in the data (Table 3.1) shows that new exporters have the largest increase in sales and training per worker from 2004 to 2007. Hence, to estimate the equation (3.21), we select the sample of firms who do not export in 2004 but become exporters in 2007.

$$\Delta EX^1_{ijst} = \sum_{n=1}^{4} \beta^1_{\tau^{ex},n}(\Delta\tau^{ex}_{jt} \times Q_{ij,n}) + \sum_{n=1}^{4} \delta^1_n Q_{ij,n} + \beta^1_{\tau^{im}} \Delta\tau^{im}_{jt} + \Delta\alpha^1_{st} + \Delta\epsilon^1_{ijst} \qquad (3.21)$$

Table 3.6: Entry in the India Export Market and Investment in On-the-Job Training

| Sample | Selected sectors New exporters | | | Selected sectors New exporters to India | | |
|---|---|---|---|---|---|---|
| Dependent variable | Change in status of exporting to India | | | log(training spending) | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| ΔIndia's tariffs | | | | | | |
| × first size quartile | -1.721*** | -1.723*** | -1.888*** | -8.054 | -15.082* | -3.714 |
| | (0.484) | (0.562) | (0.522) | (6.791) | (8.015) | (5.745) |
| × second size quartile | -1.533*** | -1.555*** | -1.676*** | -13.208* | -20.970** | -10.031 |
| | (0.486) | (0.555) | (0.512) | (7.591) | (8.530) | (6.596) |
| × third size quartile | -1.450*** | -1.471** | -1.522*** | -11.187 | -18.683** | -7.860 |
| | (0.485) | (0.565) | (0.502) | (7.409) | (8.943) | (5.681) |
| × fourth size quartile | -1.600*** | -1.615*** | -1.654*** | -15.153 | -23.569** | -12.200* |
| | (0.533) | (0.603) | (0.549) | (9.079) | (10.006) | (7.083) |
| ΔChina's tariffs w.r.t. world | | yes | | | yes | |
| ΔChina's tariffs w.r.t. India | | | yes | | | yes |
| Sector FE | yes | yes | yes | yes | yes | yes |
| Year FE | yes | yes | yes | yes | yes | yes |
| Firm-level controls | yes | yes | yes | yes | yes | yes |
| Industry controls | | yes | yes | | yes | yes |
| Observations | 2,475 | 2,475 | 2,475 | 489 | 489 | 489 |
| $R^2$ | 0.025 | 0.026 | 0.028 | 0.035 | 0.061 | 0.054 |

Notes: (1) Standard errors are clustered at the 4-digit CIC industry level. (2) Δ denotes the year-over-year change in a variable during the period 2004–2007. (3) Firm-level controls include dummies for the second, third, and fourth quartile of the firm-size distribution in the initial year (2004). (4) Industry controls include import demand elasticity, export supply elasticity, skill intensity, and capital intensity in the United States. (5) ***, **, * denote significance level at 1%, 5%, and 10%, respectively. (6) New exporters and new exporters to India are defined in Table 3.1. (7) Selected sectors include 14 manufacturing sectors mentioned in footnote 1.

Columns 1–3 of Table 3.6 present the estimated results of equation (3.21). Coefficients in each quartile of size distribution are statistically significant. In particular, the 15 percentage point decline in India's tariffs increases the probability of entering the Indian market for new exporters in the high productivity group (fourth quartile) by 24.23 percentage points. Furthermore, the empirical results of exporting to India in Table 3.6 are consistent with the extended model prediction as there are three shaded areas located in different ranges of productivity levels in Figure 3.3, representing that firms in each range of productivity levels could find it more profitable to export to India.

Next, recall that some firms switch export destinations from country $m$ to $o$ and start

to produce unskilled or skilled goods after trade liberalization (Figure 3.3). To investigate the skill upgrading decisions made by new exporters to India, we estimate equation (3.20) again with the sample of non-exporters in 2004 and exporters to India in 2007.

In column 5 of Table 3.6, trade liberalization is shown to have significantly positive effects on skill upgrading of new exporters targeting India in the last three quartiles. Although firm size may not be a perfect measure of productivity, this reflects that new exporters targeting India with a size above the medium level actually start to increase investment in on-the-job training when trade costs are lower. The absolute value of the coefficient of the fourth quartile is the largest potentially due to the fact that the most productive new exporters to India are more capable of increasing labor training. Precisely, column 5 of Table 3.6 presents that a 15 percentage point reduction in trade costs from 2004 to 2007 leads to a 3.54 log point increase in labor training provided by the new exporters to India in the fourth quartile. The absolute value of the coefficient of second quartile is smaller, but still larger than that of the third quartile. These results are in line with the pattern in Figure 3.3 as some firms in the middle range of productivity levels continue to produce unskilled products or reduce spending on labor training.

### *3.5.6 Mechanism*

Empirically, the reduction in the tariffs induces more firm entry in the export market and increase spending on labor training in the second and third quartiles of the firm size distribution (in selected sectors) based on the previous tables, which indicates that firms in the middle range of the productivity distribution greatly benefit from trade liberalization. The model mechanism implies that firms gain higher revenues when trade costs are lower, so they find it more profitable to export and have greater incentives in skill upgrading, mirroring the reduction in productivity cutoffs of export and skill upgrading. In this section, we provide evidence that how trade integration between China and India affect China's export sales to India and domestic sales.

Table 3.7: Export Sales to India and Domestic Sales of New Exporters to India in Selected Sectors

| Dependent variable | Change in log(export sales to India) | | | Change in log(domestic sales) | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| ΔIndia's tariffs | -40.851*** | -41.207*** | -43.108** | | | 0.016 |
| | (13.352) | (13.727) | (17.650) | | | (1.577) |
| ΔChina's tariffs w.r.t. India | | | | | | |
| Output | | yes | | 0.302 | 0.521 | 0.518 |
| | | | | (0.862) | (1.995) | (1.984) |
| Input | | yes | | | -1.515 | -1.517 |
| | | | | | (14.819) | (14.862) |
| ΔChina's tariffs w.r.t. world | | | yes | | | |
| Sector FE | yes | yes | yes | yes | yes | yes |
| Year FE | yes | yes | yes | yes | yes | yes |
| Firm-level controls | yes | yes | yes | yes | yes | yes |
| Industry controls | | yes | yes | | yes | yes |
| Observations | 489 | 489 | 489 | 489 | 489 | 489 |
| $R^2$ | 0.141 | 0.154 | 0.153 | 0.063 | 0.071 | 0.071 |

Notes: (1) Standard errors are clustered at the 4-digit CIC industry level. (2) Δ denotes the year-over-year change in a variable during the period 2004–2007. (3) Firm-level controls include number of employees and sales per worker, all measured in the initial year (2004). (4) Industry controls include import demand elasticity, export supply elasticity, skill intensity, and capital intensity in the United States. (5) ***, **, * denote significance level at 1%, 5%, and 10%, respectively. (6) New exporters to India are defined in Table 3.1. (7) Selected sectors include 14 manufacturing sectors mentioned in footnote 1.

*Export Sales to India*

We select the sample of firms who do not export in 2004 but become new exporters to India in 2007. We find that the reduction in India's tariffs increases China's export sales to India as reported by columns 1–3 of Table 3.7 following the previous estimation method. When we control for the change of China's import tariffs with respect to India, the 15 percentage point reduction in India's tariffs leads to an increase in export sales to India by about 6.18 log points. This exercise produces consistent results, and mainly reflects changes of export sales of new exporters to India. The coefficients are large in magnitude because we analyze the export sales at the firm level instead of the industry level. The firm level data can help emphasize how a small group of firms respond to changes in trade costs.

*Domestic Sales*

Our theoretical model also shows that the reduction in trade costs leads to a decline in domestic sales and causes more low productivity firms to exit the market. However, trade costs are symmetric for two countries in the model, which cannot be easily matched to the data. In fact, India's tariffs differ from China's tariffs. The empirical evidence under columns 4–6 of Table 3.7 suggests that the decline in China's tariffs with respect to India could result in lower domestic sales with point estimates from 0.30–0.52. However, these results are not significant. This is probably due to the fact that China as a developing country has a rapid growth rate, as well as a large population. For instance, GDP annual growth rate in China increased from 10% in 2004 to 14% in 2007.[17] Chinese firms increase export sales a lot following trade openness, and also can maintain a great amount of domestic sales even if there are more imported varieties.

### *3.5.7 Discussion*

In this section, we attempt to understand why firms in certain sectors would enter in the export markets and invest more in on-the-job training following trade liberalization, while others would not. The theoretical benchmark model can only explain some empirical findings, while others can result from special sector characteristics or policies targeting certain industries. Due to sector heterogeneity in productivity and policy, trade liberalization has positive impacts on firms' decisions of export entry and skill upgrading in only some selected sectors, including manufacturers of "ship and floating devices".

First, some large-scale manufacturing companies in China such as pharmaceutical, home appliances and electronics manufacturers, have their own universities for labor training, so their labor training procedures are canonical and their training decision might not be sensitive to changes in trade costs. Second, industry characteristics and certain policies could determine whether some industries have comparative advantage or sustainable competition in the foreign market. We pick the shipbuilding industry from selected sectors and the textile industry from other sectors to understand their different responses to the regional

[17]Data source: World Bank national accounts data, and OECD National Accounts data files.

trade liberalization.

*Textile Industry.* According to an investigation report from the CNBS[18], the labor-intensive textile industry used to have a strong advantage in terms of labor costs, but it is offset by low labor productivity. Compared with China's main competitors in Asia, its labor costs gradually lose the advantage. In 2002, the average wage level in China's textile industry reached 1.12 times that of India. Moreover, the production technology of spinning machinery is relatively mature, while the production technology of weaving and sewing machinery is relatively backward. Due to low productivity levels and less advanced technologies, some firms in the textile industry are less competitive in the export market, and find it not profitable to participate in exports or increase labor training even if trade costs are very low. This explains why they are not sensitive to changes in tariffs. Instead, they could have a higher investment in on-the-job training when tariffs are high and when they could receive protection or subsidies from the government.

*Shipbuilding Industry.* China engages in foreign trade further after joining the WTO and APTA in 2001. In particular, China's total volume of imports and exports increased by 23.2% and the export of mechanical and electrical products and high-tech products increased by 32.0% and 31.8% respectively, from 2004 to 2005.[19] Meanwhile, the Eleventh Five-Year-Plan was announced in 2006, which encourages large-scale enterprises such as shipbuilding or auto-car firms to enter the export market. In 2020, China was still the world's largest shipbuilding market, accounting for 43.1% of total shipbuilding volume in the world. Shipbuilding industry is one of the selected sectors in the empirical studies. Shipbuilding firms are more productive and likely to fall in the second or the third quartiles of the firm size distribution. They should be sensitive to changes in trade costs. Besides, they actually receive supports in the export market after the government implements the Eleventh Five-Year-Plan. Thus, the reduction in tariffs still have significantly positive impacts on their export entry decisions.

---

[18]The first China Industrial Security Forum was held in 2006. It reported the domestic environment of China's textile industry from 1997 to 2005.

[19]These data were reported at the Fourth Session of China's Tenth National Congress in 2006.

### *3.6 Conclusion*

The evidence from selected sectors reported in this chapter suggests that a reduction of trade costs can help reduce the export productivity cutoff and increases profits for exporters, resulting in more export participation and more spending in labor training. According to the model implication and empirical results, the positive impact of trade liberalization on firms in the middle range of productivity levels is the largest. Due to sector heterogeneity and specific policies targeting to some other sectors, the empirical findings in other sectors revert. As expanded export opportunities positively affect firm performance in selected sectors, it is important to implement trade policies targeting to certain industries, including regional or multilateral trade liberalizations. More export participation leads to more investment in on-the-job training, which increases labor skill levels and improves product quality in the long term.

# BIBLIOGRAPHY

E. Kathleen Adams, Robert Houchens, George E. Wright, and James Robbins. Predicting hospital choice for rural Medicare beneficiaries: the role of severity of illness. *Health Services Research*, 26(5):583–612, 1991.

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

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

George A. Akerlof. The market for "lemons". *Quarterly Journal of Economics*, 84(3):488–500, 1970.

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

Mary Amiti and Jozef Konings. Trade liberalization, intermediate inputs, and productivity: Evidence from Indonesia. *American Economic Review*, 97(5):1611–1638, December 2007. doi: 10.1257/aer.97.5.1611.

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

Kenneth J. Arrow. Uncertainty and the welfare economics of medical care: Reply (the implications of transaction costs and adjustment lags). *American Economic Review*, 55 (1/2):154–158, 1965.

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

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

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

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

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

Simona Baldi and Davide Vannoni. The impact of centralization on pharmaceutical procurement prices: the role of institutional quality and corruption. *Regional Studies*, 51(3): 426–438, 2017.

Maria Bas. Technology adoption, export status, and skill upgrading: Theory and evidence. *Review of International Economics*, 20(2):315–331, 2012.

Samuel Bazzi, Arya Gaduh, Alexander D. Rothenberg, and Maisy Wong. Skill transferability, migration, and development: Evidence from population resettlement in Indonesia. *American Economic Review*, 106(9):2658–2698, 2016.

Gloria J. Bazzoli, Richard C. Lindrooth, Romana Hasnain-Wynia, and Jack Needleman. The Balanced Budget Act of 1997 and US hospital operations. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 41(4):401–417, 2004.

Andrew B. Bernard, Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum. Plants and productivity in international trade. *American Economic Review*, 93(4):1268–1290, 2003.

Ernst R. Berndt, Linda Bui, David R. Reiley, and Glen L. Urban. Information, marketing, and pricing in the US antiulcer drug market. *American Economic Review*, 85(2):100–105, 1995.

Ernst R. Berndt, Linda Bui, David H. Lucking-Reiley, and Glen L. Urban. The roles of marketing, product quality, and price competition in the growth and composition of the U.S. antiulcer drug industry. In Timothy F. Bresnahan and Robert J. Gordon, editors, *The Economics of New Goods*, pages 277–328. University of Chicago Press, Chicago, 1996.

Ernst R. Berndt, Margaret Kyle, and Davina Ling. The long shadow of patent expiration: Generic entry and Rx-to-OTC switches. In Robert C. Feenstra and Matthew D. Shapiro, editors, *Scanner Data and Price Indexes*. University of Chicago Press, Chicago, 2003a.

Ernst R. Berndt, Robert S. Pindyck, and Pierre Azoulay. Consumption externalities and diffusion in pharmaceutical markets: Antiulcer drugs. *Journal of Industrial Economics*, 51(2):243–270, 2003b.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 64(4):841–890, 1995.

Steven T Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 25(2):242–262, 1994.

Steven T Berry and Panle Jia. Tracing the woes: An empirical analysis of the airline industry. *American Economic Journal: Microeconomics*, 2(3):1–43, 2010.

Sofia Berto Villas-Boas. Vertical relationships between manufacturers and retailers: Inference with limited data. *The Review of Economic Studies*, 74(2):625–652, 2007.

Sheng Bin. China's trade development strategy and trade policy reforms: overview and prospect. *International Institute for sustainable development. Draft Paper*, 2015.

Jonas Björnerstedt and Frank Verboven. Does merger simulation work? evidence from the Swedish analgesics market. *American Economic Journal: Applied Economics*, 8(3): 125–164, 2016.

Åke Blomqvist. Optimal non-linear health insurance. *Journal of Health Economics*, 16(3): 303–321, 1997.

Céline Bonnet and Pierre Dubois. Inference on vertical contracts between manufacturers and retailers allowing for nonlinear pricing and resale price maintenance. *The RAND Journal of Economics*, 41(1):139–164, 2010.

Loren Brandt, Johannes Van Biesebroeck, Luhang Wang, and Yifan Zhang. WTO accession and performance of chinese manufacturing firms. *American Economic Review*, 107(9): 2784–2820, 2017.

Christian Broda and David E. Weinstein. Globalization and the gains from variety. *The Quarterly Journal of Economics*, 121(2):541–585, 2006.

Christian Broda, Nuno Limao, and David E. Weinstein. Optimal tariffs and market power: the evidence. *American Economic Review*, 98(5):2032–2065, 2008.

Zarek C. Brot-Goldberg, Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad. What does a deductible do?: The impact of cost-sharing on health care prices, quantities, and spending dynamics. *Quarterly Journal of Economics*, 132(3):1261–1318, 2017.

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

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

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

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

Ariel Burstein and Marc J. Melitz. Trade liberalization and firm dynamics. *Advances in Economics and Econometrics Tenth World Congress. Applied Economics, Econometric Society Monographs*, 2, 2013.

Paula Bustos. The impact of trade liberalization on skill upgrading: Evidence from Argentina. Barcelona GSE Working Paper 559, Barcelona School of Economics, 2011a.

Paula Bustos. Trade liberalization, exports, and technology upgrading: Evidence on the impact of MERCOSUR on Argentinian firms. *American Economic Review*, 101(1):304–340, 2011b.

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

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

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

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

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

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

CPC Central Committee. Opinions of the CPC Central Committee and the State Council on deepening the reform of the medical and healthcare system [In Chinese]. http://www.gov.cn/jrzg/2009-04/06/content_1278721.htm, March 2009.

Gregory S. Crawford and Matthew Shum. Uncertainty and learning in pharmaceutical demand. *Econometrica*, 73(4):1137–1173, 2005.

Gregory S. Crawford and Ali Yurukoglu. The welfare effects of bundling in multichannel television markets. *American Economic Review*, 102(2):643–685, 2012.

Janet Currie, Wanchuan Lin, and Wei Zhang. Patient knowledge and antibiotic abuse: Evidence from an audit study in China. *Journal of Health Economics*, 30(5):933–949, 2011.

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

Avinash K. Dixit and Joseph E. Stiglitz. Monopolistic competition and optimum product diversity. *American Economic Review*, 67(3):297–308, 1977.

Julie M. Donohue and Ernst R. Berndt. Effects of direct-to-consumer advertising on medication choice: The case of antidepressants. *Journal of Public Policy and Marketing*, 23 (2):115–127, 2004.

Julie M. Donohue and Ernst R. Berndt. Information content of advertising: Empirical evidence from the OTC analgesic industry. *International Journal of Industrial Organization*, 31(5):355–367, 2013.

Pierre Dubois and Laura Lasio. Identifying industry margins with price constraints: Structural estimation on pharmaceuticals. *American Economic Review*, 108(12):3685–3724, 2018.

Pierre Dubois, Ashvin Gandhi, and Shoshana Vasserman. Bargaining and international reference pricing in the pharmaceutical industry. Working paper, Havard University, 2019a.

Pierre Dubois, Yassine Lefouili, and Stéphane Straub. Pooled procurement of drugs in low and middle income countries. Working Paper 508, Center for Global Development, 2019b.

Mark Duggan and Fiona M. Scott Morton. The distortionary effects of government procurement: Evidence from Medicaid prescription drug purchasing. *Quarterly Journal of Economics*, 121(1):1–30, 2006.

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

Eric V. Edmonds, Nina Pavcnik, and Petia Topalova. Trade adjustment and human capital investments: Evidence from Indian tariff reform. *American Economic Journal: Applied Economics*, 2(4):42–75, 2010.

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

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

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

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

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

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

José J. Escarce and Kanika Kapur. Do patients bypass rural hospitals?: Determinants of

inpatient hospital choice in rural California. *Journal of Health Care for the Poor and Underserved*, 20(3):625–644, 2009.

Rod Falvey, David Greenaway, and Joana Silva. Trade liberalisation and human capital adjustment. *Journal of International Economics*, 81(2):230–239, 2010.

Xueshan Feng, Shenglan Tang, Gerald Bloom, Malcolm Segall, and Xingyuan Gu. Cooperative medical schemes in contemporary rural China. *Social Science & Medicine*, 41(8): 1111–1118, 1995.

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

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

Nicole Fortin, Thomas Lemieux, and Sergio Firpo. Decomposition methods in economics. In *Handbook of Labor Economics*, volume 4, pages 1–102. Elsevier, 2011.

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

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

Fei Gao, Yu Jie Zhou, Da Yi Hu, Ying Xin Zhao, Yu Yang Liu, Zhi Jian Wang, Shi Wei Yang, and Xiao Li Liu. Contemporary management and attainment of cholesterol targets for patients with dyslipidemia in China. *PLoS One*, 8(4):e47681, 2013.

Philip G Gayle. On the efficiency of codeshare contracts between airlines: is double marginalization eliminated? *American Economic Journal: Microeconomics*, 5(4):244–73, 2013.

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

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

Fabio Ghironi and Marc J. Melitz. International trade and macroeconomic dynamics with heterogeneous firms. *Quarterly Journal of Economics*, 120(3):865–915, 2005.

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

Gautam Gowrisankaran, Aviv Nevo, and Robert Town. Mergers when prices are negotiated: Evidence from the hospital industry. *American Economic Review*, 105(1):172–203, 2015.

Matthew Grennan. Price discrimination and bargaining: Empirical evidence from medical devices. *American Economic Review*, 103(1):145–177, 2013.

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

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

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

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

Elhanan Helpman, Marc J. Melitz, and Yeaple Stephen. Export versus FDI with heterogeneous firms. *American Economic Review*, 94:300–316, 2004.

Hanns Günther Hilpert. China's trade policy: dominance without the will to lead. SWP Research Paper, German Institute for International and Security Affairs, 2014.

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

Kate Ho and Robin S. Lee. Insurer competition in health care markets. *Econometrica*, 85 (2):379–417, 2017.

Kate Ho and Ariel Pakes. Hospital choices, hospital prices, and financial incentives to physicians. *American Economic Review*, 104(12):3841–84, 2014a.

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

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

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

Henrick Horn and Asher Wolinsky. Bilateral monopolies and incentives for merger. *The RAND Journal of Economics*, 19(3):408–419, 1988.

Zhiyuan Hou, Ellen Van de Poel, Eddy Van Doorslaer, Baorong Yu, and Qingyue Meng. Effects of NCMS on access to care and financial protection in China. *Health Economics*, 23(8):917–934, 2014.

Chang-Tai Hsieh and Peter J. Klenow. Misallocation and manufacturing TFP in China and India. *Quarterly Journal of Economics*, 124(4):1403–1448, 2009.

Qifan Huang and Castiel Chen Zhuang. Training, productivity and wages: An investigation of China's manufacturing enterprises in a privatization era. *Economics of Transition and Institutional Change*, 00:1–20, 2021. doi: https://doi.org/10.1111/ecot.12285.

Peter J. Huckfeldt, Neeraj Sood, José J. Escarce, David C. Grabowski, and Joseph P. Newhouse. Effects of Medicare payment reform: Evidence from the home health interim and prospective payment systems. *Journal of Health Economics*, 34:1–18, 2014.

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

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

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

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

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

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond, 2020.

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

Sean P. Keehan, Gigi A. Cuckler, Sisko andrea M., Madison andrew J., Sheila D. Smith, Devin A. Stone, John A. Poisal, Christian J. Wolfe, and Joseph M. Lizonitz. National health expenditure projections, 2014–24: spending growth faster than recent trends. *Health Affairs*, 34(8):1407–1417, 2015.

Myung-Hwa Kim and Soon-Man Kwon. The effect of outpatient cost sharing on health care utilization of the elderly. *Journal of Preventive Medicine and Public Health*, 43(6): 496–504, 2010.

Patrick Kline. Oaxaca-blinder as a reweighting estimator. *American Economic Review*, 101 (3):532–37, 2011.

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

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

Hyo Jung Lee, Sung-In Jang, and Eun-Cheol Park. The effect of increasing the coinsurance rate on outpatient utilization of healthcare services in South Korea. *BMC Health Services Research*, 17(1):152, 2017.

Christy Harris Lemak, Tammie A. Nahra, Genna R. Cohen, Natalie D. Erb, Michael L. Paustian, David Share, and Richard A. Hirth. Michigan's fee-for-value physician incentive program reduces spending and improves quality in primary care. *Health Affairs*, 34(4): 645–652, 2015.

Hongbin Li, Prashant Loyalka, Scott Rozelle, and Binzhen Wu. Human capital and china's future growth. *Journal of Economic Perspectives*, 31(1):25–48, 2017.

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

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

Qing Liu and Ruosi Lu. On-the-job training and productivity: Firm-level evidence from a large developing country. *China Economic Review*, 40:254–264, 2016.

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

Fangwen Lu. Insurance coverage and agency problems in doctor prescriptions: evidence from a field experiment in China. *Journal of Development Economics*, 106:156–167, 2014.

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

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

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

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

Alice Mannocci, Gabriella De Carli, Virginia Di Bari, Rosella Saulle, Brigid Unim, Nicola Nicolotti, Lorenzo Carbonari, Vincenzo Puro, and Giuseppe La Torre. How much do needlestick injuries cost? A systematic review of the economic evaluations of needlestick

and sharps injuries among healthcare personnel. *Infection Control & Hospital Epidemiology*, 37(6):635–646, 2016.

Zongfu Mao, Xiao Shen, and Quan Wang. Retrospective study on drug's centralized purchasing system of healthcare institutions in China [In Chinese]. *Chinese Journal of Health Policy*, 7(10):5–10, 2014.

Victoria R. Marone and Adrienne Sabety. When should there be vertical choice in health insurance markets? *American Economic Review*, 112(1):304–342, 2022.

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

Marc J. Melitz. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6):1695–1725, 2003.

Zhaolin Meng, Min Zhu, Yuanyi Cai, Xiaohong Cao, and Huazhang Wu. Effect of a typical systemic hospital reform on inpatient expenditure for rural population: the Sanming model in China. *BMC Health Services Research*, 19(1):231, 2019.

Ministry of Health. Notice on the issuance of the centralized drug procurement in health facilities [In Chinese]. http://www.nhc.gov.cn/yaozs/s3573/201007/ea413230b3714b45b5b724f7bae84884.shtml, July 2010.

Edward R. Morey, Vijaya R. Sharma, and Anne Mills. Willingness to pay and determinants of choice for improved malaria treatment in rural Nepal. *Social Science & Medicine*, 57 (1):155–165, 2003.

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

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

National Health Commission. Circular on fully carrying out the work of promoting the comprehensive reform of public hospitals [In Chinese]. http://www.gov.cn/xinwen/2017-05/01/content_5189918.htm#allContent, April 2017.

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

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

Ariel Pakes. Empirical tools and competition analysis: Past progress and current problems. *International Journal of Industrial Organization*, 53:241–266, 2017.

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

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

David Powell. A new framework for estimation of quantile treatment effects: Nonseparable disturbance in the presence of covariates. Working Paper Series WR-824-1, RAND, 2013.

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

John A. Rizzo. Advertising and competition in the ethical pharmaceutical industry: The case of antihypertensive drugs. *Journal of Law and Economics*, 42(1):89–116, 1999.

John A. Rizzo. Do pharmaceutical sales respond to scientific evidence? *Journal of Economics and Management Strategy*, 11(4):551–594, 2002.

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

Corina C. Ros, Peter P. Groenewegen, and Diana MJ Delnoij. All rights reserved, or can we just copy?: Cost sharing arrangements and characteristics of health care systems. *Health Policy*, 52(1):1–13, 2000.

Li Ruiqin, Yipeng Liu, and Oscar F. Bustinza. FDI, service intensity, and international marketing agility: The case of export quality of Chinese enterprises. *International Marketing Review*, 36(2):213–238, 2019.

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

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

Zu-dong Shi. Institutional change and prediction of drugs group procurement [In Chinese]. *Chinese Journal of Medical Management Sciences*, 4(1):25–27, 2014.

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

Kenneth A Small and Harvey S Rosen. Applied welfare economics with discrete choice models. *Econometrica: Journal of the Econometric Society*, pages 105–130, 1981.

Minjae Song, Sean Nicholson, and Claudio Lucarelli. Mergers with interfirm bundling: a case of pharmaceutical cocktails. *The RAND Journal of Economics*, 48(3):810–834, 2017.

Zheng Song, Kjetil Storesletten, and Fabrizio Zilibotti. Growing like China. *American Economic Review*, 101(1):196–233, 2011.

Michael Spence and Richard Zeckhauser. Insurance, information, and individual action. *American Economic Review*, 61(2):380–387, 1971.

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

State Council's General Office. State council office's notice on establishing and standardizing essential drug procurement in government-sponsored primary health facilities [In Chinese]. http://www.gov.cn/xxgk/pub/govpublic/mrlm/201012/t20101208_63095.html, July 2010.

State Council's General Office. Guiding opinions of the general office of the state council on urban public hospital comprehensive reform pilot [In Chinese]. http://www.gov.cn/zhengce/content/2015-05/17/content_9776.htm, May 2015.

State Council's Healthcare Reform Committee. Opinion on the implementation of the "two invoices" system in the procurement of pharmaceutical products by public medical institutions (trial) [In Chinese]. http://www.nmpa.gov.cn/WS04/CL2196/324173.html, December 2016.

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

Ngo Van Long, Raymond Riezman, and Antoine Soubeyran. Trade, wage gaps, and specific human capital accumulation. *Review of International Economics*, 15(1):75–92, 2007.

Marco Varkevisser and Stéphanie A. van der Geest. Why do patients bypass the nearest hospital?: An empirical analysis for orthopaedic care and neurosurgery in the Netherlands. *The European Journal of Health Economics*, 8(3):287–295, 2003.

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

Yanling Wang. Trade, human capital, and technology spillovers: An industry-level analysis. *Review of International Economics*, 15(2):269–283, 2007.

Thomas G Wollmann. Trucks without bailouts: Equilibrium product characteristics for commercial vehicles. *American Economic Review*, 108(6):1364–1406, 2018.

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

Nobuaki Yamashita. The impact of production fragmentation on skill upgrading: New evidence from Japanese manufacturing. *Journal of the Japanese and International Economies*, 22(4):545–565, 2008.

Miaoqing Yang. Demand for social health insurance: evidence from the Chinese New Rural Cooperative Medical Scheme. *China Economic Review*, 52:126–135, 2018.

Yan Yang, Jia hui Zhou, Xuan Zou, Xin yu Liu, Min xing Chen, Jiang jiang He, Li xuan Gong, and Chun lin Jin. Study on the effectiveness of GPO in reducing drug costs in Shenzhen [In Chinese]. *Chinese Journal of Health Policy*, 13(1):57–61, 2020.

Hongmei Yi, Grant Miller, Linxiu Zhang, Shaoping Li, and Scott Rozelle. Intended and unintended consequences of China's Zero Markup Drug Policy. *Health Affairs*, 34(8): 1391–1398, 2015.

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

Zhongliang Zhou, Yanfang Su, Benjamin Campbell, Zhiying Zhou, Jianmin Gao, Qiang Yu, Jiuhao Chen, and Yishan Pan. The financial impact of the 'Zero-Markup Policy for Essential Drugs' on patients in county hospitals in western rural China. *PLoS ONE*, 10 (3):e0121630, 2015.

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

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

Appendix A

# ADDITIONAL MATERIALS FOR "UNDERSTANDING DEDUCTIBLE AND REIMBURSEMENT MAXIMUM IN A TIERED MEDICAL SYSTEM"

## *A.1 Estimation of Hospital Cost-Sharing Rules*

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

The out-of-pocket cost function in our application is defined by three parameters: a deductible, a reimbursement rate, and a reimbursement maximum. We take the true deductibles (mostly publicly available from each hospital) as given because they correspond very well to our observed data. As far as we learn from local officials, the reimbursement maximum is 100 thousand CNY per patient-year[1] during our study period. Cases with an annual reimbursement of over 100 thousand CNY do not occur in our data. Then, we are left with the reimbursement rate to estimate.

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

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

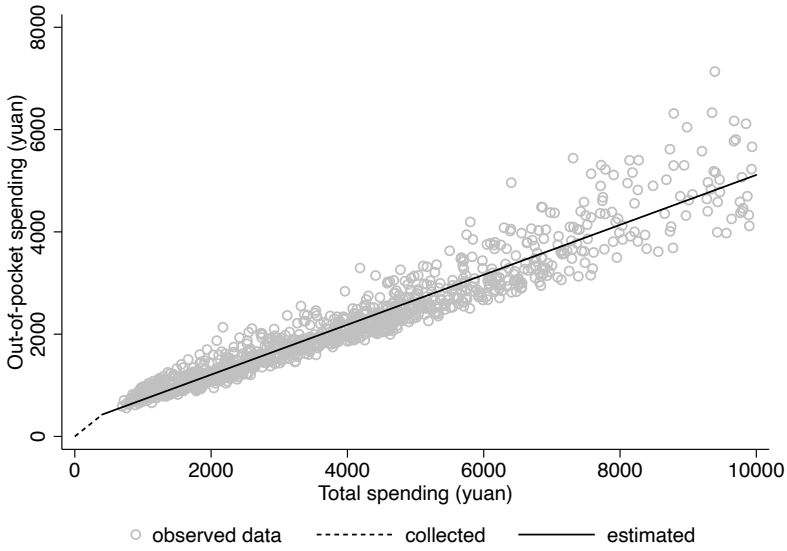



Figure A.1: An Example of Hospital Cost-Sharing Rules Estimation

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

## *A.2 Calculation of Individual Health Risk Predictors*

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

Next, we adjust the risk score by running a linear regression. Before running the regression, we take a natural log of the rescaled risk score[2] to deal with its high skewness. Then, we regress the natural log of actual total spending in thousands of CNY on the log rescaled risk score, its interactions with each of the percentile indicators, the education level indicators, the indicator for a married person, and the hospital dummies, besides the integer age and gender indicators, and the ICD 10 code indicators. Finally, we predict the log spending using this linear model, and the predicted values are our re-scaled risk predictors (range: -2.526 to 4.896). The main reasons for this adjustment are two-fold. On the one hand, the Johns Hopkins ACG system is mainly based on the United States (although it has also been implemented internationally in the United Kingdom, Europe, Singapore, Vietnam, and Australia according to the sales staff), while our data is from rural China, and thus adjusting the risk coefficient may improve the accuracy of the cost prediction in our context, which can then improve our model fit. On the other hand, the risk coefficient from the ACG system does not contain information on a patient's educational attainment, marital status, and the hospital chosen; therefore, by running this additional regression, we can incorporate additional information that we expect to play a role in determining health status.

Table A.1: Spending Distributions by Risk Quartile

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

Notes: This table is based on the estimation sample from 2012 to 2014, the same as the first column of Table 1.2.

The health risk predictors are different from log total spending, although they are highly (positively) correlated. To show how different but correlated they are, we summarize the

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

total spending distributions by quartile of the risk predictor. As shown in Table A.1, a patient in a higher risk quartile does not necessarily have a higher total spending than a patient in a lower risk quartile.

### *A.3 Variation in Hospital Menu Generosity*

Hospital menu generosity is measured by the weighted average of the reimbursement rates in the hospitals available to each community each year. It is calculated for each disease, as the hospitals available for treating each disease can be different. The weights are the proportion of patients going to each hospital from each community. By using the weights, we incorporate the likelihood that an individual would choose a generous hospital when presented with such a menu, as if the individual had been acting like the average individual in the community.

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

Table A.2: Hospital Menu Generosity and Individual Health

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

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

## *A.4 Additional Figures*

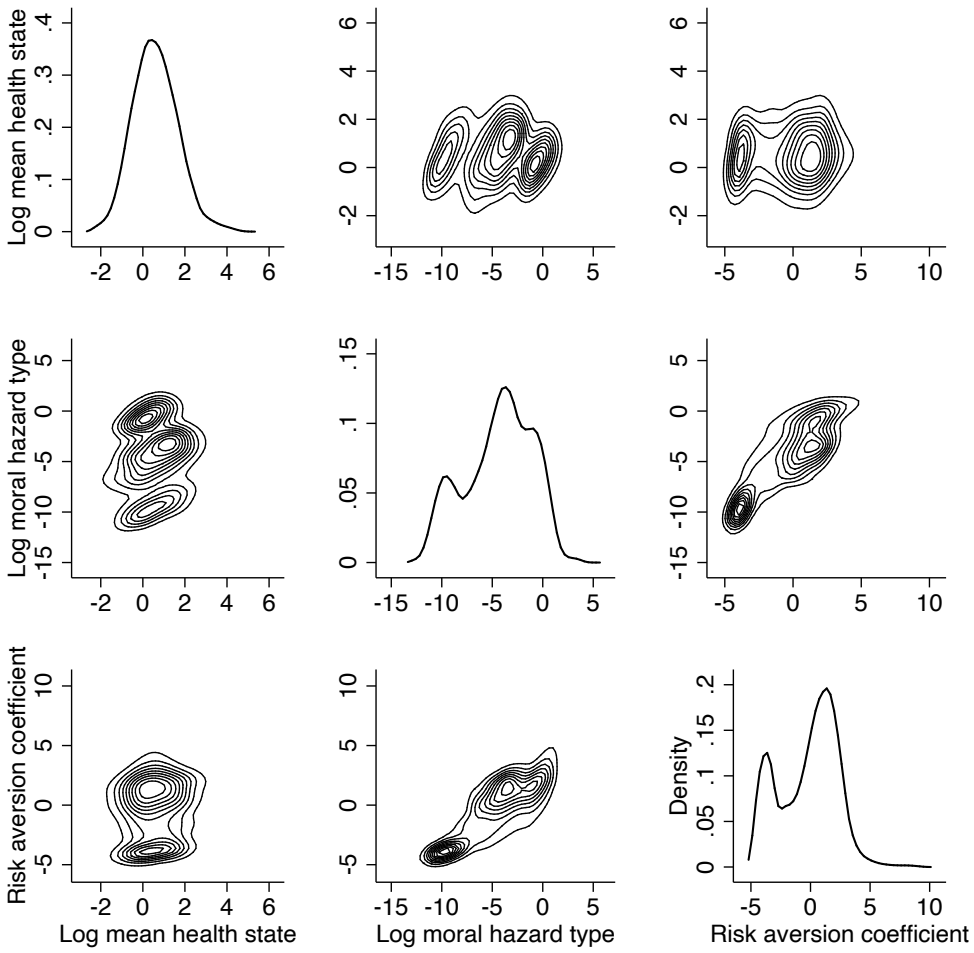



Figure A.2: Joint Distribution of Individual Types

Notes: This figure presents the joint distribution of individual types implied by the estimates in Table 1.3. The diagonals are the one-way distributions of each parameter across individuals (with the vertical axis being the density), while the off-diagonals show bivariate distributions (with both axes being the values).

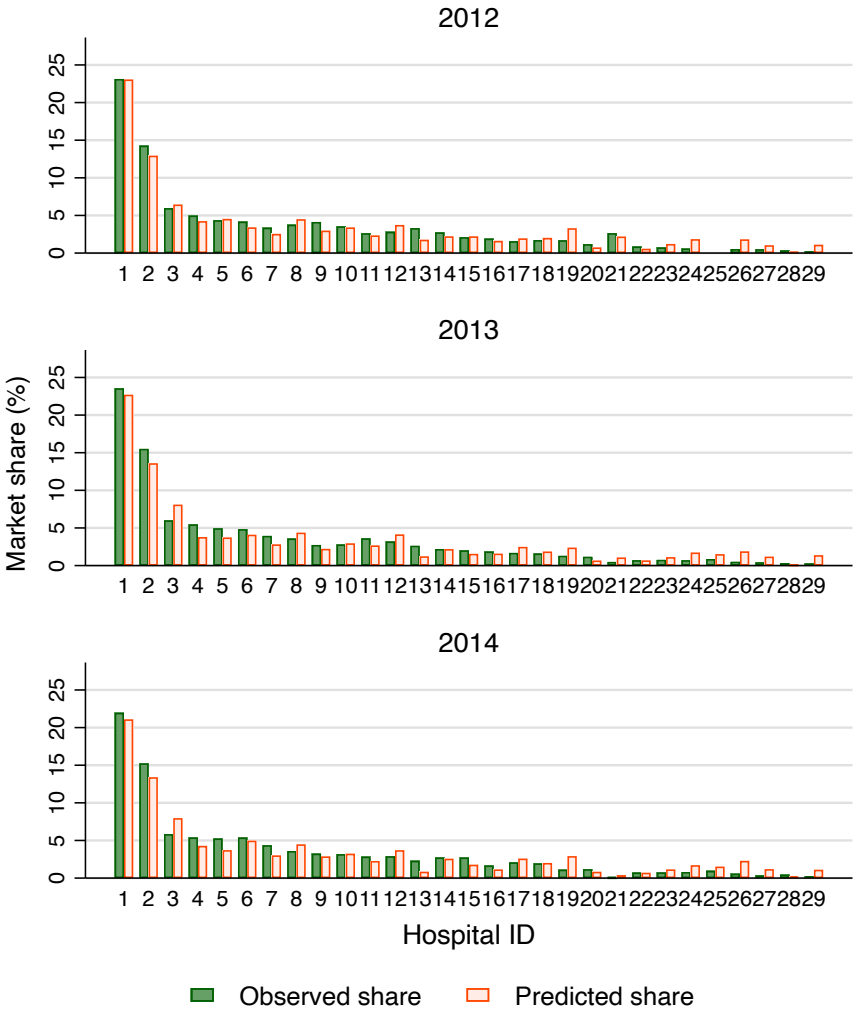
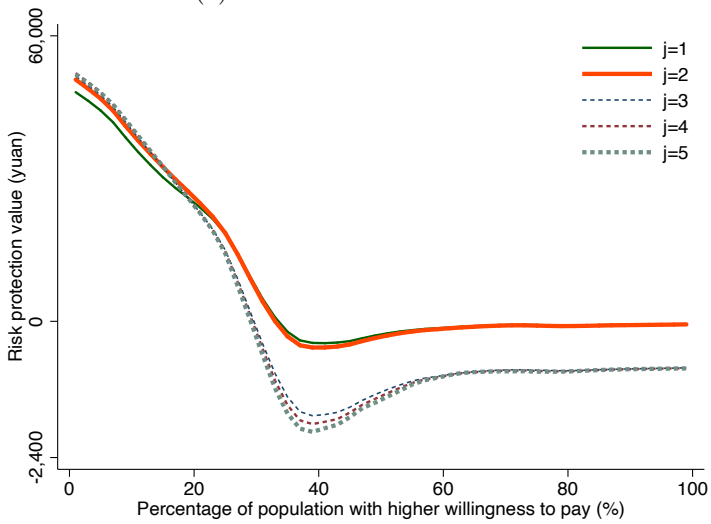



Figure A.3: Model Fit: Hospital Choices

Notes: The figure shows the observed and predicted market shares at the hospital level. An observation is a person-visit in each year. Predicted shares are calculated based on Table 1.3.

(a) Value of Risk Protection




(b) Social Cost of Moral Hazard

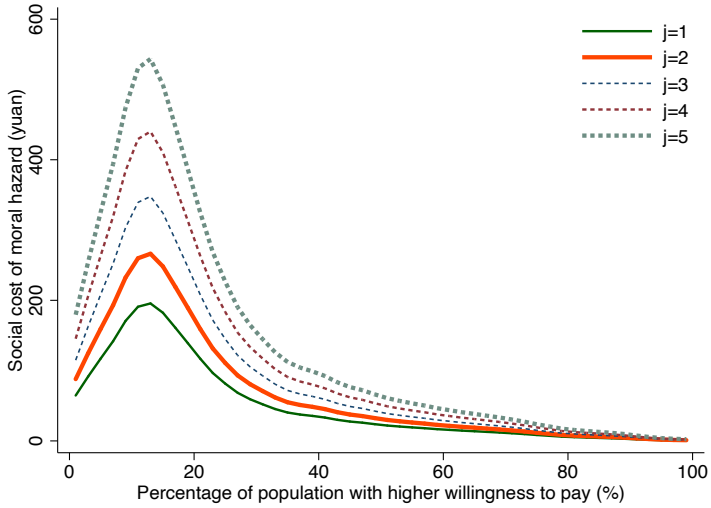



Figure A.4: Decompose Social Surplus

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

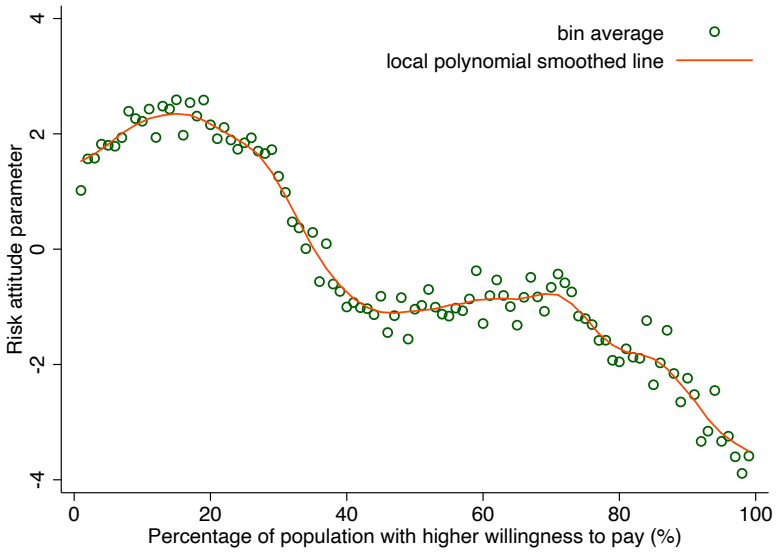



Figure A.5: Risk Attitude Parameter by Willingness to Pay

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

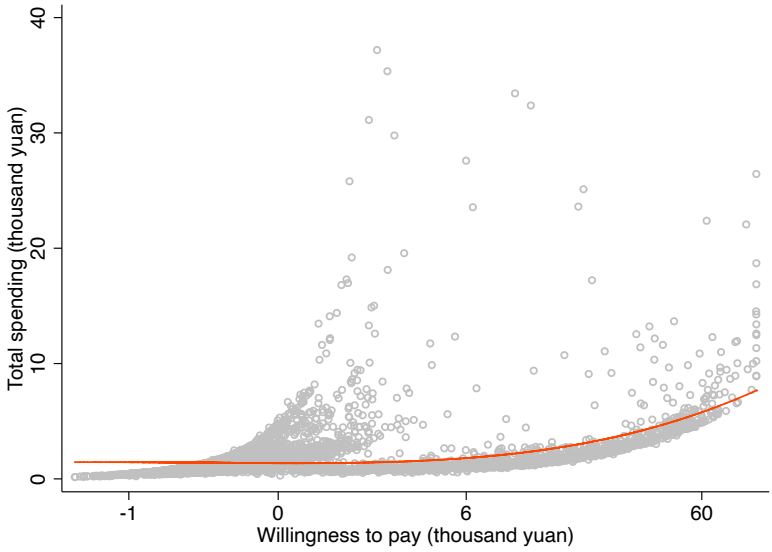
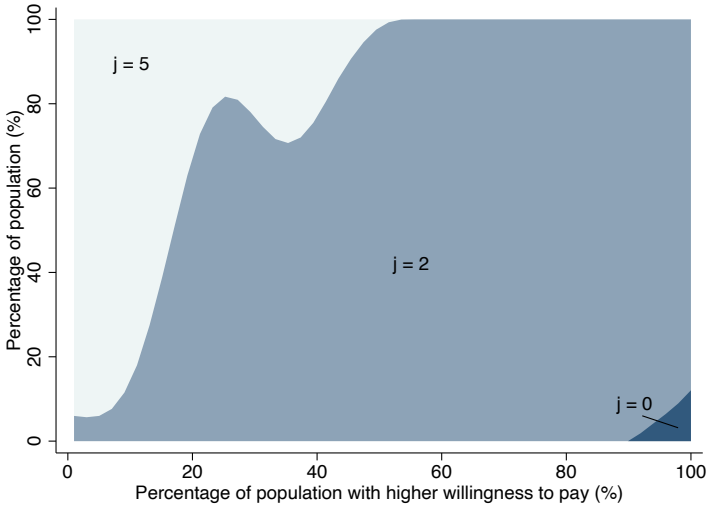



Figure A.6: Willingness to Pay and Spending

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

(a) Privately Efficient Hospitals




(b) Socially Efficient Hospitals

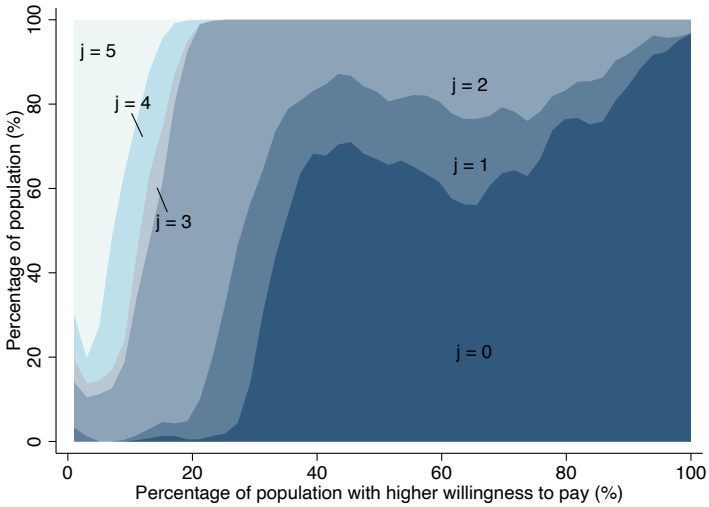



Figure A.7: Efficient Hospital by Willingness to Pay

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

## *A.5 Proofs and Derivations*

### *A.5.1 Derive and Decompose Willingness to Pay*

Consider the expected utility function specified by Equation (1.1):

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

Then, the certainty equivalent is

$$\begin{aligned} CE(j,p,\theta) =& v_\psi^{-1}(U(j,p,\theta)) \\ =& EV(j,\theta) + \hat{y} - p + l_j + v_\psi^{-1}(U(j,p,\theta)) - EV(j,\theta) - \hat{y} + p - l_j \\ =& EV(j,\theta) + \hat{y} - p + l_j - RP(j,p,\theta) \end{aligned}$$

where $EV(j,\theta) + \hat{y} - p + l_j$ is the expected payoff and $RP(j,p,\theta)$ is the risk premium. Note that,

$$\begin{aligned} EV(j,\theta) &= \mathbb{E}[x_j^*(\lambda,\omega,j)|\lambda \sim F] \\ &= \mathbb{E}[x_0^*(\lambda,\omega,j) + \Delta x_j^*(\lambda,\omega,j)|\lambda \sim F] \end{aligned}$$

and

$$RP(j,p,\theta) = EV(j,\theta) + \hat{y} - p + l_j - v_\psi^{-1}(U(j,p,\theta)) \tag{A.1}$$

Next, to find the willingness to pay for hospital $j$ relative to the null hospital 0, we have

$$\begin{aligned} U(j,p,\theta) &= U(0,p_0,\theta) \\ \Rightarrow CE(j,p,\theta) &= CE(0,p_0,\theta) \\ \Rightarrow EV(j,\theta) + \hat{y} - p + l_j - RP(j,p,\theta) &= EV(0,\theta) + \hat{y} - p_0 + l_0 - RP(j,p_0,\theta) \\ \Rightarrow WTP = p - p_0 &= EV(j,\theta) - EV(0,\theta) + RP(0,p_0,\theta) - RP(j,p,\theta) + l_j - l_0 \end{aligned}$$

Then, we can obtain a closed-form expression by assuming CARA, as $\hat{y} - p + l_j$ in Equation (A.1) will cancel out with that from $v_\psi^{-1}(U(j,p,\theta))$. That is, we have $RP(j,p,\theta) = RP(j,\theta)$. Thus, we have

$$WTP = \mathbb{E}[c_0^*(\lambda,\omega,0) - c_0^*(\lambda,\omega,j) + \Delta x_j^*(\lambda,\omega,j)|\lambda \sim F] + RP(0,\theta) - RP(j,\theta) + l_j - l_0.$$

### *A.5.2 Prove Propositions*

To prove the two propositions, consider the model in our empirical setting. Suppose hospitals $j \in J$ are characterized by increasing and continuous OOP cost functions $c_j : \mathbb{R}_+ \to \mathbb{R}_+$, where $c_j(m) \leq m \ \forall m$ and are differentiable almost everywhere with derivative $c_j' \in [0,1]$ when it exits. Patients are characterized by types $\theta = (F, \omega, \psi) \in \Delta^c(\mathbb{R}) \times \mathbb{R}_+^* \times \mathbb{R} =: \Theta$, where $\Delta^c(\mathbb{R})$ is the set of continuous probability measures on the Borel $\sigma$-algebra of $\mathbb{R}$. Given the realized health state $\lambda \in \mathbb{R}_+$, hospital admission cost $p$, average quality premium $l_j$, and initial income $\hat{y}$, suppose consumers value healthcare spending $m \in \mathbb{R}_+$ according to $v_\psi(\hat{y} - p + l_j + b(m; \lambda, \omega) - c_j(m))$ where $v_\psi(x) = \frac{1-\exp(-\psi x)}{\psi}$ when $\psi \neq 0$ and $v_\psi(x) = x$ otherwise, and $b(m; \lambda, \omega) = (m - \lambda) - \frac{1}{2\omega}(m - \lambda)^2$.

Then, we can write the social surplus as

$$SS(j, \theta) = \underbrace{[RP(0, \theta) - RP(j, \theta)]}_{VRC(j,\theta)} - \underbrace{\mathbb{E}_\lambda \left[ \frac{\omega}{2} \left(1 - c_j'(m^*(\lambda, \omega, j))\right)^2 \right]}_{SCMH(j,\theta)}.$$

where $RP(j, \theta) = \psi^{-1} \log \left( \mathbb{E}_\lambda \left[ \exp \left( -\psi(u_j^*(\lambda, \omega) - \overline{u}_j) \right) \right] \right)$. Note that, $u_j^*(\lambda, \omega) = \hat{y} - p + l_j + b(m^*(\lambda, \omega, j); \lambda, \omega) - c_j(m^*(\lambda, \omega, j))$, and $\overline{u}_j = \mathbb{E}_\lambda[u_j^*(\lambda, \omega)]$. We have solved for $m^*(\lambda, \omega, j)$ in Equation (1.14). Note that, $m^*$ falls on a kink when the spending needed to treat a disease is close to the reimbursement maximum, so $c_j'(m^*)$ does not always exist. The indirect benefit from the privately optimal spending is $b(m^*(\lambda, \omega, j); \lambda, \omega) = \frac{\omega}{2} \left(1 - c_j'(m^*(\lambda, \omega, j))^2\right)$. The willingness to pay is $WTP(j, \theta) = \overline{u}_j - \overline{u}_0 + VRC(j, \theta)$.

Let's also recall the cost-sharing structure illustrated by Figure 1.5, where $d$ denotes deductible and $z$ denotes the maximum reimbursable spending. Now we can rewrite the propositions as follows:

**Proposition 1.** A higher deductible (i.e., increasing current deductibles of all hospitals by $\Delta d > 0$) can increase $SS(j^*, \theta)$ for patients with small medical needs $\lambda$ and negative $\psi < 0$.

**Proposition 2.** A higher reimbursable spending (i.e., multiplying current reimbursable amounts of spending of all hospitals by $1 + \Delta z$) may not affect the expected insurer costs $\mathbb{E}_\lambda[k(m^*, j^*)] = \mathbb{E}_\lambda[m^* - c(m^*, j^*) + l_{j^*} - l_0]$ much when moral hazard degree $\omega$ is small.

Proof of Proposition 1:

For the analysis, fix $\theta \in \Theta$. Suppose $j_1 \succsim j_2$ for a patient with low $\lambda$ and $\psi < 0$ before the increase in deductible. Let's denote these two hospitals as $j_1'$ and $j_2'$ after the increase in deductible. Note that, when $\lambda$ is low, a patient does not obtain reimbursements until $\lambda \geq d - \frac{\omega a}{2}$ before the increase, and $\lambda \geq d + \Delta d - \frac{\omega a}{2}$ after the increase.

To prove that social surplus "can" increase, we simply need to give one example. Consider a case when deductibles $d_1 = d_2 = d$ but the reimbursement rates $a_1 < a_2$. Hospitals have the same cap for reimbursement. To make the example even simpler, we can also assume $l_{j_1} = l_{j_2} = l$.

If $d + \Delta d - \frac{\omega a_1}{2} < \lambda \leq d + z_2 - \omega a_2$, we have $j_1' \succsim j_2'$ (the patient maintains her hospital choice), as the expected total spending does not depend on the deductible and thus does not change, leading to same indirect benefits in both hospitals, and the expected OOP cost increases by $\Delta d > 0$ in both hospitals. Thus, $u^*_{j_1'}(\lambda, \omega) - u^*_{j_1}(\lambda, \omega) = u^*_{j_2'}(\lambda, \omega) - u^*_{j_2}(\lambda, \omega)$. Since $u^*_{j_1}(\lambda, \omega) \geq u^*_{j_2}(\lambda, \omega)$, we have $u^*_{j_1'}(\lambda, \omega) \geq u^*_{j_2'}(\lambda, \omega)$.

If $\lambda \leq d - \frac{\omega a_2}{2}$, we can find that $u^*_{j_1'}(\lambda, \omega) = u^*_{j_1}(\lambda, \omega) \geq u^*_{j_2}(\lambda, \omega) = u^*_{j_2'}(\lambda, \omega)$, as the expected spending and OOP cost of the patient do not change before and after the change in deductible. As a result, the patient also maintains her hospital choice.

In both situations above, the social cost of moral hazard (SCMH) term does not change. However, due to a higher deductible, the value of risk change (VRC) increases as the patient's risk attitude parameter $\psi < 0$. The complication comes from the situation when $d - \frac{\omega a_2}{2} < \lambda \leq d + \Delta d - \frac{\omega a_1}{2}$.

To study this scenario, we should note that it is possible that the patient would spend $m_1^* = \lambda$ but $m_2^* = \lambda + \omega a_2$ initially when $d - \frac{\omega a_2}{2} < \lambda \leq d - \frac{\omega a_1}{2}$. If we have $\lambda \leq d + \Delta d - \frac{\omega a_2}{2}$, then after the deductible increase, the patient will spend $\lambda$ in either hospital. Thus, $u^*_{j_1'}(\lambda, \omega) = u^*_{j_1}(\lambda, \omega) = \hat{y} - p + l - \lambda = u^*_{j_2'}(\lambda, \omega) \geq u^*_{j_2}(\lambda, \omega)$. That is, $j_1' \succsim j_2'$. If we have $d + \Delta d - \frac{\omega a_2}{2} < \lambda$, then after the deductible increase, the patient maintains her

spending decision, and

$$u^*_{j'_1}(\lambda,\omega) = u^*_{j_1}(\lambda,\omega) \geq u^*_{j_2}(\lambda,\omega) = \hat{y} - p + l + b(m^*_2;\lambda,\omega) - d - (1-a_2)(m^*_2 - d)$$
$$> \hat{y} - p + l + b(m^*_2;\lambda,\omega) - d - \Delta d - (1-a_2)(m^*_2 - d - \Delta d) = u^*_{j'_2}(\lambda,\omega)$$

suggesting that $j'_1 \succsim j'_2$. As we can see, under this scenario, the patient still maintains her hospital choice. The SCMH term is always 0, as $j'_1 \succsim j'_2$ and $m^*_1 = \lambda$. The VRC however increases as the patient's risk attitude parameter $\psi < 0$.

Since the possibility of $d + z_2 - \omega a_2 < \lambda$ is extremely low, the impact of $\Delta d$ on $SS(j^*, \theta)$ in this scenario is negligible. If $d + \Delta d + z_1 < \lambda$, we still have $j'_1 \succsim j'_2$ and $m^*_1 = m^*_2 = m^*_{1'} = m^*_{2'} = \lambda$, and thus the SCMH term is always 0 and the VRC increases. If $\lambda \leq d + \Delta d + z_1$, the situations are more complicated. We can make $\Pr\{d + z_2 - \omega a_2 < \lambda \leq d + \Delta d + z_1\}$ small enough to not outweigh the above effects.

As a result, we can see that, when $\lambda$ tends to be low and $\psi < 0$, the impact of $\Delta d$ on $SS(j^*, \theta)$ can be positive. If there are enough "mistrustful" patients in the population who experience an increase in the social surplus, then the social welfare will be higher. Q.E.D.

Proof of Proposition 2:

To prove that insurer costs "may" not increase much, we also give one example. Consider a patient with $\psi > 0$. Suppose $j_1 \succsim j_2$ and $a_1 > a_2$. Following the proof of Proposition 1, we also simplify our example by assuming that $d_1 = d_2 = d$ and $l_{j_1} = l_{j_2} = l$. Hospitals again have the same cap for reimbursement, i.e., $z_1 a_1 = z_2 a_2$. As a result, $z_1 < z_2$.

If $\lambda \leq d + z_1 - \omega a_1$, the patient maintains her spending and thus hospital choices. Spending will not be affected by $\Delta z$ under this situation, as we must have $\lambda \leq d + z_1(1 + \Delta z) - \omega a_1$, and hence spending will either be $\lambda$ or $\lambda + \omega a_1$ in hospital $j_1$, which are both independent of $\Delta z$. Similarly, we also must have $\lambda < d + z_2 - \omega a_2 \leq d + z_1(1 + \Delta z) - \omega a_1$, and thus spending will either be $\lambda$ or $\lambda + \omega a_2$ in hospital $j_2$. In summary, if $\lambda$ is within this

range, insurer costs remain unchanged. Let's consider the expected insurer costs below.

$$\begin{aligned}\mathbb{E}_\lambda[\Delta k(m^*, j^*)] = \\ & 0 \times \Pr\{0 \le \lambda \le d + z_1 - \omega a_1\} \\ & + \mathbb{E}_\lambda[\Delta k(m^*, j^*) d + z_1 - \omega a_1 < \lambda] \times \Pr\{d + z_1 - \omega a_1 < \lambda\}\end{aligned}$$

Note that, when $z_1$ is reasonably large and $\omega$ is small, $\Pr\{0 \le \lambda \le d + z_1 - \omega a_1\}$ is reasonably large, and thus $\Pr\{d + z_1 - \omega a_1 < \lambda\}$ becomes reasonably small. Therefore, the magnitude of the change in expected insurer costs should be reasonably small.

Nevertheless, let's still consider $\mathbb{E}_\lambda[\Delta k(m^*, j^*) | d + z_1 - \omega a_1 < \lambda]$. We consider a few sub-regions below.

First, if $d + z_2(1 + \Delta z) < \lambda$, then the patient maintains her spending choices $m_1^* = m_2^* = m_{1'}^* = m_{2'}^* = \lambda$ and thus her hospital choice $j_1' \gtrsim j_2'$. Insurer costs will increase by $a_1 z_1 \Delta z$. As either $z_1$ or $\Delta z$ increases, however, $\Pr\{d + z_2(1 + \Delta z) < \lambda\} \to 0$, although $a_1 z_1 \Delta z$ increases linearly. We can assume that $\Pr\{d + z_2(1 + \Delta z) < \lambda\} = 0$ for a sufficiently large $z_1$ or $\Delta z$.

Second, if $d + z_2 < d + z_1(1 + \Delta z) < \lambda \le d + z_2(1 + \Delta z)$, then $m_1^* = m_2^* = m_{1'}^* = \lambda$, and $m_{2'}^*$ will either be $d + z_2(1 + \Delta z)$ or $\lambda + \omega a_2$, depending on whether $d + z_2(1 + \Delta z) - \omega a_2 < \lambda$ or not. In either case, the patient maintains her hospital choice, as

$$\begin{aligned}u_{j_1'}^*(\lambda, \omega) &= \hat{y} - p + l - \lambda + z_1(1 + \Delta z) a_1 \\ &\ge \underbrace{\hat{y} - p + l - \lambda + (\lambda + \omega a_2 - d) a_2 - \frac{\omega}{2} a_2^2}_{u_{j_2'}^*(\lambda, \omega) \text{ if } \lambda \le d + z_2(1 + \Delta z) - \omega a_2} \\ &> \underbrace{\hat{y} - p + l - \lambda + z_2(1 + \Delta z) a_2 - \frac{1}{2\omega}(d + z_2(1 + \Delta z) - \lambda)^2}_{u_{j_2'}^*(\lambda, \omega) \text{ if } d + z_2(1 + \Delta z) - \omega a_2 < \lambda}.\end{aligned}$$

Given that $d + z_1(1 + \Delta z) < \lambda$, insurer costs will increase by $a_1 z_1 \Delta z$. As $z_1$ increases, we also have $\Pr\{d + z_1(1 + \Delta z) < \lambda \le d + z_2(1 + \Delta z)\} \to 0$; as $\Delta z$ increases, although the range of $\lambda$ becomes wider, we can also let $\Pr\{d + z_1(1 + \Delta z) < \lambda \le d + z_2(1 + \Delta z)\} \to 0$ as this range is moving to the right. Let's assume that $\Pr\{d + z_1(1 + \Delta z) < \lambda \le d + z_2(1 + \Delta z)\} = 0$ as well for a sufficiently large $z_1$ or $\Delta z$.

Third, if $d+z_1-\omega a_1 < \lambda \leq d+z_1(1+\Delta z)-\omega a_1$, for a reasonably large $\Delta z$, we will have $d+z_1 < d+z_1(1+\Delta z)-\omega a_1$ and show that $j_1' \succsim j_2'$, $m_1^* = d+z_1$ or $\lambda$, and $m_{1'}^* = \lambda+\omega a_1$. Clearly, as $\omega$ becomes small, $m_{1'}^* - m_1^*$ is small. If $\Delta z$ is reasonably small, we will instead have $d+z_1(1+\Delta z)-\omega a_1 \leq d+z_1$, but then $\Pr\{d+z_1-\omega a_1 < \lambda \leq d+z_1(1+\Delta z)-\omega a_1\} \to 0$ when $\Delta z \to 0$. Therefore, in this scenario, we either have a small change in insurer costs or a small weight.

Finally, if $d+z_1(1+\Delta z)-\omega a_1 < \lambda \leq d+z_1(1+\Delta z)$, when $\omega$ is small, the weight can also be negligible.

To conclude, we have either $\Pr\{d+z_1-\omega a_1 < \lambda\} \to 0$ or $\mathbb{E}_\lambda[\Delta k(m^*, j^*)|d+z_1-\omega a_1 < \lambda] \to 0$, and thus $E_\lambda[\Delta k(m^*, j^*)] = \mathbb{E}_\lambda[\Delta k(m^*, j^*)|d+z_1-\omega a_1 < \lambda] \times \Pr\{d+z_1-\omega a_1 < \lambda\} \to 0$. Q.E.D.

### *A.5.3 Derive Optimal Spending*

Define $x(m) = b(m) - c(m) = (m-\lambda) - \frac{1}{2\omega}(m-\lambda)^2 - c(m)$, where

$$c(m) = \begin{cases} m & \text{if } 0 \leq m \leq d \\ d+(m-d)(1-a) & \text{if } d < m \leq d+z \\ m-az & \text{if } m > d+z \end{cases}$$

Therefore, we have

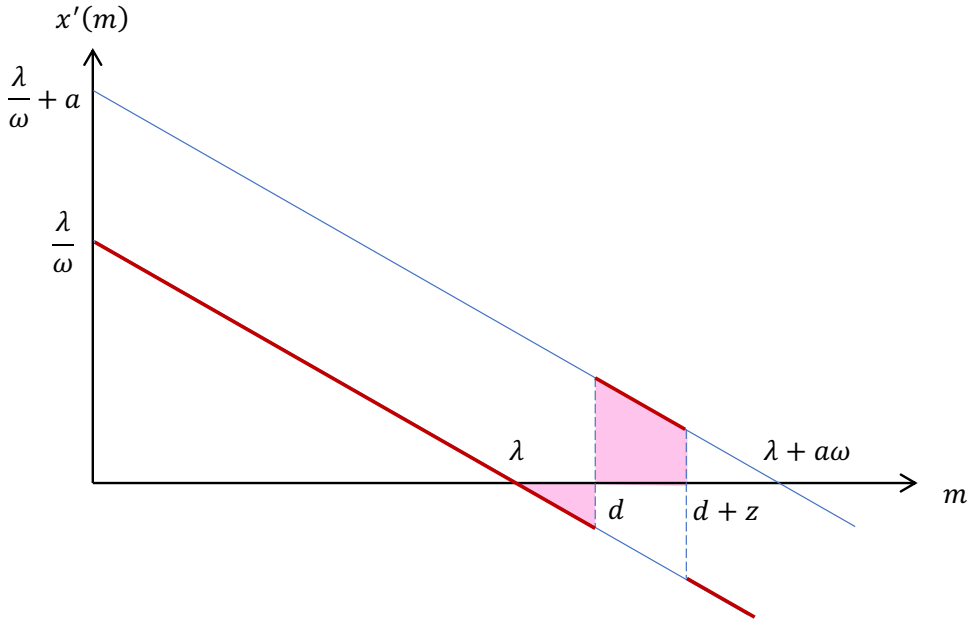
$$x'(m) = \begin{cases} -\frac{1}{\omega}(m-\lambda) & \text{if } 0 \leq m \leq d \\ a-\frac{1}{\omega}(m-\lambda) & \text{if } d < m \leq d+z \\ -\frac{1}{\omega}(m-\lambda) & \text{if } m > d+z \end{cases}$$

To find the optimal $m^*$ that maximizes $x(m)$, consider

$$x(m^*) = \begin{cases} \int_0^{m^*} x'(m)dm & \text{if } 0 \leq m^* \leq d \\ \int_0^d x'(m)dm + \int_d^{m^*} x'(m)dm & \text{if } d < m^* \leq d+z \\ \int_0^d x'(m)dm + \int_d^{d+z} x'(m)dm + \int_{d+z}^{m^*} x'(m)dm & \text{if } m^* > d+z \end{cases}$$

Clearly, we need to discuss the value of $\lambda$.

**Case 1:** $\lambda < d < d+z \leq \lambda + a\omega$ (small $z$, i.e., $z < a\omega$).




For this case, we need to compare the "triangle" and "trapezoid" in the graph. When the "triangle" is larger (smaller $\lambda$), obviously $m^* = \lambda$; when the "trapezoid" is larger (larger $\lambda$), $m^* = d + z$. Let's solve for the condition under which these two areas are equal.

$$A_{\text{triangle}} = \frac{1}{2}(d-\lambda)\frac{1}{\omega}(d-\lambda)$$

$$A_{\text{trapezoid}} = \frac{1}{2}z\left(a - \frac{1}{\omega}(d-\lambda) + a - \frac{1}{\omega}(d+z-\lambda)\right)$$

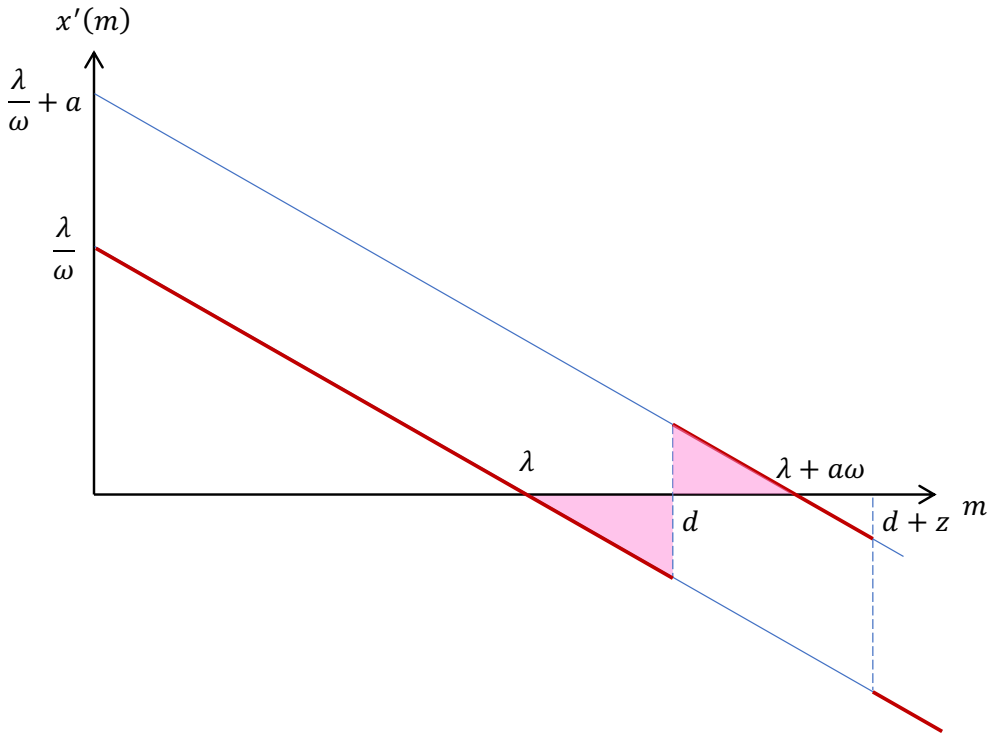
$$A_{\text{triangle}} = A_{\text{trapezoid}} \Rightarrow d - \lambda = \frac{-2z + \sqrt{8a\omega z}}{2} = \sqrt{2a\omega z} - z > 0$$

$$\Rightarrow \lambda_{\text{balance}} = d + z - \sqrt{2a\omega z}$$

Therefore, we have

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

**Case 2:** $\lambda < d \leq \lambda + a\omega < d + z$.




For this case, we compare two triangles. When $\lambda$ is smaller, $m^* = \lambda$; when $\lambda$ is larger, clearly $m^* = \lambda + a\omega$. Let's solve for the balance condition.

$$A_{\text{triangle 1}} = \frac{1}{2}(d-\lambda)\frac{1}{\omega}(d-\lambda)$$

$$A_{\text{triangle 2}} = \frac{1}{2}\left(a - \frac{1}{\omega}(d-\lambda)\right)(\lambda + a\omega - d)$$

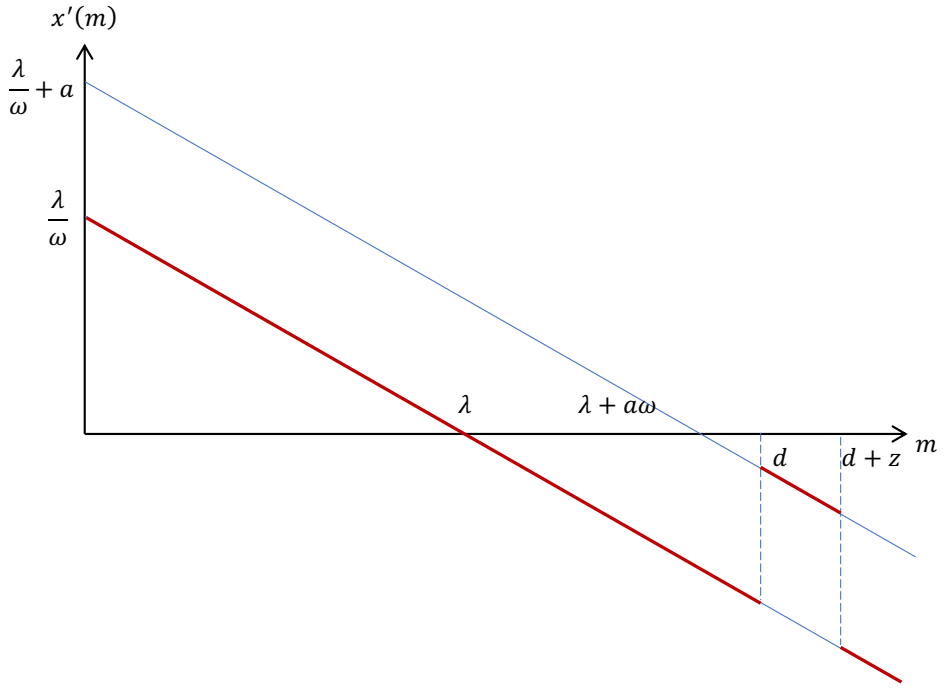
$$A_{\text{triangle 1}} = A_{\text{triangle 2}} \Rightarrow \lambda_{\text{balance}} = d - \frac{a\omega}{2}$$

Therefore, we have

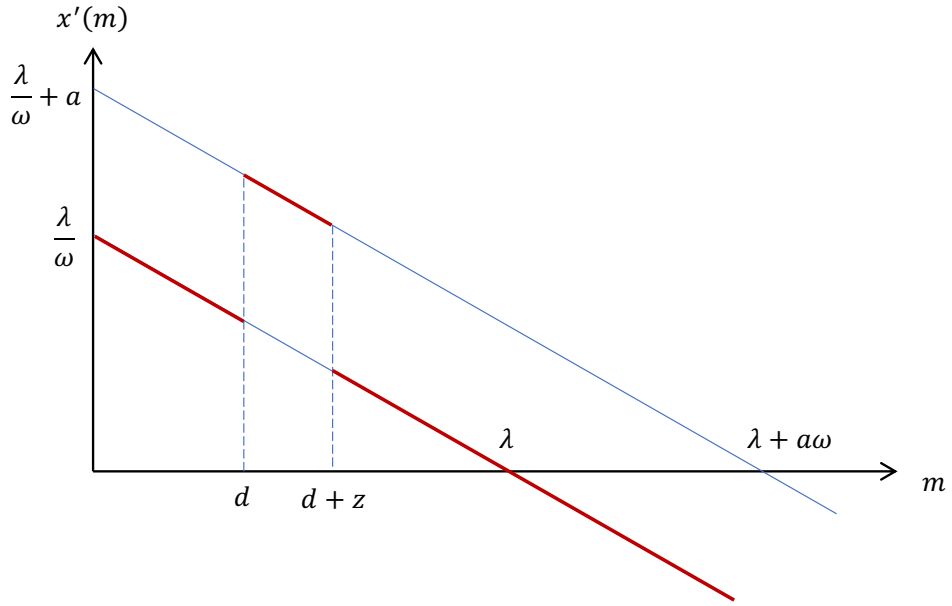
$$m^* = \begin{cases} \lambda & \text{if } d - a\omega \leq \lambda < d - \frac{a\omega}{2} \\ \lambda + a\omega & \text{if } d - \frac{a\omega}{2} \leq \lambda < d \end{cases}$$

**Case 3:** $\lambda < d - a\omega$.

Clearly, as shown in the figure, for this case, $m^* = \lambda$ when $\lambda < d - a\omega$. Any larger $m$ will lead to a portion of negative integral.



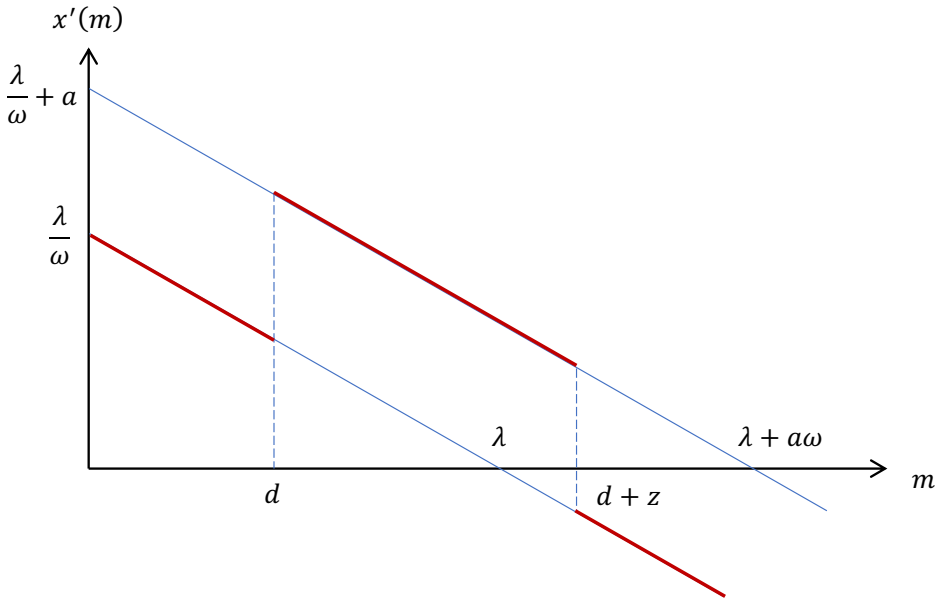

**Case 4:** $d \leq d+z < \lambda$.



Obviously, as shown in the figure, for this case, $m^* = \lambda$ when $\lambda > d + z$.

**Case 5:** $d \leq \lambda \leq d + z < \lambda + a\omega$.

Note that, since $d + z < \lambda + a\omega$, we have $d + z - a\omega < \lambda \leq d + z$. From the figure below, we can see that, $m^* = d + z$.




**Case 6:** $d \leq \lambda < \lambda + a\omega \leq d + z$ (large $z$, i.e., $z \geq a\omega$).

Finally, we consider the situation when $d \leq \lambda \leq d + z - a\omega$. From the figure below, we can conclude that $m^* = \lambda + a\omega$ in this case.

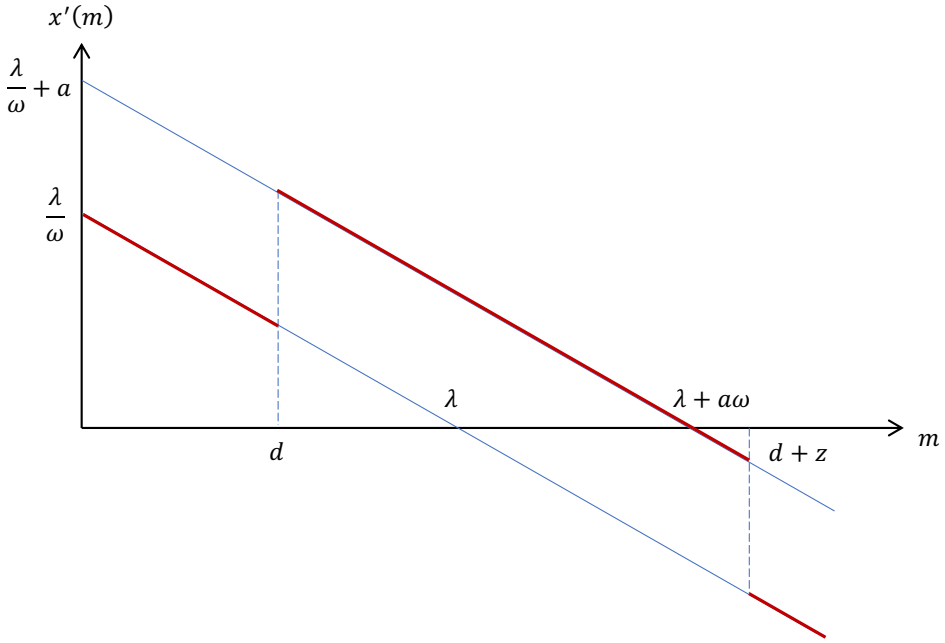
**Summary:**

(1) Clearly, if $0 \leq \lambda \leq d - \frac{a\omega}{2}$, $m^* = \lambda$ (based on cases 2 and 3).

(2) If $z \geq a\omega$ and $d - \frac{a\omega}{2} < \lambda \leq d + z - a\omega$, we have $m^* = \lambda + a\omega$ (based on cases 2 and 6). If $z < a\omega$ and $d - \frac{a\omega}{2} < \lambda < d$, we have $m^* = \lambda + a\omega$ (based on case 2 alone).

(3) If $\lambda > d + z$, we have $m^* = \lambda$ (based on case 4).

(4) Finally, it's important to combine cases 1 and 5. We need to consider the range of $z$ for this. When $z \geq a\omega$, case 1 will not exist. Thus, for large $z$, we only use case 5, which




says that if $d+z-a\omega<\lambda\leq d+z$, we have $m^*=d+z$. Now, let's consider the situation when $z<a\omega$. In case 1, $m^*=d+z$ requires that $\lambda>d+z-\sqrt{2a\omega z}$, while case 5 requires that $\lambda>d+z-a\omega$.

If $\frac{a\omega}{2}<z<a\omega$, then $\sqrt{2a\omega z}>a\omega$. As a result, this coincides with $z\geq a\omega$:

$$(\lambda>d+z-a\omega)\bigcup(\lambda>d+z-\sqrt{2a\omega z})=(\lambda>d+z-a\omega).$$

If $z\leq\frac{a\omega}{2}$, then $\sqrt{2a\omega z}\leq a\omega$, and thus

$$(\lambda>d+z-a\omega)\bigcup(\lambda>d+z-\sqrt{2a\omega z})=(\lambda>d+z-\sqrt{2a\omega z}).$$

Based on this summary, we can obtain Equations (1.14) and (1.15).

Appendix B

# ADDITIONAL MATERIALS FOR "EXAMINING THE ZERO-MARKUP DRUG POLICY IN CHINA: A STRUCTURAL APPROACH"

## *B.1 Standard Error*

To calculate the standard errors of our estimated demand parameters, we need the derivatives of the unobserved drug quality with respect to the parameters, $\partial\xi_t/\partial\theta_2$. According to the implicit function theorem, we have

$$\frac{\partial \xi_t}{\partial \theta_2} = -\left(\frac{\partial s_t}{\partial \xi_t}\right)^{-1} \frac{\partial s_t}{\partial \theta_2}.$$

Let's denote

$$\sigma^r_{j|g}\left(P_t, M_t, \delta_t, \theta_2\right) = \frac{\exp\left\{\frac{\delta_{jt}-\alpha^r P_{jt}+\gamma^r M_{jt}}{1-\lambda}\right\}}{\sum_{j\in g}\exp\left\{\frac{\delta_{jt}-\alpha^r P_{jt}+\gamma^r M_{jt}}{1-\lambda}\right\}},$$

$$\sigma^r_{g}\left(p^R_t, M_t, \delta_t, \theta_2\right) = \frac{\left(\sum_{j\in g}\exp\left\{\frac{\delta_{jt}-\alpha^r P_{jt}+\gamma^r M_{jt}}{1-\lambda}\right\}\right)^{1-\lambda}}{\sum_{g\in G}\left(\sum_{j\in g}\exp\left\{\frac{\delta_{jt}-\alpha^r P_{jt}+\gamma^r M_{jt}}{1-\lambda}\right\}\right)^{1-\lambda}},$$

and

$$\kappa_t = \begin{cases} 1 & \text{before 2015} \\ \phi & \text{between 2015 and 2017Q3} \\ 0 & \text{after 2017Q3} \end{cases}.$$

If we denote $\kappa^1_t = \kappa_t$ and $\kappa^2_t = 1-\kappa_t$, then

$$\frac{\partial s_{jt}}{\partial \xi_{jt}} = \sum_{r=1}^{2}\kappa^r_t\left(\frac{1-\sigma^r_{j|g}}{1-\lambda} + \sigma^r_{j|g}(1-\sigma^r_g)\right)\sigma^r_{j|g}\sigma^r_g,$$

$$\frac{\partial s_{jt}}{\partial \xi_{j't}} = \begin{cases} \sum_{r=1}^{2}\kappa^r_t\left(-\frac{\sigma^r_{j'|g}}{1-\lambda} + \sigma^r_{j'|g}(1-\sigma^r_g)\right)\sigma^r_{j|g}\sigma^r_g & \text{if } j'\in g(j) \\ -\sum_{r=1}^{2}\kappa^r_t\sigma^r_{j'|g(j')}\sigma^r_{g(j')}\sigma^r_{j|g}\sigma^r_g & \text{if } j'\notin g(j) \end{cases},$$

$$\frac{\partial s_{jt}}{\partial \theta_{1,k}} = \sum_{r=1}^{2} \kappa_t^r \left( \frac{1-\sigma_{j|g}^r}{1-\lambda} + \sigma_{j|g}^r (1-\sigma_g^r) \right) \sigma_{j|g}^r \sigma_g^r x_t^k \text{ for } k = 1, ..., 22,$$

$$\begin{aligned}\frac{\partial s_{jt}}{\partial \alpha} =& \kappa_t \left( \frac{1-\sigma_{j|g}^1}{1-\lambda} + \sigma_{j|g}^1 (1-\sigma_g^1) \right) \sigma_{j|g}^1 \sigma_g^1 (-p_{jpt}^R) \\ &+ (1-\kappa_t) \left( \frac{1-\sigma_{j|g}^2}{1-\lambda} + \sigma_{j|g}^2 (1-\sigma_g^2) \right) \sigma_{j|g}^2 \sigma_g^2 (-p_{jpt}^W),\end{aligned}$$

$$\frac{\partial s_{jt}}{\partial \gamma} = \kappa_t \left( \frac{1-\sigma_{j|g}^1}{1-\lambda} + \sigma_{j|g}^1 (1-\sigma_g^1) \right) \sigma_{j|g}^1 \sigma_g^1 m_{jpt},$$

$$\frac{\partial s_{jt}}{\partial \lambda} = \sum_{r=1}^{2} \kappa_t^r \left( \frac{1-\sigma_{j|g}^r}{1-\lambda} + \sigma_{j|g}^r (1-\sigma_g^r) \right) \sigma_{j|g}^r \sigma_g^r \frac{\delta_{jt} - \alpha^r P_{jt} + \gamma^r m_{jt}}{(1-\lambda)^2},$$

and

$$\frac{\partial s_{jt}}{\partial \phi} = \frac{\partial \kappa_t}{\partial \phi} \left( \sigma_{j|g}^1 \sigma_g^1 - \sigma_{j|g}^2 \sigma_g^2 \right)$$

where

$$\frac{\partial \kappa_t}{\partial \phi} = \begin{cases} 1 & 2015 \leq t \leq 2017\text{Q3} \\ 0 & \text{otherwise} \end{cases}.$$

Given $\partial \xi_t / \partial \theta_2$, the standard errors of our parameters are

$$\text{Std. Err.}(\theta_2) = \sqrt{\frac{1}{n} \left( \hat{Q}' \hat{W} \hat{Q} \right)^{-1} \hat{Q}' \hat{W} \hat{\Omega} \hat{W}' \hat{Q} \left( \hat{Q}' \hat{W} \hat{Q} \right)^{-1\prime}}$$

where

$$\hat{Q} = \frac{1}{n} \sum_{j,p,t} h(z_{jpt}^d) \left. \frac{\partial \xi_{jpt}}{\partial \theta_2} \right|_{\hat{\theta}_d},$$

$$\hat{\Omega} = \frac{1}{n} \sum_{j,p,t} \left( h(z_{jpt}^d) \xi_{jpt} - \bar{g} \right) \left( h(z_{jpt}^d) \xi_{jpt} - \bar{g} \right)'$$

in which $\bar{g} = \frac{1}{n} \sum_{j,p,t} h(z_{jpt}^d) \xi_{jpt}$; note that, $\hat{W} = \mathbb{I}$, and $n$ is the full sample size.

### *B.2 Elasticity*

Demand semi-elasticity is given by

$$\begin{aligned}
\frac{\partial \ln s_{jt}\left(\mathbf{p}_t^W\right)}{\partial p_{jt}^W} &= \frac{\partial \ln\left(\kappa_t s_{jt}^1 + (1-\kappa_t)s_{jt}^2\right)}{\partial p_{jt}^W} \\
&= \frac{1}{s_{jt}}\sum_{r=1}^{2}\kappa_t^r \frac{\partial s_{jt}^r}{\partial p_{jt}^W} \\
&= \frac{1}{s_{jt}}\sum_{r=1}^{2}\kappa_t^r \left(\frac{\partial \sigma_{j|g}^r}{\partial p_{jt}^W}\sigma_g^r + \sigma_{j|g}^r \frac{\partial \sigma_g^r}{\partial p_{jt}^W}\right) \\
&= \frac{1}{s_{jt}}\sum_{r=1}^{2}\kappa_t^r \left(\frac{\sigma_{j|g}^r\left(1-\sigma_{j|g}^r\right)}{1-\lambda}\sigma_g^r + \sigma_g^r(1-\sigma_g^r)(\sigma_{j|g}^r)^2\right)\eta^r \\
&= \frac{1}{s_{jt}}\sum_{r=1}^{2}\kappa_t^r \left(\frac{1-\sigma_{j|g}^r}{1-\lambda} + \sigma_{j|g}^r(1-\sigma_g^r)\right)\sigma_{j|g}^r\sigma_g^r\eta^r
\end{aligned}$$

where

$$\eta^1 = \begin{cases} -1.15\alpha + 0.15\gamma & p_{jt}^{Highest} \geq 1.15p_{jt}^W \\ -\gamma & p_{jt}^{Highest} < 1.15p_{jt}^W \end{cases}, \ \eta^2 = \begin{cases} -\alpha & p_{jt}^{Highest} \geq p_{jt}^W \\ 0 & p_{jt}^{Highest} < p_{jt}^W \end{cases}$$

and welfare semi-elasticity is

$$\begin{aligned}
\frac{\partial \ln \Delta_j w_t\left(\mathbf{p}_t^W\right)}{\partial p_{jt}^W} &= \frac{1}{\Delta_j w_t\left(\mathbf{p}_t^W\right)}\frac{\partial w_t\left(p_{jt}^W, \mathbf{p}_{-jt}^W\right)}{\partial p_{jt}^W} \\
&= \frac{1}{\Delta_j w_t\left(\mathbf{p}_t^W\right)}\sum_{r=1}^{2}\kappa_t^r \frac{\partial\left[\delta_{jt} - \alpha^r P_{jt} + \gamma^r m_{jt} - (1-\lambda)\ln\sigma_{j|g}^r - \ln\sigma_g^r\right]}{\partial p_{jt}^W} \\
&= \frac{1}{\Delta_j w_t\left(\mathbf{p}_t^W\right)}\sum_{r=1}^{2}\kappa_t^r\eta^r\left[1 - \left(1-\sigma_{j|g}^r\right) - \sigma_{j|g}^r(1-\sigma_g^r)\right] \\
&= \frac{1}{\Delta_j w_t\left(\mathbf{p}_t^W\right)}\sum_{r=1}^{2}\kappa_t^r\eta^r\sigma_{j|g}^r\sigma_g^r
\end{aligned}$$

Note that, the (wholesale) price elasticity of demand is

$$p_{jt}^W \frac{\partial \ln s_{jt}\left(\mathbf{p}_t^W\right)}{\partial p_{jt}^W},$$

and the cross-price elasticity of demand is

$$
\begin{aligned}
& p_{j't}^{W} \frac{\partial \ln s_{jt}\left(\mathbf{p}_t^{W}\right)}{\partial p_{j't}^{W}} \\
&= \frac{\partial \ln\left(\kappa_t s_{jt}^{1} + (1-\kappa_t) s_{jt}^{2}\right)}{\partial p_{j't}^{W}} \\
&= \frac{p_{j't}^{W}}{s_{jt}} \sum_{r=1}^{2} \kappa_t^r \frac{\partial s_{jt}^r}{\partial p_{j't}^{W}} \\
&= \frac{p_{j't}^{W}}{s_{jt}} \sum_{r=1}^{2} \kappa_t^r \left( \frac{\partial \sigma_{j|g}^r}{\partial p_{j't}^{W}} \sigma_g^r + \sigma_{j|g}^r \frac{\partial \sigma_g^r}{\partial p_{j't}^{W}} \right) \\
&= \begin{cases} \frac{p_{j't}^{W}}{s_{jt}} \sum_{r=1}^{2} \kappa_t^r \left( -\frac{\sigma_{j'|g}^r}{1-\lambda} + \sigma_{j'|g}^r (1-\sigma_g^r) \right) \sigma_{j|g}^r \sigma_g^r \eta^r & \text{if } j' \in g(j) \\ -\frac{p_{j't}^{W}}{s_{jt}} \sum_{r=1}^{2} \kappa_t^r \sigma_{j'|g(j')}^r \sigma_{g(j')}^r \sigma_{j|g}^r \sigma_g^r \eta^r & \text{if } j' \notin g(j) \end{cases}
\end{aligned}
$$

### *B.3 Dealing with Unobserved Package Sizes*

The highest price regulation varies by package size (i.e, the number of units in each package, such as 12 tablets versus 14 tablets per package), which are not observed from our data. To deal with this issue, for each aggregated drug product (a molecule-firm pair), we calculate the price caps based on several package sizes, and use the highest per unit price cap to calculate the highest possible retail price (upper bound), and the lowest to calculate the lower bound. Our calculations suggest that they are very close to each other, compared to the magnitude of the prices themselves.

Although the lower bound and upper bound are fairly close to each other, it creates a small problem in defining a "binding constraint" because in some cases $1.15p^W$ falls in the middle of the range created by the lower bound and upper bound of $p^{Highest}$. Table B.1 illustrates this potential issue by calculating the portion of observations that encounter a binding highest price (i.e., $p^{Highest} \leq 1.15p^W$) between 2012 and 2014. The binding rates calculated based on the lower bounds and upper bounds are different but in general they are fairly close to each other. Our main results will be based on the lower bound (highest per unit price cap and thus highest retail price or markup). The analysis based on the upper

Table B.1: Proportion of Retail Prices Equal to the Highest Prices (2012–2014)

| | Lower bound (%) | | | Upper bound (%) | | |
|---|---|---|---|---|---|---|
| | 2012 | 2013 | 2014 | 2012 | 2013 | 2014 |
| **Molecule:** | | | | | | |
| Ezetimibe | 82.61 | 98.78 | 92.77 | 82.61 | 98.78 | 92.77 |
| Atorvastatin | 85.23 | 76.33 | 75.31 | 86.91 | 81.00 | 78.75 |
| Inositol Nicotinate | 66.67 | 66.67 | 66.67 | 66.67 | 66.67 | 66.67 |
| Probucol | 73.81 | 70.68 | 67.16 | 73.81 | 70.68 | 67.16 |
| Fluvastatin | 69.83 | 73.41 | 72.02 | 80.45 | 80.35 | 79.17 |
| Fenofibrate | 56.83 | 53.19 | 51.84 | 60.43 | 57.60 | 55.81 |
| Pravastatin | 66.06 | 59.57 | 57.30 | 82.48 | 80.14 | 73.03 |
| Simvastatin | 56.33 | 55.20 | 60.51 | 58.77 | 58.28 | 64.23 |
| Bezafibrate | 45.55 | 46.11 | 40.31 | 63.35 | 62.69 | 54.08 |
| Acipimox | 61.38 | 49.60 | 45.38 | 66.90 | 59.20 | 58.46 |
| Lovastatin | 48.91 | 49.58 | 53.23 | 77.37 | 75.63 | 72.58 |
| Rosuvastatin | 56.82 | 67.95 | 56.59 | 59.09 | 81.79 | 72.20 |
| Gemfibrozil | 50.57 | 38.27 | 37.93 | 75.86 | 67.90 | 65.52 |
| Pitavastatin | 65.17 | 56.44 | 55.65 | 96.63 | 87.13 | 86.96 |
| Jiaogulan | 49.40 | 58.90 | 58.70 | 77.11 | 84.93 | 71.74 |
| Zhibituo | 50.00 | 35.71 | 50.00 | 53.85 | 53.57 | 50.00 |
| Xuezhikang | 48.96 | 52.17 | 48.94 | 84.38 | 83.70 | 89.36 |
| **Overall** | 60.66 | 59.86 | 59.35 | 70.20 | 70.62 | 68.81 |

Notes: (1) These rates are conditional frequencies calculated using the sub-sample with highest price regulations. (2) The highest price regulations can vary by drug form and size, and in our data only about 0.1% of the cases aggregate the forms and/or sizes without highest prices and the ones with highest prices, and we assume that those standardized drugs are under highest price regulations. (3) The highest price regulation also varies by package size, which is not observed from the data, and so we use the highest per unit price cap to calculate the lower bounds of the binding rates and the lowest per unit price cap to calculate the upper bounds of the binding rates.

bound is similar. One more takeaway from Table B.1 is that the binding rates are fairly high and vary a lot by molecule, which suggests that there should be a good variation in retail price and hospital markup, although the retail price and markup are likely correlated.

### *B.4 Market Structure at the National Level*

Table B.2 illustrates the structure of the lipid-lowering prescription drug market in China between 2012 and 2019, where we regard the whole country as a market and weight each province by the number of actual hospitals versus that of sampling hospitals. The market shares are relative to the total potential sales of lipid-lowering drugs and we assume that the treatment rate is 39 percent, based on Gao et al. (2013), for every year (so outside goods account for 69 percent of the market shares). We leave the question of the consequences of having time-varying (and later market-varying) treatment rates for future research. We also do not assume any variation in how each drug is used in each hospital.

As we can see from Table B.2, the average wholesale price of each standard unit (weighted based on a fixed "basket" of firms in 2012) is declining over time. We include the firms that appear in each year at least once in the "basket" and weight the simple average wholesale price across different quarters and provinces for each firm by its corresponding yearly market share. While the decline was modest before 2016, it became quite significant after 2017 as the ZMDP kicked in and the CDP was enhanced. Entry was observed in Atorvastatin and Rosuvastatin, the top two most popular molecules in recent years.

Table B.2: Lipid-Lowering Prescription Drug Market Structure in China (2012–2019)

| | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|
| *Wholesale price (CNY)/std unit:* | 3.25 | 3.19 | 3.16 | 3.13 | 3.08 | 2.96 | 2.85 | 2.63 |
| *Market shares of top brands:* | | | | | | | | |
| Lipitor (Pfizer) | 11.17% | 11.99% | 12.47% | 11.79% | 11.95% | 12.77% | 13.40% | 13.21% |
| Zocor (MSD) | 5.73% | 4.31% | 3.62% | 3.00% | 2.70% | 2.16% | 1.46% | 0.90% |
| Crestor (AstraZeneca) | 4.00% | 4.97% | 6.43% | 7.03% | 7.36% | 8.14% | 8.07% | 5.90% |
| A Le (Jialin) | 2.56% | 2.81% | 3.16% | 3.72% | 4.02% | 3.99% | 3.96% | 5.39% |
| Lescol (Novartis) | 2.30% | 2.21% | 1.96% | 2.04% | 1.77% | 1.12% | 1.03% | 0.90% |
| Jing Bi Shu Xin (Jingxin) | 1.36% | 0.94% | 0.69% | 0.46% | 0.44% | 0.27% | 0.21% | 0.19% |
| You Jia (Topfond/Sinopharm) | 1.08% | 1.20% | 1.11% | 1.10% | 0.81% | 0.68% | 0.48% | 0.30% |
| *Market shares of selected molecules:* | | | | | | | | |
| Atorvastatin | 14.82% | 16.09% | 16.86% | 16.82% | 17.39% | 18.13% | 18.85% | 20.22% |
| Simvastatin | 11.06% | 8.86% | 7.21% | 6.29% | 5.29% | 4.47% | 3.24% | 2.44% |
| Rosuvastatin | 5.59% | 7.35% | 9.01% | 9.95% | 10.77% | 11.85% | 12.44% | 11.46% |
| Fluvastatin | 2.66% | 2.51% | 2.18% | 2.23% | 2.01% | 1.24% | 1.10% | 0.96% |
| Xuezhikang | 0.53% | 0.46% | 0.35% | 0.37% | 0.34% | 0.29% | 0.24% | 0.24% |
| Acipimox | 0.32% | 0.28% | 0.18% | 0.18% | 0.17% | 0.18% | 0.17% | 0.25% |
| Probucol | 0.22% | 0.18% | 0.17% | 0.18% | 0.20% | 0.19% | 0.16% | 0.17% |
| *# firms of selected molecules:* | | | | | | | | |
| Atorvastatin | 4 | 6* | 6 | 6 | 6 | 6 | 6 | 7 |
| Simvastatin | 42 | 39 | 42 | 39 | 35 | 38 | 37 | 36 |
| Rosuvastatin | 5 | 5 | 5 | 7 | 7 | 7 | 7 | 7 |
| Fluvastatin | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Xuezhikang | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Acipimox | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 3 |
| Probucol | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

Notes: (1) Wholesale price per standard unit is calculated using the "basket" of firms that appear in every year, weighted by their market shares in 2012. (2) Market shares are relative to the total potential sales of lipid-lowering drugs (the treatment rate is 39%). (3) The reported mean values of prices and market shares are weighted according to the hospital distribution in China versus in the sample. * In 2003, Topfond merged into Sinopharm, so we merge them into one firm here.

Appendix C

# ADDITIONAL MATERIALS FOR "TRADE LIBERALIZATION AND SKILL UPGRADING: EVIDENCE ON THE IMPACT OF APTA ON CHINESE MANUFACTURERS"

## *C.1 Additional Figures*

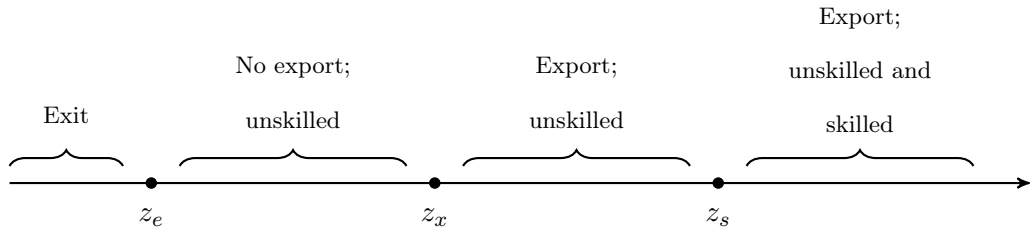



Figure C.1: Productivity Cutoffs under the Benchmark Model

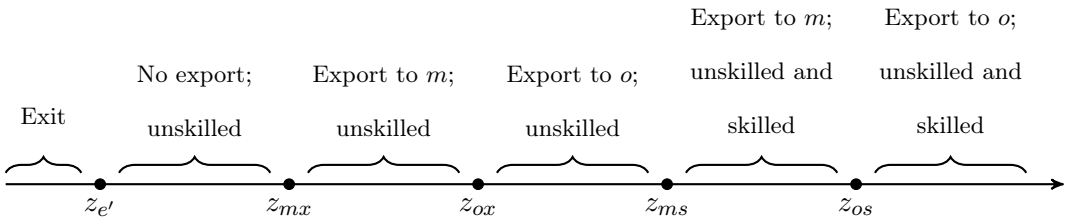



Figure C.2: Productivity Cutoffs under the Extended Model

## *C.2 More Details of the Theoretical Model*

### *C.2.1 Benchmark Model*

*Total Cost Function*

Per-period fixed export cost $f_x$ and iceberg trade cost $\tau$ are required for exporters, thus the total costs for firms that export under the unskilled and skilled sector are respectively

$$TC_u(z) = f_u + f_x + \frac{w_l}{z} y_u^d(z) + \tau \frac{w_l}{z} y_u^x(z)$$

$$TC_s(z) = f_s + f_x + \frac{w_h^\beta w_l^{1-\beta}}{\gamma z} y_s^d(z) + \tau \frac{w_h^\beta w_l^{1-\beta}}{\gamma z} y_s^x(z)$$

*Average Profit and Revenue*

The average profit $\tilde{\pi} = \tilde{\pi}_u^d + n_x \tilde{\pi}_x^d + n_s \tilde{\pi}_s^x$, where

$$\tilde{\pi}_u^d = \frac{1}{1 - G(z_e)} \int_{z_e}^{z_x} z^{\theta-1} g(z) dz$$
$$\tilde{\pi}_u^x = \frac{1}{1 - G(z_e)} \int_{z_x}^{z_s} z^{\theta-1} g(z) dz$$
$$\tilde{\pi}_s^x = \frac{1}{1 - G(z_e)} \int_{z_s}^{\infty} z^{\theta-1} g(z) dz$$

The average profit also can be describe in this way: $\tilde{\pi} = \frac{\tilde{r}}{\theta} - f_u - n_x f_x - n_s(\phi - 1) f_u$.

The average revenues of surviving firms is

$$\begin{aligned}
\tilde{r} &= \int_{z_e}^{z_x} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz + \int_{z_x}^{z_s} r_u^x(z)\frac{g(z)}{1-G(z_e)}dz + \int_{z_s}^{\infty} r_s^x(z)\frac{g(z)}{1-G(z_e)}dz \\
&= \int_{z_e}^{z_x} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz + (1+\tau^{1-\theta})\int_{z_x}^{z_s} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&\quad + \lambda^{\theta-1}(1+\tau^{1-\theta})\int_{z_s}^{\infty} r_d^x(z)\frac{g(z)}{1-G(z_e)}dz \\
&= \int_{z_e}^{z_x} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz + \int_{z_x}^{z_s} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz + \int_{z_s}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&\quad + \tau^{1-\theta}\int_{z_x}^{z_s} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz + \tau^{1-\theta}\int_{z_s}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&\quad + \left[\lambda^{\theta-1}(1+\tau^{1-\theta})\left(\frac{\rho_2}{\rho_1}\right)^{\theta-\rho} - 1 - \tau^{1-\theta}\right]\int_{z_s}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&= \int_{z_e}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&\quad + \tau^{1-\theta}\int_{z_x}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&\quad + (1+\tau^{1-\theta})\left(\lambda^{\theta-1}\left(\frac{\rho_2}{\rho_1}\right)^{\theta-\rho} - 1\right)\int_{z_s}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz
\end{aligned}$$

Using the zero profit condition, we get $r_u^d(z) = \theta f_u\left(\frac{z}{z_e}\right)^{\theta-1}$. We define $\tilde{z}_j^{\theta-1} = \int_{z_j}^{\infty} z_j \frac{g(z)}{1-G(z_j)}dz$ where $j \in (e,x,s)$, and we derive that both $z_x$ and $z_s$ can be a function of $z_e$, so

$$\tilde{r} = \theta f_u\left(\frac{\tilde{z}_e}{z_e}\right)^{\theta-1} + n_x\theta f_x\left(\frac{\tilde{z}_x}{z_x}\right)^{\theta-1} + n_s\theta f_u(\phi-1)\left(\frac{\tilde{z}_s}{z_s}\right)^{\theta-1}$$

Since

$$\begin{aligned}
\left(\frac{\tilde{z}_j}{z_j}\right)^{\theta-1} &= \int_{z_j}^{\infty}\left(\frac{z}{z_j}\right)^{\theta-1}\frac{g(z)}{1-G(z_j)}dz \\
&= z_j^{\kappa+1-\theta}\frac{\kappa z_j^{\theta-1-\kappa}}{\kappa-(\theta-1)} \\
&= \frac{\kappa}{\kappa-(\theta-1)}
\end{aligned}$$

We get $\tilde{r} = \frac{\theta\kappa}{\kappa-(\theta-1)}\left[f_u + n_x f_x + n_s f_u(\phi-1)\right]$. After substituting $\tilde{r}$ into the free entry condition, we obtain

$$f_e = \frac{z_e^{-\kappa}}{\delta}\frac{\theta-1}{\kappa-(\theta-1)}\left[f_u + n_x f_x + n_s f_u(\phi-1)\right]$$

Substituting $n_x$ and $n_s$,

$$\begin{aligned}
z_e =& \left(\frac{1}{f_e\delta}\frac{\theta-1}{\kappa-(\theta-1)}\right)^{\frac{1}{\kappa}}\left[f_u + n_x f_x + n_s f_u(\phi-1)\right]^{\frac{1}{\kappa}} \\
=& \left(\frac{1}{f_e\delta}\frac{\theta-1}{\kappa-(\theta-1)}\right)^{\frac{1}{\kappa}} \times \\
& \left[f_u + f_x\tau^{-\kappa}\left(\frac{f_x}{f_u}\right)^{\frac{-\kappa}{\theta-1}} + f_u(\phi-1)\left(\frac{\phi-1}{(1+\tau^{1-\theta})(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right)^{\frac{-\kappa}{\theta-1}}\right]^{\frac{1}{\kappa}} \\
=& \Lambda\left[1 + \frac{f_x}{f_u}\left(\frac{f_x}{f_u}\right)^{\frac{-\kappa}{\theta-1}}\tau^{-\kappa} + (\phi-1)\left(\frac{\phi-1}{(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right)^{\frac{-\kappa}{\theta-1}}(1+\tau^{1-\theta})^{\frac{\kappa}{\theta-1}}\right]^{\frac{1}{\kappa}} \\
=& \Lambda\Phi
\end{aligned}$$

where $\Lambda \equiv \left(\frac{f_u}{f_e\delta}\frac{\theta-1}{\kappa-(\theta-1)}\right)^{\frac{1}{\kappa}}$ and

$$\Phi = \left[1 + \frac{f_x}{f_u}\left(\frac{f_x}{f_u}\right)^{\frac{-\kappa}{\theta-1}}\tau^{-\kappa} + (\phi-1)\left(\frac{\phi-1}{(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right)^{\frac{-\kappa}{\theta-1}}(1+\tau^{1-\theta})^{\frac{\kappa}{\theta-1}}\right]^{\frac{1}{\kappa}}.$$

*Trade Liberalization*

**Skill Premium**:

$$\begin{aligned}
\frac{R_u}{R_s} =& \frac{\int_{z_e}^{z_x} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz + \int_{z_x}^{z_s} r_u^x(z)\frac{g(z)}{1-G(z_e)}dz}{\int_{z_s}^{\infty} r_s(z)\frac{g(z)}{1-G(z_e)}dz} \\
=& \frac{1}{\lambda^{\theta-1}(1+\tau^{\theta-1})}
\end{aligned}$$

As $\tau$ has a positive effect on $\lambda$, $\frac{\partial R_u}{\partial R_s} > 0$. The reduction in trade costs increase the relative revenues of firms producing skilled goods, so the demand for skilled labor increases. This leads to a higher equilibrium skill premium.

Next, a reduction in trade costs increases the share of firms producing skilled good, $\frac{\partial n_s}{\partial \tau} < 0$.

*Proof:*

$\tau$ has a direct negative effect on $n_s$, but an indirect positive effect through $\lambda$ since the reduction in tariffs increases the skill premium reducing the cost advantage of firms producing

skilled goods. Nevertheless, the direct effect must dominate. To prove this: we suppose it was not the case, $n_s$ falls as trade costs fall. However, we derive that $\frac{\partial R_s}{\partial R_u} > 0$, which is a contraction.

**Average Profit**:

Given $\tilde{\pi} = \frac{\theta-1}{\kappa-(\theta-1)}\left[f_u + n_x f_x + n_s f_u(\phi-1)\right] = \frac{\theta-1}{\kappa-(\theta-1)}\Phi^{\kappa}$, $\frac{\partial\tilde{\pi}}{\partial\tau} = \frac{\theta-1}{\kappa-(\theta-1)}\frac{\partial\Phi^{\kappa}}{\partial\tau}$.

$$\begin{aligned}\frac{\partial\Phi^{\kappa}}{\partial\tau} =& -\kappa\frac{f_x}{f_u}\left(\frac{f_x}{f_u}\right)^{\frac{-\kappa-1}{\theta-1}}\tau^{-\kappa} \\ &-(\theta-1)\frac{\kappa}{\theta-1}(\phi-1)\left(\frac{\phi-1}{(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right)^{\frac{-\kappa}{\theta-1}}(1+\tau^{1-\theta})^{\frac{\kappa-\theta+1}{\theta-1}}\tau^{-\theta} < 0\end{aligned}$$

Since $\theta > 1$ and $\kappa > (\theta-1)$, we get $\frac{\partial\tilde{\pi}}{\partial\tau} < 0$.

**Exit productivity cutoff**:

Since $z_e = \Lambda\Phi$ and $\frac{\partial\Phi^{\kappa}}{\partial\tau} < 0$, we get $\frac{\partial z_e}{\partial\tau} < 0$.

**Export productivity cutoff**:

Since $z_x = \tau\Lambda\Phi$,

$$\frac{\partial z_x}{\partial\tau} = \left(\frac{f_x}{f_u}\right)^{\frac{1}{\theta-1}}\Lambda\Phi + \left(\frac{f_x}{f_u}\right)^{\frac{1}{\theta-1}}\Lambda\frac{\partial\Phi\tau}{\partial\tau}$$

Given $\Phi\tau = \left[\tau^{\kappa} + \frac{f_x}{f_u}\left(\frac{f_x}{f_u}\right)^{\frac{-\kappa}{\theta-1}} + \tau^{\kappa}(\phi-1)\left(\frac{\phi-1}{(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right)^{\frac{-\kappa}{\theta-1}}(1+\tau^{1-\theta})^{\frac{\kappa}{\theta-1}}\right]^{\frac{1}{\kappa}}$

$$\frac{\partial\Phi\tau}{\partial\tau} = (\Phi\tau)^{1/\kappa-1}\left[\tau^{\kappa-1} + \Xi(\phi-1)\left(\frac{\phi-1}{(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right)^{\frac{-\kappa}{\theta-1}}\right]$$

where $\Xi \equiv \tau^{\kappa-1}(1+\tau^{1-\theta})^{\frac{\kappa}{\theta-1}} - \tau^{\kappa-\theta}(1+\tau^{1-\theta})^{\frac{\kappa}{\theta-1}-1} = \tau^{\kappa-1}(1+\tau^{1-\theta})^{\frac{\kappa}{\theta-1}}\left(1-\frac{\tau^{1-\theta}}{(1+\tau^{1-\theta})}\right)$

As $\frac{\tau^{1-\theta}}{(1+\tau^{1-\theta})} < 1$, $\Xi > 0$. Then, $\frac{\partial\Phi\tau}{\partial\tau} > 0$ and all other terms are positive, thus $\frac{\partial z_x}{\partial\tau} > 0$.

**Skill upgrading productivity cutoff**:

Since $z_s = z_e\left[\frac{\phi-1}{(1+\tau^{1-\theta})(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right]^{\frac{1}{\theta-1}} = \Lambda\Phi\left[\frac{\phi-1}{(1+\tau^{1-\theta})(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right]^{\frac{1}{\theta-1}}$,

$$\begin{aligned}\Phi^{\kappa}(1+\tau^{1-\theta})^{\frac{-\kappa}{\theta-1}} =&(1+\tau^{1-\theta})^{\frac{-\kappa}{\theta-1}} \\ &+(1+\tau^{1-\theta})^{\frac{-\kappa}{\theta-1}}\left(\frac{f_x}{f_u}\right)^{1-\frac{\kappa}{\theta-1}}\tau^{\kappa} \\ &+(\phi-1)\left(\frac{\phi-1}{(\lambda^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-1)}\right)^{\frac{-\kappa}{\theta-1}}\end{aligned}$$

$$\begin{aligned}\frac{\partial\Phi^{\kappa}(1+\tau^{1-\theta})^{\frac{-\kappa}{\theta-1}}}{\partial\tau} =&\kappa(1+\tau^{1-\theta})^{\frac{-\kappa}{\theta-1}}\times \\ &\left[(1+\tau^{1-\theta})^{-1}\tau^{-\theta}\left(1+\left(\frac{f_x}{f_u}\right)^{1-\frac{\kappa}{\theta-1}}\tau^{\kappa}\right)-\left(\frac{f_x}{f_u}\right)^{1-\frac{\kappa}{\theta-1}}\tau^{-\kappa-1}\right]\end{aligned}$$

As $\kappa(1+\tau^{1-\theta})^{\frac{-\kappa}{\theta-1}} > 0$, the sign of $\frac{\partial\Phi^{\kappa}(1+\tau^{1-\theta})^{\frac{-\kappa}{\theta-1}}}{\partial\tau}$ depends on the second term. We can derive that $\frac{\partial\Phi^{\kappa}(1+\tau^{1-\theta})^{\frac{-\kappa}{\theta-1}}}{\partial\tau} > 0$ as long as $\tau^{\theta-1}f_x > f_u$. Then $\frac{\partial z_s}{\partial\tau} > 0$ as all terms are positive.

### *C.2.2 Extended Model*

*Total cost and Price*

$$TC_u^d(z) = f_u + \frac{w_l}{z}y_u^d(z)$$

$$TC_u^{mx}(z) = f_u + f_{mx} + \frac{w_l}{z}y_u^d(z) + \tau_m\frac{w_l}{z}y_u^{mx}(z)$$

$$TC_u^{ox}(z) = f_u + f_{ox} + \frac{w_l}{\gamma_u z}y_u^d(z) + \tau_o\frac{w_l}{\gamma_u z}y_u^{ox}(z)$$

$$TC_s^{mx}(z) = f_{ms} + f_{mx} + \frac{w_h^{\beta}w_l^{1-\beta}}{\gamma_m z}y_s^{md}(z) + \tau_m\frac{w_h^{\beta}w_l^{1-\beta}}{\gamma_m z}y_s^{mx}(z)$$

$$TC_s^{ox}(z) = f_{os} + f_{ox} + \frac{w_h^{\alpha}w_l^{1-\alpha}}{\gamma_o z}y_s^{od}(z) + \tau_o\frac{w_h^{\alpha}w_l^{1-\alpha}}{\gamma_o z}y_s^{ox}(z)$$

where $f_{ox} > f_{mx}$, $f_{os} > f_{ms} > f_u$, $\alpha > \beta$ and $\gamma_o > \gamma_m > \gamma_u > 1$.

The profit maximization of both sectors yields the following pricing rules of domestic sales:

$$\rho_u^{md}(z) = \frac{\theta}{\theta-1}\frac{w_l}{z}$$
$$\rho_u^{od}(z) = \frac{\theta}{\theta-1}\frac{w_l}{\gamma_u z}$$
$$\rho_s^{md}(z) = \frac{\theta}{\theta-1}\frac{w_h^{\beta} w_l^{1-\beta}}{\gamma_m z}$$
$$\rho_s^{od}(z) = \frac{\theta}{\theta-1}\frac{w_h^{\alpha} w_l^{1-\alpha}}{\gamma_o z}$$

The four pricing rules of exporting are $\rho_u^{mx}(z) = \tau_m \rho_u^{md}(z)$, $\rho_u^{ox}(z) = \tau_o \rho_u^{od}(z)$, $\rho_s^{mx}(z) = \tau_m \rho_s^{md}(z)$, $\rho_s^{ox}(z) = \tau_o \rho_s^{od}(z)$. Hence, $\rho_s^{md}(z) = \rho_u^{md}(z)/\lambda_m$ where $\lambda_m = \gamma_m \left(\frac{w_l}{w_h}\right)^{\beta}$; $\rho_s^{od}(z) = \rho_u^{od}(z)/\lambda_o$ where $\lambda_o = \gamma_o \left(\frac{w_l}{w_h}\right)^{\alpha}$.

*Average Profit and Revenue*

The average profit $\tilde{\pi}' = \tilde{\pi}_u^{d'} + n_{mx}\tilde{\pi}_u^{mx} + n_{ox}\tilde{\pi}_u^{ox} + n_{ms}\tilde{\pi}_s^{mx} + n_{os}\pi_s^{ox}$, where

$$\tilde{\pi}_u^{d'} = \frac{1}{1-G(z_e')}\int_{z_e'}^{z_{mx}} z^{\theta-1} g(z)dz$$
$$\tilde{\pi}_u^{mx} = \frac{1}{1-G(z_e')}\int_{z_{mx}}^{z_{ox}} z^{\theta-1} g(z)dz$$
$$\tilde{\pi}_u^{ox} = \frac{1}{1-G(z_e')}\int_{z_{ox}}^{z_{ms}} z^{\theta-1} g(z)dz$$
$$\tilde{\pi}_s^{mx} = \frac{1}{1-G(z_e')}\int_{z_{ms}}^{z_{os}} z^{\theta-1} g(z)dz$$
$$\tilde{\pi}_s^{ox} = \frac{1}{1-G(z_e')}\int_{z_{os}}^{\infty} z^{\theta-1} g(z)dz$$

The average profit also can be describe in this way:

$$\begin{aligned}\tilde{\pi}' =& \frac{\tilde{r}'}{\theta} - (1-n_{ms})f_u - n_{mx}f_{mx} - n_{ox}(f_{ox}-f_{mx}) - n_{ms}(\phi_m f_u + f_{mx} - f_{ox}) \\ & - n_{os}(\phi_o f_u - \phi_m f_u + f_{ox} - f_{mx}) \\ =& \frac{\tilde{r}'}{\theta} - f_u - n_{mx}f_{mx} - n_{ox}(f_{ox}-f_{mx}) - n_{ms}((\phi_m-1)f_u + f_{mx} - f_{ox}) \\ & - n_{os}((\phi_o-\phi_m)f_u + f_{ox} - f_{mx})\end{aligned}$$

The average revenues of surviving firms is

$$\begin{aligned}
\tilde{r}' =& \int_{z_e}^{z_{mx}} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz + \int_{z_{mx}}^{z_{ox}} r_u^{mx}(z)\frac{g(z)}{1-G(z_e)}dz + \int_{z_{ox}}^{z_{ms}} r_u^{ox}(z)\frac{g(z)}{1-G(z_e)}dz \\
&+ \int_{z_{ms}}^{z_{os}} r_s^{ms}(z)\frac{g(z)}{1-G(z_e)}dz + \int_{z_{os}}^{\infty} r_s^{os}(z)\frac{g(z)}{1-G(z_e)}dz \\
=& \int_{z_e}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&+ \tau_m^{1-\theta}\int_{z_{mx}}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&+ ((1+\tau_o^{1-\theta})\gamma_u^{\theta-1} - 1 - \tau_m^{1-\theta})\int_{z_{ox}}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&+ \left[(1+\tau_m^{1-\theta})\lambda_m^{\theta-1}\left(\frac{\rho_2}{\rho_1}\right)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u^{\theta-1}\right]\int_{z_{ms}}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz \\
&+ \left[\lambda_o^{\theta-1}(1+\tau_o^{1-\theta})\left(\frac{\rho_2}{\rho_1}\right)^{\theta-\rho} - \lambda_m^{\theta-1}(1+\tau_m^{1-\theta})\left(\frac{\rho_2}{\rho_1}\right)^{\theta-\rho}\right]\int_{z_{os}}^{\infty} r_u^d(z)\frac{g(z)}{1-G(z_e)}dz
\end{aligned}$$

We derive that $z_{mx}$, $z_{ox}$, $z_{ms}$ and $z_{ox}$ can be a function of $z_e^{'}$, so

$$\begin{aligned}
\tilde{r}' =& \theta f_u\left(\frac{\tilde{z}_e^{'}}{z_e^{'}}\right)^{\theta-1} + n_{mx}\theta f_{mx}\left(\frac{\tilde{z}_{mx}}{z_{mx}}\right)^{\theta-1} + n_{ox}\theta(f_{ox}-f_{mx})\left(\frac{\tilde{z}_{ox}}{z_{ox}}\right)^{\theta-1} \\
&+ n_{ms}\theta(f_u(\phi_m-1)+f_{mx}-f_{ox})\left(\frac{\tilde{z}_{ms}}{z_{ms}}\right)^{\theta-1} \\
&+ n_{os}\theta(f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u)\left(\frac{\tilde{z}_{os}}{z_{os}}\right)^{\theta-1}
\end{aligned}$$

Given $\left(\frac{\tilde{z}_j}{z_j}\right)^{\theta-1} = \frac{\kappa}{\kappa-(\theta-1)}$ and the free entry condition, we get

$$z_e^{'} = \left(\frac{\theta-1}{\kappa-(\theta-1)}\frac{1}{\delta f_e}\Psi\right)^{1/\kappa}$$

$$
\begin{aligned}
\Psi^{\kappa} =& f_u + n_{mx} f_{mx} + n_{ox}(f_{ox} - f_{mx}) \\
&+ n_{ms}(f_u(\phi_m - 1) + f_{mx} - f_{ox}) + n_{os}(f_{ox} - f_{mx} + (\phi_o - \phi_m) f_u) \\
=& f_u + f_{mx}\tau_m^{-\kappa}\left(\frac{f_{mx}}{f_u}\right)^{\frac{-\kappa}{\theta-1}} + (f_{ox} - f_{mx})\left(\frac{f_{ox} - f_{mx}}{((1+\tau_o^{1-\theta})\gamma_u^{\theta-1} - 1 - \tau_m^{1-\theta}) f_u}\right)^{\frac{-\kappa}{\theta-1}} \\
&+ (f_u(\phi_m - 1) + f_{mx} - f_{ox}) \\
&\times \left[\frac{((\phi_m - 1) f_u + f_{mx} - f_{ox}}{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u^{\theta-1}) f_u}\right]^{\frac{-\kappa}{\theta-1}} \\
&+ (f_{ox} - f_{mx} + (\phi_o - \phi_m) f_u) \\
&\times \left[\frac{f_{ox} - f_{mx} + (\phi_o - \phi_m) f_u}{(\lambda_o^{\theta-1}(1+\tau_o^{1-\theta}) - \lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho} f_u}\right]^{\frac{-\kappa}{\theta-1}} \\
=& f_u + f_{mx}\tau_m^{-\kappa}\left(\frac{f_{mx}}{f_u}\right)^{\frac{-\kappa}{\theta-1}} + (f_{ox} - f_{mx})^{\frac{-\kappa}{\theta-1}} A + (f_u(\phi_m - 1) + f_{mx} - f_{ox}) B^{\frac{-\kappa}{\theta-1}} \\
&+ (f_{ox} - f_{mx} + (\phi_o - \phi_m) f_u) C^{\frac{-\kappa}{\theta-1}}
\end{aligned}
$$

where $A \equiv (1+\tau_o^{1-\theta})\gamma_u^{\theta-1} - 1 - \tau_m^{1-\theta}$, $B \equiv \frac{((\phi_m-1)f_u + f_{mx} - f_{ox}}{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u^{\theta-1}) f_u}$ and $C \equiv \frac{f_{ox} - f_{mx} + (\phi_o - \phi_m) f_u}{(\lambda_o^{\theta-1}(1+\tau_o^{1-\theta}) - \lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho} f_u}$.

*Changes in Trade Costs to Country m*

**Average Profit:**

Given $\tilde{\pi}' = \frac{\theta-1}{\kappa-(\theta-1)}\Psi^{\kappa}$, $\frac{\partial \tilde{\pi}'}{\partial \tau_m} = \frac{\theta-1}{\kappa-(\theta-1)}\frac{\partial \Psi^{\kappa}}{\partial \tau_m}$.

$$
\begin{aligned}
\frac{\partial \Psi^{\kappa}}{\partial \tau_m} = -\kappa f_{mx}\left(\frac{f_{mx}}{f_u}\right)^{\frac{-\kappa}{\theta-1}} \tau_m^{-\kappa-1} + \kappa A^{\frac{\theta-1-\kappa}{\theta-1}} f_u \tau_m^{-\theta} - \kappa B^{\frac{\theta-1-\kappa}{\theta-1}} f_u \lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}\tau_m^{-\theta} \\
+ \kappa C^{\frac{\theta-1-\kappa}{\theta-1}} f_u \lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}\tau_m^{-\theta}
\end{aligned}
$$

Since $f_{mx} > f_u$, $\left(\frac{f_{mx}}{f_u}\right)^{\frac{-\kappa}{\theta-1}} \tau_m^{-\kappa} > A$, $\tau_m^{-1} > \tau_m^{-\theta}$ and $B > C$, $\frac{\partial \Psi^{\kappa}}{\partial \tau_m} < 0$. Thus, $\frac{\partial \tilde{\pi}'}{\partial \tau_m} < 0$.

**Exit productivity cutoff:**

$$
z_e' = \left(\frac{\theta-1}{\kappa-(\theta-1)}\frac{1}{\delta f_e}\Psi\right)^{1/\kappa}
$$

Thus, $\frac{\partial z_e^{'}}{\partial \tau_m} < 0$

**Export Productivity Cutoff, Country m:**

$$z_{mx} = \left(\frac{\theta-1}{\kappa-(\theta-1)}\frac{1}{\delta f_e}\right)^{1/\kappa}\left[\frac{1}{n_{mx}}f_u + f_{mx} + \frac{n_{ox}}{n_{mx}}(f_{ox}-f_{mx}) + \frac{n_{ms}}{n_{mx}}(f_u(\phi_m-1) + f_{mx} - f_{ox}) + \frac{n_{os}}{n_{mx}}(f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u)\right]$$

*Proof:*

(1)

$$\frac{1}{n_{mx}} = \left(\frac{z_{mx}}{z_e^{'}}\right)^{\kappa} = \left(\tau_m\left(\frac{f_{mx}}{f_u}\right)^{\frac{1}{\theta-1}}\right)^{\kappa}$$

When $\tau_m$ falls, $\frac{1}{n_{mx}}$ goes down.

(2)

$$\frac{n_{ox}}{n_{mx}} = \left(\frac{z_{mx}}{z_{ox}}\right)^{\kappa} = \left(\frac{f_{mx}}{f_{ox}-f_{mx}}\right)^{\frac{\kappa}{\theta-1}}\left[\tau_m^{\theta-1}(\gamma_u^{\theta-1}(1+\tau_o^{1-\theta})-\tau_m^{1-\theta}-1)\right]^{\frac{\kappa}{\theta-1}}$$

Let $D \equiv \tau_m^{\theta-1}(\gamma_u^{\theta-1}(1+\tau_o^{1-\theta})-\tau_m^{1-\theta}-1)$

$$\frac{\partial \frac{n_{ox}}{n_{mx}}}{\partial \tau_m} = \left(\frac{f_{mx}}{f_{ox}-f_{mx}}\right)^{\frac{\kappa}{\theta-1}}\kappa D^{\frac{\kappa}{\theta-1}-1}\tau_m\theta\left(\gamma_u^{\theta-1}(1+\tau_o^{1-\theta})-1\right) > 0$$

Since $\gamma_u^{\theta-1}(1+\tau_o^{1-\theta}) > 1$, $\frac{n_{ox}}{n_{mx}}$ falls when $\tau_m$ drops.

(3)

$$\begin{aligned}\frac{n_{ms}}{n_{mx}} &= \left(\frac{z_{mx}}{z_{ms}}\right)^{\kappa} \\ &= \left(\frac{f_{mx}}{(\phi_m-1)f_u+f_{mx}-f_{ox}}\right)^{\frac{\kappa}{\theta-1}} \times \\ &\quad \left[\tau_m^{\theta-1}((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-(1+\tau_o^{1-\theta})\gamma_u^{\theta-1})\right]^{\frac{\kappa}{\theta-1}} \\ &= \left(\frac{f_{mx}}{(\phi_m-1)f_u+f_{mx}-f_{ox}}\right)^{\frac{\kappa}{\theta-1}} \times \\ &\quad \left[\frac{1+\tau_m^{1-\theta}}{\tau_m^{1-\theta}}\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-\frac{1+\tau_o^{1-\theta}}{\tau_m^{1-\theta}}\gamma_u^{\theta-1}\right]^{\frac{\kappa}{\theta-1}}\end{aligned}$$

When $\tau_m$ falls, $\tau_m^{1-\theta}$ increases, $\frac{1+\tau_m^{1-\theta}}{\tau_m^{1-\theta}}$ decreases and $\lambda_m$ also decreases. Thus, $\frac{n_{ms}}{n_{mx}}$ falls.

(4)

$$\begin{aligned}\frac{n_{os}}{n_{mx}} &= \left(\frac{z_{mx}}{z_{os}}\right)^{\kappa} \\ &= \left(\frac{f_{mx}}{f_{ox} - f_{mx} + (\phi_o - \phi_m) f_u}\right)^{\frac{\kappa}{\theta-1}} \times \\ &\quad \left[(\tau_m^{\theta-1}(\lambda_o^{\theta-1} - \lambda_m^{\theta-1}) + \tau_m^{\theta-1}\lambda_o^{\theta-1}\tau_o^{\theta-1} - \lambda_m^{\theta-1})(\rho_2/\rho_1)^{\theta-\rho}\right]^{\frac{\kappa}{\theta-1}}\end{aligned}$$

Since $(\lambda_o^{\theta-1} - \lambda_m^{\theta-1}) > 0$, $\frac{n_{os}}{n_{mx}}$ decreases with a lower $\tau_m$.

Therefore,

- $\tau_m \downarrow \Rightarrow \frac{1}{n_{mx}} \downarrow$
- $\tau_m \downarrow \Rightarrow \frac{n_{ox}}{n_{mx}} \downarrow$
- $\tau_m \downarrow \Rightarrow \frac{n_{ms}}{n_{mx}} \downarrow$
- $\tau_m \downarrow \Rightarrow \frac{n_{os}}{n_{mx}} \downarrow$,

We show that $\frac{\partial z_{mx}}{\partial \tau_m} > 0$.

**Skill Upgrading Productivity Cutoff, Country m**

$$\begin{aligned}z_{ms} = \left(\frac{\theta-1}{\kappa-(\theta-1)}\frac{1}{\delta f_e}\right)^{1/\kappa} \Bigg[&\frac{1}{n_{ms}} f_u + \frac{n_{mx}}{n_{ms}} f_{mx} + \frac{n_{ox}}{n_{ms}}(f_{ox} - f_{mx}) + (f_u(\phi_m - 1) \\ &+ f_{mx} - f_{ox}) + \frac{n_{os}}{n_{ms}}(f_{ox} - f_{mx} + (\phi_o - \phi_m) f_u)\Bigg]\end{aligned}$$

*Proof:*

(1)

$$\begin{aligned}\frac{1}{n_{ms}} &= \left(\frac{z_{ms}}{z_e}\right)^{\kappa} \\ &= \left[\frac{(\phi_m - 1) f_u + f_{mx} - f_{ox}}{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u) f_u}\right]^{\frac{\kappa}{\theta-1}}\end{aligned}$$

Since $\theta > 1$ and $\kappa > \theta - 1$, $\frac{1}{n_{ms}}$ falls when $\tau_m$ decreases.

(2)

When $\tau_m$ falls, $\frac{n_{ms}}{n_{mx}}$ decreases; thus $\frac{n_{mx}}{n_{ms}}$ increases.

(3)

$$\frac{n_{ox}}{n_{ms}} = \left(\frac{z_{ms}}{z_{ox}}\right)^{\kappa}$$
$$= \left[\frac{(\phi_m - 1)f_u + f_{mx} - f_{ox}}{f_{ox} - f_{mx}} \frac{((1+\tau_o^{1-\theta})\gamma_u^{\theta-1} - \tau_m^{1-\theta} - 1)}{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u)}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\tau_m \downarrow \Rightarrow \tau_m^{1-\theta} \uparrow$, $\frac{n_{ox}}{n_{ms}}$ falls with a lower $\tau_m$.

(4)

$$\frac{n_{os}}{n_{ms}} = \left(\frac{z_{ms}}{z_{os}}\right)^{\kappa}$$
$$= \left[\frac{(\phi_m - 1)f_u + f_{mx} - f_{ox}}{f_{ox} - f_{mx} + (\phi_o - \phi_m)f_u} \frac{(\lambda_o^{\theta-1}(1+\tau_o^{1-\theta}) - \lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho}}{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u)}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\tau_m \downarrow \Rightarrow \tau_m^{1-\theta} \uparrow$, $\frac{n_{os}}{n_{ms}}$ falls with a lower $\tau_m$.

Therefore,

- $\tau_m \downarrow \Rightarrow \frac{1}{n_{ms}} \downarrow$
- $\tau_m \downarrow \Rightarrow \frac{n_{mx}}{n_{ms}} \uparrow$
- $\tau_m \downarrow \Rightarrow \frac{n_{ox}}{n_{ms}} \downarrow$
- $\tau_m \downarrow \Rightarrow \frac{n_{os}}{n_{ms}} \downarrow$,

We can get $\frac{\partial z_{ms}}{\partial \tau_m} > 0$ when the second effect $\frac{n_{mx}}{n_{ms}}$ is dominated by the other three effects.

*Changes in Trade Costs to Country o*

**Average Profit:**

Given $\tilde{\pi}' = \frac{\theta-1}{\kappa-(\theta-1)}\Psi^{\kappa}$, $\frac{\partial\tilde{\pi}'}{\partial\tau_o} = \frac{\theta-1}{\kappa-(\theta-1)}\frac{\partial\Psi^{\kappa}}{\partial\tau_o}$.

$$\frac{\partial\Psi^{\kappa}}{\partial\tau_o} = -\kappa A^{\frac{\theta-1-\kappa}{\theta-1}} f_u \gamma_u^{\theta-1}\tau_o^{-\theta} + \kappa B^{\frac{\theta-1-\kappa}{\theta-1}} f_u \gamma_u^{\theta-1}\tau_o^{-\theta} - \kappa C^{\frac{\theta-1-\kappa}{\theta-1}} f_u \lambda_o^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}\tau_o^{-\theta}$$

As $A > B$, $\frac{\partial\Psi^{\kappa}}{\partial\tau_o} < 0$ and then, $\frac{\partial\tilde{\pi}'}{\partial\tau_o} < 0$.

**Exit Productivity Cutoff:**

Similarly, $\frac{\partial z_e'}{\partial\tau_o} < 0$.

**Export Productivity Cutoff, Country o:**

$$z_{ox} = \left(\frac{\theta-1}{\kappa-(\theta-1)}\frac{1}{\delta f_e}\right)^{1/\kappa}\left[\frac{1}{n_{ox}}f_u + \frac{n_{mx}}{n_{ox}}f_{mx} + (f_{ox}-f_{mx}) + \frac{n_{ms}}{n_{ox}}(f_u(\phi_m-1)+f_{mx}-f_{ox}) + \frac{n_{os}}{n_{ox}}(f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u)\right]$$

*Proof:*

(1)

$$\frac{1}{n_{ox}} = \left(\frac{z_{ox}}{z_e}\right)^{\kappa} = \left[\frac{f_{ox}-f_{mx}}{((1+\tau_o^{1-\theta})\gamma_u^{\theta-1}-\tau_m^{1-\theta}-1)f_u}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\theta > 1$ and $\kappa > \theta - 1$, $\frac{1}{n_{ox}}$ falls when $\tau_o$ decreases.

(2)

$$\frac{n_{mx}}{n_{ox}} = \left(\frac{z_{ox}}{z_{mx}}\right)^{\kappa} = \left(\frac{f_{ox}-f_{mx}}{f_{mx}}\right)^{\frac{\kappa}{\theta-1}}\left[\frac{(\gamma_u^{\theta-1}(1+\tau_o^{1-\theta})-\tau_m^{1-\theta}-1)}{\tau_m^{\theta-1}}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\tau_o \downarrow \Rightarrow \tau_o^{1-\theta} \uparrow$, $\frac{n_{mx}}{n_{ox}}$ increases with a lower $\tau_o$.

(3)

$$\frac{n_{ms}}{n_{ox}} = \left(\frac{z_{ox}}{z_{ms}}\right)^{\kappa}$$
$$= \left[\frac{f_{ox} - f_{mx}}{(\phi_m - 1)f_u + f_{mx} - f_{ox}} \frac{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u)}{((1+\tau_o^{1-\theta})\gamma_u^{\theta-1} - \tau_m^{1-\theta} - 1)}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\tau_o \downarrow \Rightarrow \tau_o^{1-\theta} \uparrow$, $\frac{n_{ms}}{n_{ox}}$ falls with a lower $\tau_o$.

(4)

$$\frac{n_{os}}{n_{ox}} = \left(\frac{z_{ox}}{z_{os}}\right)^{\kappa}$$
$$= \left[\frac{f_{ox} - f_{mx}}{f_{ox} - f_{mx} + (\phi_o - \phi_m)f_u} \frac{(\lambda_o^{\theta-1}(1+\tau_o^{1-\theta}) - \lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho}}{((1+\tau_o^{1-\theta})\gamma_u^{\theta-1} - \tau_m^{1-\theta} - 1)}\right]^{\frac{\kappa}{\theta-1}}$$

Let $E \equiv (\lambda_o^{\theta-1}(1+\tau_o^{1-\theta}) - \lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho}$ and $F \equiv ((1+\tau_o^{1-\theta})\gamma_u^{\theta-1} - \tau_m^{1-\theta} - 1)$.

$$\frac{\partial \frac{n_{os}}{n_{ox}}}{\tau_o} = \left[\frac{f_{ox} - f_{mx}}{f_{ox} - f_{mx} + (\phi_o - \phi_m)f_u}\right]^{\frac{\kappa}{\theta-1}} \kappa \frac{-\lambda_o^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}\tau_o^{-\theta}F - \gamma_u^{\theta-1}\tau_o^{-\theta}E}{F^2}$$

Since $-\lambda_o^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}\tau_o^{-\theta}F - \gamma_u^{\theta-1}\tau_o^{-\theta}E < 0$ and other terms are positive,
$\frac{\partial \frac{n_{os}}{n_{ox}}}{\tau_o} < 0$.

Hence,

- $\tau_o \downarrow \Rightarrow \frac{1}{n_{ox}} \downarrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{mx}}{n_{ox}} \uparrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{ms}}{n_{ox}} \downarrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{os}}{n_{ox}} \uparrow$,

$\frac{\partial z_{ox}}{\partial \tau_o} > 0$ if and only if the total effects of $\tau_o$ on $\frac{1}{n_{ox}}$ and $\frac{n_{ms}}{n_{ox}}$ dominate the other two effects.

**Skill Upgrading Productivity Cutoff, Country o:**

$$z_{os} = \left(\frac{\theta-1}{\kappa-(\theta-1)}\frac{1}{\delta f_e}\right)^{1/\kappa}\left[\frac{1}{n_{os}}f_u + \frac{n_{mx}}{n_{os}}f_{mx} + \frac{n_{ox}}{n_{os}}(f_{ox}-f_{mx}) + \frac{n_{ms}}{n_{os}}(f_u(\phi_m-1)+f_{mx}-f_{ox}) + (f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u)\right]$$

*Proof:*

(1)

$$\frac{1}{n_{os}} = \left(\frac{z_{os}}{z_e}\right)^{\kappa} = \left[\frac{f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u}{(\lambda_o^{\theta-1}(1+\tau_o^{1-\theta})-\lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho}f_u}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\tau_o \downarrow \Rightarrow \tau_o^{1-\theta} \uparrow$, $\frac{1}{n_{os}}$ falls with a lower $\tau_o$.

(2)

$$\frac{n_{mx}}{n_{os}} = \left(\frac{z_{os}}{z_{mx}}\right)^{\kappa} = \left[\frac{f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u}{f_{mx}}\frac{\tau_m^{1-\theta}}{(\lambda_o^{\theta-1}(1+\tau_o^{1-\theta})-\lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho}f_u}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\tau_o \downarrow \Rightarrow \tau_o^{1-\theta} \uparrow$, $\frac{n_{mx}}{n_{os}}$ falls with a lower $\tau_o$.

(3)

As $\frac{\partial \frac{n_{os}}{n_{ox}}}{\tau_o} < 0$, $\frac{n_{ox}}{n_{os}}$ decreases if $\tau$ falls.

(4)

$$\frac{n_{ms}}{n_{os}} = \left(\frac{z_{os}}{z_{ms}}\right)^{\kappa} = \left[\frac{f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u}{(\phi_m-1)f_u+f_{mx}-f_{ox}}\frac{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-(1+\tau_o^{1-\theta})\gamma_u)}{(\lambda_o^{\theta-1}(1+\tau_o^{1-\theta})-\lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho}f_u}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\tau_o \downarrow \Rightarrow \tau_o^{1-\theta} \uparrow$, $\frac{n_{ms}}{n_{os}}$ falls with a lower $\tau_o$.

Hence,

- $\tau_o \downarrow \Rightarrow \frac{1}{n_{os}} \downarrow$

- $\tau_o \downarrow \Rightarrow \frac{n_{mx}}{n_{os}} \downarrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{ox}}{n_{os}} \downarrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{ms}}{n_{os}} \downarrow$,

We prove that $\frac{\partial z_{os}}{\partial \tau_o} > 0$.

**Export Productivity Cutoff, Country m:**

$$z_{mx} = \left(\frac{\theta-1}{\kappa-(\theta-1)}\frac{1}{\delta f_e}\right)^{1/\kappa}\left[\frac{1}{n_{mx}}f_u + f_{mx} + \frac{n_{ox}}{n_{mx}}(f_{ox}-f_{mx}) + \frac{n_{ms}}{n_{mx}}(f_u(\phi_m-1) + f_{mx} - f_{ox}) + \frac{n_{os}}{n_{mx}}(f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u)\right]$$

*Proof:*

(1)

$$\frac{1}{n_{mx}} = \left(\frac{z_{mx}}{z_e^{'}}\right)^{\kappa} = \left(\tau_m\left(\frac{f_{mx}}{f_u}\right)^{\frac{1}{\theta-1}}\right)^{\kappa}$$

$\tau_o$ has no effects on $\frac{1}{n_{mx}}$.

(2)

$$\frac{n_{ox}}{n_{mx}} = \left(\frac{z_{mx}}{z_{ox}}\right)^{\kappa} = \left(\frac{f_{mx}}{f_{ox}-f_{mx}}\right)^{\frac{\kappa}{\theta-1}}\left[\tau_m^{\theta-1}(\gamma_u^{\theta-1}(1+\tau_o^{1-\theta})-\tau_m^{1-\theta}-1)\right]^{\frac{\kappa}{\theta-1}}$$

When $\tau_o$ falls, $\tau_0^{1-\theta}$ increases and $\frac{n_{ox}}{n_{mx}}$ goes up.

(3)

$$\begin{aligned}\frac{n_{ms}}{n_{mx}} &= \left(\frac{z_{mx}}{z_{ms}}\right)^{\kappa} \\ &= \left(\frac{f_{mx}}{(\phi_m-1)f_u+f_{mx}-f_{ox}}\right)^{\frac{\kappa}{\theta-1}} \times \\ &\quad \left[\tau_m^{\theta-1}((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-(1+\tau_o^{1-\theta})\gamma_u^{\theta-1})\right]^{\frac{\kappa}{\theta-1}} \\ &= \left(\frac{f_{mx}}{(\phi_m-1)f_u+f_{mx}-f_{ox}}\right)^{\frac{\kappa}{\theta-1}} \times \\ &\quad \left[\frac{1+\tau_m^{1-\theta}}{\tau_m^{1-\theta}}\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-\frac{1+\tau_o^{1-\theta}}{\tau_m^{1-\theta}}\gamma_u^{\theta-1}\right]^{\frac{\kappa}{\theta-1}}\end{aligned}$$

When $\tau_o$ falls, $\tau_o^{1-\theta}$ increases; thus $\frac{n_{ms}}{n_{mx}}$ falls.

(4)

$$\begin{aligned}\frac{n_{os}}{n_{mx}} &= \left(\frac{z_{mx}}{z_{os}}\right)^{\kappa} \\ &= \left(\frac{f_{mx}}{f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u}\right)^{\frac{\kappa}{\theta-1}} \times \\ &\quad \left[(\tau_m^{\theta-1}(\lambda_o^{\theta-1}-\lambda_m^{\theta-1})+\tau_m^{\theta-1}\lambda_o^{\theta-1}\tau_o^{\theta-1}-\lambda_m^{\theta-1})(\rho_2/\rho_1)^{\theta-\rho}\right]^{\frac{\kappa}{\theta-1}}\end{aligned}$$

When $\tau_o$ falls, $\tau_o^{\theta-1}$ falls and $\frac{n_{os}}{n_{mx}}$ goes down.

Therefore,

- $\tau_o \downarrow \Rightarrow \frac{n_{ox}}{n_{mx}} \uparrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{ms}}{n_{mx}} \downarrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{os}}{n_{mx}} \downarrow$.

A reduction in $\tau_o$ decreases $z_{mx}$ when the impact of $\tau_o$ on $\frac{n_{ox}}{n_{mx}}$ is dominated by the other two effects.

**Skill Upgrading Productivity Cutoff, Country m:**

$$z_{ms} = \left(\frac{\theta-1}{\kappa-(\theta-1)}\frac{1}{\delta f_e}\right)^{1/\kappa}\left[\frac{1}{n_{ms}}f_u + \frac{n_{mx}}{n_{ms}}f_{mx} + \frac{n_{ox}}{n_{ms}}(f_{ox}-f_{mx}) + (f_u(\phi_m-1) + f_{mx} - f_{ox}) + \frac{n_{os}}{n_{ms}}(f_{ox}-f_{mx}+(\phi_o-\phi_m)f_u)\right]$$

*Proof:*

(1)

$$\begin{aligned}\frac{1}{n_{ms}} &= \left(\frac{z_{ms}}{z_e}\right)^{\kappa} \\ &= \left[\frac{(\phi_m-1)f_u+f_{mx}-f_{ox}}{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho}-(1+\tau_o^{1-\theta})\gamma_u)f_u}\right]^{\frac{\kappa}{\theta-1}}\end{aligned}$$

When $\tau_o$ falls, $\tau_o^{1-\theta}$ increases and $\frac{1}{n_{ms}}$ goes up.

(2)

When $\tau_o$ falls, $\frac{n_{ms}}{n_{mx}}$ decreases; thus $\frac{n_{mx}}{n_{ms}}$ increases.

(3)

$$\frac{n_{ox}}{n_{ms}} = \left(\frac{z_{ms}}{z_{ox}}\right)^{\kappa}$$
$$= \left[\frac{(\phi_m - 1)f_u + f_{mx} - f_{ox}}{f_{ox} - f_{mx}} \frac{((1+\tau_o^{1-\theta})\gamma_u^{\theta-1} - \tau_m^{1-\theta} - 1)}{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u)}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\tau_o \downarrow \Rightarrow \tau_o^{1-\theta} \uparrow$, $\frac{n_{ox}}{n_{ms}}$ increases with a lower $\tau_o$.

(4)

$$\frac{n_{os}}{n_{ms}} = \left(\frac{z_{ms}}{z_{os}}\right)^{\kappa}$$
$$= \left[\frac{(\phi_m - 1)f_u + f_{mx} - f_{ox}}{f_{ox} - f_{mx} + (\phi_o - \phi_m)f_u} \frac{(\lambda_o^{\theta-1}(1+\tau_o^{1-\theta}) - \lambda_m^{\theta-1}(1+\tau_m^{1-\theta}))(\rho_2/\rho_1)^{\theta-\rho}}{((1+\tau_m^{1-\theta})\lambda_m^{\theta-1}(\rho_2/\rho_1)^{\theta-\rho} - (1+\tau_o^{1-\theta})\gamma_u)}\right]^{\frac{\kappa}{\theta-1}}$$

Since $\tau_o \downarrow \Rightarrow \tau_o^{1-\theta} \uparrow$, $\frac{n_{os}}{n_{ms}}$ rises with a lower $\tau_o$.

Therefore,

- $\tau_o \downarrow \Rightarrow \frac{1}{n_{ms}} \uparrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{mx}}{n_{ms}} \uparrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{ox}}{n_{ms}} \uparrow$
- $\tau_o \downarrow \Rightarrow \frac{n_{os}}{n_{ms}} \uparrow$,

We derive that $\frac{z_{ms}}{\tau_o} < 0$.

## *C.3 Data Description*

### *C.3.1 Computation of Input Tariffs*

We computed input tariffs for each 4-digit CIC industry in a similar way as Amiti and Konings (2007) and Bustos (2011b). The input tariff for each industry is computed as

weighted average of the tariffs of all inputs used, where the weights are based on the cost share of each input, according to the following formula:

$$\tau_{jt}^{im} = \sum_i w_{ij} \times \tau_{it}^{im} \text{ where } w_{ij} = \frac{a_{ij}}{\sum_i a_{ij}} \tag{C.1}$$

where $j$ indexes the 4-digit CIC industry for which the input tariff is computed; $i$ indexes the 4-digit CIC industry producing the input, and $t$ indexes time. $w_{ij}$ denotes the cost share of each input $i$ in the production of output $j$, and $a_{ij}$ is total expenditure in input $i$ by industry $j$. These expenditure shares include both domestic and imported inputs. We estimated $a_{ij}$ based on China's input-output (I-O) table in 2007. The data are aggregated at the sector level, and we use the same value for all the industries in the same sector.

### *C.3.2 Proxy for Initial Productivity*

In the model, heterogeneity is given by labor productivity holding skill level constant, which is not directly observed in the data. As a proxy for initial productivity, we use initial firm size in terms of employment relative to the 4-digit industry average.

### *C.3.3 Measures of Capital and Skill Intensity*

Average capital and skill intensity in the industry in the United States in the 1980s is obtained from the NBER productivity database. The measure of capital intensity is capital (real equipment plus real structures) per worker, and the measure of skill intensity is the ratio of non-production to production workers in the industry.

### *C.3.4 Summary Statistics*

Table C.1: Summary Statistics of Variables of Interest in 2004

| Variables | All | Exporters | Non-exporters | Observations | Firms |
|---|---|---|---|---|---|
| Employment | 361.055 | 691.605 | 190.219 | 131,460 | 110,632 |
| | [1806.685] | [2867.222] | [785.306] | | |
| Total sales | 160.362 | 373.104 | 50.412 | 131,460 | 110,632 |
| | [1826.812] | [3061.618] | [426.793] | | |
| Export share of sales, Exports>0 | | 0.028 | | 23,964 | 44,792 |
| | | [0.048] | | | |
| 1{Export to India}, Exports>0 | | 0.090 | | 23,964 | 44,792 |
| | | [0.286] | | | |
| 1{TS>0} | 0.443 | 0.466 | 0.432 | 131,460 | 110,632 |
| | [0.497] | [0.499] | [0.495] | | |
| Total training spending | 43.139 | 82.403 | 22.846 | 131,460 | 110,632 |
| | [432.506] | [639.843] | [266.359] | | |
| Total training spending, TS>0 | 97.279 | 176.890 | 52.899 | 58,296 | 48,525 |
| | [645.421] | [928.517] | [403.346] | | |
| Training spending per worker | 111.912 | 117.462 | 109.044 | 131,460 | 110,632 |
| | [591.856] | [561.849] | [606.766] | | |
| Training spending per worker, TS>0 | 252.367 | 252.149 | 252.488 | 58,296 | 48,525 |
| | [868.612] | [802.307] | [903.476] | | |
| Observations | 131,460 | 44,792 | 86,668 | | |
| Firms | 110,632 | 23,964 | 86,668 | | |

Notes: Standard deviations in brackets. Employment in number of workers, sales in millions of 2004 CNY, total training spending in thousands of 2004 CNY, and training spending per capita in 2004 CNY.