DEALING WITH HEALTH AND HEALTH CARE SYSTEM CHALLENGES IN CHINA: ASSESSING HEALTH DETERMINANTS AND HEALTH CARE REFORMS

ISBN: 978 90 3610 490 6

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 696 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Dealing with Health and Health Care System Challenges in China: Assessing health determinants and health care reforms

Uitdagingen voor de volksgezondheid en gezondheidszorg in China:
Een evaluatie van determinanten en hervormingen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 5 oktober 2017 om 11:30 uur

door
Hao Zhang
geboren te Henan, China

Erasmus University Rotterdam

Erasmus

Promotiecommissie:

Promotor:	Prof.dr. E. K. A. van Doorslaer
Overige leden:	Prof.dr. O. O'Donnell Prof.dr. M. Lindeboom Prof.dr. W. Yip
Copromotor:	Dr. T. M. Bago d'Uva

Acknowledgements

Born to a physician mother, I have fond childhood memories of doctors in white coats as my playmates and medicines as my toys, and have always been interested in topics related to health and health care. It appeared that this interest would develop into nothing when my mother discouraged me from being a physician and I chose International Economics and Trade as my undergraduate major. However, after a series of "coincidences", health and health care found their way into the title of my PhD thesis. Looking back, I realized that I might have been following a deterministic trend, just with very large disturbances.

Despite my interest in health economics, doing a PhD has not always been pleasant. The feeling of frustration and dissatisfaction with the papers came up from time to time. Fortunately, many people helped me manage through those challenging periods. I am most indebted to Eddy van Doorslaer and Teresa Bago d'Uva, who have been my supervisors since the second year of my MPhil. I have benefited tremendously from your expertise and encouragement, and I cannot thank you enough for being so kind and available. Deep thanks also go to Darjusch Tafreschi for never failing to provide a solution to our problems, to Joris van de Klundert for involving me in the Rural Health Project and offering generous help and advice, to Ling Xu and Yaoguang Zhang for answering my many questions and providing valuable insights, to Zhiyuan Hou, Qingyue Meng, Yang Sun and Jian Wu for generously sharing your knowledge and previous work on the payment reform in Henan, and to Adam Wagstaff, Winnie Yip, Yanhua Chi and her many colleagues at the National Health and Family Planning Commission for helpful insights and comments.

As a rather shy and quite person (even by Chinese standards), I was lucky to have joined an extremely welcoming group – the Health Economics group. Many thanks go to Anne, Bastian, Ellen, Hale, Igna, Kim, Max, Pieter, Pilar, Sven, Tom, and Wameq for helping me with various academic and non-academic issues. Even for questions like whether to have an ACL reconstruction, two of you were able to provide first-hand experience of treatment and recovery following the same injury. What more can one ask for? Special thanks go to Hans for inviting me to Hong Kong and being such a pleasant person to talk and work with, to Owen for always providing sharp and helpful comments with my favorite Scottish accent, and to Tingting for being a sweet colleague for one year and remaining a caring friend afterwards.

My PhD journey has been blessed with amazing new friends outside the Health Economics circle as well – from Tong on my first flight to Schiphol to Eugina and Heather in my last days in the Netherlands. Words cannot express how lucky I feel and how grateful I am, but still, my heartfelt thanks go to Siyu and Tong for always being there for me from the very beginning and through all my ups and downs; to my cherished WeChat group members Mengyang, Rui, Yawen, Yingjie and Zhihua for your continual supply of personal attacks and yet deep (and apparently blind) fondness of me even after you have left Rotterdam one by one; to my roommates, Uyanga, for the best borscht, beetroot salad and spaghetti I have ever tasted and could ever imagine, and Xinzhu, for being my cheerful encyclopedia of everything I

need to know to not look like a cavewoman when back in China; to my officemates, Charlie, Elio, and Vitalie, for not reporting on my whimsical absence; and to so many other friends with whom I had unforgettable holiday trips in Europe and/or memorable chats at the dinner table, along the beautiful canals, or on a comfortable couch.

Thanks also go to old friends: Di for staying emotionally and geographically close for 14 years, Shanshan and Zhengrong for making Beijing feel like home, Tingyi for everything, and other former classmates and schoolmates for lovely chats and reunions, and Can and Yaoyao for our everlasting comradeship and wonderful Christmas trips in the U.S. and Mexico.

Last and most importantly, Mom and Dad, thank you for your unlimited and unconditional love and support. I am eternally grateful for having you as my parents, my superheroes and my best friends.

Contents

CHAPTER 1

Introduction

It is generally acknowledged that China has seen enormous health improvements in the first three decades following her founding in 1949. For example, average life expectancy at birth reached 68 years by the early 1980s, a remarkable increase from about 35 in the early 1950s (Hesketh & Zhu 1997). This was possible because of major investments in public health measures such as immunization, improved sanitation, and control of disease vectors (e.g., mosquitoes for malaria) through a highly centralized governmental agency (Blumenthal & Hsiao 2005). However, since the 1980s, mortality reduction became harder as chronic diseases (e.g., heart diseases, cancer and stroke) started to replace infectious diseases as leading causes of death (Ministry of Health 2004). Health inequalities also increased as China's economy took off after the economic reform around 1979 (Wagstaff et al. 2009). Moreover, this transition from a planned economy to a market one greatly reduced government funding in health care, leaving the vast majority of the population uninsured and turning health care providers into profit-seeking entities (Blumenthal & Hsiao 2015).

In the next three decades, measures have been taken to address the consequences of these demographic, epidemiological and socioeconomic changes. Nevertheless, health inequalities, non-communicable diseases, and the low performance and high cost of the health care system have become prominent social issues. This dissertation addresses the challenges faced by China around 2010 in both the population health domain and the health care system. Specifically, the first two chapters are devoted to health challenges, explaining the *female health disadvantage* in later life and assessing the effect of only children on their elderly parents' *mental wellbeing*. The next two chapters are devoted to health care system challenges, assessing in rural China if bottom-up results-based reforms could improve the health system performance under limited funding and if a simplified diagnosis-related group (DRG)-based hospital payment system could contain the fast-growing health expenditure.

While female health disadvantage has been observed around the world (Nathanson 1975), a national survey of the Chinese elderly (45+) found it to be larger in China than in developed countries (CHARLS Research Team 2013). The survey revealed a female disadvantage in all health measures collected, both subjective and objective. It was particularly pronounced for needing help with basic daily activities (27.5% versus 19.8%), body pain (39.1% versus 27.5%), overweight (31.7% versus 24.3%), and hypertension (58.6% versus 49.1%). As this contrasts with women's general longer life expectancy, questions then emerge as to whether the observed female health disadvantage is simply a reporting artifact (Spiers et al. 2003; Verbrugge 1982). If, on the other hand, it is real, to what extent is it due to biological

differences or to the culturally embedded discrimination against females? Clearly, different answers would have different policy implications.

Using the elderly Chinese sample (50+) from the WHO SAGE survey, Chapter 2 of the dissertation assesses the aforementioned three potential explanations of the female health disadvantage. This chapter first examines gender differences in reporting exploiting information on anchoring vignettes. As vignettes represent fixed health states, systematic differences by gender in the ratings can thus be viewed as indicative of gender-specific reporting. Since no such systematic differences are found in eight domains of health functioning, gender-specific reporting is ruled out as a possible explanation of the gender health gap in the China context. Next, a non-linear extension of the Oaxaca-Blinder decomposition is used to assess the proportion of the gender health gap attributable to discrimination (reflected in female education disadvantage) and biological differences (reflected in gender differences in chronic conditions and health functioning). The baseline specification including only education and other basic socio-demographic variables assesses the gross contribution of the female education disadvantage to the gender health gap. Richer specifications including chronic conditions and health functioning assess whether gender differences in these factors can (1) explain the health gap left unexplained by socio-demographic characteristics, and (2) explain the gross contribution of the female education disadvantage.

Another health challenge relates to the mental wellbeing of the elderly. It has received insufficient attention from the general public, policy makers, and even health professionals and researchers. For one better-studied outcome – depression/depressive symptoms, a meta-analysis found the rate of depressive symptoms to be about 15% among the elderly in the 1980s and 1990s (Chen et al. 1999). However, it more than doubled in about two decades and reached 40% in 2011 (CHARLS Research Team 2013). More alarmingly, the already high rate may have been underestimated due to people's unwillingness to admit negative emotions.[1] In addition to the problem of stigma, professional help is not always easily accessible. By 2016, some of the highest level of hospitals (level A tertiary hospitals) still did not have a department of psychiatry and, for those that did, the department was often poorly staffed and its value not recognized by other departments.[2] Given the barriers in diagnosis and treatment, identification of the factors contributing to depressive symptoms may provide an alternative way to ameliorate the problem.

One potential candidate is reduced fertility. Following the one-child policy introduced around 1980, China saw the emergence of a generation of parents with only one child. Given the strong cultural and economic motivations for multiple children, especially sons, the one-child restriction has led to widespread public discontent. It is thus straightforward to expect affected parents to be less happy. In addition, reduced fertility may imply reduced parent-child interactions, which can also bring negative emotions and contribute to cognitive decline among parents. Chapter 3 aims to assess if having only one child negatively affects parents'

[1] A survey in Shanghai in 2016 found that the proportion of local residents with negative attitudes towards mental health patients remained unchanged at 42% after five years. Source: http://www.wsjsw.gov.cn/wsj/n473/n1995/u1ai138621.html.

[2] Source: http://paper.people.com.cn/smsb/html/2016-10/14/content_1718235.htm.

mental wellbeing in later life, as measured by depressive symptoms, cognitive skills and life satisfaction. Identifying causal effects is challenging because we do not observe some important community-level and household-level characteristics that are correlated with both the number of children and the mental wellbeing of the elderly parents. Examples include local pension generosity and economic conditions in the 1980s and couples' preferences for more children. To overcome this endogeneity problem, an instrument for having only one child is constructed exploiting variation in regional enforcement of the one-child restriction (yes/no) and in women's age at the time of the enforcement.

The aforementioned two population health challenges not only pose threats to a satisfying and active life in old age, but also create heavy burdens for the health care system. Proper incentives need to be in place for health care providers so that diseases can be detected early, treated properly and managed effectively. However, the current health care system in China is highly hospital-centric, fragmented and wasteful (World Bank 2016). Because government funding is limited and basic services are underpriced, providers of all levels have a strong incentive to overprescribe profitable drugs and medical tests. Over time, this has contributed to escalating costs and large-scale public discontent. In response, a national health reform was initiated in 2009 with a comprehensive set of objectives including removing the perverse incentives (e.g., drug markups), promoting health prevention and disease management, and directing patients away from tertiary hospitals to primary health facilities (Z. Chen 2009).

Given the vast regional variation in health needs and constraints in China, different places implemented diverse reforms to achieve the reform goals. However, after five years, progress has been limited in redressing the perverse incentives and curbing cost escalation on the national level. For example, insurance data showed that average inpatient expenditure for urban employees[3] increased by 147% (from 2865 Yuan in 2009 to 7083 Yuan in 2014), and drugs and medical tests still accounted for 86% of the total expenditure (Ministry of Human Resources and Social Security 2015). Recent national statistics on patient flow also paint a disappointing picture. From January to October in 2016, while total visits to health facilities increased by 1.8% year-on-year, hospital visits (41% of total) increased by 5.1% and tertiary hospital visits (20% of total) increased by 7.3%.[4] Sound empirical analysis is needed to identify the successes and failures of local reforms so that future interventions can be better designed to achieve the intended goals.

Chapters 4 and 5 make a contribution to this aim by evaluating two important reforms, one being a six-year system-wide reform project and the other an inpatient provider payment reform. Specifically, Chapter 4 assesses if a bottom-up results-based reform project can improve health care system performance in China. This Rural Health Project (or Health XI) was initiated in 2008 when China faced challenges in improving the financing and delivery of

[3] There are three major medical schemes in China. Urban employees are covered by Urban Employee Basic Medical Insurance, unemployed urban residents by Urban Resident Basic Medical Insurance for the unemployed, and rural residents by New Rural Cooperative Medical Scheme.

[4] Source: http://www.nhfpc.gov.cn/mohwsbwstjxxzx/s7967/201612/6bebbc7c1187481ca98bd4049c825116.shtml.

medical care and public health services under limited funding. To find the most appropriate interventions under different conditions, Health XI involved 40 counties across eight provinces that vary substantially in geography and socioeconomic development. It allowed participating counties to design their own interventions according to local conditions (bottom-up), but the interventions needed to be assessed and approved for their substance. Effective management was achieved by pre-specifying a set of common project targets (results-based), and by providing tight supervision and timely feedback and assistance.

Like many other social programs, Health XI did not adopt a randomized design. To construct an appropriate counterfactual, we use the non-Health-XI counties from an existing national survey. These counties are matched to Health XI counties based on county socioeconomic characteristics over eight years (four years before and four years after the project initiation). This long time span gives confidence that control counties were comparable to Health XI counties in important aspects (e.g., per capita GDP) not only before, but also during Health XI. A difference-in-differences (DID) model is used to assess the effects of Health XI in the domains of medical care, public health services provision and self-rated health by 2013. To further examine treatment effect heterogeneity, a triple difference (DDD) model is used to assess if participating counties benefit from previous experience with a similar reform project or better fiscal conditions.

While there is plenty of scope for improvement in the current health care system, inpatient payment reform is possibly the most promising because it addresses the perverse incentives to overprescribe drugs and medical tests. It is also the most challenging because it reforms the way in which providers obtain the bulk of their income. One such payment reform was implemented in 2011 as part of the Health XI project in two counties of the Henan province. These two counties piloted a simplified diagnosis-related group (DRG)-based payment system that (i) set the payment rates for a selection of inpatient conditions, and (ii) imposed clinical pathways. Chapter 5 assesses if this payment reform achieved its intended goal of containing cost without compromising quality. As providers can respond strategically and bring about unintended consequences, particular attention is paid to whether and how providers selected patients. Outcomes examined include service volume, case mix proxies, treatment intensity, expenditures and patient satisfaction. These outcomes are also examined by provider level, as room for strategic response is different at township health centers than at the higher-level county hospitals.

Two non-project counties of Henan are selected from the aforementioned national survey to serve as controls because of their comparable socioeconomic conditions. First, general effects of Health XI are identified using a DID model. Next, effects of the new payment system are isolated by additionally taking the difference between DRG-eligible and DRG-ineligible conditions. This DDD approach provides valid effect estimates under the assumption that, in the absence of the reform, the difference between DRG-eligible and DRG-ineligible conditions in treatment counties would have followed parallel trends with that in control counties. Last, a particular DRG-eligible condition – delivery – is examined in isolation with a DID model to assess if the magnitude of provider response can indeed be restricted by certain features of a condition.

Chapters 2-5 in this dissertation examine population health and health care system challenges. The first two chapters address the determinants of female health disadvantage and mental wellbeing among the elderly. The results provide some insights on which areas are important and which are not when designing interventions to address the problems at an early stage. The last two chapters assess attempts at improving the health care system in China. The results shed light on the successes and failures of the piloted reforms and provide valuable lessons for the design of future reforms. Causal inference is used in all chapters except for Chapter 2, where a decomposition analysis of an exploratory nature is more suited. While the validity of the identification strategies is carefully substantiated in each chapter, it does not answer the question "to what circumstances can the results be extrapolated". This external validity question is addressed by explaining in detail the context of the interventions (i.e. the one-child policy, the Health XI project, and the provider payment reform), and by exploring potential mechanisms with data (Chapter 3) or with reference to previous more descriptive studies (Chapters 4 and 5).

CHAPTER 2

The Gender Health Gap in China: A Decomposition Analysis

Joint work with Teresa Bago d'Uva and Eddy van Doorslaer

Around the world, and in spite of their higher life expectancy, women tend to report worse health than men. Explanations for this gender gap in self-assessed health may be different in China than in other countries due to the traditional phenomenon of son preference. We examine several possible reasons for the gap using the Chinese SAGE data. We first rule out differential reporting by gender as a possible explanation, exploiting information on anchoring vignettes in eight domains of health functioning. Decomposing the gap in general self-assessed health, we find that about 31% can be explained by socio-demographic factors, most of all by discrimination against women in education in the 20th century. A more complete specification including chronic conditions and health functioning fully explains the remainder of the gap (about 69%). Adding chronic conditions and health functioning also explains at least two thirds of the education contribution, suggesting how education may affect health. In particular, women's higher rates of arthritis, angina and eye diseases make the largest contributions to the gender health gap, by limiting mobility, increasing pain and discomfort, and causing sleep problems and a feeling of low energy.

This chapter has been published as:
Zhang, H., Bago D'Uva, T. & Van Doorslaer, E., 2015. The gender health gap in China: A decomposition analysis. *Economics & Human Biology*, 18, pp.13–26.

2.1 Introduction

Gender inequality has long been a topic of academic research and a target for policy interventions. Female disadvantage is still apparent in various aspects of life, such as educational attainment and labor market outcomes. Studies also consistently show that, compared to men, women report more illnesses, worse health, and higher health care utilization despite their higher life expectancy (see e.g., Nathanson (1975)). There have been many attempts to explain this phenomenon.

One strand of literature has looked at epidemiological reasons (Case & Paxson 2005; Malmusi et al. 2012; Verbrugge 1989). Case and Paxson (2005) provide the most convincing evidence using 14 years of data from the US National Health Interview Survey (1986-2001). Using a decomposition method, they demonstrate that female disadvantage in self-assessed health in the US is entirely explained by differences in prevalence rates of chronic conditions between men and women: females have significantly higher rates of degenerative but non-fatal conditions like arthritis and other pains, most respiratory conditions (excluding cancer), hypertension, vision problems and depression.

Another possible explanation that is often put forward – but not examined in Case and Paxson (2005) – is that women are less stoical than men. Given objective health, women may be more likely to report health problems (Verbrugge 1982), or to factor less serious ailments into their assessment of own health (Spiers et al. 2003). While not implausible, such claims were mostly supported by suggestive evidence. It also contrasts the findings by MacIntyre et al. (1999), who ask women and men open-ended questions about health problems, followed by a series of probes for specific conditions. They find that men provide more information to the open-ended questions and women are not more likely to report trivial health conditions.

More recently, vignette methods have been used to formally test the claim of women's higher tendency to report health problems but have produced mixed results. Peracchi and Rossetti (2012) do find that European females are more likely to report difficulties in six health domains and that correcting for this reduces – but does not entirely eliminate - health differences by gender. By contrast, Grol-Prokopczyk et al. (2011) find American women to be more optimistic in their health assessment. For Chinese respondents, Bago d'Uva et al. (2008) find gender-specific reporting in only two out of six health domains – mobility and pain – but do not analyze the direction of the bias.

So far, most of the research has focused on western countries whereas relatively little is known for China. Figure 2.1 shows gender differences in self-assessed health (SAH) using Chinese elderly (50+) from WHO SAGE[1] Wave 1 data (described in Section 2.2 below). In contrast to what is shown in western countries, where female disadvantage in SAH disappears at older ages, e.g. in the early sixties in the US (Case & Paxson 2005, Figure 1), the disadvantage persists into very old age (80+) in China. This is likely to be related to the son preference long embedded in Chinese traditions, of which a particularly worrisome aspect is

[1] World Health Organization Study on global AGEing and adult health (www.who.int/healthinfo/sage/en/).

the fact that female education was given little importance, if not even opposed.[2] As a result, previous generations of Chinese women have suffered from very unequal opportunities to obtain an education while such gender gap in education has been eradicated - and sometimes even reversed - in western countries.

The current generation of elderly women in China is especially disadvantaged in education as they were born and raised during a time of poverty and social instability. Research has shown that, faced with the hardship of having more children than they can provide for, parents often invested more in sons at the expenses of daughters (Greenhalgh 1985; Parish & Willis 1993).[3] Due to such institutional and financial barriers, Chinese women obtained much less education than men.

To put the Chinese gender gap in perspective, we compare it to that in the US. This is done by first comparing data on the abovementioned Chinese elderly to data on American elderly from the first wave of AHEAD cohort, of the Health and Retirement Study (HRS).[4] We deliberately choose the oldest cohort in the HRS, born in 1923 or earlier (and aged 70+ at the time of the survey), while our Chinese dataset covers individuals born before 1961, because the rise in female education in the US long preceded a similar rise in China.

Gender differences (women-men) in the proportions of each education category are presented in Figure 2.2.[5] The female disadvantage in education is striking in China. The proportion of Chinese women with no formal education is more than 20 percentage points higher than that of men. In sharp contrast, the second bar in each education category shows little gender difference even though these Americans were born decades earlier than the majority of the Chinese sample. Figure 2.2 also shows the distribution of education for the original HRS cohort (born between 1931 and 1941 and aged 50+ at the time of the survey). This illustrates a shift to female advantage in education in the US, even for people who are on average still much older than the Chinese cohort.

Education is known to be a protective factor of people's health. It is also potentially very important because the advantage/disadvantage can accumulate over the years and affect health through various mechanisms (see e.g., Cutler & Lleras-Muney 2008 for a review). While female disadvantage in education is no longer considered an important explanation in western countries, it might still make a substantial contribution to the health gender gap in China. The contribution to the health gap not only depends on the gender gap in education but also on the relative benefits of education for women and men. If - as was found in the US (Ross et al. 2012; Ross & Mirowsky 2010) – education in China would benefit women's health more than men's, then this would dampen the gap. This is essentially an empirical question and

[2] Under the influence of the traditional doctrine that having no education was a virtue for women (*nvzi wucai bian shi de*), it was not until 1907 that females were officially permitted to enter the national education system (Lei et al. 1993, p. 261). But even so, girls still had to go to separate schools, received fewer years of education, and had to focus more on etiquette and needlework rather than modern sciences (Du 1995, p. 340).

[3] The one-child policy was only introduced in 1980. Before that, parents were allowed to have multiple children.

[4] Asset and Health Dynamics among the Oldest Old (AHEAD).

[5] Sampling weights are applied.

decomposition analysis can help to assess its importance in dampening (or widening) the gender health gap.

We further try to unpack the (potential) contribution of female disadvantage in education to the gender gap by investigating the roles of chronic conditions and impaired health functioning. Though partially biologically determined, these have also been found to be important pathways through which education affects health. For example, Herd et al. (2007) find that education positively affects health through postponing the onset of functional limitations and chronic conditions. Goldman and Smith (2002) find that education improves patients' subsequent self-assessed health through better management of illnesses, namely stronger adherence to treatments for diabetes and HIV. Elo (2009) summarizes the accumulating evidence in both sociology and medicine on how education affects health through one's exposure to – and ability to cope with – negative emotions (e.g., stress and anxiety).

We will consider contributions of chronic conditions and impaired health functioning separately because they are conceptually different, albeit interrelated. Chronic conditions are typically long-lasting diseases which often lead to impaired functioning, but not always. For example, arthritis and stroke may reduce mobility for some people but not for others. Similarly, functioning in daily life activities can be impaired for reasons that may or may not be related to underlying diseases. Diminished eyesight, for example, and reduced sleep quality can be due to cataracts and asthma, or simply a consequence of aging. Information on impaired health functioning over and above the commonly used chronic conditions provides supplemental health information, and reflects the heterogeneity in given conditions. While the diagnosis alone can reduce a person's assessment of own health, it is conceivable that the reduction also depends on the resulting functional limitations. The same lung disease is likely to affect SAH more adversely if it prevents a person from climbing a few flights of stairs without rest. Given the complementarity between disease diagnosis and functioning information, the WHO recommends using both to obtain a more meaningful picture of individuals' health.[6]

Our paper starts with a comparison of the characteristics of elderly Chinese men and women. Female disadvantage is clearly observed in SAH, in all domains of health functioning and in most conditions. Secondly, we test for gender-specific health reporting using anchoring vignettes for a wide range of health functioning domains, and find that homogeneous reporting by gender cannot be rejected. We run an additional test using a generalized ordered probit model for SAH which allows the estimated cut-points to differ between women and men, and find again no evidence of heterogeneous reporting by gender. Therefore, we conclude that the observed health gap in China is not simply a result of different reporting between men and women.

Next, we examine the contribution of education to the gender gap in SAH by applying an Oaxaca-Blinder type decomposition to a standard ordered probit model with reporting homogeneity, controlling for basic socio-demographic variables. We find that gender differences in socio-demographics explain about one third of the total gap. Female

[6] http://www.who.int/classifications/icf/training/icfbeginnersguide.pdf

disadvantage in education, in particular, makes the single largest contribution. Although women are found to benefit more from education, this advantage is not sufficient to narrow the gap.

Finally, indicators of chronic conditions and impaired health functioning are added sequentially. Our decomposition results show that the gender gap in health is then fully explained by gender differences in educational level, prevalence rates of chronic conditions and impaired functioning. At least two thirds of the original education contribution can be explained by gender differences in chronic conditions and functioning. Among the chronic conditions, women's higher rates in arthritis and angina make the largest contributions to the gap. These seem to work mostly through limiting mobility, increasing pain and discomfort, and causing sleep problems and a feeling of low energy.

We believe we make four important contributions to the current literature. First, China offers an interesting case study, not only because of its different level and pace of social and economic development, but also because of its deeply rooted son preference. Due to such preference, women were discriminated in the access to education in early years, which is associated with gender inequality in health in their later life.[7] Our study reveals this harmful component of the inequality with potentially important policy implications. Second, we formally test for gender-specific health reporting by using anchoring vignettes ratings in a HOPIT model to rule out reporting differences between women and men as an explanation of the female disadvantage in health. Third, the finding of homogeneous health reporting allows us to restrict the cut-points of the standard ordered probit to be identical for women and men, which simplifies the Oaxaca-Blinder decomposition based on ordered probit models for categorical outcome measures. To the best of our knowledge, this is the first paper to apply this type of decomposition to gender differences using an ordered response model. Fourth, we quantitatively assess the contribution of gender differences in education, chronic conditions and health functioning. In doing so, we reveal their complementarity in explaining the health gap, including the extent to which the contribution of education acts through chronic conditions and health functioning.

2.2 Data and Variables

Our data are obtained from SAGE, a multi-country survey conducted by the WHO to collect comprehensive longitudinal information on the health and wellbeing of adult populations in six countries.[8] In each country, it targets a nationally representative cohort of people aged 50 and above, with a smaller cohort of people aged 18 to 49 for comparison purposes. This

[7] Our findings are not inconsistent with those in Mu and Zhang (2011), where the authors find that son preference is likely to be a major contributing factor to the gender differences in the famine impacts on education, but not to the gender differences in the famine impacts on health. Additionally, it is unlikely that the famine cohort (those born between 1958-1961 and living mostly in rural areas before age 10) complicates our analysis as they only account for 2.6% of the sample.

[8] The six countries are China, Ghana, India, Mexico, Russian Federation and South Africa.

paper uses the 50+ sample of the Chinese data from Wave 1, which was conducted during 2007-2010. A multi-stage cluster sampling strategy was used to draw households from eight provinces that vary substantially in geography and socioeconomic level: Guangdong, Hubei, Jilin, Shaanxi, Shandong, Shanghai, Yunnan, and Zhejiang.[9]

The data provide an unusually broad range of health measures. Overall health (SAH) is assessed using a five-point scale with 1 corresponding to "Very good", 2 to "Good", 3 to "Fair", 4 to "Poor" and 5 to "Very poor". Detailed information is collected on a series of health conditions. In the main questionnaire, respondents are asked if they have ever been diagnosed with arthritis, stroke, angina, diabetes, lung disease, asthma, depression, hypertension and cataracts. For each condition, except for diabetes and hypertension, respondents are further asked if they have experienced some very typical symptoms of that condition. To minimize the possibility of under-diagnosis, we treat those who do not report the condition but do report at least one typical symptom of it as having the condition. The diagnosis of hypertension is supplemented with a three-time average of measured blood pressure being at or above 140/90 mmHg. We further incorporate information from the interviewer assessments on respondents' vision and hearing problems. The final list contains 10 conditions – arthritis, stroke, angina, diabetes, lung disease, asthma, hypertension, vision problems, hearing loss and depression (details on the definitions are included in Appendix Table 2.A.1). In descriptive statistics, we report the proportion of individuals with any chronic condition and with each particular condition.

Health functioning is assessed by asking respondents the level of difficulty they have in the following eight domains: mobility, self-care, pain and discomfort, cognition, interpersonal activities, sleep and energy, affect, and vision. Difficulty in each domain is mostly assessed on two aspects[10] using a five-point scale with 1 for "None", 2 for "Mild", 3 for "Moderate", 4 for "Severe" and 5 for "Extreme/cannot". For each domain, we use the two aspects that have corresponding vignettes, which are also the most important aspects. Dummies for having at least mild difficulty are created and used in descriptive statistics and the decomposition analysis in Section 2.3.2, while complete categorical variables are used in the test for gender-specific reporting in Section 2.3.1.

Socio-demographic variables including age, ethnicity, marital status, education and wealth quartiles are coded into categories. All non-Han ethnic groups are classified as minorities, which account for 8.49% of the total population.[11] Wealth quartiles are derived from the household ownership of durable goods, dwelling characteristics, and access to services such as improved water, sanitation, and cooking fuel.[12] After dropping observations with missing

[9]Further details about sampling method can be found in the China national report downloadable at http://www.who.int/healthinfo/sage/national_reports/en/.

[10] Out of eight domains, only three – self-care, interpersonal activities and vision –– have more than two aspects. For example, self-care domain is assessed on three aspects: self-care, grooming and staying by oneself for a few days.

[11] Communiqué of the National Bureau of Statistics of People's Republic of China on Major Figures of the 2010 Population Census [1] (No. 1). Source: http://www.stats.gov.cn/tjfx/jdfx/t20110428_402722253.htm

[12] See (He et al. 2012) for a more detailed explanation of the wealth variable.

values in any of the variables used in the analysis, we are left with 11855 observations.[13] Table 2.1 gives the weighted sample averages of variables used in the analysis.

In the sample, women are slightly older than men and more likely to be widowed. This is consistent with the lower mortality of women. Educational attainment is shown to be lower for women than for men. While this is not rare in developing countries, the magnitude of the difference is striking: as much as 33% of women have no education at all, while the rate for men is significantly lower at 13%. A female disadvantage of about 7 percentage points exists at the primary school level and persists to the level of high school or above.

Table 2.1 provides strong evidence of a female health disadvantage. Women report a lower average level of SAH and are more likely to rate their health as poor or very poor. They are also more likely to report having at least one chronic condition and, conditional on that, report having more conditions on average. Female excess appears in six of the ten chronic conditions (arthritis, angina, diabetes, hypertension, vision problems and depression), while the opposite is observed for stroke, lung disease and hearing loss. The pattern of female disadvantage is most consistent in health functioning – women are significantly more likely to have difficulty in every health domain.

2.3 Possible Explanations

2.3.1 Testing for gender differences in health reporting

To test if gender differences in reporting explain women's worse reported health, one would ideally want to use anchoring vignettes for SAH. However, because the SAH question is so broad, it seems impossible to design such vignettes. As an alternative, we use the anchoring vignettes for the eight health domains distinguished in SAGE. These offer a broad coverage of health functioning that ranges from physical mobility to cognition and mental wellbeing, and are likely to capture most of the underlying domains that a general health assessment is based on. Testing reporting differences by gender in these aspects provides an indirect test of reporting differences in SAH.

The vignette section in SAGE asks respondents to use the five response categories to evaluate hypothetical health conditions in the same way as they evaluate their own. Each health domain has five corresponding vignettes, and each respondent is randomly assigned to answer a set of 10 vignettes for two domains. Appendix 2.B provides an example of the questions for mobility domain and the corresponding vignette.

Observed female disadvantage in health functioning may be (partially) explained by gender-specific reporting if the mapping of true latent health to self-assessments differs systematically between women and men. In the context of ordered probit analysis, such

[13] Except for vignettes data that, as explained in Section 2.3.1, is only available for about one quarter of the sample for each domain. Missing values in vignette ratings are only dropped in the respective analysis presented there.

differences are reflected in gender-specific cut-points. While the cut-points are assumed constant in a standard ordered probit model, we can allow them to depend on observed personal characteristics using vignette ratings and the hierarchical ordered probit (HOPIT) model proposed by King et al. (2004).

The HOPIT model has two components. The vignette component uses information from the vignettes to reflect reporting behavior by modeling the cut-points as functions of individual characteristics. The rationale behind the use of vignettes to model reporting behavior is that, as vignettes represent fixed health states, any systematic correlation between the ratings and personal characteristics indicates heterogeneous reporting. This is then purged of self-reports in the second – own health – component by fixing the respective cut-points to be the same as those determined by the vignette component. In this own health component, one is then able to model the relationship between true latent health and observables.

Clearly, some assumptions are implied. The first one is vignette equivalence: there is no systematic variation in the perceived level of health represented by the vignettes (King et al. 2004). As in many other surveys, SAGE matches the gender of the vignette person to that of the respondent, by including gender specific names in the vignette description. Kapteyn et al. (2007) find that in the context of work disability the gender of the vignette person affects vignette rating. If this is true also in the context of health domains, the vignette equivalence may not hold. However, while the same health condition was rated as less work limiting for a female vignette person in Kapteyn et al. (2007) possibly because people assume less demanding work for women (as confirmed in Vermeer et al. 2016), it is less likely that the same description of a certain degree of health functioning (e.g. difficulties in moving around) is systematically considered to be worse for one gender than for the other. Nevertheless, we re-run the test using the subsample of vignettes 1, 2 and 5 of each health domain, where the assigned names are not gender-specific by pronunciation in Chinese.[14] Our conclusion holds. The second assumption is response consistency: respondents rate the vignettes according to the same criteria used when rating their own cases. There has been little and mixed formal evidence on these assumptions.[15]

Under the vignette equivalence assumption, the vignette component of the HOPIT model specifies V_{ik}^{*}, the latent health level of the condition described in vignette *k*, *k*=1,…,5, as perceived by individual *i,* as an exogenously determined true level, α_k, plus a random error term:

$$V_{ik}^{*} = \alpha_k + \varepsilon_{ik}^{v}, \qquad \varepsilon_{ik}^{v} \sim N(0, \sigma_v^2) \tag{2.1}$$

The observed rating for this vignette, v_{ik}, is then determined by its inclusion in one of the five intervals:

[14] The vignettes were read to the respondents. Names that are not gender-specific by pronunciation are Li Min, Wang Wei, Sun Xin and Yang Jie.

[15] On vignette equivalence, see Bago d'Uva et al. (2011), Kristensen and Johansson (2008), Murray et al. (2003), and Rice et al. (2011); on response consistency, see Bago d'Uva et al. (2011), Datta Gupta et al. (2010), and Van Soest et al. (2011).

$$v_{ik} = j \; if \; \tau_i^{j-1} \leq V_{ik}^* \leq \tau_i^j, j = 1, \ldots, 5, \tau_i^0 \leq \tau_i^1 \leq \cdots \leq \tau_i^5 \text{ and } \tau_i^0 = -\infty, \tau_i^5 = \infty, \forall, i, k \tag{2.2}$$

The four cut-points[16] are modeled as:

$$\tau_i^j = \beta_f^j Female_i + X_i \beta^j, \qquad j = 1, 2, 3, 4 \tag{2.3}$$

where $Female_i$ is a female dummy, and X_i is a set of variables including a constant term (normalized to zero for *j*=1), 5-year age groups (reference category: 50-54 years), a dummy for belonging to an ethnic minority, a dummy for being married, a dummy for living in urban areas, education categories (reference category: no education), wealth quartiles (reference category: the lowest wealth quartile), and province fixed effects. By allowing reporting behavior to depend additionally on socio-demographic variables and provinces, we avoid confounding the effect of gender on reporting with effects of other potential related factors. What we then test is the existence of reporting differences between women and men with the same socio-demographic characteristics and living in the same province.[17]

Using the individual specific cut-points determined by the vignette component, the self-assessment component is similar to an interval regression:

$$Y_i^* = \gamma_f Female_i + X_i \gamma + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$
$$y_i = j \; if \; \tau_i^{j-1} \leq Y_i^* \leq \tau_i^j \tag{2.4}$$

where X_i is defined as above (including a constant term), and σ^2 is normalized to 1 for identification. The vignette component and the self-assessment component of the HOPIT model are estimated jointly for efficiency (see e.g., Kapteyn et al. 2007). Since the cut-points depend on gender, and other individual characteristics, health effects are "purged" of any difference in reporting behavior. By comparing estimated effects of gender on health from the HOPIT and the standard ordered probit models, we can assess the role of gender-specific reporting.

Coefficient estimates of the female dummy in the cut-points are shown in Appendix Table 2.A.2, with test statistics for gender homogeneity in reporting, i.e., of the null hypothesis: $\beta_f^1 = \beta_f^2 = \beta_f^3 = \beta_f^4 = 0$. In general, the results show that gender homogeneity cannot be rejected for all but one case (*p*-value=0.051, depression).

Table 2.2 gives the coefficient estimates of the latent own health in Equation (2.4 from both models and their differences (ordered probit-HOPIT). After taking into account gender-specific reporting (i.e. moving from ordered probit to HOPIT), the coefficient estimate of the female dummy is reduced in 12 out of 16 cases. However, the reduction in magnitude is small,

[16] Unreported estimation results show that monotonicity is satisfied.

[17] We also estimate a model with only the female dummy in the cut-points to test for gender differences in reporting unconditional on socio-demographic and provincial controls. The results (available upon request) are similar.

and it does not take away the significance, except for learning and recognizing objects. When significant at 1% in the ordered probit (8 out of 16 cases), the coefficient remains significant at 1% in the HOPIT model. Overall, the comparison shows that gender-specific reporting does not appear to affect many aspects of health in China and is thus unlikely to be an explanation for the female disadvantage in health.

It is possible that the results for the eight health domains do not translate into overall SAH reporting. We therefore conduct an additional test using a generalized ordered probit model for SAH, which allows the cut-points to depend on gender to capture the potential differences in reporting. We are able to test a necessary, but not sufficient, condition for reporting heterogeneity. As we cannot make use of external (vignette) information on SAH, identification is achieved by excluding the gender dummy from one of the cut-points (Terza 1985). This means that the effect of gender is identified by the distances of each of the other cut-points to the first one.

The generalized ordered probit model takes the following form:

$$H_i^* = Z_i \times Female_i \psi + Z_i \times Male_i \xi + \varepsilon_i \tag{2.5}$$

$$H_i = j \; if \; \tau_i^{j-1} \leq H_i^* \leq \tau_i^j, j = 1, \ldots, 5 \tag{2.6}$$

where Z_i is set to be X_i as defined previously (including a constant term), and $Z_i \times Female_i$ and $Z_i \times Male_i$ are vectors of interaction terms of gender dummies with Z_i. This is done to allow for all the coefficients to differ between males and females. The first cut-point, τ_i^1 is normalized to zero for identification, and the next three cut-points, τ_i^2, τ_i^3, and τ_i^4, are modeled as a constant plus a female dummy.[18] Standard errors are clustered at the county level. Table 2.3 gives the cut-point estimates. The coefficients for the female dummy are all very small and highly insignificant, confirming homogeneous reporting of SAH by gender.

2.3.2 Decomposing the gap: gender differences in education, chronic conditions and health functioning

This section examines whether and how education explains the female disadvantage in SAH, with special attention paid to chronic conditions and health functioning. This is done by first regressing SAH on education and other socio-demographics controls, and then sequentially adding chronic conditions and health functioning to the explanatory variable list. The ordered probit model takes the form of Equations (2.5(2.6 with homogeneous reporting imposed by fixing all cut-points in Equation (2.6, i.e. by dropping the insignificant female dummies. Z_i is sequentially set to be i) X_i as defined previously (including a constant term), ii) (X_i, C_i) where

[18] This joint generalized ordered probit model with gender-specific health effects and cut-points corresponds to two separate standard ordered probits for males and females. It has however the advantage of making it easier to test, and impose, the restriction of common cut-points.

C_i is the set of ten chronic conditions,[19] and iii) (X_i, C_i, F_i) where F_i is the set of dummies for having at least mild difficulty in the eight health domains.[20] Using results from this model, the probability of reporting poor health (SAH=4 or 5) for women and men respectively is calculated as follows:

$$\Pr(\text{Poor Health}) = 1 - \Phi(\tau^3 - Z_i \times Female_i\psi - Z_i \times Male_i\xi) \tag{2.7}$$

where $\boldsymbol{\Phi(\cdot)}$ denotes the cumulative normal distribution function.

An extension of Oaxaca-Blinder decomposition method devised for non-linear models by Yun (2004) is applied to assess the contribution of explanatory variables to the female excess in the probability of reporting poor health. Decomposing a non-linear model is less straightforward as the mean value of the dependent variable is generally not equal to the value predicted at the mean value of regressors. Representing the probability of reporting poor health as $Y=F(Z\theta)$, the formula for decomposition at an aggregate level with women as the reference group becomes:[21]

$$\begin{aligned} \overline{Y^w} - \overline{Y^m} &= \overline{\Phi(Z^w\theta^w)} - \overline{\Phi(Z^m\theta^m)} \\ &= \underbrace{\left\{\overline{\Phi(Z^w\theta^w)} - \overline{\Phi(Z^m\theta^w)}\right\}}_{E} + \underbrace{\left\{\overline{\Phi(Z^m\theta^w)} - \overline{\Phi(Z^m\theta^m)}\right\}}_{C} \end{aligned} \tag{2.8}$$

The first difference in the last line of Equation (2.8 measures the *endowment effect* (labeled *E*).[22] A positive number shows the reduction of gender difference in SAH that would have occurred if women had men's characteristics. The second difference measures the *coefficient effect* (labeled *C*). A positive number shows the gap reduction that would occur if men had women's coefficients. Reporting homogeneity simplifies the calculation of counterfactual probabilities because men and women can be assumed to use the same cut-points.

Understanding the unique contribution made by each explanatory variable requires a detailed decomposition. One way of obtaining the individual endowment (coefficient) effect of a single variable is to replace its value for one group with that of the other group one by one. However, this sequential replacement is not only tedious, but in non-linear cases also

[19] We also use an alternative specification where C_i includes dummies for having a specific chronic condition or any combination of two conditions. This is to allow the effect of having two conditions to differ from a simple sum of the effects of these two conditions. The results (available upon request) are similar.

[20] Gender-specific reporting is tested with the latter two specifications of Z_i using the original Equations (2.5) and (2.6) before imposing homogeneous reporting. The coefficients for the female dummy are again all very small and highly insignificant.

[21] We use women as the reference because the counterfactual scenario where women have men's characteristics, e.g., higher educational level, seems more interesting than that where men have women's. Nevertheless, aware of the index problem, we also perform decomposition with men as reference (results available upon request). Our conclusions hold except that chronic conditions and health functioning fully explain the education contribution.

[22] We stick to the terminology used in the literature on decomposition methods, although our results should not be interpreted as causal effects.

sensitive to the order of switching.[23] We adopt a different procedure proposed by Yun (2004), where after a first-order Taylor linearization of the non-linear function at the regressor means, the weights of individual endowment and coefficient effects turn out to be simple and easy to implement. Applied to Equation (2.8, it results in the following individual weights for *E*:

$$W_{\Delta Z_k} = \frac{\theta_k^w(\overline{Z_k^w} - \overline{Z_k^m})}{\sum_{k=1}^{K}(\overline{Z_k^w} - \overline{Z_k^m})} \tag{2.9}$$

and the following individual weights for *C*:

$$W_{\Delta \theta_k} = \frac{\overline{Z_k^m}(\theta_k^w - \theta_k^m)}{\sum_{k=1}^{K}\overline{Z_k^m}(\theta_k^w - \theta_k^m)} \tag{2.10}$$

where $\sum_{k=1}^{K} W_{\Delta Z_k} = \sum_{k=1}^{K} W_{\Delta \theta_k} = 1$.

In essence, this method assigns each covariate a weight that is equal to its proportional contribution to the total endowment or coefficient effect in a linear regression. Weights obtained in this way are free from the path dependence problem and are invariant to a change in the scale of the covariates. The gender difference in the predicted probabilities of reporting poor SAH can then be expressed as a sum of individual contributions of all covariates.

One problem with the detailed decomposition is that the coefficient effects are not invariant to the choice of omitted category when categorical variables are present. This is because the group difference in the constants captures both the difference between true group membership and the difference between the omitted categories of the two groups. It is impossible to distinguish the two parts. Consequently, a change in the omitted category of one variable almost always leads to a reallocation of coefficient effects between this variable and the constant.[24] For example, if another age category is omitted instead of the current 50-54 years, the total coefficient effects of age and the coefficient effect of the constant will change. This is not desirable as our choice of the omitted age category, ethnic group, marital status, urban/rural residence, educational level, wealth quartile and province is rather arbitrary. To "solve" this problem, different approaches have been proposed (Gardeazabal & Ugidos 2004; Suits 1984; Yun 2005).

We adopt the intuitive and convenient method suggested by Yun (2005), which expresses the coefficients of a categorical variable as deviations from the coefficients' grand mean. Specifically, the grand mean is calculated as $\bar{\vartheta} = \sum_{k=1}^{j} \vartheta_j / j$, where ϑ_j are the coefficients of the categorical variable with $\vartheta_1 = 0$ for the omitted category. This mean is first deducted from the coefficient of each category including the omitted one, and then added to the constant to maintain mathematical equality. After this normalization, the individual coefficient

[23] See page 1137 of Ham et al. (1998) for a discussion of this path-dependency problem.
[24] It should be noted that the coefficient effects of other variables are not affected.

effect of each category can be calculated and no longer depends on the choice of omitted category.[25]

Table 2.4 gives the results from the ordered probit model with three sets of explanatory variables and Table 2.5 gives the corresponding gross and detailed decompositions.[26] The upper panel of Table 2.5 shows that the predicted probability of reporting poor health for women and men are stable across different specifications: about 24.1% for women and 18.7% for men, giving a total difference of roughly 5.4 percentage points.

In specification (1) of Table 2.5, differences in endowments explain 31.4% of the total gender gap. Among all socio-demographic variables, education makes the single largest contribution. If women would have the same level of educational attainment as men, the female excess in the probability of reporting poor SAH would be reduced by 1.3 percentage points (23.2%). Moreover, although the education coefficient estimates for women in specification (1) of Table 2.4 indicate that women benefit *more* from education than men in China (as in the US), the coefficient effect of education does not reduce the female disadvantage in health by much: the gender gap would be only 1.9% larger if men and women obtained the same health benefit from education. Recall that the coefficient effect depends on both the difference between coefficients and the value of the variable, which in this case is the educational level of men. The small coefficient effect of education is partly a result of the overall low educational attainment among Chinese elderly.

Specification (2) adds chronic conditions to specification (1) of Table 2.5. This results in a reduction of the education contribution from 23.2% to 15.0%. This is not surprising since education is likely to affect health through the onset of chronic conditions, as explained in the introduction. The addition also increases the total endowment effect to 54.9%, largely by reducing the "unexplained" contribution by the constant. One reason for this is that chronic conditions also reflect biological endowments that are not affected by education. Before controlling for chronic conditions, only socio-demographic sub-groups of men and women were defined. Biological differences between men and women within the same sub-group were then captured in the constants, and appear as unexplained in the detailed decomposition. Adding chronic conditions divides men and women into finer sub-groups with higher gender homogeneity, leaving a smaller unexplained gender difference in the constants.

In specification (2), gender differences in the prevalence rates of chronic conditions together explain 42.7% of the total gap, with considerable variation across conditions. The largest individual contribution is made by the much higher prevalence of arthritis among women, 1.2 percentage points (22.2%), followed by angina, 0.7 percentage points (12.3%). With some offsetting effects, the gender-specific health impacts of these conditions explain a negligible portion of the total female health disadvantage.

---

[25] Note that the omitted categories of chronic conditions and health functioning are not chosen arbitrarily. Thus, normalization is not needed.

[26] To apply sample weights in the decomposition, we follow the strategy in Pylypchuk and Selden (2008) to approximate an unweighted sample by generating 100 observations for the highest-weighted case and proportionately fewer replications for less weighted individuals given the weights are largely continuous over the range from 1485.735 to 27216.64.

Health conditions obviously have an effect on an individual's capability to function. Adding health functioning in specification (3) halves the education contribution and more than halves the contribution of chronic conditions.[27] At the same time, female disadvantage in health is now fully explained by endowment effects.[28] Among the eight health functioning domains, mobility, pain and discomfort, and sleep and energy are the top three in terms of contribution, accounting for 61% of the total gender gap. Correspondingly, a sizable reduction in the endowment effects is observed for conditions such as arthritis and angina. This is not surprising because pain and discomfort are typical symptoms of arthritis and angina. They do not only cause suffering, but can also reduce a person's mobility and the amount of quality sleep, which makes it difficult for a person to feel refreshed and energetic.

2.4 Conclusion

Female disadvantage in health appears to be particularly problematic for Chinese elderly women. They have been exposed to unequal opportunities in education since their childhood due to the widespread preference for sons. Their lower socioeconomic status, often accompanied by a lack of effective pension and health care arrangements, also leaves them with little protection against the causes and consequences of poor health (Baeten et al. 2013). Moreover, women lose access to the support and resources provided by their husbands when they become widowed. Adequate policy response requires a better understanding of the factors explaining the observed difference. Ideally, this would be based on the identification of the causal effects of education on later life health and on the mechanisms linking the two. In the absence of exogenous variation in education that could be exploited to do this, we examine whether it is primarily the differing characteristics rather than the different partial associations between these characteristics and health that explain the female disadvantage in health.

After ruling out gender-specific reporting as an explanation, we find that female disadvantage in education indeed plays an important role in explaining the gender gap in health in China. While women are found to obtain greater health benefits from completing primary school education and above, this does not lead to any substantial reduction in the health gap, mainly because of the large population share with less than primary school education. The decomposition analysis brings out the substantial quantitative importance of gender differences in education as an explanation for the female disadvantage in health.

Controlling for chronic conditions reduces the gap left unexplained by socio-demographic variables only (i.e. the coefficient effect of the constant) from about 70% to about 30%. Substantial additional explanatory power derives from biological male-female

[27] Effects of health aspects are summed at the domain level to improve readability.

[28] [In unreported specification (4)] Including only health functioning and not chronic conditions leaves the female health disadvantage still fully explained by endowment effects, with the education contribution slightly larger than in specification (3) and the health functioning contribution increased.

differences, such as women's higher susceptibility to conditions like arthritis. Further controlling for health functioning leads to almost full explanation of the gender gap. This may derive from the discomfort and impaired functioning consequences of under-diagnosis and unlisted health conditions, which impair self-perceived health. Adding functioning also absorbs part of the contribution of chronic conditions. This illustrates how conditions affect health, namely by limiting mobility, increasing pain and discomfort, and causing sleep problems and a feeling of low energy. But the sizable remaining contribution does suggest that the awareness of the presence of chronic conditions can lower one's assessment of own health even without impaired functioning.

The inclusion of both chronic conditions and health functioning reduces, and thereby explains, the contribution of female disadvantage in education by at least two thirds.[29] This is consistent with findings from previous literature that education affects health through the onset of chronic conditions and health functioning. However, one third of the education contribution still remains in the full model with women as reference, suggesting that for women education also operates through other channels not considered here.

In sum, our results suggest that the gender health gap in China does not merely originate from either reporting or biological male-female differences. An important additional contribution derives from the female disadvantage in education. While one can expect the male-female education differences to shrink with China's rapid economic and social development, it is likely that women's health disadvantage will persist for some time into the near future, especially in rural areas, given the relatively slow development of rural education and the persistence of son preference.[30] While our results should not be interpreted as causal, if such evidence could be brought to bear in the future, it would suggest that greater investment in rural education and positive policy discrimination in favor of girls are worth considering. In fact, there is still much scope for improvement in this area. For example, Ningxia province released a regulation in 2014 demanding that girls from rural areas be given priority in high school enrollment, and in getting tuition and accommodation fee waivers.[31] Such policies seem well guided if China hopes to reduce or even eliminate the large female health disadvantage in the future.

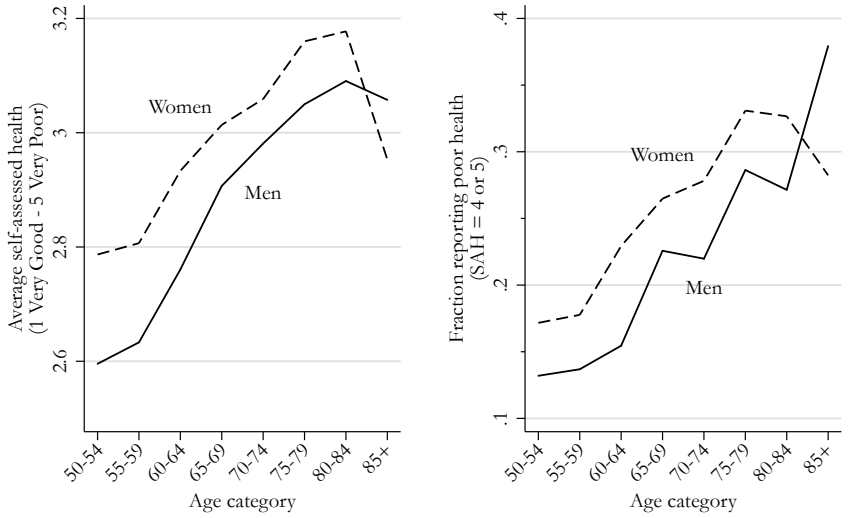
[29] The education contribution is fully explained by chronic conditions and health functioning with men as reference group in the decomposition.

[30] According to the speech given by the former premier Wen Jiabao at a national conference, 431 (out of 2861) counties still had not achieved universal nine-year compulsory education by 2003, and son preference was still a problem in promoting female education. Source: http://news.xinhuanet.com/zhengfu/2003-10/30/content_1150774.htm.

[31] http://www.nxfp.gov.cn/fpxw/fpyw/12544.htm

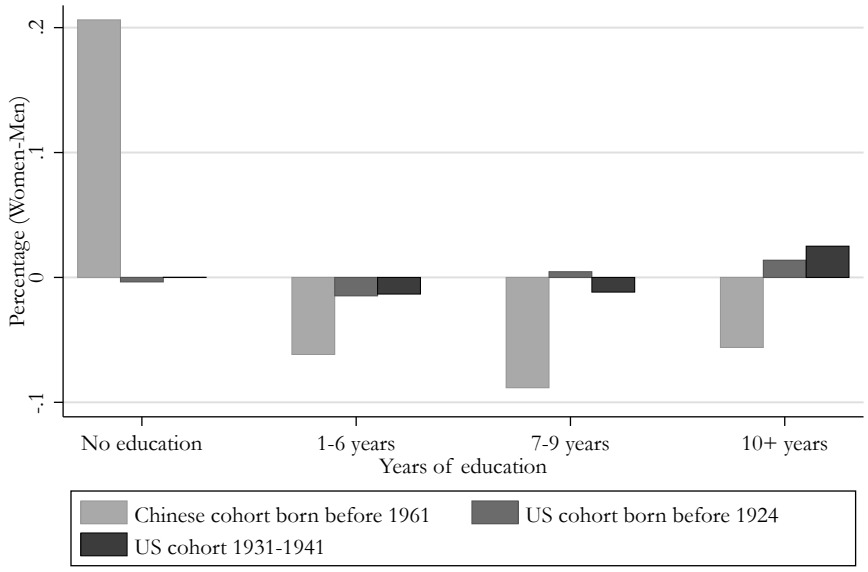
Figures

Figure 2.1. Self-assessed health (SAH) for men and women in China.



Data Source: WHO SAGE Wave 1.

Figure 2.2. China-US comparison of gender differences in education.



Data Source: Chinese data from SAGE Wave 1, US data from RAND HRS Data, Version L: cohort born before 1924 from AHEAD 1993, and cohort 1931-1941 from HRS 1992.

Tables

Table 2.1. Weighted sample means for women and men.

	Women	Men	Difference
Age	62.92	62.07	0.855***
50-54 years	0.21	0.23	-0.016*
55-59 years	0.22	0.24	-0.018*
60-64 years	0.17	0.18	-0.012
65-69 years	0.15	0.15	-0.001
70-74 years	0.12	0.10	0.023***
75+ years	0.13	0.11	0.025***
Minority	0.01	0.01	0.002
Married	0.80	0.91	-0.107***
Widowed	0.18	0.06	0.115***
Urban	0.51	0.43	0.083***
No education	0.33	0.13	0.204***
Less than primary school	0.20	0.19	0.005
Primary school	0.18	0.26	-0.076***
Secondary school	0.17	0.24	-0.070***
High school or above	0.13	0.19	-0.063***
Lowest wealth quartile	0.21	0.21	-0.001
2nd wealth quartile	0.24	0.24	0.008
3rd wealth quartile	0.27	0.29	-0.015
Highest wealth quartile	0.28	0.27	0.009
SAH	2.94	2.80	0.144***
Poor SAH (SAH=4 or 5)	0.24	0.19	0.050***
Chronic conditions	0.82	0.76	0.060***
No. of chronic conditions	2.30	2.13	0.176***
Arthritis	0.51	0.39	0.117***
Stroke	0.04	0.05	-0.009**
Angina	0.17	0.12	0.053***
Diabetes	0.08	0.05	0.021***
Lung disease	0.12	0.14	-0.027***
Asthma	0.15	0.14	0.010
Hypertension	0.41	0.39	0.023**
Vision problems	0.36	0.26	0.094***
Hearing loss	0.05	0.07	-0.018***
Depression	0.03	0.02	0.009***
Moving around	0.24	0.18	0.053***
Vigorous activity	0.71	0.60	0.108***
Self-care	0.09	0.07	0.018***
Grooming	0.08	0.07	0.012**
Bodily pain	0.55	0.44	0.106***
Bodily discomfort	0.56	0.47	0.096***
Remembering	0.53	0.45	0.087***
Learning	0.63	0.54	0.089***
Interpersonal relations	0.10	0.09	0.018**
Dealing with conflicts	0.12	0.09	0.026***
Sleep	0.45	0.32	0.129***
Feeling refreshed	0.43	0.30	0.126***
Depression	0.21	0.16	0.052***
Anxiety	0.21	0.16	0.051***
Recognizing objects	0.41	0.33	0.089***
Recognizing people	0.62	0.57	0.047***
N	6,335	5,520	

Note: Difference=Women-Men. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 2.2. Coefficients of the female dummy in latent health equation of ordered probit and HOPIT models.

	Ordered probit	HOPIT	Difference	*N*
Moving around	0.064	0.045	0.018	2888
	(0.071)	(0.074)		
Vigorous activities	0.181***	0.173***	0.008	2888
	(0.055)	(0.063)		
Self-care	0.121	0.093	0.028	2908
	(0.090)	(0.093)		
Grooming	0.024	0.001	0.023	2908
	(0.087)	(0.088)		
Bodily pains	0.221***	0.238***	-0.016	3064
	(0.044)	(0.048)		
Bodily discomfort	0.228***	0.244***	-0.016	3064
	(0.039)	(0.041)		
Remembering	0.048	0.052	-0.005	2906
	(0.054)	(0.054)		
Learning	0.091*	0.082	0.009	2906
	(0.054)	(0.056)		
Personal relationships	-0.022	-0.014	-0.008	3064
	(0.098)	(0.093)		
Dealing with conflicts	0.038	0.025	0.012	3064
	(0.091)	(0.086)		
Sleep	0.276***	0.262***	0.013	2950
	(0.058)	(0.056)		
Not feeling refreshed	0.274***	0.233***	0.041	2950
	(0.065)	(0.065)		
Depression	0.225***	0.184***	0.041	2888
	(0.065)	(0.069)		
Anxiety	0.219***	0.187***	0.032	2888
	(0.060)	(0.062)		
Recognizing people	0.296***	0.280***	0.016	2950
	(0.053)	(0.053)		
Recognizing objects	0.122**	0.073	0.049	2950
	(0.054)	(0.057)		

Notes: Difference=Ordered Probit - HOPIT. Models estimated using maximum likelihood with standard errors in parentheses clustered at the county level. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 2.3. Cut-points from the generalized ordered probit model.

	Cut-point 2	Cut-point 3	Cut-point 4
Constant	1.512***	2.819***	4.123***
	(0.049)	(0.064)	(0.083)
Female	-0.016	0.013	-0.012
	(0.046)	(0.055)	(0.080)

Notes: Model estimated using maximum likelihood with standard errors in parentheses clustered at the county level. *** p<0.01.

Table 2.4. Estimation results from probit models.

	(1)		(2)		(3)	
	Women	Men	Women	Men	Women	Men
55-59 years	0.030	0.044	-0.003	0.019	-0.043	-0.030
60-64 years	0.131***	0.149**	0.034	0.058	-0.076	-0.069
65-69 years	0.224***	0.333***	0.035	0.201***	-0.082	0.047
70-74 years	0.274***	0.441***	0.087*	0.273***	-0.159**	0.015
75+ years	0.354***	0.478***	0.094	0.243***	-0.282***	-0.146*
Minority	0.085	-0.027	0.002	-0.074	-0.015	-0.016
Married	-0.072	-0.034	-0.048	-0.043	-0.038	-0.022
Urban	-0.093*	0.081*	-0.191***	0.010	-0.029	0.099*
Less than primary school	0.013	-0.078	0.057	-0.032	0.087*	-0.018
Primary school	-0.147***	-0.083	-0.092*	-0.011	-0.065	0.037
Secondary school	-0.151***	-0.094	-0.113**	-0.039	-0.028	0.021
High school or above	-0.298***	-0.211***	-0.218***	-0.149**	-0.139**	-0.067
2nd wealth quartile	-0.162***	-0.160***	-0.216***	-0.159***	-0.165***	-0.112**
3rd wealth quartile	-0.187***	-0.294***	-0.231***	-0.315***	-0.158***	-0.212***
Highest wealth quartile	-0.339***	-0.449***	-0.355***	-0.422***	-0.190***	-0.218***
Arthritis			0.364***	0.492***	0.119***	0.204***
Stroke			0.455***	0.614***	0.202**	0.373***
Angina			0.444***	0.365***	0.272***	0.263***
Diabetes			0.476***	0.492***	0.423***	0.447***
Lung disease			0.333***	0.362***	0.277***	0.282***
Asthma			0.169**	0.167**	0.101	0.070
Hypertension			0.167***	0.104**	0.158***	0.070*
Vision problems			0.165***	0.090*	0.098**	0.023
Hearing loss			0.366***	0.217***	0.162	0.139*
Depression			0.522***	0.407***	0.166*	0.159
Moving around					0.420***	0.426***
Vigorous activity					0.491***	0.478***
Self-care					0.314**	0.176
Grooming					0.091	0.173
Bodily pain					-0.071	0.182***
Bodily discomfort					0.476***	0.402***
Remembering					0.074	0.127***
Learning					0.098	0.078
Interpersonal relations					0.052	0.020
Dealing with conflicts					0.028	-0.105
Sleep					0.002	0.015
Feeling refreshed					0.187***	0.182***
Depression					0.047	0.048
Anxiety					0.262***	0.113
Recognizing objects					-0.028	-0.074
Recognizing people					0.064	0.079
Constant	2.266***	2.040***	2.023***	1.792***	1.520***	1.429***
Provincial controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Models estimated using maximum likelihood with standard errors in parentheses clustered at the county level. *** p<0.01; ** p<0.05; * p<0.1.

Table 2.5. Decomposition of female excess in the probability of reporting poor SAH.

	(1)		(2)		(3)	
Probability of reporting poor SAH		Percent		Percent		Percent
Predicted mean in women	0.242		0.240		0.240	
Predicted mean in men	0.187		0.186		0.187	
Female excess	0.054		0.054		0.052	
Decomposition into						
endowment effects	0.017	31.4%	0.030	54.9%	0.053	100.5%
coefficient effects	0.037	68.6%	0.024	45.1%	0.000	-0.5%
Endowment effects	Absolute	Percent	Absolute	Percent	Absolute	Percent
Age	0.004	7.2%	0.001	2.0%	-0.002	-3.9%
Ethnicity	0.000	0.1%	0.000	0.0%	0.000	0.0%
Marital status	0.002	4.4%	0.001	2.6%	0.001	1.8%
Residence	-0.002	-4.4%	-0.004	-8.1%	-0.001	-1.0%
Education	0.013	23.2%	0.008	15.0%	0.004	7.1%
Wealth	0.000	-0.7%	0.000	-0.6%	0.000	-0.2%
Arthritis			0.012	22.2%	0.003	6.1%
Stroke			-0.001	-2.1%	0.000	-0.8%
Angina			0.007	12.3%	0.003	6.4%
Diabetes			0.003	5.1%	0.002	3.9%
Lung disease			-0.003	-4.7%	-0.002	-3.3%
Asthma			0.000	0.9%	0.000	0.5%
Hypertension			0.001	2.0%	0.001	1.6%
Vision problems			0.004	8.0%	0.002	4.1%
Hearing loss			-0.002	-3.4%	-0.001	-1.3%
Depression			0.001	2.4%	0.000	0.7%
Total effect of conditions			0.023	42.7%	0.009	17.8%
Mobility					0.017	33.3%
Self-care					0.002	3.0%
Pain and discomfort					0.009	16.9%
Cognition					0.004	6.7%
Interpersonal activities					0.000	0.7%
Sleep and energy					0.005	10.5%
Affect					0.004	7.0%
Vision					0.000	0.2%
Total effect of health domains					0.041	78.3%
Province	0.001	1.8%	0.001	1.2%	0.000	0.7%
Coefficient effects						
Age	0.014	25.5%	0.013	23.3%	0.000	-0.5%
Ethnicity	-0.011	-19.4%	-0.005	-9.5%	0.000	0.0%
Marital status	-0.003	-5.5%	0.000	-0.5%	0.000	0.0%
Residence	0.003	4.6%	0.002	3.8%	0.000	-0.1%
Education	-0.001	-1.9%	-0.001	-1.8%	0.000	0.1%
Wealth	0.001	2.2%	0.001	1.3%	0.000	0.0%
Arthritis			-0.007	-12.7%	0.000	0.2%
Stroke			-0.001	-2.0%	0.000	0.1%
Angina			0.001	2.3%	0.000	0.0%
Diabetes			0.000	-0.2%	0.000	0.0%
Lung disease			-0.001	-1.0%	0.000	0.0%
Asthma			0.000	0.1%	0.000	0.0%
Hypertension			0.003	6.2%	0.000	-0.2%
Vision problems			0.003	5.0%	0.000	-0.1%
Hearing loss			0.001	2.5%	0.000	0.0%
Depression			0.000	0.6%	0.000	0.0%
Total effect of conditions			0.000	0.7%	0.000	-0.1%
Mobility					0.000	0.0%
Self-care					0.000	0.0%
Pain and discomfort					0.000	0.5%
Cognition					0.000	0.1%
Interpersonal activities					0.000	-0.1%
Sleep and energy					0.000	0.0%
Affect					0.000	-0.2%
Vision					0.000	0.0%
Total effect of health domains					0.000	0.2%
Province	-0.003	-5.4%	-0.002	-3.5%	0.000	0.1%
Constant	0.037	68.5%	0.017	31.3%	0.000	-0.2%

Appendix 2.A

Table 2.A.1. Definition of chronic conditions.

Condition	Definition
Arthritis	Ever diagnosed with arthritis; pain, aching, stiffness or swelling in or around the joints which were not related to an injury and lasted for more than a month; stiffness in the joint in the morning after getting up from bed, or after a long rest of the joint without movement; back pain during the last 30 days
Stroke	Ever diagnosed with stroke; ever sudden onset of paralysis or weakness in arms or legs on one side of body for more than 24 hours; ever sudden onset of loss of feeling on one side of body for more than 24 hours, without anything having happened immediately before
Angina	Ever diagnosed with angina; pain or discomfort in chest when walking uphill or hurry; pain or discomfort in chest when walking at an ordinary pace on level ground
Diabetes	Ever diagnosed with diabetes
Lung disease	Ever diagnosed with chronic lung disease (emphysema, bronchitis, COPD); shortness of breath at rest; coughing or wheezing for ten minutes or more at a time; coughing up sputum or phlegm for most days of the month for at least 3 months
Asthma	Ever diagnosed with asthma; attacks of wheezing or whistling breathing; attack of wheezing that came on after stopping exercising or some other physical activity; feeling of tightness in chest; waking up with a feeling of tightness in chest in the morning or any other time; an attack of shortness of breath that came on without obvious cause when not exercising or doing some physical activity
Hypertension	Ever diagnosed with hypertension; three-time average measured blood pressure above 140/90 mmHg
Vision problems	Diagnosed with cataract in the last 5 years; cloudy or blurry vision; vision problems with light, such as glare from bright lights, or halos around lights; other vision problem observed by interviewer
Hearing loss	Hearing problem observed by interviewer
Depression	Ever diagnosed with depression; sad, empty or depressed for more than 2 weeks; lost interest in most things usually enjoy such as personal relationships, work or hobbies/recreation for more than 2 weeks; feeling energy decreased or tired all the time for more than 2 weeks

Note: 12-month reference period unless noted otherwise.

Table 2.A.2. Coefficients of the female dummy and test for reporting homogeneity.

	Coefficient of the female dummy in				*p*-values for
	Cut-point 1	Cut-point 2	Cut-point 3	Cut-point 4	homogeneity
Moving around	-0.025	-0.003	0.014	0.073*	0.286
	(0.028)	(0.023)	(0.030)	(0.044)	
Vigorous activities	-0.017	-0.020	0.023	0.028	0.705
	(0.031)	(0.026)	(0.026)	(0.030)	
Self-care	-0.031	-0.010	-0.008	0.019	0.659
	(0.026)	(0.018)	(0.018)	(0.020)	
Grooming	-0.026	-0.021	-0.020	-0.009	0.821
	(0.023)	(0.026)	(0.024)	(0.021)	
Bodily pains	0.032	-0.000	-0.043	-0.044	0.379
	(0.026)	(0.025)	(0.036)	(0.034)	
Bodily discomfort	0.019	0.025	-0.043	-0.019	0.224
	(0.026)	(0.026)	(0.036)	(0.051)	
Remembering	0.011	-0.012	0.001	0.004	0.976
	(0.039)	(0.029)	(0.032)	(0.037)	
Learning	-0.000	-0.024	-0.026	-0.001	0.815
	(0.037)	(0.025)	(0.031)	(0.035)	
Personal relationships	0.008	-0.003	0.012	-0.058	0.550
	(0.029)	(0.030)	(0.034)	(0.054)	
Dealing with conflicts	-0.016	-0.015	0.021	-0.030	0.663
	(0.029)	(0.030)	(0.031)	(0.055)	
Sleep	-0.029	0.025	0.030	-0.036	0.120
	(0.030)	(0.029)	(0.026)	(0.047)	
Not feeling refreshed	-0.059*	0.023	0.022	-0.056	0.103
	(0.033)	(0.030)	(0.030)	(0.043)	
Depression	-0.053*	0.006	0.049*	0.011	0.051
	(0.028)	(0.034)	(0.028)	(0.039)	
Anxiety	-0.041*	0.006	0.038	0.015	0.198
	(0.025)	(0.033)	(0.031)	(0.036)	
Recognizing people	-0.017	-0.024	-0.029	-0.022	0.877
	(0.033)	(0.036)	(0.030)	(0.033)	
Recognizing objects	-0.056	-0.052	-0.007	-0.008	0.572
	(0.040)	(0.044)	(0.040)	(0.032)	

Note: Model estimated using maximum likelyhood with standard errors in parentheses clustered at the county level. * $p<0.1$.

Appendix 2.B

Questions on Mobility Domain

Overall in the last 30 days, how much difficulty did you have…

… with moving around?

… in vigorous activities ('vigorous activities' require hard physical effort and cause large increases in breathing or heart rate)?

Corresponding Vignette

[name] is able to walk distances of up to 200 meters without any problems but feels tired after walking one kilometer or climbing up more than one flight of stairs. He has no problems with day-to-day physical activities, such as carrying food from the market.

Overall in the last 30 days, how much difficulty did…

[name] have with moving around?

[name] have in vigorous activities ('vigorous activities' require hard physical effort and cause large increases in breathing or heart rate)?

CHAPTER 3

Only Children and Their Long-term Effects on Parental Mental Wellbeing in China

Population control policies are often employed by developing countries with the aim of reducing poverty. However, little is known about how reduced fertility affects the mental wellbeing of parents in later life. We estimate such long-term causal effects of having only one child using nationally representative data from the China Health and Retirement Longitudinal Study (CHARLS). As fertility level is likely endogenous, we construct an instrument exploiting variation in regional enforcement of the one-child policy (yes/no) and in women's age at the time of the enforcement. Our IV results show that having only one child did not negatively affect parents' mental wellbeing in later life, as measured by depressive symptoms, cognitive skills and life satisfaction. Further analysis on potential mechanisms suggests that having only one child did not reduce parents' chances of having a child living in the same household or close by, or receiving transfers from a child. Instead, it significantly increased parents' chances of seeing a child at least once a month.

3.1 Introduction

The Chinese population is aging at a faster pace than most of the world due to the combination of an increasing life expectancy and reduced fertility. The share of the population aged 65 and above reached 7.5% in 2005, and is projected to grow to 14.2% in 2025, and to 27.6% by 2050 (United Nations Population Division 2015). The mental wellbeing of the elderly thus becomes a salient social question. Recent research has documented a high proportion of people suffering from depressive symptoms (30% of men and 43% of women) among people aged 45 and older in China (Lei et al. 2014). A meta-analysis showed that the prevalence of mild cognitive impairment, a pre-dementia syndrome, was 12.7% among the 60+ population (Nie et al. 2011). Depressive symptoms and impaired cognitive skills later in life can have serious economic and health consequences. This is especially true in a developing country such as China where healthcare and pension systems are immature and institutional support in decision-making is lacking for the elderly. Thus, a better understanding of the driving forces behind these trends is much needed.

Reduced fertility is one of the potential forces that drove up the prevalence of depressive symptoms and impaired cognition. Since 1980, the one-child policy has created a multitude of parents with only one child in China. There are several economic and cultural reasons to expect such a restriction to make parents less happy. Chinese parents traditionally rely on children for old-age support[1] and Chinese culture prefers a large family. A variety of fertility preference surveys conducted in the 1980s confirmed that few people preferred only one child in both urban and rural areas (Whyte & Gu 1987).[2] Besides the dissatisfaction from not being allowed to have the desired number of children, reduced fertility may also contribute to depression and cognitive decline by reducing parent-child interactions. Much research has shown that social engagement (including interactions with children) is associated with less depressive symptoms (e.g., Glass et al. 2006) and better cognitive skills (e.g., Fratiglioni et al. 2004). In particular, good social support (mainly through intergenerational co-residence) has been offered as one possible explanation for the much lower prevalence of depression among older Chinese in the 1980s and 1990s (3.86%) than in Western Europe (12%) (Chen et al. 1999).[3]

While the one-child restriction seems likely to exert some immediate negative effects on parents, household members can consider using various coping strategies to mitigate these effects. For example, happiness research suggests that income relative to that of the reference group is more important in predicting wellbeing than absolute income (Knight et al. 2009). It is possible that parents with only one child use other one-child families as the reference and

[1] Old-age support to parents is not only dictated by the Confucian tradition, but also by the Marriage Law of China.

[2] The majority preference was two children, but preference for three or more children was also common among people belonging to minorities, living in mountainous regions or engaged in high-risk professions (e.g., fishermen in coastal regions).

[3] The prevalence of depressive mood was 14.81% in Chen et al. (1999), much lower than that in Lei et al. (2014). Another meta-analysis also documented an increasing prevalence of depressive symptoms among older Chinese from 1987 to 2012 (Li et al. 2014).

do not therefore experience a substantial reduction in life satisfaction. Additionally, some papers provide evidence of only children receiving more education as a result of the quality-quantity trade-off (Li et al. 2008; Rosenzweig & Zhang 2009), and some others show only children to be more responsive to parents' emotional needs (Deutsch 2006). Both of these mechanisms could lead to more cognitively engaging or more satisfying parent-child interactions, offering more – not less – protection against depression, and cognitive decline (Berkman 2000).

Therefore, empirical analysis is needed to determine the net impact of having only one child on parents' mental wellbeing and to explore the potential coping strategies. It is also relevant to analyze mothers and fathers separately because they can be differentially affected. For example, failing to have a son to carry on the family name is likely to be less dissatisfying for the mother than the father as it is the father's lineage that is at stake. On the other hand, the mother enjoys personal health benefits as reduced fertility directly reduces her risk of maternal mortality and postpartum depression. Reduced fertility has also been credited with contributing to gender equality in social life, e.g., better job opportunities and career advancement for women (Zeng & Hesketh 2016). Such social benefits could lead to an improvement in mothers' mental wellbeing (Allen et al. 2014). As some research found fertility to affect marriage breakdown (Cáceres-Delpiano & Simonsen 2012),[4] we additionally examine the subsample of mothers whose husbands are present as a comparison to the full sample of mothers.

In this paper, we examine specifically how having only one child (compared to two or more children) affects parents' depressive symptoms, cognitive skills, and life satisfaction after the age of 45. Identification of causal effects is challenging because both the number of children and parents' mental wellbeing in later life are affected by unobserved community-level and household-level characteristics, such as pension generosity, economic conditions, and the preference for more children. To overcome the endogeneity problem, we exploit the variation in the number of children, specifically the occurrence of only children, induced by regional variation in the enforcement of the one-child restriction. As will be elaborated on in Section 3.2, whether a community enforced the restriction can be predicted to a large extent from its location and administrative characteristics (e.g., whether an autonomous region or not, whether urban or rural). Enforcement also depended on local cadres'[5] effort in implementation, which is arguably exogenous to parents' mental wellbeing in later life. Additionally, we are able to exploit the variation in women's age at the time of enforcement, which implies differential policy exposure. In particular, our instrument for having only one child is constructed as a function of whether the local community enforced the one-child restriction and mother's accumulated years of exposure to the restriction.

We use nationally representative data from the China Health and Retirement Longitudinal Study (CHARLS). Our IV estimates show very small and mixed effects of only children on

[4] Increased fertility was found to increase marriage breakdown in Cáceres-Delpiano & Simonsen (2012). However in China, having only one child, specifically a daughter, is more likely to *increase* marriage breakdown given the son preference.

[5] Cadre (*ganbu*) refers to a public official holding a responsible or managerial position in party and government.

parents' depressive symptoms, cognitive skills and life satisfaction in later life. This leads us to a general conclusion of no effect. As discussed above, this might be due to coping from household members. IV estimates on parent-child interactions indeed show that having only one child did not reduce parents' chances of having a child living in the same household or close by, nor did it reduce their chances of receiving any transfer or receiving a net transfer[6] from a child. Moreover, having only one child significantly increased parents' chances of seeing a child at least once a month. Overall, our results suggest that coping from children mitigated the potential negative effects of having only one child on parents' mental wellbeing in later life.

Our paper contributes to a poorly studied branch of a large body of economic literature examining how fertility affects household coordinated choices (see Schultz 2007 for an overview). While most studies focused on outcomes such as mothers' labor supply, household saving behaviors, and children's human capital, little research has examined the effects of fertility on parents' health in the long run, and even fewer on elderly parents' mental wellbeing. Among the few papers examining physical health, Cáceres-Delpiano & Simonsen (2012) used U.S. data and found that a shock in fertility from multiple births increased mothers' likelihood of having high blood pressure and becoming obese; and Joshi & Schultz (2013) examined the Matlab program in Bangladesh and found no effect of reduced fertility on mothers' self-assessed health or ADLs, but positive effect on weight gain and the probability of BMI reaching 18.

To the best of our knowledge, only two papers have touched upon parents' mental health. Using data on elderly Europeans, Kruk & Reinhold (2014) found mixed evidence regarding the effects of additional children on mothers' depressive symptoms and no evidence for causal effects on fathers'. Using data from two provinces from CHARLS, Islam & Smyth (2015) found no effect of reduced fertility on a composite mental health index among elderly parents. Our paper contributes to this small literature using a nationally representative sample from China and exploiting the variation both in the quasi-random local enforcement of the one-child restriction and in women's policy exposure. In addition to estimating the net impact of having only one child, we also examine potential coping strategies from household members through which only children affect parents' mental wellbeing.

The remainder of the paper is organized as follows. Section 3.2 provides an overview about the recent history of family planning in China and, in particular, about the de-facto enforcement of the one-child restriction. Section 3.3 describes the data and the construction of key variables. Section 3.4 explains our identification strategy. Section 3.5 presents the results. Section 3.6 concludes.

[6] Parents receive a net transfer if the amount of transfer from children is larger than that to children.

3.2 Family Planning in China

In the first two decades after the founding of the People's Republic of China in 1949, there was in practice little control over fertility. Fertility remained high and stable at about six births per woman, except during famine years (Bongaarts & Greenhalgh 1985). Effective measures took shape in the early 1970s with the Later-Longer-Fewer (*wan xi shao*) policy, of which specific terms differed across regions but mostly promoted later marriage, longer birth spacing and fewer children among Han people (Bongaarts & Greenhalgh 1985; Tien 1980). Family planning committees and offices were set up at various levels of government and local cadres were directed to promote the family planning knowledge and policy. Fertility control tightened in 1978 with an explicit call from the Communist Party of China (CPC) Central Committee that one child was the preferred option and two children the maximum. In the following year, a national meeting was held among officials to announce the call for all Han couples to limit themselves to only one child (Banister 1991). Enforcement hardened in 1980 after the CPC Central Committee issued an open letter to further emphasize the necessity of one child per couple to reduce poverty, and call for cooperation across the nation.[7]

Though a second child was allowed for a small proportion of couples in "extenuating circumstances",[8] the one-child restriction was poorly enforced in rural areas. This was because central government control was weak and economic and cultural motivations for more children (especially sons) were strong. In some regions, higher-level cadres' excessive ambition in fertility control irrespective of local conditions actually led to more disobedience and consequently no reduction in second births (Greenhalgh 1986). Acknowledging the practical difficulty, CPC Central Committee issued Document 7 in 1984 to "open a small hole to close up a large one" (Greenhalgh 1986). This document officially allowed substantial policy relaxation to meet local conditions. A common relaxation was to allow rural couples to have a second child following a female firstborn (i.e. the one-and-a-half child policy). This was adopted by all provinces, except for the four direct-controlled municipalities (Beijing, Chongqing, Shanghai, Tianjin), Jiangsu province and Sichuan province.

The one-child restriction was never put into a law. Instead, a birth quota system was established to control and monitor the national fertility level. The State Family Planning Commission first set a national population target every year and birth quotas were then allocated top-down by administrative level until the point of implementation (Greenhalgh 1994). In the last step, couples were required to obtain an official birth permit from the authority at the place of their Hukou[9] registration before having a child. To incentivize local cadres, political pressure and rewards were employed. Cadres who exceeded their quotas or

---

[7] The original Open Letter in Chinese can be found at http://cpc.people.com.cn/GB/64184/64186/66677/4493829.html.

[8] Extenuating circumstances could include, e.g., a firstborn with severe disability. The proportion was set at 5 percent for urban couples and 10 percent for rural couples in 1980 (Bongaarts & Greenhalgh 1985).

[9] Hukou is a record in the Chinese government household registration system. It identifies a person as a resident of an area irrespective of where this person lives and entitles this person to local social benefits.

sabotaged the policies were criticized or punished, and their family planning work was one area in which their job performance was evaluated (Hardee-Cleaveland & Banister 1988). Since 1991, family planning work was further given "veto power" in cadres' evaluation and promotion nationwide.[10] This means failing to meet family planning targets would render a cadre's all other achievements null and void. On the other hand, policies forbade coercion and some cadres exploit this to shirk from meeting their targets (Hardee-Cleaveland & Banister 1988).

To encourage compliance among the general public, ideological education was used and family planning services were provided conveniently at local family planning stations and other health facilities. Barefoot doctors in rural areas were trained to perform intrauterine device insertions, early abortions and sterilizations so that policy implementation was feasible even in remote villages (Chen & Kols 1982). This system was further supplemented by economic incentives and disincentives, which differed by Hukou type or, more fundamentally, the extent of government control. One of the two Hukou types was non-agricultural (or, colloquially, urban), which guaranteed its holders formal employment. People then received income, food ration, housing, medical services, children's education, and pension from the government through their work units, which were predominantly public. Couples who complied with the one-child restriction were to enjoy benefits in the aforementioned aspects,[11] and those who disobeyed were to face severe punishment (e.g., losing one's job). The other Hukou type is agricultural (or colloquially rural). Its holders fed mainly on agricultural production, and were not guaranteed food, cash income, medical services or pension from the government. Benefits for having only one child were less uniform as funding came from local collectives, but often involved preferential treatment in food and land allocation. Punishment also varied, and sometimes involved forced abortion.

Though intended for all Han couples, the one-child restriction was, in practice, not uniformly enforced across the country in practice. The enforcement can be predicted to some extent from a region's characteristics such as location and administrative type (confirmed in Section 3.3.3). For example, due to the policy design that was more lenient towards ethnic minorities and people living in harsh conditions, the one-child restriction is less likely to be enforced in regions with a high concentration of minorities or in mountainous areas. Regions in the former case are often designated as autonomous and include autonomous in their names. Examples include the *five autonomous provinces* – Inner Mongolia, Xinjiang, Guangxi, Ningxia, and Tibet – and autonomous regions in provinces like Yunnan and Guizhou. Given the unpopularity of the restriction, enforcement is also more likely to be absent in regions where government control is weak. Examples include areas that were rural (in 1980s), areas in which most residents held agricultural Hukou, areas at the juncture of jurisdictions, and areas where the local cadre did not have a strong career ambition.

[10] This practice first appeared in Changde city in 1982 and was promoted to other places afterwards. Source: http://cpc.people.com.cn/pinglun/n/2013/1107/c78779-23460246.html.

[11] However, the amount of cash payments was small (Short & Fengying 1998), and preferential treatment in other aspects such as career advancement and housing allocation was difficult to materialize when other employees had only one child too.

Could people try to escape the one-child restriction by changing their Hukou registration place to autonomous regions (often remote and poor) or by changing their Hukou type from non-agricultural to agricultural? This is unlikely for two reasons. First, policy relaxations were consistently excluded from press coverage in fear of a baby boom (Greenhalgh 1990). For example, the government never even published the aforementioned Document 7. Without media coverage, it was difficult for people to learn about family planning policies in other places. Second, changing registration was very difficult in practice because migration controls were strictly enforced until the late 1990s.

3.3 Data and Variables

3.3.1 China Health and Retirement Longitudinal Study

We use data from the first two waves of the China Health and Retirement Longitudinal Study (CHARLS). This survey is based on the U.S. Health and Retirement Study (HRS) and related aging surveys such as the English Longitudinal Study of Aging (ELSA) and the Survey of Health, Aging and Retirement in Europe (SHARE). CHARLS collected comprehensive information on social, economic, and health conditions from a nationally representative sample of Chinese residents aged 45 and older and their spouses. The baseline was conducted in 2011-2012, among about 10000 households from 450 villages/resident committees from 28 provinces through multistage probability sampling.[12] The follow-up was conducted in 2013 and tracked not only the baseline respondents, but also the non-response sample and the refreshment sample, giving a net addition of about 900 households.

Our main outcome variables are obtained from the follow-up for a larger sample and with better data quality.[13] After dropping respondents with missing key demographic and socioeconomic information, we further drop (1) women born after 1966 as they did not reach age 45 at the time of the baseline survey in 2011, or before 1940 as they have generally finished childbearing by 1980; (2) households without a mother[14] (as information needed to construct the instrument is missing); and (3) households without a child or with a non-biological child of the mother (as our instrument is less relevant in these cases). The final sample consists of 1676 mother- only households and 3944 households with both the mother and the father. We examine all mothers, mothers who have a husband present in the sample, and fathers, respectively in the following analysis because they may be differentially affected by only children.

[12] Detailed description of the sampling method is available on the CHARLS website at http://charls.ccer.edu.cn/en/page/about-sample-2011.

[13] For the attrition sample, child and other household member loop datasets (constructed from the looping questions on household members) are not available and accurate matching of children's non-demographic information in the first wave is difficult because child identifiers are not made available.

[14] These are households with a never-married, separated, divorced or widowed male.

3.3.2 Outcome variables

We measure parents' depressive symptoms using the 10-question version of the Center for Epidemiologic Studies Depression Scale, CES-D 10. The validity of this measure has been established for the elderly Chinese by Boey (1999). Respondents were asked to indicate how frequently they experienced the following ten items during the week before the interview: (1) I was bothered by things that don't usually bother me; (2) I had trouble keeping my mind on what I was doing; (3) I felt depressed; (4) I felt everything I did was an effort; (5) I felt hopeful about the future; (6) I felt fearful; (7) My sleep was restless; (8) I was happy; (9) I felt lonely; (10) I could not get "going." The answers have four frequency categories, from rarely or none of the time, to some or a little of the time (1-2 days), to occasionally or a moderate amount of the time (3-4 days) and to most or all of the time (5-7 days).

We score the CES-D 10 answers using the metric suggested by Radloff (1977). It assigns values from 0 to 3 to frequencies from rarely to most of the time for negative emotions such as "I felt depressed". For positive emotions such as "I was happy", the scoring is reversed assigning 0 to most of the time and 3 to rarely. Summing over the 10 components gives the total score ranging from 0 to 30, where a higher number reflects a higher level of depression. A threshold of 12 has been suggested to indicate the likely presence of clinically significant depression among 60+ Chinese (Cheng & Chan 2005). Results are similar using this alternative outcome, and are thus omitted to save space. A question on life satisfaction was answered on a five-point scale with 1 for "completely satisfied", 2 for "very satisfied", 3 for "somewhat satisfied", 4 for "not very satisfied" and 5 for "not at all satisfied". The constructed indicator for being satisfied with life is equal to one for ratings below 4.

Given the subjective nature of the CES-D 10 and life satisfaction questions, we first conduct an indirect check on whether one-child parents use different standards when evaluating mental wellbeing. To make it possible to test for reporting heterogeneity, existing surveys often use vignettes, i.e. descriptions of the condition of hypothetical persons (King et al. 2004). The rationale behind it is that, as vignettes represent fixed states, any systematic correlation between respondents' characteristics and their ratings of the vignettes indicates heterogeneous reporting. While vignettes are not available directly for CES-D 10 or life satisfaction, they are available for self-reported health in the related affect domain in CHARLS baseline. Technical details of the test and results are provided in

Appendix 3.A. In general, we do not find evidence for reporting heterogeneity even though the vignette descriptions covered life satisfaction and many of the CES-D 10 components, and the severity rating – a five-point scale ranging from no problem to extreme problem – is arguably more subjective than the frequency rating used in CES-D 10 questions. This lends support to the use of CES-D 10 and life satisfaction for comparing parents with only one child and those with more children.

Cognitive skills are measured using episodic memory and mental intactness, constructed from questions that are similar to those used in HRS. Specifically, episodic memory is measured by the number of words (0-10) recalled in any order four minutes after they were read to the respondent; mental intactness by the number of correct answers to 12 questions taken from the Telephone Interview of Cognitive Status (TICS) battery. In CHARLS, these 12 questions include naming the date of the survey (month, day, year, and season), the day of the week, serial 7 subtraction from 100 (up to five times), whether the respondent used paper and pencil or other aid for calculation, and the ability to redraw a picture of two overlapping pentagons shown to the respondent.

Figure 3.2 and Figure 3.3 plot for mothers and fathers respectively the CES-D 10 score, episodic memory score, mental intactness score, and the proportion satisfied with life against age and by the number of children (one versus two or more). The two figures show that in general the CES-D 10 score, episodic memory, and mental intactness deteriorate over time, except that fathers' cognitive skills experience an improvement during early 60s. The pattern in life satisfaction is less clear, except that mothers with two or more children appear to be increasingly satisfied with life over time. For both mothers and fathers and in most age groups, those with only one child have distinctly better outcomes than parents with two or more children.

To explore the mechanisms through which fertility affects parents' mental wellbeing, we obtain information on three types of parent-child interactions. The first type relates to children's residence. Traditionally, children have provided old-age support to parents through co-residence or living close by. Such living arrangement remains to be a popular way to care for elderly parents, especially in rural China (Cai et al. 2012). Thus, we construct an indicator for co-residence, which is equal to one if at least one child lives in the same household. Another indicator for having a child living close by is equal to one if at least one child lives in the same village/neighborhood or in another village/neighborhood but within 10 kilometers.[15]

The second type of parent-child interactions relates to how frequent parents are contacted by non-co-resident child(ren), which cannot directly be inferred from geographical proximity. Parents were asked about how often they saw each non-resident child, and if less frequent than once a week, they were further asked about the frequency they receive phone calls, text messages, mails or emails from that child. Communication with children is an important way through which elderly parents remain connected to the society and emotional intimacy with children can bring much joy and satisfaction. To capture this aspect of parent-child interactions, we combine the information from all children and construct indicators for

[15] Using alternative thresholds such as 5km and 15km does not change the conclusion.

parents receiving any form of contact, in-person contact, and non-in-person contact, respectively from any child at least once a month.

The last type of interaction concerns money and in-kind transfers from children. In addition to care-giving and emotional support, financial help is also an important way to improve the quality of life for elderly parents. Moreover, the gesture of making a transfer in itself is considered a sign of filial piety. It may bring happiness and satisfaction even when the amount is offset by that from parents back to children. Therefore, two transfer variables are constructed. The indicator for any transfer is equal to one if any child provided any transfer. The other indicator for net transfer is equal to one if the monetary amount of transfers from children was larger than that from parents back to children.

3.3.3 Local enforcement of the one-child restriction

Information on local family planning policies is obtained from the community history section of the questionnaire. Interviewers were instructed to summon 2-3 elderly people familiar with the local history to complete this section. A series of questions were asked about the year in which the community started to execute the family planning policy performed nationwide in the late 1970s, the specific policy terms for Han couples (i.e., only one child was allowed, second child was allowed only when the first was a girl, two children were allowed, or more than two children were allowed), whether there was a policy change, and if so, the year of change and the new terms. We determine whether a community enforced the one-child restriction mainly based on these questions.[16] The one-child dummy is equal to one if the county reported having the one-child restriction no later than 1989 (to allow recall error) and not relaxing it within five years.

As is clear from Section 3.2, the location and the administrative type of an area are highly predictive of the enforcement of the restriction. CHARLS offers administrative types that are more refined than the rough urban-rural division defined by the National Bureau of Statistics (NBS). It identifies an urban community as city district, combined urban-rural area, town center area, combined town-township area or special district, and a rural community as township center area or village. It is important to distinguish between city districts and combined urban-rural areas because the latter are often rural villages previously on the outskirts of a city and are incorporated into the city only later through urbanization. These combined urban-rural areas often preserve many rural features and are very different from city districts. In particular, combined areas are subject to weaker government control and have more agricultural Hukou holders. As the one-child restriction is known to be well enforced in

[16] Though the question on the starting year of the family planning policy performed nationwide in the late 1970s was meant to ask for the starting year of the one-child policy, in some communities it was answered with the starting year of more relaxed policies implemented earlier and the one-child policy was reported as a later policy change. Thus, questions on both the policy and the later change are used to identify a long-lasting one-child restriction.

city centers, the one-child dummy is automatically set to one for communities identified as city districts since 1980.[17]

Relying on survey data for local enforcement of the one-child restriction may give rise to measurement error concerns. However, there are reasons to believe that the data quality in CHARLS is fairly good and, more importantly, free from systematic misreporting. First, in more than 90% of the communities, at least one person aged above 53 (20+ in 1978) was consulted when filling in the community history section. While it might be difficult to accurately remember the implementation year of a policy, it is difficult *not* to know about the restriction on the number of children, especially for people of childbearing age around 1980. Therefore, we have confidence in the quality of data on policy terms. Second, the reported information is unlikely to be systematically biased due to political reasons because (1) questions such as starting year and specific terms are not particularly sensitive, (2) the questionnaire does not appear to have a focus on evaluating family planning policies, and (3) the interviewers are not associated with state or local family planning organizations.

Table 3.1 presents in the first two columns results from regressing the dummy for the one-child restriction on the characteristics of the communities (Panels A & B) and community history respondents (Panel C). The first row of Panel A shows, as expected, a clear negative effect of autonomous region status on the enforcement of the one-child restriction. Next, estimates for administrative types confirm that enforcement was less stringent in combined areas (i.e. at the juncture of jurisdictions) than in center areas and in urban regions than in rural regions. However, enforcement in combined town-township areas in the urban region was comparable to that in rural villages. This reveals substantial heterogeneity within the rough categorization of urban/rural. The last two variables on local office status further refine the administrative categorization with village committee offices (*cunweihui*) indicating a more rural region and community office (*juweihui*) indicating a more urban one.[18] While being a rural community before 1980 still has a negative effect on the enforcement of the one-child restriction, the effect is no longer significant after adding other controls. Having a community office (versus a village committee office) has a significant positive effect on local enforcement, but loses significance after conditioning on other community characteristics.

Panel B of Table 3.1 examines four community characteristics that can potentially affect the enforcement of the one-child restriction. However, after controlling for detailed location and administrative characteristics, mountainous terrain, a proxy for harsh living conditions, is no longer significant. Neither is the number of barefoot doctors, a proxy for accessibility and capacity of family planning services in rural regions. We also assess the predictive power of the implementation of the Later-Longer-Fewer (LLF) policy, a more voluntary and relaxed fertility control policy promoted before the one-child policy. Given the lack of strong pressure and

[17] However, we do not automatically assume that the one-child restriction was strictly enforced in the four direct-controlled municipalities, Jiangsu province and Sichuan province even though they did not have the one-and-a-half child policy. This is because there might be other common extenuating circumstances permitting a second child. For example, Greenhalgh (1986) reported that in Sichuan province a city allowed two children in mountainous towns at high sea levels.

[18] In the CHARLS community questionnaire, administration of urban/rural specific questions was based on whether the community has a village committee office or a committee office, rather than the NBS urban/rural classification.

incentives on cadres and the general public, implementation of LLF can be considered as a proxy of local acceptance of fertility control. The presence of big surname(s) is included as an indication of strong influences of the traditional culture, in particular the preference for more children. Neither of these two variables appears to be significant. Together, Panel B suggests that the one-child restriction was indeed pushed forward forcefully as a political task.

Panel C of Table 3.1 examines the quality of reporting under the assumptions that (1) people during prime childbearing age (20-35) in 1980 are more likely to recall policy terms accurately, (2) reporting quality is likely to be higher when more respondents were consulted, and (3) women are more likely to recall policy terms accurately. Results show that the reported enforcement of the one-child restriction is not correlated with characteristics of the community history respondents, including whether there was at least one person aged 20-35 in 1980, the number of respondents and the proportion of female respondents. Assuming the abovementioned three assumptions are valid, our results suggest good knowledge of local family planning policy terms among the general population and good reporting quality.

One may worry that the enforcement of the one-child restriction was associated with other community or individual characteristics that affect parents' mental wellbeing in old age. We examine possibly the two strongest candidates – local income level and parents' childhood health. Though historical income level is not available, the community questionnaire asked about average per capita net income in 2010, which should be highly correlated with that in the late 1970s.[19] Childhood (younger than 15) health was obtained from CHARLS baseline for follow-up respondents and from the follow-up wave for non-response and refreshment sample. For each community, we calculate the proportion of people reporting below good (i.e. fair or poor) health during childhood. The last three columns of Table 3.1 report the results from adding these two variables to the basic specification (Column 1). None of the estimates are significant after conditioning on location and administrative characteristics, relieving the concern that they may invalidate the exclusion restriction assumption underlying our instrumental variable approach.

3.4 Identification Strategy

We aim to estimate the causal effect of having only one child on parents' mental wellbeing in later life. Using OLS would yield biased results because both variables are likely correlated with community-level and household-level characteristics that are unobserved or difficult to measure. Previous papers using Chinese data have tried to address this endogeneity problem using the natural occurrence of twins (Rosenzweig & Zhang 2009; Li et al. 2008). However, this strategy requires either a dedicated twins dataset or a very large sample of the general population. Moreover, the benefits (e.g., joy) and costs (e.g., mothers' forgone labor income, stress) of having twins are likely to be different from having two sequential births.

[19] As Chinese provinces experienced differential growth since the economic reform in 1978, we also test the effect of average per capita net income in the subsample of provinces whose per capita GDP ranking did not change by more than 5 places since 1978. The magnitude and the significance of the income coefficient are reduced compared to that in the full sample.

Other frequently used strategies exploit variations in the implementation of the one-child policy, e.g., exemption of ethnic minorities (Li & Zhang 2007; Li & Zhang 2009), or the one-and-a-half-child policy in rural areas (Qian 2009; Islam & Smyth 2015). However, ethnic minorities comprised only 8% of the total population by 1990[20] and differ in various ways from Han people. Instruments exploiting the one-and-a-half-child policy were not ideal either because of the use of the rough NBS urban-rural definition, which ignores the substantial heterogeneity within urban and rural regions and raises concerns about the monotonicity assumption. Furthermore, sex of the first child was subject to selection given the widespread availability of ultrasound since the 1980s (Yi et al. 1993), which casts doubt on the validity of instruments derived from it.

We try to improve on existing measures by incorporating more accurate enforcement information and allowing the instrument to reflect different policy impacts on women at different stages of childbearing.[21] Formally, we first define a binary measure of policy exposure for the mother from household *i* in community *j* at age *t* as

$$z_{i,j,t} = \mathbb{I}\big(OCP_{j,t} = 1\big), \qquad t = 15, 16, \dots, 40 \tag{3.1}$$

where $z_{i,j,t} = 1$ indicates that the mother from household *i* was subject to the one-child restriction at age *t*, and $z_{i,j,t} = 0$ otherwise. For each household, we then obtain the accumulated years of exposure by mother's age *T*, calculated as

$$Z_{i,j,T} = \sum_{t=15}^{T} z_{i,j,t}\,, \qquad T = 15, 16, \dots, 40 \tag{3.2}$$

We first test the relevance of these policy exposure measures to having only one child, a condition for their use as instruments. An indicator for having only one children is regressed on each of the 36 binary instruments, $z_{i,j,t}$, and the 36 continuous measures, $Z_{i,j,T}$, respectively, controlling for the characteristics of the mothers, fathers and the communities (explained after Equation (3.3). Figure 3.1 plots the first-stage F-statistics of our binary and continuous measures separately against mothers' age. Beyond the age of 21, the F-statistics remain above the rule of thumb of 10 in the test for weak instruments. The binary measure displays, as expected, an increasing impact of policy exposure on fertility during early childbearing years and a decreasing impact during later years. Maximum F-statistic is achieved at age 28 for the binary measure, and a few years later at age 31 for the continuous measure. We choose $Z_{i,j,31}$ as our preferred instrument for its strength (F-statistics 28.0) and also its ability to better capture differential policy impacts on women of different ages (compared to a binary instrument).

[20] Source: http://www.stats.gov.cn/tjsj/tjgb/rkpcgb/qgrkpcgb/200204/t20020404_30320.html

[21] Though Wang (2012) also used the accumulated years of policy exposure to account for the differential policy impacts on women of different ages, the author did not have the more refined urban-rural classification.

The second condition – the exclusion restriction – requires that the instrument can only affect parents' mental wellbeing through the induced changes in fertility. This assumption is not directly testable. However, a careful examination in Section 3.3.3 of factors associated with the enforcement of the one-child restriction provided evidence against two most potential alternative pathways – local income level and parents' childhood health. Furthermore, in addition to controlling for the important predictors of the one-child restriction, we allow for differential time trends by location and Hukou type. This reduces the possibility that our instrument affects parents' mental wellbeing through its correlation with other social and welfare policies, of which the impact also varies by location and Hukou type, *and* is linearly correlated with women's age.

Last, the monotonicity assumption requires that people who are affected by the instrument be affected in the same way. This assumption is likely to hold because there is little reason to suspect that (1) people who had only one child in communities without one-child restriction would have more children in communities with one-child restriction, and (2) mothers would have given birth to more children had they been exposed to the one-child restriction for more years.

Our second-stage model is specified as follows:

$$Y_i = \alpha + \beta OneChild_i + \sum_k \gamma_k X_{i,k} + \sum_{\substack{c=prov,loc\ type,vil\ stat\ urb\ stat,\ autonomous}} \zeta_{j,c} + \sum_{\substack{m=agri\ Hukou,loc\ type}} \eta_{t,m} + \varepsilon_i \quad (3.3)$$

where outcome $Y_{i,j}$ includes mothers' and fathers' CES-D 10 score (range 0-30) and its 10 components (range 0-3) for their individual interest, episodic memory score (range 0-10), mental intactness score (range 0-12) and satisfaction with life (binary); $OneChild_i$ is a dummy for having only one child; $X_{i,k}$ includes a set of mother characteristics (5-year age categories, ethnicity, Hukou type, five education levels and dummies for being married and widowed), a set of father characteristics (5-year age categories, ethnicity, Hukou type, and five education levels), household per capita expenditure quintiles, household size; $\zeta_{j,c}$ is a set of community-level fixed effects (province, administrative type, village status before 1980, urban status in 2011, and the autonomous region status) and $\eta_{t,m}$ allows the time trend to differ between agricultural and non-agricultural Hukou and between different administrative types of the communities.

3.5 Results

In our analysis, we separate mothers and fathers as reduced fertility may bring mothers personal health benefits and may be more dissatisfying for fathers due to cultural reasons. It is worth noting that some of the effects of only children could also work through marriage

dissolution. Thus, we additionally examine the subsample of mothers where husbands are present to see if findings from the full sample still hold for the subsample.

Table 3.2 presents descriptive statistics by the one-child family status for mothers (full sample and subsample) and fathers. A comparison between columns 1 and 2, between 4 and 5, and between 7 and 8 in Panel A shows significant differences between parents with one and more children in all demographic and socioeconomic characteristics, especially in Hukou status. Both mothers and fathers with only one child are younger, less likely to be ethnic minority or to have agricultural Hukou. They are also better educated, more likely to be married and less likely to be widowed (not applicable to fathers), and richer (proxied by expenditure), though partly because they are younger and more likely to be Han. Panel B on community administrative characteristics suggests that one-child parents are less likely to be found in autonomous regions, rural regions (township center areas and villages), or in communities that were villages before 1980, but are more likely to be found in city districts and urban communities (i.e. with a community office as opposed to a village committee office).

Comparing the first and the second four columns of Panel A shows that restricting the mother sample to those with husbands leads to only modest change in sample demographic and socioeconomic characteristics. Specifically, the husband-present subsample has a similar age structure as the full mother sample, but is slightly less educated, and by definition more likely to be married. The income difference between mothers with one and more children is smaller in the husband-present subsample. In contrast, Panel B shows little change in both the averages among mothers with one and more children and the differences between these averages.

Table 3.3 presents IV estimates of the effect of having only one child on mothers' and fathers' CES-D 10 score and its 10 components (higher score indicating worse outcome), cognitive skills including episodic memory and mental intactness, and life satisfaction. To facilitate comparison, OLS estimates are first provided in columns 1, 5 and 9 for the full sample of mothers, the subsample of husband-present mothers, and the sample of fathers, respectively. The OLS results suggest that for both mothers and fathers, ignoring the endogeneity of fertility decisions, having only one child is associated with better outcomes in only a few components of CES-D 10 and cognitive skills, and is not associated with life satisfaction.

Columns 3, 7, 11 of Table 3.3 provide the IV estimates in which having only one child is instrumented using accumulated years of exposure to the one-child restriction by mother's age of 31. For mothers, in both the full sample and the husband-present subsample, accounting for endogeneity shifts most coefficient estimates in the direction of a protective effect of having only one child against depressive symptoms. However, most IV estimates remain insignificant, except for the last CES-D 10 component. The estimates from the full sample and the subsample suggest that having only one child reduces mothers' frequency of feeling that they could not get "going" by 0.48 point and 0.53 point respectively, which are about half of the score (1 point) needed to move to a lower frequency category. In contrast, results on cognitive skills in the full sample suggest that having only one child reduces mothers' episodic memory by 1.13 points, which translates to about one fewer word recalled in the test. The

estimate in the subsample is also negative, though smaller and insignificant. The last row suggests no effect of having only one child on mothers' life satisfaction in both the full sample and the subsample.

The last four columns in Table 3.3 show that, unlike for mothers, accounting for endogeneity shifts the coefficient estimates in the direction of worse scores in CES-D 10 and many of its components for one-child fathers. While OLS results suggest that having only one child is associated with a smaller CES-D 10 score and a lower frequency of feeling troubled or depressed for fathers, the IV estimates do not confirm any causal effects. Similarly, no causal effect of having only one child is found for cognitive skills, despite the large magnitude of the IV estimate for mental intactness. The last row again shows that having only one child does not reduce life satisfaction. In general, our results do not provide evidence for differences in mental wellbeing between parents with only one child and those with more children in later life, even after accounting for endogeneity of the fertility decision.

Why having only one child, which goes against the preference of almost all couples, did not noticeably reduce parents' mental wellbeing? To shed light on this question, we examine three potential mechanisms through which only children can mitigate the potential negative effects. Table 3.4 presents results on children's living arrangements, frequency of parents being contacted by children and transfer from children estimated using Equation (3.3. In both samples, OLS results suggest less parent-child interactions in the form of living arrangements and contact frequency. In the couple subsample (the right half), one-child parents are also significantly less likely to receive any transfers and receive net transfers from children.

Accounting for endogeneity leads to better parent-child interactions among one-child parents. Specifically, having only one child no longer reduces the likelihood of parents co-residing with a child, or having a child living close by. Moreover, conditional on only children being not less likely to co-reside with parents, one-child parents are significantly more likely to be visited by a non-co-resident child. The probability of one-child parents being contacted by phone, text message, mail/email, etc., is also higher, though not significantly so, leading to a higher overall probability of one-child parents being contacted by children (significant in the couple subsample). The last two rows show that having only one child does not reduce the probability of parents receiving any transfers or receiving net transfers from children. The flipping of signs from OLS to IV estimates reflects the bias from omitting more detailed information on parents' socioeconomic status as a higher status is correlated with both having only one child and better mental wellbeing in later life. For example, people with better jobs face higher costs of a second birth and are more likely to have only one child. They are less dependent on children for emotional and financial support and their children are typically more capable of migrating to bigger cities for better job opportunities. Failing to correct for such biases would produce a misleading picture of parent-child interactions.

3.6 Conclusion

Population control policies are often employed by developing countries with the aim of reducing poverty. However, little is known about how reduced fertility affects the mental wellbeing – an important aspect of the quality of life – of people later in life. Identification is often challenging because of the endogeneity of fertility decisions. We overcome this problem by exploiting the one child policy in China. We show that local enforcement of the one-child restriction is largely determined by county characteristics such as location and administrative type. Local cadres' characteristics, such as career ambition, also play a role, but are arguably exogenous. After conditioning on county characteristics, enforcement of the one-child restriction is no longer associated with a wide range of potential confounding factors such as local economic condition and parents' health during childhood. Our instrument is constructed for each mother (and her husband) as a function of whether the one-child restriction was enforced locally and her age (to reflect differential policy exposure).

Our IV estimates reveal small and mixed effects of having only one child on parents' depressive symptoms, cognitive skills and life satisfaction, leading to a general conclusion of no effect on elderly parents' mental wellbeing. This finding might be due to coping within the household. Further analysis on parent-child interactions indeed suggests that only children might have mitigated the negative effects by increasing contacts with their parents. As a result, having only one child did not decrease parents' chance of having a child living in the same household or close by, or receiving transfers. On the contrary, it even significantly increased their chance of seeing a child at least once a month. Coping strategies from household members in response to fertility shocks have also been documented by Li et al. (2015). The study exploited naturally occurring twins and found no evidence of a negative effect of fertility on parental labor supply. However, co-residence of grandparents was found to increase significantly, suggesting that the negative effects of fertility were mitigated by the childcare provided by grandparents.

Our finding of no effect of reduced fertility on elderly parents' mental wellbeing is somewhat in agreement with the limited literature on this subject. For example, it is consistent with Kruk & Reinhold (2014) in that no effect is found for elderly fathers. An even more relevant reference is Islam & Smyth (2015), which used the pilot data from CHARLS. The study found no effects of reduced fertility on parents' mental health, as measured by a composite score. While our finding is not as expected given the strong cultural and economic motivations for more children in China, it is important to note the rapid social and economic development in China since the 1980s. More children may not necessarily mean more economic and emotional old-age support because some pre-conditions no longer hold. For example, capital and skills are replacing labor as the main constraint on income so that more children do not always generate higher economic return. While migration was not common in the past, rural migrant workers have reached 169 million in 2015.[22] Last, cultural pressure from the local community on children to care for their parents has lessened nowadays (B. Chen

[22] Source: http://www.stats.gov.cn/tjsj/zxfb/201604/t20160428_1349713.html.

2009). Therefore, it is perhaps not entirely surprising that the quantity of children does not seem to determine elderly parents' mental wellbeing nowadays.

While this paper shows that parents with only one child were not at a disadvantage in mental wellbeing, this cohort is still relatively young (mostly younger than 65), and in general active and healthy. It is possible that only children will become less capable of providing needed support when parents become more help-dependent or severely ill in ten years. Follow-up research can shed light on the longer-term effects of having only one child. Alternatively, a larger dataset at present can also enable us to assess how well only children are able to cope with less favorable situations, such as when parents become widowed or develop difficulties with daily activities.

Figures

Figure 3.1. IV relevance - first stage F-statistics by mother's age.

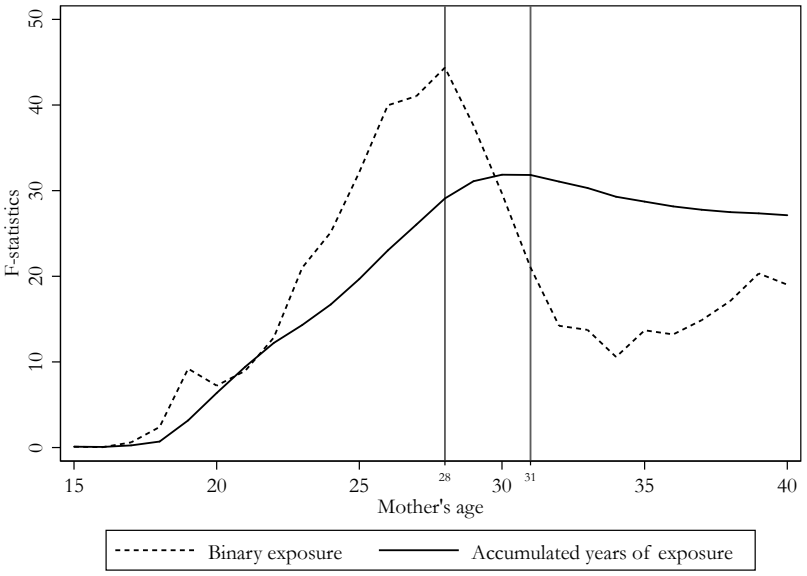


Figure 3.2. Mothers' mental wellbeing measures.

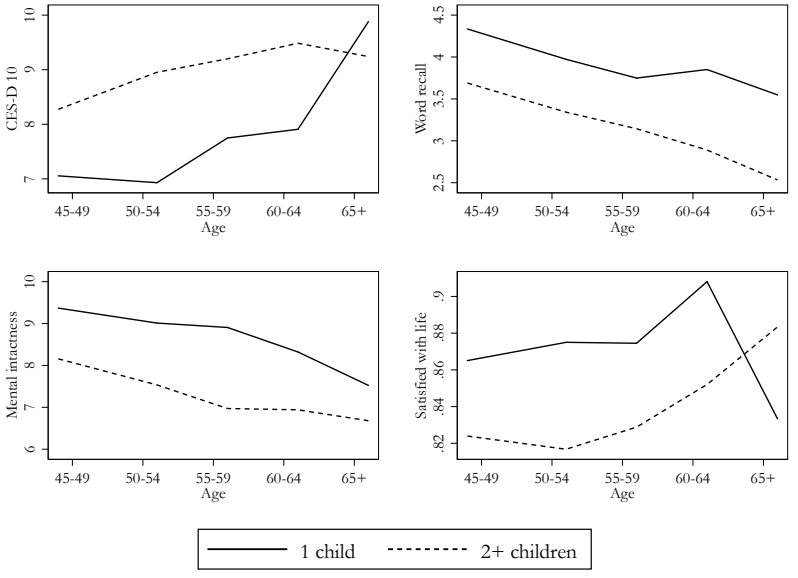
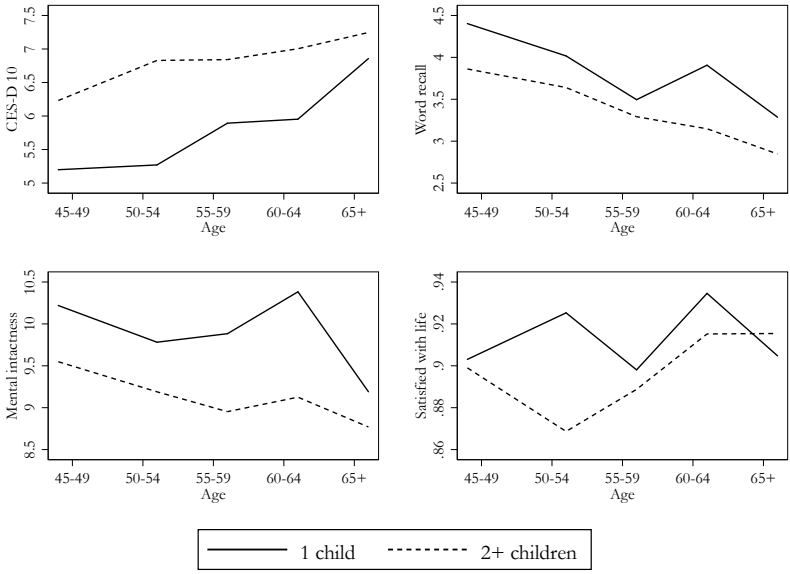


Figure 3.3. Fathers' mental wellbeing measures.



Tables

Table 3.1. Potential factors affecting the enforcement of the one-child restriction.

	(1)	(2)	(3)	(4)	(5)
Panel A. Administrative characteristics					
Autonomous region	-0.349***	-0.359***	-0.346***	-0.353***	-0.350***
	(0.095)	(0.098)	(0.095)	(0.096)	(0.096)
Located in (city district omitted)					
combined urban-rural area	-0.423***	-0.410***	-0.437***	-0.424***	-0.437***
	(0.102)	(0.103)	(0.103)	(0.102)	(0.103)
town center area	-0.331***	-0.333***	-0.328***	-0.331***	-0.328***
	(0.063)	(0.064)	(0.063)	(0.063)	(0.063)
combined town-township area	-0.510***	-0.495***	-0.515***	-0.510***	-0.514***
	(0.094)	(0.096)	(0.094)	(0.094)	(0.094)
special district	-0.392**	-0.466**	-0.423**	-0.394**	-0.425**
	(0.195)	(0.203)	(0.196)	(0.195)	(0.196)
township center area	-0.708***	-0.705***	-0.717***	-0.707***	-0.716***
	(0.121)	(0.124)	(0.121)	(0.121)	(0.121)
village	-0.504***	-0.491***	-0.500***	-0.503***	-0.499***
	(0.082)	(0.086)	(0.082)	(0.082)	(0.082)
With a village committee office before 1980	-0.107	-0.118	-0.108	-0.107	-0.108
	(0.075)	(0.076)	(0.075)	(0.075)	(0.075)
With a community office in 2011	0.179**	0.143	0.168*	0.179**	0.168*
	(0.088)	(0.112)	(0.088)	(0.088)	(0.088)
Panel B. Other community characteristics					
Mountainous terrain		-0.001			
		(0.046)			
Number of barefoot doctors in the 1970s		0.013			
		(0.011)			
Implemented Later-Longer-Fewer policy		-0.023			
		(0.034)			
Big surname		-0.070			
		(0.053)			
Panel C. Community history respondents' characteristics					
At least one person aged 20-35 in 1980		-0.002			
		(0.048)			
Number of respondents		0.002			
		(0.029)			
Proportion of female respondents		-0.081			
		(0.101)			
Panel D. Potential confounding factors (community level)					
Per capita income in 2010			0.004		0.004
			(0.003)		(0.003)
Proportion with below good childhood health				-0.028	-0.026
				(0.112)	(0.112)
Province dummies	Yes	Yes	Yes	Yes	Yes
Dummies for missing values	NA	Yes	Yes	Yes	Yes
Constant	0.640***	0.652***	0.643***	0.655***	0.646***
	(0.130)	(0.158)	(0.158)	(0.160)	(0.160)
N	447	447	447	447	447
Adjusted R-squared	0.614	0.623	0.624	0.623	0.624

Notes: Missing values in per capita income in 2010 are replaced with the average per capita income of other communities in the same city. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 3.2. Descriptive statistics for mothers and fathers.

	Mothers						Fathers		
	Full sample			Husband-present subsample					
	1 child	2+ Children	Difference	1 child	2+ Children	Difference	1 child	2+ Children	Difference
Panel A. Individual characteristics									
Age	53.8	58.1	-4.3***	53.6	57.4	-3.8***	55.5	59.2	-3.8***
44-50	0.38	0.22	0.16***	0.37	0.23	0.15***	0.28	0.16	0.12***
51-55	0.26	0.16	0.10***	0.26	0.17	0.09***	0.25	0.16	0.09**
56-60	0.24	0.23	0.01	0.27	0.26	0.02	0.28	0.23	0.04
61-65	0.08	0.21	-0.14***	0.07	0.22	-0.15***	0.14	0.23	-0.09***
66+	0.04	0.18	-0.14***	0.02	0.12	-0.10***	0.06	0.22	-0.16***
Minority	0.05	0.08	-0.04***	0.04	0.08	-0.04***	0.04	0.07	-0.03*
Agricultural Hukou	0.45	0.78	-0.33***	0.50	0.81	-0.32***	0.48	0.76	-0.28***
Illiterate	0.15	0.33	-0.18***	0.16	0.34	-0.18***	0.03	0.09	-0.06***
Can read or write	0.11	0.19	-0.08***	0.12	0.20	-0.07***	0.11	0.17	-0.05***
Primary school	0.15	0.23	-0.07***	0.17	0.21	-0.03	0.18	0.27	-0.09***
Junior high school	0.27	0.18	0.09***	0.26	0.18	0.09***	0.36	0.30	0.06*
High school or above	0.32	0.08	0.24***	0.28	0.08	0.20***	0.32	0.18	0.14***
Married	0.91	0.87	0.03**	1.00	1.00	0.00	1.00	1.00	0.00
Widowed	0.07	0.12	-0.05***						
Household expendiure per capita	15521	11163	4358***	13805	11231	2575***	13805	11231	2575***
Panel B. community administrative characteristics									
Autonomous region	0.08	0.15	-0.07***				0.08	0.15	-0.07***
Located in									
City district	0.53	0.16	0.37***				0.49	0.14	0.36***
Combined urban-rural area	0.04	0.05	-0.01				0.04	0.03	0.01
Town center area	0.12	0.13	-0.01				0.12	0.14	-0.01
Combined town-township area	0.06	0.07	-0.01				0.07	0.08	-0.01
Special district	0.00	0.00	0.00				0.00	0.00	0.00
Township center area	0.01	0.04	-0.04***				0.01	0.05	-0.04***
Village	0.24	0.54	-0.30***				0.27	0.57	-0.30***
With a village committee office before 1980	0.47	0.82	-0.35***				0.53	0.84	-0.31***
With a community office in 2011	0.63	0.28	0.35***				0.58	0.24	0.35***
N	1047	4573		725	3219		725	3219	

Notes: The first three columns present statistics for all mothers, the next three for mothers who have a husband, and the last three for fathers. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 3.3. Effects of only children on parents' mental wellbeing.

	Mothers								Fathers			
	Full sample				Husband-present subsample				OLS	*p*-value	2SLS	*p*-value
	OLS	*p*-value	2SLS	*p*-value	OLS	*p*-value	2SLS	*p*-value				
CES-D 10	-0.08	[0.793]	-0.68	[0.707]	-0.15	[0.656]	-0.58	[0.780]	-0.57**	[0.033]	-0.05	[0.976]
	(0.29)		(1.80)		(0.33)		(2.05)		(0.27)		(1.79)	
Bothered	-0.07	[0.160]	-0.13	[0.697]	-0.07	[0.185]	-0.14	[0.713]	-0.06	[0.158]	0.01	[0.977]
	(0.05)		(0.34)		(0.05)		(0.38)		(0.04)		(0.29)	
Troubled	0.05	[0.324]	0.54	[0.109]	0.05	[0.393]	0.55	[0.191]	-0.08*	[0.061]	0.12	[0.720]
	(0.05)		(0.33)		(0.05)		(0.42)		(0.04)		(0.34)	
Depressed	0.02	[0.691]	-0.33	[0.287]	0.04	[0.479]	-0.19	[0.593]	-0.07*	[0.080]	0.01	[0.985]
	(0.04)		(0.31)		(0.05)		(0.36)		(0.04)		(0.29)	
Effort	0.05	[0.260]	0.04	[0.903]	-0.01	[0.889]	-0.13	[0.743]	-0.09	[0.104]	-0.14	[0.675]
	(0.05)		(0.35)		(0.06)		(0.38)		(0.05)		(0.34)	
Hopeful	-0.05	[0.478]	-0.01	[0.989]	-0.06	[0.466]	-0.45	[0.397]	-0.07	[0.498]	0.23	[0.609]
	(0.08)		(0.46)		(0.09)		(0.53)		(0.11)		(0.46)	
Fearful	-0.05	[0.115]	-0.31	[0.154]	-0.06*	[0.063]	-0.33	[0.194]	0.00	[0.918]	-0.21	[0.292]
	(0.03)		(0.22)		(0.03)		(0.26)		(0.03)		(0.19)	
Restless	0.06	[0.403]	0.67	[0.119]	0.06	[0.397]	0.63	[0.190]	-0.07	[0.201]	0.26	[0.529]
	(0.07)		(0.43)		(0.08)		(0.48)		(0.06)		(0.42)	
Happy	-0.05	[0.420]	-0.35	[0.379]	-0.10	[0.157]	-0.04	[0.931]	-0.12	[0.242]	-0.33	[0.415]
	(0.06)		(0.40)		(0.07)		(0.42)		(0.10)		(0.40)	
Lonely	0.02	[0.588]	-0.32	[0.200]	0.04	[0.369]	0.06	[0.837]	-0.01	[0.728]	-0.26	[0.241]
	(0.04)		(0.25)		(0.04)		(0.27)		(0.03)		(0.22)	
Can't get going	-0.05*	[0.052]	-0.48**	[0.022]	-0.03	[0.372]	-0.53**	[0.015]	0.00	[0.947]	0.24	[0.240]
	(0.03)		(0.21)		(0.03)		(0.22)		(0.02)		(0.21)	
Episodic memory	0.20**	[0.024]	-1.13*	[0.089]	0.10	[0.307]	-0.67	[0.392]	0.30**	[0.035]	-0.29	[0.649]
	(0.09)		(0.67)		(0.10)		(0.78)		(0.14)		(0.63)	
Mental intactness	0.14	[0.284]	0.30	[0.720]	0.32*	[0.072]	0.21	[0.832]	0.13	[0.409]	1.03	[0.298]
	(0.13)		(0.84)		(0.18)		(0.98)		(0.16)		(0.99)	
Satisfied with life	0.03	[0.162]	0.10	[0.505]	0.01	[0.433]	0.22	[0.147]	0.03	[0.273]	-0.14	[0.251]
	(0.02)		(0.15)		(0.02)		(0.15)		(0.03)		(0.12)	
N	5,620				3,944				3,944			

Notes: The first three columns present results for all mothers, the next three for mothers who have a husband, and the last three for fathers. CES-D 10 ranges from 0 to 30 and each of its components ranges from 0 to 3, with a higher score indicating a worse outcome. Episodic memory ranges from 0 to 10 and mental intactness from 0 to 12. Respondents reporting completely satisfied, very satisfied and somewhat satisfied with life are considered as being satisfied with life. Standard errors in parentheses are clustered at the community level. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 3.4. Effects of only children on parent-child interactions as potential mechanisms.

	Mothers				Fathers			
	OLS	*p*-value	2SLS	*p*-value	OLS	*p*-value	2SLS	*p*-value
Child living at home	-0.10***	[0.000]	0.06	[0.709]	-0.13***	[0.000]	0.09	[0.651]
	(0.02)		(0.16)		(0.03)		(0.19)	
Child living close by	-0.29***	[0.000]	0.27	[0.151]	-0.27***	[0.000]	0.29	[0.174]
	(0.02)		(0.19)		(0.03)		(0.21)	
Contacted by child at least once a month	-0.03**	[0.022]	0.15	[0.141]	-0.03**	[0.044]	0.35**	[0.021]
	(0.01)		(0.10)		(0.01)		(0.15)	
Meeting in person	-0.15***	[0.000]	0.58*	[0.074]	-0.18***	[0.000]	0.87*	[0.081]
	(0.04)		(0.33)		(0.04)		(0.50)	
By phone, text message, mail/email, etc.	-0.02*	[0.057]	0.04	[0.704]	-0.02	[0.110]	0.15	[0.297]
	(0.01)		(0.11)		(0.01)		(0.14)	
Transfer from child(ren)	-0.05	[0.217]	-0.09	[0.736]	-0.08**	[0.042]	0.29	[0.368]
	(0.04)		(0.28)		(0.04)		(0.32)	
Net transfer from child(ren)	-0.07	[0.124]	0.00	[0.988]	-0.09**	[0.015]	0.44	[0.200]
	(0.04)		(0.28)		(0.04)		(0.34)	
N	5,620				3,944			

Notes: Parents receive a net transfer if the amount of transfer from children is larger than that to children. Standard errors in parentheses are clustered at the community level. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Appendix 3.A

Three vignettes are available for self-reported health in the affect domain in CHARLS baseline. They are:

(1) Wang Dong/Tang Jing enjoys his/her work and social activities and is generally satisfied with his/her life. He/She gets depressed every 3 weeks for a day or two and loses interest in what he/she usually enjoys but is able to carry on with his/her day-today activities. Overall in the last month, how much of a problem did Wang Dong/Tang Jing have with feeling sad, low, or depressed?

(2) Li Feng/Zhang Yan feels nervous and anxious. He/She worries and thinks negatively about the future, but feels better in the company of people or when doing something that really interests him/her. When he/she is alone he/she tends to feel useless and empty. Overall in the last month, how much of a problem did Li Feng/Zhang Yan have with feeling sad, low, or depressed?

(3) Zheng Bo/Wu Na feels depressed most of the time. He/She weeps frequently and feels hopeless about the future. He/She feels that he/she has become a burden on others and that he/she would be better dead. Overall in the last month, how much of a problem did Zheng Bo/Wu Na have with feeling sad, low, or depressed?

A randomly selected sample of respondents was asked to rate the severity of the problem experienced by the vignette person as (1) none, (2) mild, (3) moderate, (4) severe, or (5) extreme. The gender of the vignette person was randomly selected, but remained the same within a household.

Formally, we test reporting heterogeneity between parents with one and more children using the hierarchical ordered probit (HOPIT) model proposed by King et al. (2004).[23] The vignette component of the HOPIT model is specified as follows:

$$V_{ih}^{*} = \mu_h + \epsilon_{ih}^{v}, \qquad \epsilon_{ih}^{v} \sim N(0, \sigma_v^2)$$

$$v_{ih} = l \; if \; \tau_i^{l-1} < V_{ih}^{*} \leq \tau_i^{l} \;\; , \;\; l = 1, \ldots, 5 \;\; , \;\; \tau_i^{0} \leq \tau_i^{1} \leq \cdots \leq \tau_i^{5} \;\; , \;\; \tau_i^{0} = -\infty \;\; , \;\; \tau_i^{5} = +\infty, \forall i, h$$

where the observed rating v_{ih} for vignette *h*, *h*=1, 2, 3, by individual *i* is equal to *l* if the latent severity of the problem experienced by the vignette person, V_{ih}^{*}, falls between τ_i^{l-1} and τ_i^{l},

[23] Two assumptions are imposed. The first one is vignette equivalence: there is no systematic variation in the perceived level of health represented by the vignettes. Kapteyn et al. (2007) found vignette rating to be affected by the gender of the vignette person in the context of work disability. This is possibly because people assume less demanding work for women (as confirmed in Vermeer et al. 2016). Vignettes on health functioning do not appear to suffer from this problem, as is shown in Section 2.3.1. The second assumption is response consistency: respondents rate the vignettes according to the same criteria used when rating their own cases. In general, there has been little and mixed formal evidence on these two assumptions: on vignette equivalence, see Bago d'Uva et al. (2011), Kristensen and Johansson (2008), Murray et al. (2003), and Rice et al. (2011); on response consistency, see Bago d'Uva et al. (2011), Datta Gupta et al. (2010), and Van Soest et al. (2011).

and V_{ih}^{*} is modeled as the exogenously determined true severity, μ_h, plus a random error term. The four cut-points are modeled as:

$$\tau_i^l = \varphi_o^l OneChild_i + Z_i\varphi^l + \pi_c, \qquad l = 1, \ldots, 4$$

We first include in $\mathbf{Z_i}$ only a constant term (normalized to zero for l=1) and then obtain a richer specification by adding more controls. Specifically, we further include in $\mathbf{Z_i}$ five age categories, ethnicity, Hukou type, five education levels, dummies for being married and widowed, household per capita expenditure quintiles, and household size; and $\boldsymbol{\pi_c}$ includes administrative types, village status before 1980, urban status in 2011, and the autonomous region status. We are not able to identify all HOPIT coefficients for fathers in the rich specification, thus, only results from the basic specification are presented.

Using the individual specific cut-points determined by the vignette component, the self-assessment component is similar to an interval regression:

$$H_i^* = \delta_o OneChild_i + Z_i\delta + \pi_c + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$
$$h_i = l \; if \; \tau_i^{l-1} < H_i^* \leq \tau_i^l$$

where Z_i is defined as above (including a constant term) and σ^2 is normalized to 1 for identification. The vignette component and the self-assessment component of the HOPIT model are estimated jointly for efficiency.

Table 3.A.1 presents the coefficient estimate of the one child dummy in a standard ordered probit model, the HOPIT model and the cut-points. The last row presents the test statistics for reporting homogeneity, i.e. of the null hypothesis: $\varphi_o^1 = \varphi_o^2 = \varphi_o^3 = \varphi_o^4 = 0$. Though the one child dummy is significant in cut-point 2 for fathers, the *p*-values suggest that reporting homogeneity cannot be rejected in any case.

Table 3.A.1. One child dummy coefficient estimates and test for reporting homogeneity.

	Mothers			Fathers	
	Ordered probit	HOPIT		Ordered probit	HOPIT
		Basic spec.	Rich spec.		Basic spec.
Affect	-0.307	-0.086	-0.110	-0.160	-0.251
	(0.215)	(0.224)	(0.226)	(0.269)	(0.326)
Cut-point 1		0.104	0.064		0.170
		(0.143)	(0.134)		(0.120)
Cut-point 2		0.155	0.030		0.186**
		(0.102)	(0.108)		(0.091)
Cut-point 3		0.084	0.019		0.033
		(0.096)	(0.095)		(0.083)
Cut-point 4		0.072	0.085		-0.006
		(0.116)	(0.126)		(0.103)
p-value for reporting homogeneity		0.596	0.943		0.303
N	363			251	

Notes: The basic specification allows the cutpoints to differ only by the one child status; the richer specification allows the cutpoints to differ additionally by (1) individual characteristics, including age categories, ethnicity, Hukou type, education levels, marital status, household per capita expenditure quintiles, household size; and (2) community characteristics, including administrative types, village status before 1980, urban status in 2011, and the autonomous region status. Standard errors in parentheses. ** $p<0.05$.

CHAPTER 4

Can a Bottom-up Results-based Reform Improve Health Care System Performance? Evidence from the Rural Health Project in China

Joint work with Eddy van Doorslaer, Ling Xu, Yaoguang Zhang, Joris van de Klundert

In 2008, a six-year system-wide reform project – the Rural Health Project (Health XI) – was initiated in 40 counties of China to pilot interventions aimed at improving the financing and delivery of medical care and public health services. Due to substantial variation in local conditions, project counties were allowed to design their specific interventions but faced the same pre-specified project targets. We estimate the project effects using a difference-in-differences method, with control counties selected from a national survey based on socioeconomic conditions over eight years (four years before and four years after project initiation). Results show that, by 2013, Health XI has improved some outcomes in all three domains examined – medical care, public health services and self-rated health. We find particularly strong evidence for a decrease in outpatient expenditure, intravenous drip use and financial strain, and an increase in all four components of public health services provision. Further exploration of effect heterogeneity reveals that previous experience with a similar reform project did not bring about additional Health XI benefits, and low local financial capacity was not always a barrier to meeting project targets. In general, our results suggest that interventions adopted in the bottom-up and results-based approach generated substantial benefits given the investment.

4.1 Introduction

After the turn of the new millennium, China launched a series of major initiatives to improve the health care system. For example, the *New Rural Cooperative Medical Scheme* (NRCMS) was introduced in 2003 to provide health insurance to rural residents not formally employed.[1] At the same time, a complementary safety-net program, the *Medical Assistance* program (MA), was promoted to provide financial assistance with the NRCMS premium and other medical expenses for the poor. In 2002, the *Chinese Center for Disease Control and Prevention* was established. Soon thereafter, the public health system received major investment and reconstruction following the Severe Acute Respiratory Syndrome (SARS) outbreak in 2003. Despite all these efforts, coverage was far from complete due to the limited resources available for NRCMS, MA and the public health system. For example, the actual inpatient reimbursement rate for rural residents was only 32.9% five years after the introduction of NRCMS (Meng et al. 2012). Clearly, China faced major challenges in improving health care system performance under limited funding, especially in rural areas.

In this context, the eleventh World Bank health project in China*, Rural Health Project* (*Health XI*), was launched in October 2008 in 40 counties spreading across eight provinces (World Bank 2015a).[2, 3] It aimed at improving the financing and delivery of medical care and public health services by piloting a series of interventions over six years. However, the interventions were not uniform due to the complexity of system-wide reforms and the substantial geographic and socioeconomic variation in project counties. Instead, Health XI adopted a bottom-up approach – i.e. counties were given discretion in the specific design of interventions so as to meet local conditions. Project management was results-based using a pre-specified monitoring and evaluation framework. Participating counties were informed at the beginning about the target values of 22 monitoring and evaluation (M&E) indicators to be used for the final assessment. These indicators cover medical care use and financial burden, public health services provision (e.g., gynecological checkups, hypertension follow-up), project experience dissemination and so on (see Appendix Table 4.A.1 for a full list and the target values). To meet these targets, project counties proposed annual activity plans and executed them after receiving approval from the national Project Management Office. Funding disbursement was conditional on the completion of these approved activities.

We perform an evaluation of the early effects of Health XI using household survey data collected in the summers of 2008 and 2013. We select control counties by matching on county socioeconomic conditions and use a difference-in-differences (DID) method to estimate the

[1] It is estimated that 80% of China's rural population – approximately 640 million people – lacked health insurance by 2003 (Centre for Health Statistics and Information of Ministry of Health 2004).

[2] Project provinces are Gansu province, Heilongjiang province, Henan province, Jiangsu province, Qinghai province, Shaanxi province, Shanxi province, and Chongqing municipality. For the sake of brevity, province-level administrative regions including provinces and municipalities are all referred to as provinces; county-level administrative regions including city districts and counties are all referred to as counties in the main text.

[3] Project webpage can be found on the World Bank's website at: http://www.worldbank.org/projects/P084437/rural-health-project?lang=en&tab=overview.

project's effects. Among a multitude of potential outcomes, we use those that measure or proxy the pre-specified M&E indicators to minimize the multiple inference problem. However, a common concern with results-based projects is that they can give rise to multitasking problems, with efforts being concentrated on incentivized indicators at the expense of non-incentivized ones (Holmstrom and Milgrom 1991; Eggleston 2005; Miller and Babiarz 2013; Sherry 2016). In order to provide a more complete picture of the success and failure of project incentives, we also evaluate two non-M&E indicators that are closely related to the incentivized ones, namely outpatient expenditure and health checkups.[4] The list of outcomes then includes (1) for medical care, outpatient and inpatient utilization and expenses, two quality measures (outpatient intravenous drip use and caesarean section), and financial strain; and (2) for public health services provided by primary health care facilities, health checkups, gynecological checkups, availability of female medical staff for female patients and hypertension follow-up. Finally, we also examine self-rated health, as population health improvement was the ultimate goal of the interventions. Measures used include the EQ-5D descriptive system and the EQ visual analogue scale (EQ VAS).

Our findings suggest that Health XI achieved significant improvements in all three domains – medical care, public health services and self-rated health, though not in all individual components. In the medical care domain, we find strong evidence that Health XI significantly decreased (1) outpatient expenditure by 86 Yuan, (2) outpatient intravenous drip use by 8.7 percentage points, and (3) the incidence of catastrophic medical expenditure by 5.8 percentage points and of medical impoverishment by 3.0 percentage points. In the public health services domain, we also find strong evidence for a significant and substantial increase in all four components, namely health checkups (24.4 percentage points), gynecological checkups (24.6 percentage points), the probability of having female medical staff when desired by female patients (22.8 percentage points), and hypertension follow-up (16.5 percentage points). For self-rated health, when analyzing effects on separate health domains, we observe that difficulty in mobility and pain/discomfort decreased significantly. However, these changes did not lead to a significant rise of the overall health rating. Given the project design, we are not able to attribute the effects to specific intervention(s) operating in a well-defined context. Instead, we explore effect heterogeneity and find that counties did not benefit more from previous experience with a similar project or from being in a better fiscal condition.

Our paper adds to several lines of literature. First, it supplements the limited research on the impact of system-wide health reforms involving both supply and demand side interventions in developing countries (Wagstaff & Yu 2007; King et al. 2009; Gruber et al. 2014; Limwattananon et al. 2015; Powell-Jackson et al. 2015). The most relevant study is the evaluation of the World Bank's Health VIII project in China (Wagstaff & Yu 2007).[5] This

[4] Contextual knowledge and economic theory predict a close relationship of the non-incentivized outcomes with the incentivized ones. For example, health care providers may increase income-generating practices in outpatient care when such practices are restricted in inpatient care; outpatient reimbursement competes with inpatient reimbursement in NRCMS; and basic public health services share the same government funding. Therefore, though the addition of outcomes increases the number of tests, it is unlikely to noticeably increase the number of rejections by chance.

[5] Project webpage at http://www.worldbank.org/projects/P003566/cn-basic-health-hlth8?lang=en.

project aimed to resuscitate the old commune-based rural health insurance (the precursor of NRCMS) and to establish MA. The bulk of the project budget was devoted to supply side interventions, including upgrading township health centers, improving the medical information system, and quality improving measures such as the promotion of clinical pathways (treatment protocols). Health VIII also supported region-specific 'priority' interventions, e.g., maternal and child health or infectious disease interventions. Exploiting a pre-existing survey in one project province, Wagstaff and Yu (2007) found little change in health care use, reduced financial strain and mixed health effects.

Second, our paper relates to the literature evaluating results-based financing programs that link funding to performance. In spite of the increasing interest in using this approach in development (especially health) programs (World Bank 2015b), its effectiveness is still under debate, with one concern being the aforementioned multitasking problem. However, favorable evidence was found in an evaluation of the PNPM Generasi program in Indonesia (Olken et al. 2014). A randomized experiment was set up to isolate incentive effects from treatment effects by assigning treatment areas into one incentivized and one non-incentivized group. The study found that incentives speeded up the treatment impact on preventive health indicators, but detected no differential impacts on education, nor any evidence of multitasking problems for a wide array of measures.[6] While the link between funding and performance took a somewhat different form in Health XI than in Olken et al. (2014),[7] we similarly do not find any evidence of multitasking problems. On the contrary, our results even suggest positive spillovers on non-incentivized indicators, possibly because these non-incentivized indicators share common inputs with some of the incentivized ones.

Last, the bottom-up feature of Health XI echoes the advocacy for local participation in the design and implementation of development programs (Mansuri & Rao 2012). The two major modalities for inducing local participation are decentralization (as in Health XI) and community development (Mansuri & Rao 2012). While no consensus has been reached on their effectiveness (Mansuri & Rao 2012), favorable evidence was provided by a community development program in Sierra Leone aimed at reforming local institutions (Casey et al. 2012). The study found positive short-run effects on local public goods provision and economic outcomes, but no sustained impacts on institutional outcomes such as collective action, decision making and involvement of marginalized groups. Our results are similar in that we find Health XI to be highly successful in increasing public health services provision, but much less so in curbing the fast-growing inpatient expenditure, which results from the institutionalized practice of overprescribing drugs and medical tests to cross-subsidize underpriced medical services and generate income.

---

[6] Although the reduction in neonatal mortality did not persist in incentivized treatment areas, both multitasking problem – quantity (incentivized) of prenatal care increased at the expense of quality (non-incentivized) – and a lack of it – improved prenatal care resulting in more marginal births – can explain the phenomenon. The authors did not have enough information to distinguish between the two hypotheses.

[7] In Olken et al. (2014), performance on incentivized indicators directly determined 20 percent of the block grant in the subsequent year, whereas in Health XI, funding disbursement was indirectly determined by incentivized indicators through a county's completion of activities aimed at improving these indicators.

The remainder of the paper is structured as follows. Section 4.2 provides background information on Health XI. Section 4.3 introduces data and variables. Section 4.4 describes our empirical strategy. Section 4.5 presents the results. Section 4.6 discusses and concludes. Appendix 4.A provides additional details and results.

4.2 Institutional Background and Health XI

For reasons of space, we only briefly sketch some relevant features of the health care system in rural China.[8] Before doing so, three general features of China's health care system are worth mentioning.[9] First, public institutions are the main provider of health services and multiple government departments besides the health department play a key role in the system. For example, the Finance Bureau pays for the basic salary of health professionals in budgeted posts, the Development and Reform Commission sets prices for medical services and drugs, the Human Resources and Social Security Bureau sometimes replaces the health bureau in managing NRCMS, the Civil Affairs Bureau manages MA, and the local government appoints and dismisses the directors of public hospitals. Second, public providers receive insufficient government funding and are paid by fee-for-service. As basic health service charges are set below costs, providers cross-subsidize and generate income by overprescribing drugs and medical tests, for which markups are allowed.[10] Third, specialist services at hospitals are overused because referral is not required and quality of care is low at primary health care facilities (Bhattacharyya et al. 2011; Sylvia et al. 2015).

The health care network in rural China has three tiers: county hospitals, township health centers (THC) and village clinics. While all provides medical care, public health services are mainly provided by primary health care facilities, i.e. THC and village clinics. Financial protection against medical spending is provided by the heavily subsidized insurance scheme NRCMS[11] and the safety net MA. Both programs are highly decentralized with local governments in charge of policy design and implementation, and responsible for a large proportion of the funding. This has led to considerable variation in generosity (covered

[8] Interested readers are referred to Wagstaff et al. (2009), who reviewed the developments and reform directions of the rural health system. This paper was also used as a key reference to guide the design of Health XI.

[9] While the three features remain largely, exceptions are emerging in recent years. For example, the Development and Reform Commission has relaxed its control on the price of most drugs since late 2015.

[10] In November 2016, the central government explicitly demanded that drug markups be removed in all public hospitals and health service fees be adjusted accordingly. Source: http://www.gov.cn/zhengce/2016-11/08/content_5130271.htm.

[11] NRCMS coverage reached 91.5% of the population by the end of 2008. Source: http://www.nhfpc.gov.cn/tigs/s9664/200904/eec4490ff8484103b849abbc6b7e5d26.shtml.

services and drugs, reimbursement rate, etc.) across counties.[12] Moreover, NRCMS and MA typically undergo annual revisions in design. As local governments often determine expenditure based on revenues, increase in generosity is highly dependent on local program funds and fiscal capacity. However, despite the regional variation, coverage of NRCMS and MA has generally been low (Wagstaff et al. 2009). For example, due to the limited funding, outpatient expenses were hardly reimbursed[13] by NRCMS in early years although research showed that they could be a major cause of medical impoverishment for the chronically ill (Yip & Hsiao 2009). The situation is similar for public health services – funding is limited and dependent on local conditions, and substantial regional variations exist.

Health XI was initiated with the broad intention to learn lessons from project interventions and develop viable models that can be scaled up to strengthen the health care system in different conditions nationwide (World Bank 2015a). Thus, the eight project provinces were selected to cover a wide range of the geographic and socioeconomic spectrum. They range from Jiangsu, one of the richest provinces in east China with a per capita GDP of 39,622 Yuan ($6340)[14] in 2008, to Gansu, one of the poorest province in the west with a per capita GDP of 12,110 Yuan (Ministry of Finance 2009). The 40 project counties represented a total population of 21 million by 2010.[15] Health XI was funded by a $50 million loan from the World Bank and a £5 million grant from the Department for International Development, UK. The loan was to be repaid by the central (80%) and provincial (20%) governments, instead of project counties, as in previous World Bank health projects. Moreover, Health XI abandoned the usual requirement on counterpart funding, which posed a significant entry barrier for poor counties in Health VIII (Wagstaff & Yu 2007). It was also required that Health XI funds do not crowd out governmental fiscal funds, nor substitute it in funding NRCMS or health workers' salary. These financial arrangements made participation less selective in terms of a county's fiscal capacity. In addition, provincial governments were explicitly instructed to enroll representative counties into Health XI to increase project replicability. For these reasons, it is unlikely that only counties with high implementation capacity participated.[16]

The two objectives of Health XI were: (1) health reform innovations supported via block grants, and (2) project coordination, policy development and lesson learning. Three reform aims were identified under the first objective. The first aim was to improve rural health

[12] For an illustration, Table I in Hou et al. (2014) showed some differences in NRCMS design in six counties in 2006 and 2008; Ma et al. (2011) described the differences in MA design in four counties in 2006.

[13] Initially, outpatient care was typically covered through a medical savings account that consisted of individual contributions and involved no risk pooling. Strictly speaking, this was not considered as NRCMS reimbursement. Full transition to risk pooling for outpatient care was not called for until 2012. Source: http://www.mof.gov.cn/zhengwuxinxi/bulinggonggao/tongzhitonggao/201205/t20120528_654516.html.

[14] We use the same exchange rate, 1 Yuan = 0.16 Dollar, as in the World Bank final report (World Bank 2015a).

[15] Authors' calculation based on the 2010 Population Census.

[16] According to the internal final report, many counties also experienced changes in key officials during Health XI (due to reasons unrelated to the project). This further relieves the concern of selection bias with respect to officials' motivation and competence.

financing and the financial protection provided by NRCMS and MA. Twelve percent of the total project spending went into this area. One major supply side intervention was the provider payment reform, implemented for outpatient care in 28 project counties and for inpatient care in all (Zhang et al. 2014). Counties varied in their choices of the new payment system and the coverage of hospitals and diagnoses given their different human resources and IT infrastructure. The majority piloted a mixed payment system, the most popular one being case-mix for single diseases plus per-diem payment (Zhang et al. 2014). Across counties, coverage of the new payment system varied from the major county hospital(s) to full coverage, and from a few tens of common diagnoses to full coverage (Zhang et al. 2014). On the demand side, the NRCMS design became more generous with improved financing. Implementation of MA improved, as its information system was better linked up with that of NRCMS to cover NRCMS premium and out-of-pocket payments.

The second reform aim was to improve service delivery capacity, efficiency and quality, which absorbed 45 percent of the total spending. Relatedly, the third aim was to strengthen core public health functions, which absorbed 18 percent of the total spending. Specifically, about 30 percent of the total spending went to village clinics (1109 constructed, 599 expanded and 582 renovated) and equipping them with clinical and office equipment. To improve public health services provision, some counties refined the current capitation payment by adjusting for estimated cost of service delivery, and some switched to purchasing public health services from local providers. Better fringe benefits, e.g., health insurance and pension arrangements, were provided to recruit and retain primary health workers. Training was used to enhance capacity and clinical pathways were used, often as a part of the provider payment reform, to improve service quality.

As is clear from the above explanation, the three reform aims are closely related. One intervention can affect more than one reform aim and one M&E indicator can benefit from various interventions. Therefore, instead of relating our outcomes to reform aims, we group them using more distinct categories, namely medical care, public health services and self-rated health. In addition, we do not explicitly evaluate the second Health XI objective in this paper due to its qualitative nature. This is not to deny that it contributed to project implementation by establishing tighter supervision and more efficient management. For example, before receiving an approval, the annual activity plans were carefully examined by the Center for Project Supervision and Management (CPSM) of the National Health and Family Planning Commission (NHFPC), the agency responsible for overall coordination of Health XI at the nation level. Revisions were required and assistance provided if the plan was considered insufficient.[17] Progress monitoring was active and feedback loops were short. Technical supports, policy debates and special studies were also provided timely when necessary.

[17] This typically occurred at the early stage of Health XI when project counties were new to the bottom-up approach and proposed routine activities in their annual activity plans.

4.3 Data and Variables

Our analysis draws on pre- and post-reform household data from the 2008 and 2013 waves of China's *National Health Services Survey* (NHSS), respectively. NHSS is a nationwide repeated cross-sectional survey administered by the National Health and Family Planning Commission. It has been conducted at five-year intervals since 1993. The survey covers China's 31 provinces, and employs a multistage cluster sampling procedure, stratified by province (Meng et al. 2012). Five townships are randomly selected from each county, and then two villages from each township, and last, 60 households from each village, giving around 2000 respondents[18] per county. Health XI was rolled out shortly after the 2008 wave of NHSS so that pre- and post-reform data could be collected conveniently by including project counties into NHSS in 2008 and 2013. Our data access is restricted to the eight Health XI provinces, which leaves us with 40 project and 22 non-project counties (see Appendix Table 4.A.2 for a list).

Outcome variables draw on data obtained from five modules of the NHSS household questionnaire. The family module provides information on household composition, income, expenditure (total and by category), etc. The individual module records information on demographics, socioeconomic status, chronic conditions, health insurance type, health checkup, etc. Health is measured using the three-level version of EQ-5D (EQ-5D-3L), a widely used instrument developed by the EuroQol Group (The EuroQol Group 1990).[19] EQ-5D-3L has two components. One is an EQ-5D descriptive system comprising five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has three levels: no problems, some problems and extreme problems. The scoring, i.e. a health state, can be converted into a single index using the value set, which is a set of weights for each level in each domain elicited from population samples. Another component of EQ-5D-3L is a visual analogue scale (EQ VAS) that asks respondents to rate their own health on a scale of 0 to 100.

The outpatient module is administered to respondents who report illness in the past two weeks. It asks about the illness type (recent illness, acute illness persisting into the past two weeks, chronic condition relapse), treatment type (no treatment, self-treatment, on doctor's orders, outpatient visit), type of health care provider visited (from village clinics up to provincial hospitals), treatment details (e.g., whether had an intravenous drip), and expenditure (gross[20] in 2008 and OOP in 2013). The inpatient module is administered to respondents in need of hospitalization (judged by a physician) in the past 12 months. It asks about the type of health care provider visited (from THC up to provincial hospitals), gross expenditure and reimbursement. For comparability between 2008 and 2013, the first recorded (i.e. the most

[18] Some Health XI and control counties have around 3000 observations in 2013 due to a sample increase in certain regions in the 2013 wave. Different sample sizes are adjusted for in our analysis using analytical weights.

[19] Sun et al. (2011) validated the use of EQ-5D in China.

[20] The gross expenditure can be seen as rough approximation of the OOP amount because NRCMS reimbursement for outpatient expenditure has been very limited, and only about one quarter of the outpatients reported any reimbursement or exemption in both treatment and control counties.

serious) outpatient visit and the latest hospitalization episode of each respondent is used. The ever-married women module asks women aged 15-49 about gynecological checkups, availability of female medical staff upon request at THCs,[21] and detailed information about deliveries with a five-year recall.

Subject to data availability and suitability for individual level analysis, our list of outcomes includes half of the M&E indicators measured using either the original indicators or proxies (see Appendix Table 4.A.1 for a detailed explanation). All outcomes are categorized into three domains – medical care, public health services provision at primary health care facilities, and self-rated health. While outcomes in the latter two domains are examined using the full sample, medical care is examined using only NRCMS enrollees (90% of the full sample) due to the project design.[22] Specifically, utilization measures include outpatient and inpatient care use conditional on illness, which are then examined by provider level to reveal changes in patient flow. Outpatient, gross and OOP inpatient expenditures[23] are examined separately, and also by provider level as Health XI had no control over providers higher than county hospitals (i.e. urban hospitals). Outpatient expenditure is added because outpatient coverage is an integral part of NRCMS, but has been traded off against better inpatient coverage. This may continue to be the case during Health XI due to its absence among the incentivized M&E indicator. OOP inpatient expenditure is added because it is the actual expense borne by the patient, and it complements the M&E indicator – OOP rate of inpatient expenditure.

Quality measures include the use of outpatient intravenous drips and caesarean section rate. Because NHSS did not ask for antibiotic use, the original M&E indicator, we use outpatient intravenous drip instead as it is mostly intravenous antibiotics, and has itself been cautioned against by the government in order to reduce antibiotic abuse.[24] Reduction in both (bad) quality measures is desirable as China is known to have serious antibiotic abuse (Wang et al. 2014) and an excessively high caesarean section rate (46.2% in 2007-08 according to Lumbiganon et al. (2010)). We additionally examine intravenous drip by provider level to see if compliance with clinical pathways was different. Measures of financial strain include two original M&E indicators – OOP rate of inpatient expenditure (additionally by provider level) and catastrophic medical expenditure (annual household medical expenditure>20% of household income). Medical impoverishment is included to reflect whether the poor and the vulnerable have received better financial protection from the improved MA targeting and link-up with NRCMS.

[21] Though the question on the availability of female medical staff was asked in all project counties irrespective of their rural/urban status as defined by the National Bureau of Statistics, it was not asked in three non-project counties defined as urban. When such an urban county is selected as control, the matched pair(s) is not included in the estimation for this particular outcome.

[22] In a few counties that had integrated NRCMS and the Urban Resident Basic Medical Insurance by 2013, people covered by the unified insurance are all counted as NRCMS enrollees. For two outcomes that are calculated at the household level, i.e. catastrophic medical expenditure and medical impoverishment, we add back non-NRCMS-enrollees (3% of the full sample) from households with at least one NRCMS enrollee.

[23] Expenditures are trimmed at the 99.5 percentile.

[24] http://news.xinhuanet.com/mrdx/2016-09/09/c_135674850.htm

The public health services provision domain consists of three original M&E indicators – gynecological checkups, availability of female medical staff for female patients at THCs, and follow-up for hypertensive patients aged 35+. We supplement the list with health checkups because, like gynecological checkups and hypertension follow-up, it is one component of the public health services. As all public health services share the same government funding, health checkups is a potential victim of resource reallocation given its absence among the M&E indicators. Last but not least, the self-rated health domain includes two summary indices: EQ-5D value score and EQ VAS score. We also report the five dimensions of the EQ-5D descriptive system because the Chinese value set was elicited from university staff and students in cities (Liu et al. 2014) and therefore may not apply to rural residents.

County socioeconomic conditions include whether the county is autonomous (due to a high proportion of ethnic minorities)[25] and three county-level economic indicators. The autonomous county status can be directly determined from the county names, and economic indicators are extracted from various statistical yearbooks. Specifically, we extract GDP, fiscal expenditure and rural net income from provincial Statistical Yearbooks and China Statistical Yearbooks for Regional Economy during 2006-2013. Per capita terms are then calculated using yearly resident population imputed from the Population Census in 2000 and 2010, assuming constant population growth rate.[26] Three non-project counties[27] are dropped due to a lack of such statistics, leaving us with 19 potential controls. After dropping respondents missing key variables, our final sample consists of 255,294 observations.[28] Descriptive statistics are presented in the results section, after we have explained the empirical strategy below.

4.4 Empirical Strategy

We estimate the impacts of Health XI using a DID framework combined with matching. Due to a lack of data on multiple pre-treatment periods, it is not possible to select control counties by assessing how well they satisfy the common trend assumption. As an alternative strategy, we select controls by matching on county economic indicators during 2005-2012. We also match on the autonomous county status (yes/no) to take into account the different policy environments (e.g., government priorities) and other differences originating from the ethnic composition in autonomous counties. We do not risk reducing balance on these county characteristics by matching additionally on sample characteristics. This is because local

[25] Autonomous areas are often remote and poor, but have special rights and enjoy more favorable and policies. Appendix Table 4.A.2 shows which of the sample counties are autonomous.

[26] We first impute year-end resident population assuming constant growth rate, and then average two year-end numbers to get the yearly average.

[27] They are Beilin district of Suihua city, Jinchang district of Suzhou city, and Weibin district of Xinxiang city. They all are city districts of prefecture-level cities, whose statistics have been rarely reported in national or provincial statistical yearbooks.

[28] Dropped observations account for 1.2% of the original sample, and are not correlated with treatment status.

socioeconomic conditions are more important in determining comparability than sample characteristics, where small imbalances can occur by chance and be conveniently adjusted for in parametric models.

Economic indicators used for matching are per capita GDP, per capita fiscal expenditure, and rural net income during 2005-2012. First, they are valid matching variables because they could not predict Health XI participation[29] and were not affected by Health XI during project implementation. Second, they are important contextual variables because they reflect local economic development, government fiscal capacity, and rural people's level of consumption. As is explained in section 4.2, NRCMS and MA generosity and public health services provision are highly dependent on local economic conditions and fiscal capacity. Consumption of medical services is also heavily influenced by income level. Last, the time span of eight years consists of approximately four years before and four years after the project initiation. This affords the opportunity to assess comparability of control counties not only before, but more importantly also during Health XI because the project operated under a fast evolving economic and policy environment in China.

Three major events might confound our estimates. One is the 2008 financial crisis. It had an uneven impact across China, which we take into account by matching on per capita GDP. Another event is the new round of national health reform initiated in 2009. The guiding action plan identified the main tasks as to provide universal health insurance, establish a national essential drug system, strengthen the medical care and public health service system at grassroots level, promote basic public health services, and launch the pilot reform of public hospitals (Z. Chen 2009). As usual, design and implementation of specific reforms were largely left to local governments and critically depended on local fiscal capacity. Matching on per capita fiscal expenditure thus reduces the possibility that this round of reform differentially affected treatment and control counties. The third event is a drastic increase in the rural poverty line from 1,196 Yuan in 2009 to 2,300 Yuan in 2011.[30] Matching on rural net income therefore reduces the possibility that poverty reduction efforts were more vigorous in one group than in the other.

The next matching decision concerns the method. The idea of using a weighted control is appealing, but it has important drawbacks. As shown in Abadie et al. (2012), extrapolation bias is often difficult to avoid using traditional matching methods because weights are not restricted to be positive. Though synthetic control methods safeguard against extrapolation, interpolation bias can be severe when matching variables have a large support and a non-linear relationship with the outcome variables, which is likely in our case. Thus, preprocessing is advised to restrict the pool of potential controls to those similar to the treatment unit. However, an appropriate caliper for pruning is case-specific, which in this application makes it

[29] Regression results are available upon request. The reasoning behind this claim is that, if these variables are significantly correlated with Health XI participation, matching on them may increase the imbalance in important omitted variables that prevented the control county from participating in Health XI.

[30] Source: http://www.gov.cn/jrzg/2011-12/02/content_2009471.htm

difficult to conduct certain permutation tests. Therefore, we use 1:1 nearest neighbor matching in the main analysis and "synthetic controls" only as a robustness check.[31]

An examination of popular distance metrics shows that Euclidean distance matching works, and performs better without normalization as judged by the larger minimum *p*-value from the balance check (0.380 versus 0.284).[32] A graphical examination reveals that normalization leads to several cases of significant worsening in the comparability of per capita GDP in exchange for slight improvement in the comparability of per capita fiscal expenditure and/or rural net income.[33] This is not surprising as the standard deviation of per capita GDP is larger than that of the other two indicators. Taken together, Euclidean distance is preferred for the main analysis (see Appendix Table 4.A.3 for the matched pairs).

Table 4.1 provides a balance check. A comparison of the first three columns shows that matching leads to a considerable improvement in balance for GDP per capita and fiscal expenditure per capita, both in means and standard deviations. Though balance in rural net income worsened in some cases, the discrepancy in means and standard deviations remains reasonably small. The fourth column presents the standardized bias calculated as the difference between treatment and matched control means divided by treatment standard deviation. All are below the conventional threshold of 0.25 (Imai et al. 2008), except for fiscal expenditure per capita in 2008 and 2009. In the last column, we present *p*-values from tests of no difference in means between treatment and matched control counties.[34] None of the *p*-values come close to significance, though partly because of the large standard deviations as a result of substantial regional variation. Nevertheless, the last two columns suggest reasonably good balance in treatment and matched control counties.

The Health XI treatment effect is estimated using a standard DID model specified as the following:

---

[31] The "synthetic controls" are constructed as follows. We first use a county-specific caliper to preprocess the data. For a treatment county with the minimum value y_{min} in economic indicator Y, the caliper is set to be $1.5y_{min}$, and for a treatment county with the maximum value y_{max}, the caliper is set to be $0.5y_{max}$. The scaling factor is linearly decreasing in the value of Y to account for the fact that whether a difference of 1000 Yuan is large depends on the value of Y. The pool of potential controls for a treatment county consists of non-project counties that are of the same autonomous county status (yes/no) and whose 24 economic indicators are all within their respective calipers (except for one treatment county where we allow non-project counties with two economic indicators outside their calipers to avoid an empty pool). We then obtain "synthetic controls" by concatenating 24 economic indicators into one variable and passing it off as an outcome with 24 years of pre-treatment values (and one post-treatment value arbitrarily set to 0) to the synthetic control algorithm. We can do so because the matching is invariant to the time order of pre-treatment values. Two sets of "synthetic controls" are constructed using this procedure, one with the original values of the economic indicators and one with the standardized values.

[32] The Mahalanobis distance fails to produce good balance given the large number of covariates (Gu and Rosenbaum, 1993), and neither is propensity score matching suitable because matching variables have no predictive power for the treatment status.

[33] Figures and balance check results are available upon request.

[34] Though Imai et al. (2008) argued that hypothesis tests are irrelevant as balance is a characteristic of the sample, not some hypothetical population, we nevertheless report *p*-values as is common practice.

$$Y_{ict} = \alpha + \gamma HXI_c + \lambda Post_t + \delta(HXI_c \times Post_t) + X'_{ict}\beta + \varepsilon_{ict} \quad (4.1)$$

where i denotes an individual, c denotes a county, t denotes a year, HXI_c is a dummy for treatment counties, $Post_t$ is a dummy for year 2013, and X'_{ict} includes a series of individual characteristics, including age categories, gender, ethnicity, marital status, educational level, employment status, chronic diseases,[35] household composition and income quartiles, as well as matched pair fixed effects.[36] Relevant controls are dropped when outcomes do not apply to children, men, or never-married women. Throughout our analysis, standard errors are clustered at the county level, the level of treatment.[37] Given the absence of sampling weights, we generate analytical weights to weight samples per county per wave equally instead of in proportion to population.[38] In this way, we estimate simple average treatment effects across the 40 Health XI counties unweighted by county population.[39]

4.5 Results

Table 4.2 provides sample descriptive statistics by treatment status. The sample from treatment counties does not differ from the control counties for a long list of demographic and socioeconomic variables, except for the proportion of children under age 15. However, the difference, though significant, is not substantial and is adjusted for in the estimation model. The results for income and poverty can be seen as a *falsification test*. The change in household income per capita and poverty rate did not differ significantly between treatment and control

[35] Chronic conditions were defined so loosely by NHSS that even a persistent cold was recorded. However, we only count those commonly used as chronic diseases. They are arthritis, hypertension, diabetes, heart diseases, lung diseases, tuberculosis, liver diseases, gall diseases, nephritis, spinal disc herniation, benign tumors, malignant tumors, mental disorders, Parkinson's disease, cataract and glaucoma.

[36] Specifically, they are age categories: 0-14, 15-24 (omitted), 25-34, 35-44, 45-54, 55-64, and 65+; female; ethnic minority; marital status: not married (omitted), single, and married; educational level: no education (omitted), primary school, secondary school, and high school and above; employment status: employed (omitted), student, jobless, and retired; dummies for the number of chronic diseases: 0 (omitted), 1, 2, 3, and 4+; household composition: dummies for number of adults up to 6 and children up to 4; income quartiles (lowest 20% omitted). Variables not defined for children (aged under 15) are recoded into zero and the dummy for age 0-14 captures the effect of being children.

[37] To adjust for multiple testing with respect to the CES-D 10 components, we calculate the family-wise *p*-values based on 1000 iterations proposed in Westfall & Young (1993), following the practice of Finkelstein et al. (2012). Results are not reported, however, because they are higher than conventional levels of significance in all cases, except for the 'can't get going' component in the husband-present subsample.

[38] As the sample size per county per wave fluctuates around 2000, analytical weights are conveniently calculated as 2000 divided by the sample size.

[39] Given the broad intention of Health XI in finding replicable models to improve the health system, we are more interested in evaluating whether the majority of counties, rather than of people, benefited from the project.

counties.[40] This suggests that the matching is reasonably good and relieves the concern that the estimates may be confounded by differential economic development or poverty reduction efforts. The last six rows show that the treatment and control counties do not differ significantly in NRCMS coverage, the proportion with no health insurance, MA coverage, the proportion with chronic diseases, the proportion sick in the past two weeks (precondition for the outpatient module) and the proportion in need of hospitalization in the past 12 months (precondition for the inpatient module).

4.5.1 Main DID results

Table 4.3 presents baseline values, five-year changes and estimated program effects for the medical care domain. The first row in Panel A shows that there was an increase in the outpatient visit rate conditional on illness in both treatment and control counties, but the increase in treatment counties was significantly higher at the 10% level, by 6.9 percentage points. Recall that treatment type in the outpatient module includes no treatment, self-treatment, on doctor's orders and outpatient visit. Unreported results show that the increased rate of outpatient visit was driven by a decrease in the rate of no treatment. In addition, outpatient visits increased significantly and substantially for those with recent illness (31% of the illnesses), and also substantially but not significantly for those with a chronic condition relapse (64% of the illnesses). The next two rows of Table 4.3 show that while the majority of patients still sought care at primary health care facilities, there was a small increase in patient flow to higher-level providers in both treatment and control counties. The fourth row shows that the rate of hospitalization when needed increased in both treatment and control counties, but the difference is not significant. Again, we observe an increase in patient flow to higher-level providers but the trend was not significantly different between treatment and control counties.

Panel B of Table 4.3 displays a picture of a general increase in medical expenditure over the five years in both treatment and control counties. The first row suggests that Health XI led to a significant and substantial decrease by 88 Yuan in outpatient expenditure, which was mainly due to the fall at primary health care facilities. The fourth and eighth rows show that gross inpatient expenditure roughly doubled in both treatment and control counties, whereas the increase in the OOP amount was much more modest. However, treatment counties did not perform significantly better than control counties on either measure. Disaggregating inpatient expenditure by provider level shows a similar pattern – a substantial increase in gross expenditure and a more modest increase in OOP expenditure (except at THC) and no significant effect from Health XI.

Panel C of Table 4.3 reports results for the two quality measures. Intravenous drip was prescribed in more than one third of outpatient visits in 2008, a rate too high to be medically justifiable. The rate did, however, drop significantly in treatment counties while increasing

[40] Using Equation (4.1), the DID estimates further reduce to 456 (p-value=0.538) and -0.050 (p-value=0.513) without controlling for income quartiles, and 230 (p-value=0.727) and -0.001 (p-value=0.988) with the full set of controls.

slightly in control counties. Combined, the effect of Health XI is estimated to be a significant and substantial reduction of 8.7 percentage points. The next two rows show that the drop was significant and substantial at both THC and county-level hospitals. By contrast, caesarean section rate went up slightly in treatment counties but, since it increased even more in control counties, the treatment effect was still estimated to be (insignificantly) negative.

Panel D of Table 4.3 presents results for three measures of financial strain. The first row on OOP rate of inpatient expenditure shows an already large drop in control counties, likely due to a general increase in NRCMS generosity. Nevertheless, Health XI achieved a significant additional 6.6-percentage-point decrease. This reduction in OOP rate was more significant and substantial at county-level hospitals. While the incidence of catastrophic medical expenditure remained virtually the same in control counties, it decreased significantly in treatment counties, leading to a significant and substantial 5.8-percentage-point reduction due to Health XI. The last row focuses particularly on the poor and the vulnerable and examines the proportion of people falling below poverty line after deducting household medical expenditure from income. Results follow a similar pattern with catastrophic medical expenditure – incidence increased in control counties but decreased in treatment counties, leading to a significant and substantial reduction of 3 percentage points.

Table 4.4 presents results for public health services provision at primary health care facilities. The first four columns show that while a large increase was observed in all four measures in treatment counties, it was observed only for health checkups in control counties. Taken together, the fifth column shows that Health XI led to a significant and substantial increase of 24.4 percentage points in health checkups, 24.6 percentage points in gynecological checkups, 22.8 percentage points in the reporting of female medical staff available upon request, and 16.5 percentage points in follow-up of hypertension patients aged 35+. Unreported results show that, while increased availability of female medical staff partly derived from reduced no availability, it mostly came from fewer respondents reporting never requesting and not knowing. Women report these two answers typically because they never wanted to have gynecological checkups.[41] Therefore, the treatment effect does not necessarily mean that primary health care facilities hired more female medical staff. It could also be attributable to increased awareness and acceptance of gynecological checkups.

Table 4.5 presents results for self-rated health. The first row shows a small improvement in the EQ-5D value score in treatment counties and a small deterioration in control counties, together giving a positive but insignificant treatment effect. An examination of the respective five component dimensions of EQ-5D reveals that Health XI contributed to a significant reduction of 1.4 percentage points in self-reported difficulty in mobility and 3.5 percentage points in self-reported pain/discomfort. These effects were primarily driven by a significant worsening of these two indicators in the control counties. The bottom row presents another general health rating, constructed by the respondents based on their health conditions in various dimensions using their own weighting schemes. Results are similar for the self-rated EQ VAS score as for EQ-5D value score – the treatment effect was positive but insignificant.

[41] According to provincial proposals submitted at the beginning of Health XI, some women – especially those living in remote areas or of ethnic minorities – were still uncomfortable with – even against – the idea of gynecological checkups.

Apparently, significant improvements in mobility and pain/discomfort dimensions did not translate into a significant overall improvement, whether using the EQ-5D value set or respondents' own weighting.

4.5.2 Robustness checks

As robustness checks, we first examine whether our results were driven by one particularly well-performing treatment or one particularly poor-performing control county. The first test involves re-estimating treatment effects after dropping each of the 40 Health XI counties at a time. This procedure produces 40 new sets of estimates. The second test involves re-matching and re-estimating treatment effects after dropping each of the 13 matched control counties at a time. This produces 13 new sets of estimates. For reasons of space, we only present aggregate variables (i.e. not disaggregated by provider level) for which a significant effect was detected. It should be noted that, while the finding of a particularly well-performing Health XI county may invalidate our estimates, the finding of a particularly poor-performing control county does not necessarily do so. It could be the case that this control county is a valid one and that dropping it leads to poor matches for some treatment counties, resulting in less reliable estimates.

Figure 4.1 presents for each variable a density plot of the 40 DID estimates with corresponding *p*-values obtained after dropping one Health XI county at a time. For 10 out of 12 significant outcomes in the main analysis, variation in coefficient magnitude is limited and statistical significance maintained. The effect estimate becomes insignificant in some cases only for outpatient visit and OOP rate of inpatient expenditure, two of the 12 outcomes with a *p*-value larger than 0.05. Figure 4.2 presents results from dropping one control county at a time. As some Health XI counties are matched to second-best controls, variation in point estimates and *p*-values is larger than in Figure 4.1. Specifically, in addition to outpatient visit and OOP rate of inpatient expenditure, medical impoverishment and difficulty in mobility become insignificant in a few cases. However, most estimates remain significant with moderate variation in magnitude.

Next, we check robustness to using alternative matching methods (see Appendix Table 4.A.4 for a summary of the different sets of control counties). To save space, we only show point estimates and significance levels for aggregate variables. We deviate from our preferred matching first by removing the condition that treatment and control counties must be matched on the autonomous county status. The second column of Table 4.6 presents results using the new set of controls. As non-autonomous counties are matched to autonomous ones, falsification test results worsened as shown in Panel A. This may be because the autonomous county has greater inequality than the non-autonomous one with similar economic indicators, and has therefore put more effort into poverty reduction. As a result, the estimates for income increase and poverty reduction become larger as compared to the first column. Panel B of column 2 shows that 10 out of 12 significant outcomes remain significant, with limited changes in magnitude. Only OOP rate of inpatient expenditure and medical impoverishment lose significance. The less favorable results in financial strain outcomes, especially medical

impoverishment, are to be expected if autonomous counties devoted more attention and fiscal funds to poverty reduction, rather than to NRCMS and MA.[42]

Moving from column 2 to column 3 of Table 4.6, we add one restriction that the control must come from the same province as the treatment county. This takes better account of province-level characteristics and events at the expense of reducing balance in the economic indicators. Panel A shows that the new matching does not pass the falsification test using poverty rate. Nevertheless, Panel B shows that effects remain significant in outpatient expenditure and intravenous drip use, catastrophic medical expenditure, and all four components of the public health services provision domain. Last, we examine whether our results are robust to a substantially different matching method using "synthetic controls". The last two columns of Table 4.6 present results from using two sets of such controls, one constructed with the original values of the economic indicators and one with the normalized values, but both with the autonomous county status condition. Panel A shows reassuring results from falsification tests, and Panel B shows that all significant variables remain significant, the only exception being OOP rate of inpatient expenditure.

4.5.3 Treatment effect heterogeneity

After showing that our results are robust, a next question is whether (and how) Health XI effects differ by county characteristics. One testable hypothesis is that experience with a similar previous World Bank health project, Health VIII, may put a county at an advantage. Another relevant hypothesis is that poor counties may perform worse than richer ones because of their lower capability to innovate and implement reforms. We examine both hypotheses using a triple-difference (DDD) framework in which a dummy for Health VIII experience/state-designated poor county status is added to Equation (4.1 together with its interaction with HXI_c, $Post_t$ and $HXI_c \times Post_t$ (see Appendix Table 4.A.2 for county characteristics).[43] It is important to note that this exercise is of an exploratory nature as the validity of the identification strategy (be it experimental or observational) on the full sample does not necessarily extend to the subgroup analysis.

The first column of Table 4.7 shows that counties with Health VIII participation did not experience any significantly larger improvement in any outcomes. On the contrary, they show an incremental increase in caesarean section rate and self-reporting of pain/discomfort. The second column shows that treatment counties that have state-designated poor status performed significantly worse in containing gross and OOP inpatient expenditure growth, in reducing the caesarean section rate, and in improving self-reported health. However, they performed significantly better in providing health checkups and gynecological checkups. Admittedly, we cannot definitively attribute the heterogeneity to the characteristics examined because some other characteristics may be at play. For example, an examination of the

[42] Both poverty reduction programs and MA are managed by the Civil Affairs Bureau, and may compete for funding and human resources.

[43] As we are testing the interactive effects of the characteristic and Health XI, the value of the dummy is determined by the treatment county and assigned to the matched pair regardless of the characteristic of the control county.

baseline caesarean section rate in Health VIII treatment counties and state-designated poor treatment counties gives an average of 15% and 17% respectively. This seems to suggest that the incremental rise in caesarean section rate may be associated with a low baseline value rather than with the examined characteristics per se. Since the baseline values were close to the WHO recommended rate, the increase may not be considered as too worrisome. However, due to the lack of information on relevant clinical outcomes, we are unable to make any assessment. Last, for the self-reporting of pain/discomfort, we do not have a consistent explanation why Health VIII treatment counties and poor treatment counties did *not* enjoy the same improvement as the others.

4.6 Conclusion

In contrast to most health programs evaluated in the literature, Health XI is a particularly broad-based reform project that aimed at improving numerous aspects of both the medical care and the public health systems under limited funding. Recognizing that project counties have different needs and face different constraints, Health XI adopted a bottom-up approach. To effectively manage a project with a broad scope and decentralization, a results-based monitoring and evaluation framework was used. Using DID, we evaluate project effects on several aspects of medical care, public health services provision, and self-rated health.

Our results suggest that Health XI has led to improvements in all three domains, though not in all individual components, and the evidence is particularly strong for some outcomes in the medical care and public health services domains. Specifically, we find, first of all, a robust significant decrease in (i) outpatient expenditure (86 Yuan), (ii) outpatient intravenous drip use (8.7 percentage points), (iii) the incidence of catastrophic medical expenditure (5.8 percentage points), and (iv) medical impoverishment (3.0 percentage points). Secondly, we also find a robust significant increase in all four components of public health services domain, namely health checkups (24.4 percentage points), gynecological checkups (24.6 percentage points), availability of female medical staff (22.8 percentage points), and hypertension follow-up (16.5 percentage points). The estimated magnitude of these effects is remarkable considering that the project funding amounted to less than 1% of total government health expenditure in project counties (World Bank 2015a).

On the other hand, the insignificant effect on inpatient expenditure might reflect the difficulty in reforming the way in which inpatient care providers obtain the bulk of their income. The move from fee-for-service to more complicated payment systems is demanding in terms of human resources and IT infrastructure, and also requires cooperation from and coordination between multiple stakeholders (Zhang et al. 2014). Well-designed incentive structures and surveillance systems are required for the new payment systems to be effective, but these are difficult to accomplish in a short period of time. Poorly designed systems or systems without proper supervision are likely to have low compliance from the health care

providers. Thus, it is not completely surprising that we do not find an overall Health XI effect on inpatient expenditure.

In contrast to inpatient expenditure, we do find significant effects on outpatient expenditure. Together with the substantial increase in health checkups, our results suggest no multitasking problem but rather a positive spillover effect. One possible explanation is that Health XI interventions aimed to improve inputs to the production function of M&E indicators rather than the indicators alone, and many of these inputs are also in the production function of non-M&E indicators. For example, efforts to improve the quality of care, e.g., through the promotion of standard prescriptions for common diseases as in Huzhu county, may reduce the use of unnecessary drugs and decrease outpatient expenditure. Health checkups can benefit from improved physical infrastructure and human resources at primary health care facilities, and an incentive structure that is more refined than the current simple capitation system for public health. For example, in some counties where performance-based payment was introduced to stimulate public health service delivery, health checkups was also included as a performance measure (Zhang et al. 2014).

The broad scope of Health XI objectives was not fully captured by the 22 M&E indicators used for World Bank's final assessment. Nevertheless, we find significant project effects on a wide range of outcomes. Moreover, non-M&E indicators benefitted from improvements in M&E indicators instead of suffering from the competition for resources. This is likely attributable to the use of annual activity plans in project management. To obtain approval from CPSM, proposed activities were often more substantive than targeted specifically at one M&E indicator, e.g., a new payment system for public health providers rather than bonuses for gynecological checkups alone. The fact that funding disbursement was conditional on activity completion provided strong incentives for project counties to implement proposed interventions. Overall, the bottom-up and results-based approach appears to be a model that also holds promise for future reform projects in China. It aims to strike a delicate balance between effective management and local discretion, which appears very important in a context of substantial regional variation in needs and constraints.

More generally, the design of Health XI illustrates some of the major challenges faced by impact evaluations (Deaton 2010; Pritchett & Sandefur 2015; Vivalt 2015). One is the trade-off between attribution and the narrowness of scope. In our case, on the one hand, a single well-defined and uniformly implemented intervention is unlikely to succeed given the large variation in socioeconomic development and health needs in project counties. On the other hand, the adoption of decentralization in intervention design for pragmatic reasons renders precise attribution of effects practically impossible. Authors like Pritchett and Sandefur (2015) have questioned the use of one intervention to infer about a class of programs as any specific intervention has to make choices from a high-dimensional space of design attributes. Therefore, even though we cannot easily identify exactly which of the many Health XI interventions are driving the effects, our results do suggest that decentralization under results-based management and tight supervision can produce interventions that are effective.

Another limitation relates to the difficulty of determining what contextual variables matter most (Pritchett & Sandefur 2015). Denizer et al. (2013) analyzed data from over 6000 World Bank projects and identified the most important correlates of project performance to

be country-level macro variables (e.g., strong economic growth and better policy performance) and project-level micro variables (e.g., project size and preparation and supervision costs). In our case, while all Health XI counties shared project-level characteristics, county-level economic conditions did vary and appeared not to be a crucial determinant of project performance. Therefore, other Chinese counties might also have benefited from Health XI had they been selected as project counties. This may be particularly true for improvements in outpatient care and public health services provision at primary health care facilities, given the robustness of effects across different subgroups of treatment counties. However, it is worth stressing that the implementation of Health XI interventions, especially those that challenged vested interests, faced fewer political obstacles because of the timing overlap with the national health reform. Thus, one should be cautious in predicting the effectiveness of similar projects in less favorable political environments.

Figures

Figure 4.1. Density plot of DID estimates and scatter plot of *p*-values (one Health XI county dropped at a time).

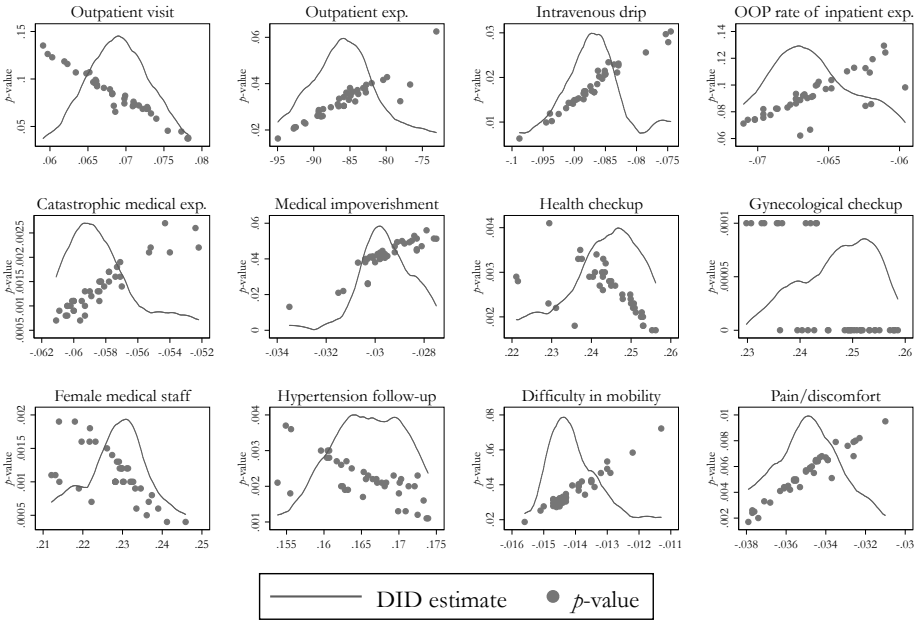
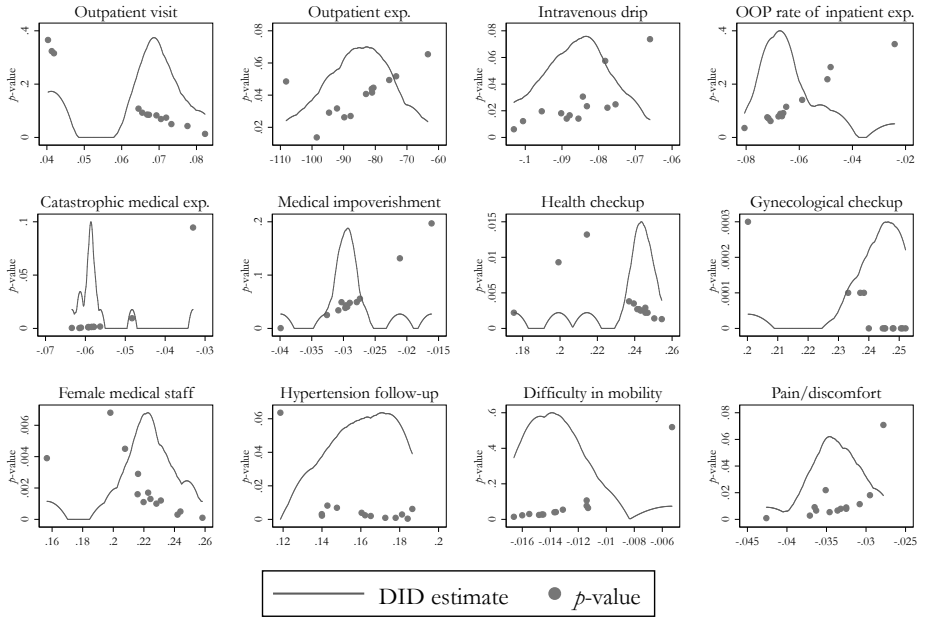


Figure 4.2. Density plot of DID estimates and scatter plot of *p*-values (one control county dropped at ta time)



Tables

Table 4.1. Macroeconomic indicators and balance check.

	Health XI counties	Unmatched controls	Matched controls		
			Mean	Standardized bias	*p*-value
GDP per capita 2005	9569	10839	9012	.08	.773
	(6798)	(8103)	(7284)		
GDP per capita 2006	11345	12648	10547	.10	.722
	(8152)	(9199)	(8211)		
GDP per capita 2007	13843	14947	12428	.15	.587
	(9719)	(10523)	(9245)		
GDP per capita 2008	16621	18313	14818	.16	.552
	(11372)	(12860)	(10679)		
GDP per capita 2009	19120	27791	17555	.12	.661
	(13095)	(34458)	(12928)		
GDP per capita 2010	23141	26067	20649	.16	.550
	(15492)	(18013)	(14811)		
GDP per capita 2011	28126	32868	25895	.12	.654
	(18353)	(23152)	(17863)		
GDP per capita 2012	31932	37916	30203	.08	.762
	(20882)	(26513)	(20656)		
Fiscal expenditure per capita 2005	913	1259	943	-.08	.768
	(364)	(1448)	(380)		
Fiscal expenditure per capita 2006	1184	1672	1189	-.01	.967
	(425)	(1767)	(434)		
Fiscal expenditure per capita 2007	1564	2093	1660	-.19	.533
	(511)	(2314)	(613)		
Fiscal expenditure per capita 2008	2112	2798	2304	-.30	.380
	(649)	(3050)	(920)		
Fiscal expenditure per capita 2009	2733	6038	2989	-.29	.387
	(895)	(14533)	(1238)		
Fiscal expenditure per capita 2010	3508	4594	3547	-.04	.915
	(1112)	(5634)	(1543)		
Fiscal expenditure per capita 2011	4508	6144	4373	.08	.770
	(1641)	(8753)	(1701)		
Fiscal expenditure per capita 2012	5413	7337	5479	-.03	.904
	(1928)	(9570)	(2087)		
Rural net income 2005	2985	3005	2940	.03	.909
	(1482)	(1328)	(1344)		
Rural net income 2006	3323	3344	3266	.03	.900
	(1700)	(1540)	(1544)		
Rural net income 2007	3858	3841	3732	.06	.806
	(1958)	(1753)	(1779)		
Rural net income 2008	4512	4505	4303	.10	.716
	(2187)	(1906)	(1997)		
Rural net income 2009	5095	5017	4761	.14	.598
	(2405)	(2102)	(2198)		
Rural net income 2010	5941	5950	5491	.16	.532
	(2739)	(2376)	(2492)		
Rural net income 2011	7133	7196	6567	.17	.514
	(3283)	(2928)	(3020)		
Rural net income 2012	8071	8191	7523	.16	.563
	(3527)	(3152)	(3368)		
N	40	22	40		

Notes: Standardized biases are calculated as (treatment mean - matched control mean)/treatment standard deviation. *P*-values are from tests of no difference in means between treatment and matched control counties that are selected using 1:1 nearest-neighbor matching with Euclidean distance and the autonomous county status condition.

Table 4.2. Descriptive statistics by treatment status.

	Year 2008		Change (2013-2008)		
	Health XI	Control	Health XI	Control	ΔT=ΔC
	(T)	(C)	(ΔT)	(ΔC)	*p*-value
Age	37	37	**4**	**2**	0.077
0-14	0.19	0.18	**-0.02**	<0.01	**0.005**
15-24	0.13	0.13	**-0.03**	**-0.02**	0.564
25-34	0.11	0.11	**-0.01**	-0.02	0.850
35-44	0.19	0.19	**-0.04**	**-0.04**	0.849
45-54	0.15	0.16	**0.02**	0.01	0.185
55-64	0.13	0.13	**0.04**	**0.03**	0.586
65+	0.10	0.10	**0.04**	**0.03**	0.376
Female	0.50	0.50	<0.01	0.01	0.360
Minority	0.10	0.50	-0.01	0.01	0.077
No schooling	0.21	0.18	**-0.03**	-0.01	0.162
Primary school	0.30	0.27	0.01	<0.01	0.445
Middle school	0.36	0.37	<0.01	-0.01	0.480
High school and above	0.13	0.18	**0.02**	0.02	0.994
Single	0.16	0.16	**-0.03**	**-0.03**	0.624
Married	0.76	0.75	**0.03**	**0.04**	0.314
Student	0.70	0.80	**-0.02**	**-0.02**	0.844
Employed	0.77	0.74	**0.03**	<0.01	0.388
Jobless	0.12	0.13	-0.01	0.02	0.406
Retired	0.30	0.50	<0.01	<0.01	0.536
Household income per capita	4949	497	**6085**	**5424**	0.336
Poverty rate	0.29	0.28	**-0.26**	**-0.20**	0.486
Household expenditure per capita	3443	3611	**3814**	**3715**	0.797
New Rural Cooperative Medical Scheme[a]	0.90	0.82	**0.04**	0.03	0.877
No health insurance	0.60	0.90	**-0.05**	**-0.07**	0.446
Medical Assistance	0.30	0.30	0.01	0.01	0.868
Any chronic disease	0.14	0.13	**0.06**	**0.07**	0.863
Sick in the past two weeks	0.18	0.17	0.01	0.03	0.376
Hospitalization needed in the past 12 months	0.07	0.06	**0.01**	**0.03**	0.201

Notes: Control counties are selected using 1:1 nearest-neighbor matching with Euclidean distance and the autonomous county status condition. In all tests of no difference in means, standard errors are clustered at the county level. Bold indicates that difference in means is significantly different from zero at the 5% level.

[a] In a few counties that had integrated NRCMS and the Urban Resident Basic Medical Insurance by 2013, people covered by the unified insurance are all counted as NRCMS enrollees.

Table 4.3. Baseline values, five-year changes and treatment effects for medical care.

	Baseline (2008)		Change (2013-2008)		DID			Sample
	Health XI	Control	Health XI	Control	Coef.	SE	*p*-value	size
Panel A. Utilization								
Outpatient visit conditional on illness	0.378	0.468	**0.076**	0.043	0.069	(0.038)	0.077	54444
proportion at primary health care facilities	0.860	0.880	-0.020	-0.023	0.009	(0.027)	0.742	23194
proportion at county-level hospitals	0.115	0.102	0.001	0.014	-0.019	(0.022)	0.387	23194
Hospitalization when needed	0.764	0.794	**0.157**	**0.117**	0.041	(0.025)	0.102	22942
proportion at township health centers	0.400	0.422	**-0.097**	**-0.153**	0.054	(0.057)	0.348	19735
proportion at county-level hospitals	0.467	0.496	**0.067**	**0.115**	-0.037	(0.048)	0.436	19735
proportion at urban hospitals	0.113	0.073	**0.040**	0.032	-0.002	(0.023)	0.934	19735
Panel B. Expenses								
Outpatient expenditure	203	162	1	74	**-86**	(38)	0.030	23021
at primary health care facilities	151	116	**-47**	7	**-65**	(25)	0.013	19742
at county-level hospitals	482	484	**210**	**487**	-249	(160)	0.127	2616
Gross inpatient expenditure	3340	2916	**3066**	**3407**	-220	(737)	0.766	19617
at township health centers	1288	1213	**595**	**835**	-182	(238)	0.449	6348
at county-level hospitals	3644	3424	**2453**	**2832**	-169	(736)	0.820	10065
at urban hospitals	8333	9318	**7731**	**9669**	-2432	(1630)	0.142	2403
Out-of-pocket inpatient expenditure	2469	2070	**782**	**1166**	-316	(438)	0.474	19617
at township health centers	809	744	-124	-71	-26	(217)	0.904	6348
at county-level hospitals	2661	2437	239	**801**	-437	(412)	0.294	10065
at urban hospitals	6770	7170	**2397**	**3275**	-1515	(1083)	0.168	2403
Panel C. Quality measures								
Intravenous drip	0.343	0.378	**-0.071**	0.024	**-0.087**	(0.035)	0.017	23217
at primary health care facilities	0.332	0.385	**-0.074**	0.007	-0.076	(0.039)	0.054	19822
at county-level hospitals	0.424	0.332	-0.042	0.164	**-0.197**	(0.074)	0.011	2661
Caesarean section rate	0.308	0.236	0.036	0.054	-0.042	(0.047)	0.385	4462
Panel D. Financial strain								
Out-of-pocket rate of inpatient expenditure	0.667	0.628	**-0.224**	**-0.151**	-0.066	(0.038)	0.090	19617
at township health centers	0.576	0.515	**-0.252**	**-0.228**	-0.032	(0.044)	0.471	6348
at county-level hospitals	0.709	0.698	**-0.249**	**-0.161**	**-0.075**	(0.035)	0.040	10065
at urban hospitals	0.780	0.779	**-0.209**	**-0.171**	-0.032	(0.030)	0.290	2403
Catastrophic medical expenditure	0.183	0.133	**-0.066**	0.005	**-0.058**	(0.017)	0.001	279068
Impoverishment due to medical expenditure	0.058	0.054	**-0.019**	0.012	**-0.030**	(0.014)	0.040	279068

Notes: Column 5 gives DID estimates from Equation 4.1 where standard errors are clustered at the county level. Bold indicates five-year changes and treatment effects in DID are significant at the 5% level. Inconsistancies in sample sizes are due to providers of unspecified level, capping of outpatient and inpatient expenditures, and the addition of non-NRCMS-enrolless from households with at least one NRCMS enrollee for two household-level outcomes - catastrophic medical expenditure and medical impoverishment.

Table 4.4. Baseline values, five-year changes and treatment effects for public health services provision at primary health care facilities.

	Baseline (2008)		Change (2013-2008)		DID			Sample
	Health XI	Control	Health XI	Control	Coef.	SE	*p*-value	size
Health checkups	0.139	0.175	**0.445**	**0.198**	**0.244**	(0.077)	0.002	275562
Gynecological checkups	0.456	0.409	**0.214**	-0.033	**0.246**	(0.054)	<0.001	66866
Female medical staff available upon request	0.678	0.729	**0.178**	-0.055	**0.228**	(0.065)	0.001	59428
Follow-up of hypertension patients aged 35+	0.780	0.750	**0.138**	-0.029	**0.165**	(0.051)	0.002	2814

Notes: Column 5 gives DID estimates from Equation (1) with standard errors clustered at the county level. Bold indicates that five-year changes and treatment effects are significantly different from zero at the 5% level.

Table 4.5. Baseline values, five-year changes and treatment effects for self-rated health.

	Baseline (2008)		Change (2013-2008)		DID			Sample
	Health XI	Control	Health XI	Control	Coef.	SE	*p*-value	size
EQ-5D value score (the higher the better)	0.922	0.927	**0.005**	-0.004	0.006	(0.004)	0.122	275562
Difficulty in mobility	0.065	0.053	-0.003	**0.015**	**-0.014**	(0.006)	0.034	275562
Difficulty in self-care	0.045	0.039	**-0.009**	0.004	-0.010	(0.007)	0.126	275562
Difficulty in usual activities	0.067	0.059	**-0.012**	0.001	-0.008	(0.010)	0.422	275562
Pain/discomfort	0.116	0.094	-0.003	**0.035**	**-0.035**	(0.012)	0.005	275562
Anxiety/depression	0.077	0.078	**-0.017**	**-0.020**	0.006	(0.011)	0.547	275562
EQ VAS (0-100, the higher the better)	80.21	79.74	**2.659**	0.510	1.992	(1.198)	0.102	275562

Notes: Column 5 gives DID estimates from equation (1) where standard errors are clustered at the county level. Bold indicates five-year changes and treatment effects in DID are significant at the 5% level.

Table 4.6. Comparison with results from alternative matching methods.

	Preferred matching	Euclidean with no conditions	Euclidean + same province	Synthetic control	Normalized synthetic control
Panel A. Falsification tests					
Household income per capita	456	624	209	99	-42
Poverty rate	-0.050	-0.090	-0.103*	-0.028	-0.014
Panel B. Outcomes					
Outpatient visit conditional on illness	0.069*	0.065*	0.036	0.067*	0.076**
Hospitalization when needed	0.041	0.055***	0.071***	0.029	0.023
Outpatient expenditure	-86**	-75*	-115**	-75**	-77**
Gross inpatient expenditure	-220	-469	-534	129	130
Out-of-pocket inpatient expenditure	-316	-488	-591*	-251	-187
Intravenous drip	-0.087**	-0.085**	-0.085**	-0.065*	-0.065*
Caesarean section rate	-0.042	-0.53	-0.052	-0.062	-0.047
Out-of-pocket rate of inpatient expenditure	-0.066*	-0.58	-0.033	-0.046	-0.048
Catastrophic medical expenditure	-0.058***	-0.051***	-0.041*	-0.071***	-0.072***
Impoverishment due to medical expenditure	-0.030**	-0.023	-0.015	-0.039***	-0.041***
Health checkup	0.244***	0.192***	0.170***	0.249**	0.263***
Gynecological checkup	0.246***	0.225***	0.212***	0.245***	0.257***
Female medical staff available upon request	0.228***	0.195***	0.176***	0.199**	0.203**
Follow-up of hypertension patients aged 35+	0.165***	0.148***	0.076*	0.111**	0.123***
EQ-5D value score	0.006	0.005	0.001	0.005	0.006*
Difficulty in mobility	-0.014**	-0.012*	-0.007	-0.015**	-0.016**
Difficulty in self-care	-0.010	-0.008	-0.006	-0.012**	-0.013**
Difficulty in usual activities	-0.008	-0.004	-0.002	-0.012	-0.014**
Pain/discomfort	-0.035***	-0.036***	-0.021	-0.032**	-0.031**
Anxiety/depression	0.006	0.004	0.026	0.014	0.016
EQ VAS	1.992	2.002*	0.222	1.668	1.768

Notes: Variables in Panel A are estimated using Equation 4.1 without controlling for income quartiles. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 4.7. Treatment effect heterogeneity by county characteristics.

	Health VIII experience	State-designated
Outpatient visit conditional on illness	-0.081	0.040
Hospitalization when needed	-0.009	-0.032
Outpatient expenditure	48	-6
Gross inpatient expenditure	451	2201*
Out-of-pocket inpatient expenditure	312	1321*
Intravenous drip	0.003	0.010
Caesarean section rate	0.131*	0.212***
Out-of-pocket rate of inpatient expenditure	-0.039	-0.022
Catastrophic medical expenditure	-0.007	-0.033
Impoverishment due to medical expenditure	-0.014	-0.011
Health checkup	0.039	0.344***
Gynecological checkup	-0.019	0.186*
Female medical staff available upon request	-0.083	0.174
Follow-up of hypertension patients aged 35+	0.023	0.147
EQ-5D value score	-0.007	-0.011**
Difficulty in mobility	<0.001	0.002
Difficulty in self-care	-0.010	<0.001
Difficulty in usual activities	-0.007	0.009
Pain/discomfort	0.064***	0.045**
Anxiety/depression	0.026	0.042**
EQ VAS	-0.96	1.76

Notes: Effects are estimated using a triple-difference (DDD) framework in which the county characteristic and its interaction with *HXI, Post* and *HXI*×*Post* are added to Equation 4.1. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Appendix 4.A

Table 4.A.1. Health XI Monitoring & Evaluation (M&E) indicators.

	M&E indicators	Baseline	Target	Reason for inclusion/exclusion
1	Ratio of hospitalization rate of MA vs. non-MA beneficiaries	Average 1.5	Positive trend	Not suitable for individual level analysis, proxied by inpatient care use when needed
2	Ratio of outpatient visit rate of the bottom 2% vs. top 40% in the income distribution in the previous year	Average 1.1; 18 counties<1	6 counties<1	Not suitable for individual level analysis, proxied by outpatient care use when needed
3	Percentage of households with annual health expenditures in excess of 20% of total income	19.2%	14.2%	Included
4	Rate of overall satisfaction with rural health services among men	63.8%	67.8%	Excluded due to a major change in the framing of the series of satisfaction questions
5	Rate of overall satisfaction with rural health services among women	66.6%	70.6%	Excluded due to a major change in the framing of the series of satisfaction questions
6	Public health system scorecard rating	61.78	80	Data not available
7	Innovation accepted and rolled out at provincial or regional level	NA	Positive trend	Data not available
8	Percentage of total annual (individual) inpatient expenses financed through out-of-pocket payments for NCMS members	74.7%	67.7%	Annual data not available, proxied by the rate from the latest hospitalization episode
9	Percentage of total annual (individual) inpatient expenses financed through out-of-pocket payments for MA beneficiaries covered by the NCMS	72.3%	65.3%	Insufficient observations, proxied by medical impoverishment
10	Total annual NCMS expenditures as percentage of total annual NCMS funds	72.7	>85	Not suitable for individual level analysis
11	NCMS enrollment rate	93.9%	Maintain>90%	Not suitable for individual level analysis
12	Average number of outpatient visits per health professional per day over last year at THC	3.9	Positive trend	Data not available
13	Rate of change in average cost per inpatient case in county hospitals	7.7%	Lower than provincial average rate of change	Not suitable for individual level analysis, proxied by the gross expenditure of the latest hospitalization episode
14	Percentage of deliveries by caesarean section	28.9%	Closer to 15%	Included
15	Percentage of outpatients at township health centers and village clinics that received two or more antibiotics	18.9% at township health centers; 19.8% at village clinics	Reduction of 4 percentage points	Data not available, proxied by outpatient intravenous drip
16	Percentage of women who have access to a female qualified health worker	28 counties>90%; 12 counties<90%	32 counties>90%	Included
17	Number of administrative villages that have been initiated the implementation of "healthy village"	NA	Implementation rate>50%; at least 1 village in each province meet healthy village	Excluded as data is unavailable
18	Percentage of women between age 15 and 49 who undergo gynecological check-up in last year	44.1%	54.1%	Included
19	Percentage of individuals over 35 with hypertension who have been followed-up in the three months	66.2%	>80%	Included
20	Annual per capita government expenditures on county level public health institutions and programs	14.6 Yuan	20.9 Yuan	Excluded due to lack of data
21	Project lessons and experiences are documented and disseminated	NA	100%	Excluded due to lack of data
22	Project experiences are extended or adopted outside project areas	NA	Implemented by 60% of the project provinces	Excluded due to lack of data

Table 4.A.2. Project and non-project counties and their characteristics.

Province	Health XI county	Autonomous	Health VIII experience	State-designated poor	Non-project county	Autonomous	Health VIII experience	State-designated poor
Shanxi	Wuxiang		✓	✓	Pingding			
	Zezhou				Yangcheng			
	Yushe		✓					
	Taigu							
Heilongjiang	Gannan			✓	Fuyu			
	Lindian			✓	Baoqing			
	Fujin							
	Linkou							
Jiangsu	Gaochun				Pizhou			
	Liyang				Jinhu			
	Haimen				Yangzhong			
	Gaoyou							
	Danyang							
	Jiangyan							
Henan	Yiyang		✓	✓	Ruyang			✓
	Ruzhou				Sui			✓
	Wuzhi				Fugou			
	Qingfeng							
	Xi		✓					
Chongqing	Jiulongpo				Wanzhou			✓
	Qianjiang	✓	✓	✓	Shapingba			
	Yongchuan				Zhong			
	Rongchang							
	Liangping							
	Shizhu	✓		✓				
Shaanxi	Mei				Jintai			
	Xunyi			✓	Hanyin		✓	✓
	Ningqiang		✓	✓				
	Hanbin		✓	✓				
	Zhenan		✓	✓				
Gansu	Gaolan				Yuzhong			✓
	Huining			✓	Jingtai			
	Gangu			✓	Lintan	✓		✓
	Jingning			✓				
	Kangle	✓	✓	✓				
Qinghai	Datong	✓	✓	✓	Haiyan	✓		
	Huangzhong		✓	✓				
	Huangyuan							
	Ledu		✓	✓				
	Huzhu	✓	✓					

Table 4.A.3. Pairs of matched treatment and control counties.

Health XI county	Control county
Wuxiang	Pizhou
Zezhou	Jinhu
Yushe	Jingtai
Taigu	Pingding
Gannan	Yuzhong
Lindian	Zhong
Fujin	Pizhou
Linkou	Pingding
Gaochun	Shapingba
Liyang	Jintai
Haimen	Shapingba
Gaoyou	Jinhu
Danyang	Yangzhong
Jiangyan	Shapingba
Yiyang	Ruyang
Ruzhou	Pizhou
Wuzhi	Pizhou
Qingfeng	Pingding
Xi	Fugou
Jiulongpo	Jintai
Qianjiang	Lintan
Yongchuan	Jinhu
Rongchang	Pizhou
Liangping	Zhong
Shizhu	Lintan
Mei	Ruyang
Xunyi	Ruyang
Ningqiang	Yuzhong
Hanbin	Fugou
Zhenan	Hanyin
Gaolan	Ruyang
Huining	Yuzhong
Gangu	Yuzhong
Jingning	Yuzhong
Kangle	Lintan
Datong	Lintan
Huangzhong	Ruyang
Huangyuan	Hanyin
Ledu	Jingtai
Huzhu	Lintan

Table 4.A.4. Control county weights by matching methods.

	Preferred matching	Euclidean with no conditions	Euclidean + same province	Synthetic control	Normalized synthetic control
Pingding	3	3	3	1.429	1.212
Yangcheng			1		
Fuyu			3	1.33	1.166
Baoqing			1	2.380	1.896
Pizhou	5	6		1.34	2.14
Jinhu	3	3	2	2.191	2.982
Yangzhong	1	1	4	2.494	1.788
Ruyang	5	6	4	1.954	1.272
Sui				1.351	1.222
Fugou	2	2	1	1.654	1.367
Wanzhou			3	1.937	2.159
Shapingba	3	3	1	3.762	4.38
Zhong	2	3	2	1.659	1.32
Jintai	2	2		0.529	0.22
Hanyin	2	3	5	1.855	1.296
Yuzhong	5	2		6.762	7.792
Jingtai	2	2	1	2.434	3.214
Lintan	5	4	4	5	5
Haiyan			5		

CHAPTER 5

Impact evaluation of a diagnosis-related group (DRG)-based hospital payment pilot in rural China

Joint work with Eddy van Doorslaer, Yang Sun, Zhiyuan Hou, Jian Wu, Qingyue Meng

As part of a health care reform project in rural China, the Henan province introduced a simplified DRG-based payment system in two of its counties. The new system (i) set the payment rates for a selection of inpatient conditions to contain costs, and (ii) imposed clinical pathways to ensure better quality of care and to encourage the use of basic services and drugs. Using household data from a national survey, we estimate the effects of this payment reform using a triple difference (DDD) method. We find that township health centers (THCs) significantly and substantially reduced their service to DRG-eligible patients by referring more severe cases to higher-level providers. While neither total nor out-of-pocket (OOP) expenditure decreased, reimbursement rates increased. At the higher-level county hospitals, treatment intensity increased noticeably given the worsened case mix. But this did not substantially increase total or OOP expenditure, or decrease the rate of reimbursement. Overall, our results show a reduced service volume, little intended effect on cost containment, and no effect on patient satisfaction. The provider response was smaller at county hospitals than at THCs, and for delivery cases than for other DRG-eligible conditions. Moreover, we find suggestive evidence of a greater use of well-reimbursed basic services and drugs.

5.1 Introduction

The economic reform in 1978 initiated China's transformation from a planned economy to a market one. In the process, the government experienced a drastic drop in revenue, reducing its capacity to fund health care. By the early 1990s, the proportion of total revenues of public health facilities that came from government subsidy had fallen from 50 percent before the reform to a mere 10 percent (Yip & Hsiao 2008). To help health care providers survive, the government allowed high profit margins for drugs and medical tests but priced basic medical services below cost to ensure affordability. To motivate high performance of the health professionals, another policy linked their income with the revenue they generated.[1] Both policies, though designed with good intentions, provided health care facilities and health professionals with an aligned incentive to engage in profit-seeking activities. These and other perverse system incentives led to an over-prescription of drugs and expensive tests, which over time has contributed to escalating costs and large-scale public discontent (World Bank 2016).

The rural population suffered in particular from their lack of health insurance.[2] To ameliorate the problem, the government introduced in 2003 the heavily subsidized *New Rural Cooperative Medical Scheme* (NRCMS) to insure rural residents. At the same time, a social safety net program, *Medical Assistance* (MA), was introduced to provide additional financial assistance for the poor. However, actual coverage was shallow because of the benefit package design (e.g., limited coverage of services and drugs, low ceilings) and the complicated reimbursement procedures (e.g., restrictions on the time and place of claim submission). Moreover, research showed that the risk of catastrophic health expenditure was even increased, with one explanation being that insurance made patients more likely to use higher-level providers (Wagstaff & Lindelow 2008). This is not surprising considering that the disease burden in China has been shifting from infectious diseases to non-communicable diseases (Yang et al. 2008), for which treatment is considered inferior at lower-level facilities.[3]

To improve financial protection and service delivery capacity, efficiency and quality under limited funding, the eleventh World Bank health project in China*, the Rural Health Project* (or *Health XI*), was launched in October 2008.[4] It involved 40 counties spreading across eight provinces, with substantial variation in geography and socio-economic development. During the six-year implementation period, Health XI piloted a series of interventions in both medical care and public health sectors (World Bank 2015a). While project counties faced the same pre-

[1] Physicians in China have very low government-funded basic salary. The majority of their income comes from bonuses linked with the hospital's and the department's profit, and yet physicians earn only marginally higher than average employees (Ran et al. 2013).

[2] It is estimated that 80% of China's rural population – approximately 640 million people – lacked health insurance by 2003 (Centre for Health Statistics and Information of Ministry of Health 2004).

[3] In 2006, only 6.8% of rural health professionals possessed a bachelor degree or above, and they were concentrated in county hospitals (as opposed to the lower-level township health centers and village clinics) (Ministry of Health 2007).

[4] The project webpage can be found on the World Bank's website at: http://projects.worldbank.org/P084437/rural-health-project?lang=en&tab=overview.

specified reform aims and final assessment indicators, they were allowed to differ in specific intervention design to suit local conditions. For example, to better incentivize the provision of public health services, some counties refined the current system of capitation payment by adjusting for estimated cost of service delivery, and some switched from capitation to purchasing public health services from local providers (Zhang et al. 2014).

In this paper, we evaluate one particular Health XI intervention, a new inpatient payment system that replaced the fee-for-service (FFS) one. It was implemented in two project counties – Xi and Yiyang – in the province of Henan. As a first step towards diagnosis-related group (DRG) payment, the new system had greatly simplified rules and limited coverage of conditions due to capacity constraints. Specifically, the payment method grouped inpatient cases based on diagnosis and procedure (e.g., with or without surgery). For each case group, it defined severity levels A, B and C, and for each severity level of each case group, it set the price that NRCMS paid to the provider. A clinical pathway was formulated for each severity level of each case group to serve as the treatment protocol. Though the new payment system was designed to contain costs without compromising quality, provider responses can lead to unintended consequences, especially during the early stages of implementation. For example, to increase the number of cases, providers may admit patients unnecessarily (supplier-induced demand); to reduce cost per case, providers may turn unprofitable patients away (cream skimming), or withhold necessary services; to increase revenue per case, providers may manipulate the classification of patients (upcoding). We investigate some of these potential provider responses for which relevant data is available.

Our analysis draws on pre- and post-reform household data from two Health XI counties, Xi and Yiyang (treatment), and two non-Health XI counties, Sui and Ruyang (control). Due to the lack of clinical and claims data, we could only examine the effects of the new payment system on service volume, case mix proxies, treatment intensity (inpatient days and whether a surgery was performed), expenditures (total and out-of-pocket payments and reimbursement rate), and patient satisfaction. We first identify the general effects of Health XI using a difference-in-differences (DID) model, and then estimate the impact of the new system using a triple difference (difference-in-difference-in-differences, DDD) model that further compares DRG-eligible with DRG-ineligible conditions.[5] Admittedly, other Health XI interventions also occurred around the time of the payment reform. However, we do not expect them to confound our DDD estimates because (1) they did not affect the inpatient sector in the short run (e.g., interventions in the public health sector or investment in village clinics), (2) they affected only a negligible proportion of inpatients (e.g., better reimbursement for MA beneficiaries), or (3) they did not have very different effects on DRG-eligible versus DRG-ineligible conditions (e.g., a general increase in NRCMS generosity).

[5] The new payment system in Henan is very different from the DRG systems in the U.S. and Europe in that (i) the case classification system did not cover all conditions, and (ii) the payment rates were determined case-by-case rather than calculated as a base rate multiplied by DRG weights and adjustment factors. Nevertheless, for the sake of simplicity, we abuse the term DRG and refer to the new system as DRG-based payment and the conditions eligible/ineligible for it as DRG-eligible/DRG-ineligible conditions.

Our results confirm that, between treatment and control counties, there was no differential change in inpatient care needs or the proportion of DRG-eligible conditions. Using the DDD framework, we find township health centers (THCs) to show a significant and substantial reduction in DRG-eligible cases, and a significant and substantial improvement in case mix. This suggests that THCs reduced service provision instead of inducing demand, and engaged in cream skimming. Consequently, county hospitals experienced a significant and substantial increase in DRG-eligible cases, and a significant and substantial deterioration in case mix. In turn, treatment intensity of DRG-eligible cases increased noticeably (albeit estimated with low precision). However, this did not lead to an increase in total expenditure or a large reduction in reimbursement rate. A possible explanation is good compliance with clinical pathways, which favor basic medicines and services that are well reimbursed by NRCMS.

Next, we examine one particular DRG-eligible condition – delivery. Compared to other DRG-eligible conditions, it can be more accurately identified and is observed for more years. Provider responses are expected to be limited in delivery cases because (1) treatment occur within a narrow time frame, (2) patients play an important role in the decision making, and (3) the already high caesarean section rates have been recognized as a national issue and have been closely monitored by supervising agencies. Using DID on the sample of delivery cases, we first confirm that mothers' age at delivery and the incidence of low birth weight did not change differentially between treatment and control counties. Then, we indeed find little effect of the new payment system on service volume, treatment intensity (i.e., caesarean section rate) and total expenditure. Last, our results show that, while patients generally report higher satisfaction in treatment counties, those with DRG-eligible conditions were not more satisfied with their inpatient care.

DRG-based payment schemes have been piloted in many locations in China since 1999 (Meng 2005), but systematic evaluation of their effectiveness has been scarce. Jian et al. (2015) recently examined the first DRG payment system piloted in some of the best tertiary hospitals in Beijing and found it to reduce expenditures and out-of-pocket (OOP) payments. However, most regions in China do not yet have the capacity to implement a real DRG payment system. This paper illustrates the challenges of implementing a DRG-based payment when system capacity is low and the purchasing function of health financing is weak. It provides a reference for other low- and middle-income countries as they move toward some form of case-based payment. More generally, our paper relates to the literature examining physician response to financial incentives. A long-standing question is how service volume responds to financial incentives (McGuire 2000). Most previous studies focused on this relationship for specific treatments such as caesarean sections (Grant 2009), coronary artery bypass graft surgeries (Yip 1998), chemotherapy drugs (Jacobson et al. 2010), and neonatal intensive care unit utilization (Shigeoka & Fushimi 2014). Though Clemens & Gottlieb (2014) examined a broad range of procedures, the setting is still a high-income country (U.S.) with a mature health care system. We add to this literature by providing new evidence from an array of common inpatient conditions in a low- and middle-income setting.

The remainder of this paper is organized as follows. Section 5.2 describes the provider payment reform in Henan, the conceptual framework and the data. Section 5.3 explains the

empirical strategy. Section 5.4 presents descriptive statistics and estimation results. Section 5.5 concludes.

5.2 Background

5.2.1 Context and the provider payment reform in Henan

Henan is a less economically developed province located in central China. By 2010, it had a population of 94 million and a per capita GDP of 24,446 Yuan[6] (National Bureau of Statistics 2011a). The average per capita income in rural households was 7293 Yuan, ranking 19th among 31 provinces in China (National Bureau of Statistics 2011b). NRCMS covered all 157 rural counties by the end of 2007, with an average enrolment rate of 91.9% (Project Management Office of Henan Health Bureau 2009). As elsewhere in China, rural inpatient care in Henan is provided by county hospitals and the lower-level THCs.[7] Each county typically has one general county hospital, one traditional Chinese medicine hospital, and one maternal and child service center, and each township of a county typically has one THC. The county general hospital provides the most comprehensive and advanced care within a county. Patients are free to seek better care at higher-level urban hospitals outside their county without a referral, but at the expense of much higher OOP payments and associated costs.[8] In 2007, average inpatient expenditure for NRCMS enrollees was 609 Yuan at THCs, 2225 Yuan at county hospitals, 6592 Yuan at city hospitals and 8562 at province hospitals, with average reimbursement rates of 47%, 31%, 23% and 23%, respectively (Project Management Office of Henan Health Bureau 2009).

Inpatient care providers in Xi and Yiyang were traditionally paid by FFS, which encourages overprovision of services. Funded and supervised by Health XI,[9] the two counties introduced a new DRG-based payment system in 2011. The main goal was to control cost escalation without compromising the quality of care. The NRCMS management office selected the common inpatient conditions to be paid by the new method. Selection criteria included large service volume, modest variation in costs, and so on. Then, clinically and economically homogenous case groups were formed based on diagnosis and procedure, and for each case group, severity levels A, B, and C were defined. Level A cases were those with

[6] We use the same exchange rate, 1 Yuan = 0.16 Dollar, as in the World Bank final report on Health XI (World Bank 2015a).

[7] For brevity, we refer to county-level, city-level, province-level hospitals as county, city, province hospitals throughout the paper.

[8] Costs such as transportation and accommodation are not reimbursed by NRCMS.

[9] Health XI funding amounted to less than 1% of government health expenditures and was used to fund mainly the designing of interventions (World Bank 2015a). It was explicitly required that project funds do not crowd out governmental fiscal funds, nor substitute it in funding NRCMS or health workers' salary. National and local Health XI project teams consisting of officials from various relevant departments provided guidance, coordination, and supervision.

no or minor comorbidities/complications, level B moderate comorbidities/complications, and level C severe comorbidities/complications. To restrict upcoding, proportion limit was set for each severity level.

To ensure quality, a clinical pathway was formulated for each severity level of each case group in accordance with national and provincial treatment guidelines. Optional service items were included to allow for some treatment flexibility. A tentative payment rate for each severity level of each case group was determined with reference to the clinical pathway and based on historical expenditure minus the cost of unnecessary drugs and services. Last, the NRCMS management office negotiated with providers to settle on the service items in clinical pathways, the payment rate for each severity level of each case group, and the proportion limits for three severity levels. The list of DRG-eligible conditions can be found in Appendix 5.B.

NRCMS funds for DRG-eligible cases were prepaid to providers, with about 10% of the budget paid at year-end based on performance. To provide further incentives, providers were allowed to retain the savings (but they were also at risk of bearing losses). A comprehensive set of performance indicators was pre-specified for monitoring and evaluation. They included compliance with various aspects of the clinical pathway, whether the proportion of each severity level was within limit, mortality rate, patient satisfaction, two-week readmission rate, and so on. Supervision and evaluation were conducted both internally by providers and externally by the NRCMS management office and the local Health XI team. Given that the reform was a part of the Health XI project, the national project team also evaluated the performance of this new payment system.

It is worth noting that the coverage of conditions, payment rates, and clinical pathways were slightly different between county hospitals and THCs due to systematic differences in medical equipment and competence of health professionals. In addition, as the reform was designed locally, some of the system characteristics also differed between the two counties. Using administrative data, Figure 5.1 gives an idea of the variation in payment reform coverage over time in Xi and Yiyang, respectively. By the time post-reform data was collected in September 2013, the number of case groups was 106 (about 45% of total inpatients) in the county general hospital and 80 (about 90% of total inpatients) in THCs in Xi, and 188 (about 50% of total inpatients) in the county general hospital and 92 (about 70% of total inpatients) in THCs in Yiyang. There was also variation between Xi and Yiyang in payment rates and clinical pathways. However, the reform was conceptually similar. Particularly, in both counties the proportions of severity levels A-C were roughly 70%, 20% and 10%, and the traditional FFS was used for severity level C.

5.2.2 Conceptual framework

As shown in McGuire & Pauly (1991) in a one-service scenario, the way in which a fee cut affects service volume depends on the relative strength of the income and substitution effects. When the substitution effect dominates, physicians will reduce service provision and increase leisure, whereas when the income effect dominates, they will increase service provision. Extension to a multiple-services scenario introduces substitution across services in addition to

substitution between work and leisure (McGuire & Pauly 1991). Consequently, the impact of a fee cut in one service may also depend on the marginal disutility of demand inducement, patient acceptance of inducement, market share, margins, and time costs for this service relative to others. Our case can be viewed as a multiple-services scenario, where DRG-eligible conditions received a fee cut because (1) the new payment rates were lower since they were calculated based on historical expenditures minus the cost of unnecessary drugs and services, (2) it took much time and effort to explain the new payment system to patients and obtain approval from them,[10] (3) the operation of the computerized system designed for the new payment system was challenging, and (4) changing practice habits to comply with clinical pathways could be undesirable.[11]

This sudden change in the relative profitability of treating different conditions created an incentive for substitution across services. Physicians were likely to substitute DRG-ineligible cases for DRG-eligible ones, with the extent depending on how easily they could refer DRG-eligible cases to higher-level providers.[12] Compared with county hospitals, THCs are in a better position to do so. This is because referring patients to county hospitals complies with general patient preferences for higher-level (perceived as better) care without inflicting an excessively higher cost. Among other DRG-eligible conditions, delivery cases are more difficult to direct away because (1) case severity and the appropriate provider level are better known in advance, and (2) treatment must occur within a narrow time frame.

Another type of physician response is to modify the case coding.[13] To avoid this problem, there was tight regulation and close monitoring of the proportion of DRG-eligible cases actually being paid by the new system and the proportion of each severity level. Therefore, it was difficult for physicians to exempt a DRG-eligible condition from the new payment method or to upcode it to a higher severity level. However, changing the case group of a condition, e.g., into one with surgical treatment, might be easier given the practical difficulty in auditing medical charts. Again, it is more difficulty to induce surgery in delivery cases than in others because (1) patients are more likely to have a preference for the type of treatment and are more influential in the decision making, and (2) caesarean section rates have been tightly controlled and monitored by other health initiatives.

The two bottom graphs in Figure 5.1 show that DRG-eligible conditions constitute a sizable proportion of inpatient cases, especially at THCs. However, a dominant income effect does not follow directly. Without information on prices and margins, it is difficult to assess the importance of DRG-eligible conditions in generating income. Furthermore, the income effect is likely to be modest given that DRG-ineligible conditions are still paid by FFS and can

[10] Because the payment reform introduced such a drastic change in the way conditions were treated and patients billed, the provider was required to obtain a signed consent form from the patient before a DRG-eligible case could actually be paid by the new payment system.

[11] Changing habits is undesirable in itself, and it could also imply less kickbacks from pharmaceutical companies as clinical pathways limit the use of unnecessary drugs and encourage the use of basic ones.

[12] The payment reform changed reimbursement to hospitals, not physicians. However, we consider hospitals and physicians as one agent instead of principal and agent because their incentives are closely aligned in China.

[13] This is not discussed in McGuire & Pauly (1991), but in other papers such as a recent one by Geruso & Layton (2015).

be used to make up for the income losses. Nonetheless, the income effect is expected to be larger at THCs, where DRG-eligible conditions account for a larger proportion of inpatient cases. DRG-eligible conditions are also a more important source of net income given THCs' incapability to treat more complicated and profitable cases.

5.2.3 Data and control counties

Our data comes from the 2008 and 2013 waves of China's *National Health Services Survey* (NHSS). NHSS is a nationally representative repeated cross-sectional household survey conducted at five-year intervals since 1993 by the National Health and Family Planning Commission. The survey employs a multistage cluster sampling procedure, stratified by province (Meng et al. 2012). In each county, five townships were randomly selected, two villages from each township, and 60 households from each village, giving around 2000 respondents per county per wave. Xi and Yiyang were added together with other Health XI counties in 2008 and 2013 using the same sampling procedure and questionnaires.[14] The two control counties Sui and Ruyang are original NHSS sample counties in Henan that did not participate in Health XI. They were selected among other candidates before the release of NHSS 2013 for their comparable socio-economic conditions with the treatment counties. Specifically, Sui was selected as the control for Xi, and Ruyang for Yiyang.

Given that only one pre-treatment wave is available, it is not possible to examine how well the controls satisfy the common trend assumption. Instead, we examine the comparability of treatment and control counties in socio-economic conditions, as measured by per capita GDP, per capita fiscal expenditure and rural net income during 2005-2012.[15] These indicators are selected because they capture local economic development, government fiscal capacity and rural people's spending power, which are important contextual factors for the outcomes we examine. Moreover, these indicators do not predict Health XI participation, nor did Health XI affect them during its implementation. The time span of eight years affords the opportunity to assess the comparability of controls for an extended period before Health XI, during Health XI and also after the payment reform. This is important in the face of a fast evolving economic and policy environment in China.

Figure 5.2 plots per capita GDP, per capita fiscal expenditure and rural net income during 2005-2012 for treatment and control counties. The controls closely matched the treatment counties, especially for rural net income. Although per capita GDP in Ruyang deviated from that in Yiyang in 2009, the discrepancy was small in relation to the variation in per capita GDP. Furthermore, Zhang et al. (2016) have shown that, among 22 candidate controls from the eight Health XI provinces in NHSS, Ruyang was consistently chosen as the control for Yiyang using various matching specifications. While a different county, Fugou, was chosen as the control for Xi, Sui is not a discernibly worse match. In fact, Sui would be a better match if we put more weight on comparability in later years considering that the payment reform was

[14] Due to a sample expansion in 2013, the number of observations increased to about 3300 in Yiyang in 2013.

[15] Data was drawn from Henan Statistical Yearbooks 2006-2013.

initiated in 2011. As is shown in Appendix Table 5.A.1, per capita GDP in Sui converged to that in Xi, whereas per capita GDP in Fugou started to diverge since 2009.

Our key outcome variables are drawn from the inpatient module of the NHSS household questionnaire. Respondents were asked whether they were hospitalized or told by a physician to be hospitalized in the past 12 months. For those reporting a hospitalization, the medical condition was asked and then coded by the medically trained interviewer according to the NHSS medical condition code table. The module further asked about the type of health institution (from THCs up to provincial hospitals) visited, whether a surgery was performed, inpatient days, total expenditure, reimbursement, and patient satisfaction with several aspects of that inpatient stay. Such information was collected for the latest hospitalization episode in the 2008 wave and all episodes in the 2013 wave. For comparability between waves, we use only the latest episode per person. Satisfaction questions also differed between the two waves. For comparability and relevance, we use only patients' satisfaction with disease explanation, satisfaction with treatment explanation, trust in the medical staff, and the overall satisfaction.

The survey does not allow us to identify the precise diagnosis of each inpatient case. However, medical condition classification used in NHSS was detailed enough to provide 133 codes, of which 23 are identified to be the broad categories of DRG-eligible conditions. Given that an important selection criterion of DRG-eligible conditions was large volume, it is likely that these broad categories contain mostly DRG-eligible conditions. Among these conditions, deliveries are particularly well identified using the code for delivery and caesarean sections can be determined if a surgery was performed. Moreover, the ever-married women module of the survey provided additional information on deliveries for women aged 15-49. It asked about childbirths within five years from the time of the survey, including the time, type and place of delivery, birth weight, and expenditures.

5.3 Empirical Strategy

We identify the effects of the new inpatient payment system on service volume, case mix proxies, treatment intensity, expenditures and patient satisfaction by comparing between treatment and control counties, between pre- and post-reform periods, and between DRG-eligible and DRG-ineligible conditions. This DDD approach is valid as long as the difference between DRG-eligible and DRG-ineligible conditions in treatment and control counties would have followed parallel trends in the absence of the reform.

The parallel trend assumption is not directly testable. Instead, we estimate the general impacts of Health XI on some contextual variables (i.e., inpatient care need and the proportion of DRG-eligible conditions) and all outcome variables to provide some indirect evidence and a context for the payment reform. Health XI impacts are estimated by comparing treatment and control counties using a difference-in-differences (DID) model specified as follows:

$$Y_{ict} = \alpha + \beta_1 HXI_c + \beta_2 Post_t + \beta_3 (HXI_c \times Post_t) + X'_{ict}\gamma + \varepsilon_{ict} \quad (5.1)$$

where *i* denotes an individual, *c* denotes a county, *t* denotes a year, Y_{ict} is a series of outcomes as mentioned previously, HXI_c is a dummy for treatment counties Xi and Yiyang, $Post_t$ is a dummy for wave 2013, and X'_{ict} includes a series of individual characteristics, including age categories, a female dummy, age category and gender interactions, marital status, educational level, employment status, number of chronic diseases, household composition and income quartiles. For outcomes relating to case mix, treatment intensity (inpatient days and whether a surgery was performed), expenditures and patient satisfaction, X'_{ict} additionally includes the medical condition codes and provider levels. All income and expenditure variables are converted into 2007 price and trimmed at the 99 percentile. Standard errors are clustered at the county level.

Next, we estimate the impacts of the new payment reform using a triple difference (difference-in-difference-in-differences, DDD) model specified as follows:

$$\begin{aligned} Y_{ictd} = {} & \delta + \lambda_1 HXI_c + \lambda_2 Post_t + \lambda_3 DRG_d + \lambda_4 (HXI_c \times Post_t) + \lambda_5 (HXI_c \times DRG_d) \\ & + \lambda_6 (Post_t \times DRG_d) + \lambda_7 (HXI_c \times Post_t \times DRG_d) + X'_{ictd}\eta + \epsilon_{ictd} \end{aligned} \quad (5.2)$$

where *d* denotes a medical condition, DRG_d is a dummy for DRG-eligible conditions and the rest are specified as in Equation (5.1.[16]

For deliveries, we supplement the general inpatient information with more years of data from the ever-married women module. We use deliveries that occurred during July 2005 and August 2013 because since July 2005 the proportion of deliveries at county hospitals and THCs stabilized in a reasonably narrow range of 70%-82%. As NHSS and the payment reform both took place in the summer, we conveniently construct eight yearly data points using July as the delimiter (except in 2013). The effects of the new payment system are estimated using a slightly modified Equation (5.1), where Y_{ict} includes provider levels, whether a caesarean section was performed and total expenditure,[17] and X'_{ict} no longer includes inapplicable controls such as gender, but includes year dummies and additionally low birth weight and provider levels when they are not used as the outcome. Standard errors are clustered at the county-year level.[18]

[16] We do not leave out deliveries as they constitute 26% of DRG-eligible cases and our sample size is already small.

[17] We do not examine (i) case mix proxies because pregnancy is not a disease, (ii) inpatient days and patient satisfaction because such information is not available in the extra years, or (iii) out-of-pocket expenditure and reimbursement rate because the results may be confounded by a general increase in NRCMS generosity.

[18] We do not correct for the serial correlation problem by collapsing data into "before" and "after", as recommended by Bertrand et al. (2004) for the case of small *N*. This is because the correction substantially diminishes power, which is even less desirable than over-rejection given our hypothesis that provider response is more limited in delivery cases.

5.4 Results

After dropping individuals/hospitalization episodes with missing key information[19] and households with no members covered by NRCMS[20], we obtain 16461 individuals and 1051 hospitalization episodes. Table 5.1 presents descriptive statistics for the treatment and control samples. Column 3 shows that, at the baseline, the treatment sample closely resembles the control sample regarding a long list of demographic, socio-economic and health characteristics. The only exceptions are that respondents from treatment counties were on average more likely to be jobless and less likely to be employed or MA beneficiaries. However, the differences were modest or small in magnitude.

Column 6 in Table 5.1 indicates no differential changes in most variables. In particular, household per capita income and spending did not change differentially between treatment and control counties. The last six rows further show no significantly different changes in chronic condition prevalence, the need for outpatient and inpatient care, the proportion of inpatients with DRG-eligible conditions, and two predictors of the severity of delivery cases – mother's age at delivery and low birth weight. Differential changes between treatment and control are observed in the sample age structure, in particular the proportions of children and young adults. This contributes to differential changes in two other socio-demographic characteristics, namely the proportion of married respondents and the proportion of student respondents. However, the magnitude of the difference in changes was small and these young age groups are not the main component of our inpatient sample. Significantly different changes are also observed in MA coverage as a result of Health XI, but MA affected only a very small proportion of inpatients.[21] In general, our treatment and control samples appear well balanced in demographic and socio-economic characteristics.

Table 5.2 presents descriptive statistics and regression estimates on inpatient care use.[22] We examine use conditional on inpatient care need after having demonstrated the comparability of the need between treatment and control counties in Table 5.1.[23] The first row of Table 5.2 shows that inpatient care use (the complement of not hospitalized when needed) was significantly lower in treatment counties at baseline, but improved significantly more – by 9.6 percentage points – after five years. The next four rows examine use by provider level.[24] At

[19] The numbers of deleted individuals and inpatient episodes are 120 and 17 respectively. For 24 observations missing only overall satisfaction, we impute the value using the satisfaction rating on specific aspects.

[20] NRCMS enrolment is at the household level and 926 individuals are deleted.

[21] Even in the post-reform sample, only 41 individuals reported being MA beneficiaries.

[22] Results on inpatient care need and the proportion of DRG-eligible cases do not show significant differences, and are omitted to save space.

[23] We reach the same conclusion after controlling for sample demographic and socioeconomic characteristics in the estimation using Equation (5.1).

[24] We do not report the statistics for cases where the level of provider is unspecified. Therefore, the proportion of patients not hospitalized and the proportion hospitalized at urban and rural providers (the first three rows of the first column of Table 5.2) do not add up to one.

baseline, inpatient care use was more concentrated at urban hospitals than at rural providers in treatment counties than in control counties. Over the five years, all counties witnessed a move of rural patients upward along the provider ladder: using more care from urban hospitals than rural providers, and more from county hospitals than THCs. The DID estimates in column 5 of Table 5.2 show that the greater improvement in care use in treatment counties translated into more utilization of all levels of providers.

Column 6 of Table 5.2 presents results from the DDD model that additionally compares between DRG-eligible and DRG-ineligible conditions. Significant and substantial changes are observed: patients with DRG-eligible conditions were 9.7 percentage points *more* likely to be hospitalized at urban hospitals, and significantly *less* likely to be hospitalized at rural providers by 20.2 percentage points. Disaggregation of rural providers shows a significant and substantial decrease (42.9 percentage points) in hospitalization at THCs, and a significant and substantial increase (22.7 percentage points) at county hospitals. Table 5.1 has demonstrated the comparability of inpatient care need and the proportion of DRG-eligible conditions between treatment and control counties. These differential changes must therefore indicate that THCs reduced their service provision to DRG-eligible patients by referring some of them to higher-level providers. In response, county hospitals increased their service provision, but not enough to offset the reduction at THCs. Lacking relevant data, we cannot determine whether this was due to capacity constraints or that county hospitals also attempted to direct some DRG-eligible cases to higher-level (urban) hospitals.

Next, we examine the recent health condition of these inpatients using their rate of morbidity in the two weeks prior to the post-reform survey. Results are shown in the Panel A of Table 5.3. The first two rows of column 5 suggest that the recent morbidity rate was in general lower in treatment counties than in control counties. In particular, the difference was significant and substantial (20.6 percentage points) for those previously hospitalized at county hospitals. Column 6 presents the effects of the payment reform estimated using DDD. It appears that the new system significantly increased the two-week morbidity by 29.2 percentage points for patients previously hospitalized at county hospitals, but significantly decreased it by 19.6 percentage points for those previously hospitalized at THCs.

Admittedly, increased morbidity can be a result of poor inpatient care. However, Meng et al. (2014) analyzed the medical charts of bronchial pneumonia, caesarean section, ischemic stroke and chronic obstructive pulmonary disease at the county general hospitals in Xi and Yiyang during January 2009 and April 2014. They found substantial improvement in compliance with the clinical pathways after the payment reform. Assuming that compliance with clinical pathways improves quality, we interpret the increase/decrease in two-week morbidity as reflecting a better/worse case-mix. Moreover, we construct a more refined morbidity indicator, which is equal to one if the reported condition was the same condition that the patient was hospitalized for. Arguably, it better reflects the quality of care than the general two-week morbidity. The last two rows of column 6 show that the DDD estimates are no longer significant. Their magnitude decreases for patients previously hospitalized at county hospitals, and increases for those at THCs, as compared to those in the first two rows. Taken together, our results suggest that cream skimming occurred at THCs and resulted in a more severe case mix at county hospitals.

Panel B of Table 5.3 presents descriptive statistics and regression results for treatment intensity. In general, we do not expect Health XI interventions other than the promotion of clinical pathways to affect treatment intensity. Column 5 indeed shows limited differential change in intensity between treatment and control counties. However, the triple difference reveals some noticeable, though not significant, effects of the new payment system, especially at county hospitals. For example, the second row shows an increase of 3.7 inpatient days, and the second to last row an increase of 19.3 percentage points in the rate of surgery. However, this finding is not surprising given the worsened case mix at county hospitals.

Panel C of Table 5.3 presents the results on inpatient expenditures. As improved financial protection was one of the major Health XI project objectives, we expect reduced OOP payment and increased reimbursement rate in treatment compared to control counties. This is indeed what we observe in column 5, and the improvement was larger at county hospitals than at THCs. Column 6 presents the effects of the new payment system. Overall, total inpatient expenditure was not reduced, but the rate of reimbursement increased significantly by 8.4 percentage points. Disaggregating the effects by provider level reveals some unexpected findings. For example, despite the worsened case mix at county hospitals, total expenditure decreased (though not significantly). In contrast, total expenditure increased (though again not significantly) at THCs, and surprisingly, the reimbursement rate also increased. Restricted by the small sample size and the survey nature of our data, we are unable to explore further what was driving these results. However, it is possible that reimbursement increased substantially at the lower end of the expenditure distribution at THCs due to the increased use of basic and well-reimbursed drugs and services.

Next, we examine deliveries at health care facilities separately. Figure 5.3 plots the time trends in two contextual variables – mother's age at delivery and low birth weight – and two delivery outcomes – caesarean section rate and total expenditure. Again, statistics for county hospitals and THCs are presented separately. Due to the small sample size, outcomes are averaged over 24 months, yielding three observations pre-reform and one post-reform. The upper left graph of Figure 5.3 shows that, while average age at delivery followed parallel trends in treatment and control counties for women giving births at THCs, it was not the case at county hospitals. However, the divergence from a parallel path was modest in magnitude and our DID result (unreported) does not show significant difference. The upper right graph of Figure 5.3 shows that the incidence of low birth weight was very volatile and followed no clear pattern. In contrast, the bottom two graphs show more modest changes in caesarean section rate and total expenditure over time.

Table 5.4 presents descriptive statistics and regression estimates on care use by provider level, caesarean section rate, and total expenditure.[25] The first two rows show little differential change in care use by provider level on top of a shift to higher-level providers. Row 3 shows increases in the caesarean section rate at county hospitals in both treatment and control counties, but these increases were neither significantly different from zero, nor from each other, leading to an insignificant DID estimate. A similar case is true for the caesarean section

[25] Results on mother's age at delivery and low birth weight do not show significant changes, and are omitted to save space.

rate at THCs shown in row 4 of Table 5.4. The bottom two rows reveal modest changes in total expenditures at county hospitals and THCs, leading to insignificant DID estimates of the reform impact. In sum, we observe no evidence of provider response such as cream skimming or induced surgery, which is not unexpected given the special features of delivery cases.

Table 5.5 presents descriptive statistics and regression results for three specific satisfaction questions – satisfaction with the explanation on the medical problems and the treatment plan, and trust in medical staff – and one overall satisfaction question. While satisfaction increased in treatment counties in general, it was exclusively driven by DRG-ineligible cases. Satisfaction ratings even decreased slightly among patients with DRG-eligible conditions. Some may find it surprising given that, under the new payment system, health professionals spent more time communicating with the patients and patients were better informed about their treatment plan and expenditure in advance. However, previous research has found that patients' satisfaction correlates with the extent to which their requests are fulfilled (Rao et al. 2000). It is conceivable that in this case patients were unsatisfied when their care demands – e.g., for advanced diagnostic tests and prescription drugs – could not be met given the restrictions imposed by clinical pathways. Moreover, under the new system, patients could no longer receive daily medical bills listing all their expenses, but only a grand total at discharge. This is conducive to distrust given the currently poor physician-patient relationship in China (Zeng et al. 2013).

5.5 Conclusion

Hospital payment schemes vary across health care systems. While there is no clear consensus on which is to be preferred, many countries are moving toward some variant of case-base payment (Langenbrunner et al. 2009). The most frequent reasons for doing so are to increase efficiency and contain costs. However, these goals can be undermined by the payment system's inherent incentives for up-coding and under-provision of needed services (Busse et al. 2011; Langenbrunner et al. 2009). Moreover, implementing such payment systems in low- and middle-income counties is particularly challenging given the requirements on coding standardization, data availability and quality, and information technology (Mathauer & Wittenbecher 2013). Therefore, it takes much effort to counteract the unwanted incentives and also many years for the payment system to mature (Langenbrunner et al. 2009).

We evaluate intended and unintended effects of a DRG-based inpatient payment system that was piloted in rural China. Although it was designed to contain costs, we did not find evidence that the new payment system significantly reduced total inpatient expenditure. We do observe a large reduction in total expenditure at county hospitals but the estimate is very imprecise and is far from reaching conventional levels of significance. It is possible that county hospitals would have managed to significantly reduce their expenditure if their case mix had not worsened. However, a more plausible explanation is that some crucial preconditions for the case-based payment to improve efficiency and reduce cost were not met.

One such precondition is competition between different providers on a playing field made level by the payment method. Researchers have cautioned that efficiency gains will be small if the case-based payment approaches hospital-specific payment (O'Dougherty et al. 2009). While all six pilot hospitals in Beijing face the same DRG rates (Jian et al. 2015) and are in close competition, the new payment method in our case is close to a hospital-specific one and not conducive to fair competition at the county level. This is not only due to the small number of county hospitals, but also because of the way fee adjustments are made: case by case, rather than across groups of cases or groups of hospitals as is done in mature DRG systems. At the township level, though the number of THCs is much larger (around 20) and the payment rates uniform, effective competition is practically nonexistent given THCs' general low quality of care and geographic dispersion. In practice, THCs are more often in competition with the higher-level county hospitals than with other low-quality THCs.[26] However, we do not find in internal documents any mention of efforts to level the playing field between county hospitals and THCs when setting the payment rates.

In contrast to the small effect on cost containment, we find a large reduction in services provided by THCs. While some recent studies have also documented a positive relationship between fee levels and service volume (Clemens & Gottlieb 2014; Shigeoka & Fushimi 2014), the magnitude of the response is more substantial in our case. This may be driven by the fact that the new system was applied to only a selection of inpatient conditions. The partial coverage left THCs with the opportunity to recoup income losses from DRG-ineligible cases, which were not subject to clinical pathways and still paid by FFS. Somewhat similarly, Jian et al. (2015) documented that 35 percent of DRG-eligible cases in the Beijing pilot were allowed to and actually did receive FFS payments. These FFS reversions cases did not experience significant reductions in expenditures or OOP payments, and were suspected as a result of hospitals shifting their financial risk of treating sicker patients to the health insurance. Previous literature discussing case-based payment has warned against such incomplete coverage of cases (Mathauer & Wittenbecher 2013; Yip et al. 2010). However, full coverage was not possible in Xi and Yiyang due to local capacity constraints and the experimental nature of the reform.

In addition to the features of the new payment system, the institutional setting also made it possible for THCs to reduce service volume for DRG-eligible conditions that experienced a fee cut. Specifically, THCs in China do not have any formal responsibility as gatekeepers, and are free to refer patients to county hospitals. Patients, on the other hand, are in general also willing to be referred to higher-level providers that are neither too far away nor too expensive (i.e. county hospitals). Combined, these two factors led to a relatively large reduction in service volume at THCs, implying a dominant substitution effect. In comparison, county hospitals are less capable of referring patients to higher-level providers because time and money costs rise sharply for rural patients if they have to seek care at urban hospitals. Consequently, county hospitals were forced to treat a larger volume of DRG-eligible cases. However, the total

[26] Townships within a county are typically better connected with the county center than with each other.

number of DRG-eligible cases decreased among rural providers, suggesting that the substitution effect dominated the income effect.

Overall, our results reveal little or no intended effect on cost containment, but an unintended reduction in service provision and cream skimming at THCs. The findings illustrate the challenges faced when implementing case-based payment in a low- and middle-income setting with, among other issues, insufficient competition between providers and partial coverage of inpatient cases. To address these problems, payment rates need to be improved with an explicit intention to level the playing field between different providers and all inpatient conditions should be covered by the new payment system. This requires high-quality clinical and costing data as well as an appropriate information technology system to link them. However, in practice, the systems for generating the necessary data are usually not set up until a case-based system is already in place (Mathauer & Wittenbecher 2013), as was the case here. Therefore, periodical data analysis and targeted supervision would be required to reduce unwanted incentives as the payment system matures.

Figures

Figure 5.1. Inpatient provider payment reform coverage over time.

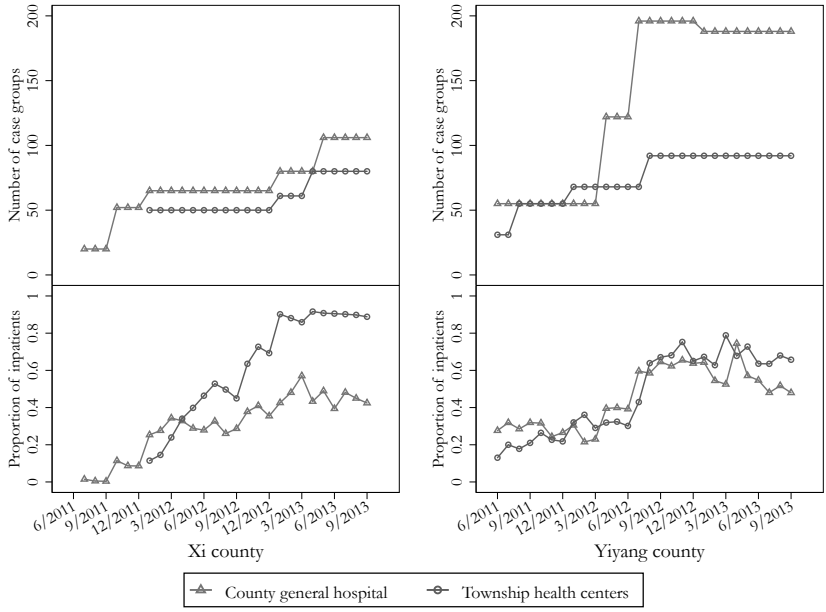


Figure 5.2. Economic conditions over eight years in treatment and control counties.

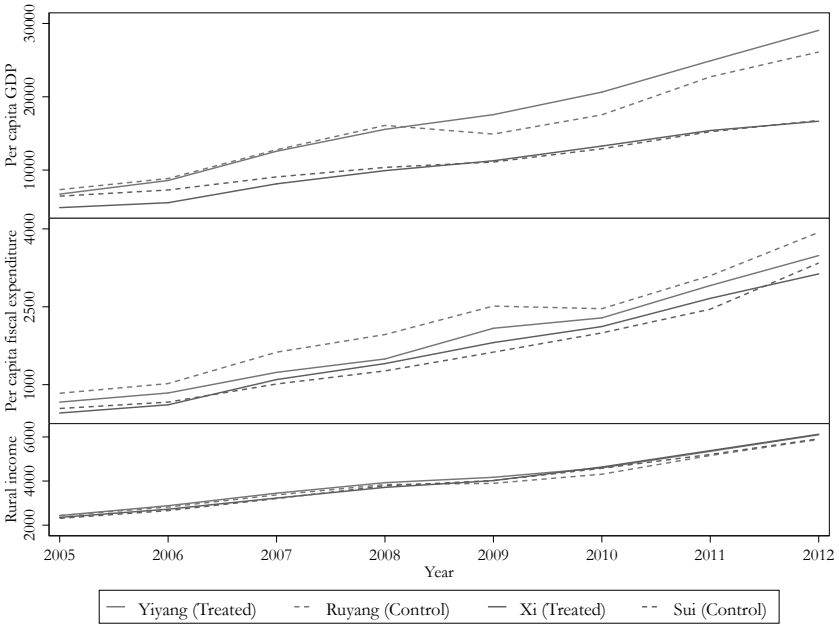
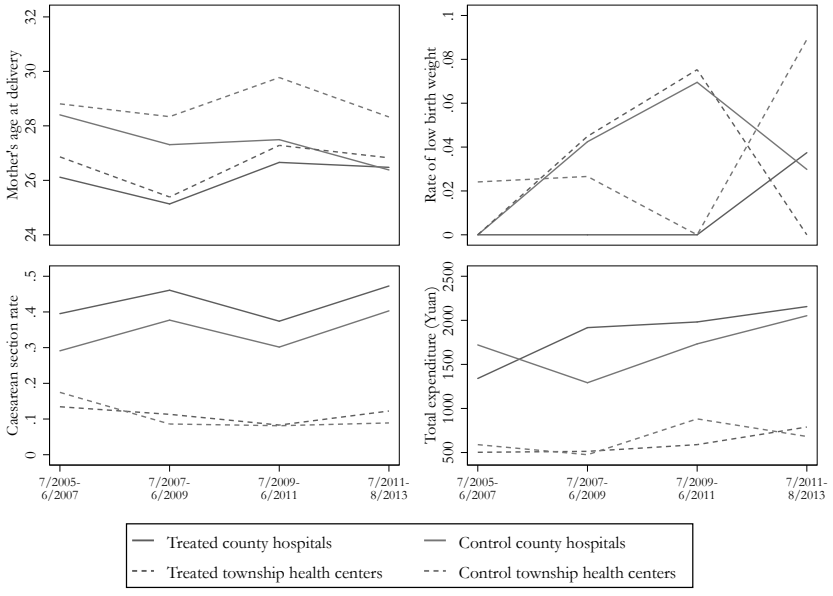


Figure 5.3. Time trends in delivery-related variables.



Tables

Table 5.1. Descriptive statistics by treatment status.

	Year 2008			Change (2013-2008)			Sample size
	Treatment (T)	Control (C)	T=C *p*-value	Treatment (ΔT)	Control (ΔC)	ΔT=ΔC *p*-value	
Age	35.4	34.9	0.715	4.6	0.8	0.147	16461
0-14	0.26	0.24	0.518	-0.04	0.00	**0.087**	16461
15-24	0.11	0.13	0.473	-0.03	0.00	0.254	16461
25-34	0.09	0.10	0.969	**-0.02**	0.01	**0.016**	16461
35-44	0.15	0.17	0.274	**-0.02**	**-0.04**	**0.012**	16461
45-54	0.14	0.15	0.575	0.04	0.00	0.145	16461
55-64	0.14	0.12	0.539	0.03	0.01	0.109	16461
65+	0.10	0.09	0.148	0.04	0.02	0.518	16461
Female	0.50	0.50	0.790	0.00	0.00	0.506	16461
No schooling	0.28	0.19	0.393	0.00	-0.02	0.782	12490
Primary school	0.30	0.28	0.572	-0.03	-0.02	0.738	12490
Middle school	0.34	0.42	0.511	0.02	0.02	0.928	12490
High school and above	0.09	0.11	0.116	0.00	0.01	0.817	12490
Single	0.15	0.17	0.411	-0.05	-0.01	0.247	12490
Married	0.77	0.76	0.711	**0.05**	0.01	**0.067**	12490
Student	0.07	0.08	0.639	-0.04	**-0.01**	**0.100**	12490
Employed	0.79	0.88	**0.056**	0.08	-0.11	0.107	12490
Jobless	0.12	0.03	**0.032**	-0.05	0.13	0.126	12490
Retired	0.01	0.01	0.521	**0.00**	0.00	0.585	12490
Household income per capita	3386	3802	0.349	2216	2621	0.556	16461
Household expenditure per capita	2822	2773	0.892	1139	1735	0.199	16461
New Rural Cooperative Medical Scheme	0.96	0.95	0.816	0.03	0.04	0.910	16461
Medical Assistance	0.02	0.04	**0.012**	0.03	0.00	**0.030**	16156
Chronically ill	0.20	0.13	0.192	0.03	0.02	0.912	12490
Two-week morbidity	0.20	0.18	0.619	-0.07	-0.04	0.497	16463
Hospitalization needed in the past 12 months	0.08	0.07	0.756	0.00	0.00	0.902	16463
Proportion of DRG-eligible inpatient cases	0.53	0.50	0.696	0.03	0.07	0.662	1051
Mother's age at delivery	27.4	28.6	0.132	-0.90	-1.90	0.23	880
Low birth weight	0.04	0.02	0.224	0.00	0.03	-0.04	880

Notes: Income and expenditure are in 2007 price. Bold indicates significance at the 10% level.

Table 5.2. Descriptive statistics and regression estimates for inpatient care use.

	Descriptive statistics				DID (Health XI)	Sample size	DDD (New payment system)	Sample size
	Baseline		Change					
	Treatment	Control	Treatment	Control				
Not hospitalized when needed	0.266	0.149**	-0.196*	-0.104	-0.096**	1225		
	(0.442)	(0.357)	(0.016)	(0.021)	(0.020)			
Hospitalized at urban hospitals	0.059	0.026**	0.065	0.044	0.022	1225	0.097	1064
	(0.236)	(0.160)	(0.081)	(0.030)	(0.070)		(0.076)	
Hospitalized at rural providers	0.641	0.801**	0.162	0.084*	0.083	1225	-0.202**	1064
	(0.480)	(0.400)	(0.040)	(0.011)	(0.039)		(0.035)	
Hospitalized at county hospital	0.293	0.389	0.231	0.220	0.025	1225	0.227*	1064
	(0.456)	(0.488)	(0.059)	(0.099)	(0.106)		(0.095)	
Hospitalized at THCs	0.348	0.412	-0.069	-0.136	0.058	1225	-0.429**	1064
	(0.477)	(0.493)	(0.019)	(0.088)	(0.086)		(0.114)	

Notes: THCs stands for township health centers. Column 5 gives the coefficient of $HXI \times Post$ in Equation 5.1. Column 9 gives the coefficent of $HXI \times Post \times DRG$ in Equation 5.2. Stars in column 2 signify significant differences in baseline values between treatment and control counties. Standard errors in brackets are clustered at the county level. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 5.3. Descriptive statistics and regression estimates for case mix, treatment intensity and expenditures.

	Descriptive statistics				DID (Health XI)	DDD (New payment system)	Sample size
	Baseline		Change				
	Treatment	Control	Treatmen	Control			
Panel A. Case mix							
Two-week morbidity for patients previously hospitalized							
at county hospitals	0.477	0.376	-0.256	-0.046	-0.206**	0.292*	554
	(0.503)	(0.487)	(0.052)	(0.015)	(0.051)	(0.104)	
at THCs	0.345	0.408	-0.011	-0.068	-0.028	-0.196*	397
	(0.478)	(0.494)	(0.145)	(0.032)	(0.157)	(0.065)	
Two-week morbidity due to the same inpatient condition for patients previously hospitalized							
at county hospitals	0.204	0.191	-0.169**	-0.049	-0.127	0.035	554
	(0.406)	(0.395)	(0.012)	(0.041)	(0.055)	(0.102)	
at THCs	0.114	0.238**	0.066	-0.153	0.140	-0.350	397
	(0.319)	(0.428)	(0.075)	(0.092)	(0.110)	(0.163)	
Panel B. Treatment intensity							
Inpatient days at rural providers	9.7	9.4	0.7	0.5	0.1	0.4	951
	(10.6)	(8.5)	(0.6)	(0.8)	(0.5)	(2.1)	
at county hospitals	13.4	12.2	-2.5	-1.4	-0.9	3.7	554
	(14.2)	(10.9)	(2.4)	(0.2)	(0.7)	(4.9)	
at THCs	6.6	6.7	2.7**	1.1	0.4	-1.2	397
	(4.6)	(3.6)	(0.2)	(0.8)	(1.8)	(1.4)	
Surgery at rural providers	0.307	0.207**	0.017*	0.060	-0.002	0.058	951
	(0.462)	(0.406)	(0.002)	(0.011)	(0.032)	(0.048)	
at county hospitals	0.370	0.343	0.026	-0.002	-0.045	0.193	554
	(0.486)	(0.477)	(0.069)	(0.016)	(0.056)	(0.122)	
at THCs	0.257	0.077**	-0.069	0.027	-0.031	0.009	397
	(0.439)	(0.268)	(0.064)	(0.036)	(0.018)	(0.129)	
Panel C. Expenditures							
Total expenditure at rural providers	2133	1707	1491*	1529	-21	-43	948
	(2565)	(2596)	(151)	(406)	(315)	(523)	
at county hospitals	3552	2700**	1010	1396	-665	-1016	552
	(3225)	(3387)	(191)	(338)	(860)	(1324)	
at THCs	1002	749	835	587	88	895	396
	(834)	(602)	(240)	(293)	(392)	(1171)	
OOP expenditure at rural providers	1531	1007	14	1014	-929**	430	948
	(2153)	(1414)	(339)	(372)	(255)	(427)	
at county hospitals	2743	1693**	-676	1003	-1778*	379	552
	(2745)	(1753)	(445)	(325)	(674)	(834)	
at THCs	566	346	-15	182	-233	433	396
	(511)	(298)	(85)	(70)	(162)	(497)	
Reimbursement rate at rural providers	0.354	0.421	0.252	0.040	0.233***	0.084*	947
	(0.271)	(0.281)	(0.049)	(0.042)	(0.017)	(0.032)	
at county hospitals	0.255	0.304	0.277	0.089*	0.246***	-0.039	551
	(0.257)	(0.250)	(0.060)	(0.009)	(0.015)	(0.061)	
at THCs	0.432	0.535**	0.315**	0.078	0.203**	0.120	396
	(0.257)	(0.262)	(0.018)	(0.056)	(0.038)	(0.093)	

Notes: THCs stands for township health centers. Column 5 gives the coefficient of $HXI \times Post$ in Equation 5.1. Column 9 gives the coefficent of $HXI \times Post \times DRG$ in Equation 5.2. Stars in column 2 signify significant differences in baseline values between treatment and control counties. Standard errors in brackets are clustered at the county level. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 5.4. Descriptive statistics and regression estimates for delivery outcomes.

	Baseline		Change		DID	Sample size
	Treatment	Control	Treatment	Control		
Delivery at county hospitals	0.185	0.287**	0.215***	0.202*	0.012	880
	(0.389)	(0.453)	(0.056)	(0.103)	(0.092)	
Delivery at THCs	0.502	0.488	-0.196*	-0.242***	0.054	880
	(0.501)	(0.501)	(0.097)	(0.076)	(0.088)	
Cesarean section rate at county hospitals	0.336	0.313	0.137	0.090	0.034	264
	(0.476)	(0.466)	(0.154)	(0.118)	(0.154)	
Cesarean section rate at THCs	0.130	0.097	-0.008	-0.008	-0.004	392
	(0.338)	(0.298)	(0.090)	(0.032)	(0.083)	
Total expenditure at county hospitals	1838	1599	318	453*	-162	264
	(1211)	(1296)	(292)	(234)	(243)	
Total expenditure at THCs	602	615	187	68	103	392
	(464)	(821)	(147)	(141)	(190)	

Notes: THCs stands for township health centers. Column 5 gives the coefficient of *HXI×Post* in Equation 5.1. Stars in column 2 signify significant difference in baseline values between treatment and control counties. Standard errors in brackets are clustered at the county level. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

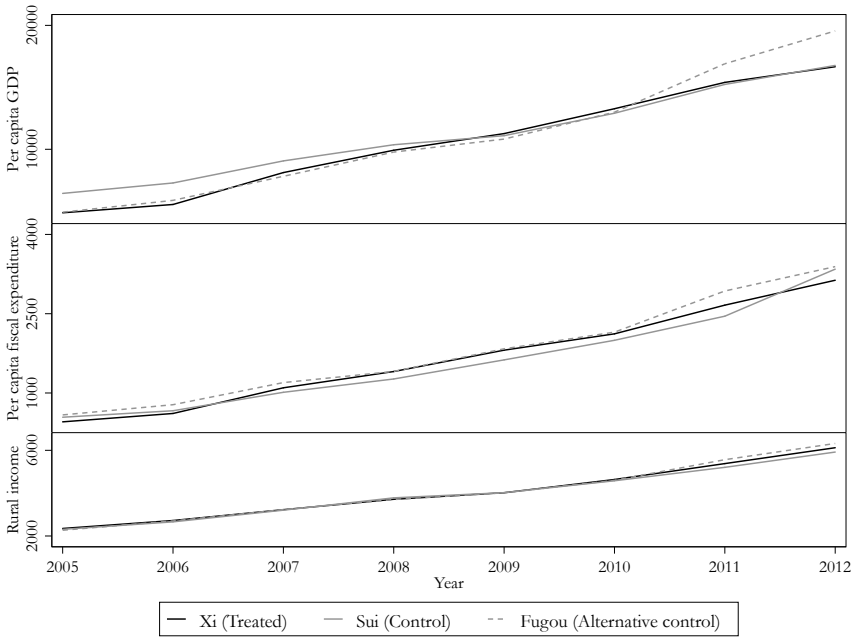
Table 5.5. Descriptive statistics and regression estimates for satisfaction.

	Descriptive statistics				DID (Health XI)	DDD (New payment system)	Sample size
	Baseline		Change				
	Treatmen	Control	Treatmen	Control			
Very satisfied with explanation on the problems							
at rural providers	0.077	0.081	0.261	0.045	0.218*	-0.145	951
	(0.268)	(0.273)	(0.120)	(0.010)	(0.092)	(0.090)	
at county hospitals	0.094	0.116	0.272	0.040	0.261*	-0.182	554
	(0.294)	(0.322)	(0.133)	(0.009)	(0.085)	(0.083)	
at THCs	0.064	0.047	0.220	0.015	0.210	0.051	397
	(0.246)	(0.213)	(0.092)	(0.025)	(0.100)	(0.175)	
Very satisfied with explanation on the treatment plan							
at rural providers	0.089	0.062**	0.263	0.064	0.188*	-0.209	951
	(0.286)	(0.242)	(0.105)	(0.023)	(0.076)	(0.092)	
at county hospitals	0.107	0.070	0.286	0.075	0.217*	-0.138	554
	(0.311)	(0.256)	(0.130)	(0.050)	(0.069)	(0.070)	
at THCs	0.075	0.055	0.201	0.030	0.194*	-0.172	397
	(0.265)	(0.229)	(0.064)	(0.020)	(0.078)	(0.186)	
Trust the medical staff very much							
at rural providers	0.103	0.084	0.237	0.116	0.097	-0.136	951
	(0.304)	(0.278)	(0.169)	(0.050)	(0.126)	(0.061)	
at county hospitals	0.069	0.088	0.311	0.116**	0.147	-0.148	554
	(0.254)	(0.285)	(0.208)	(0.007)	(0.142)	(0.078)	
at THCs	0.130	0.079	0.135	0.110	0.018	0.032	397
	(0.338)	(0.272)	(0.083)	(0.125)	(0.115)	(0.213)	
Overall satisfied							
at rural providers	0.807	0.777	0.183	0.203	-0.026	-0.055	951
	(0.396)	(0.418)	(0.132)	(0.087)	(0.116)	(0.088)	
at county hospitals	0.763	0.717	0.227	0.259	-0.012	0.061	554
	(0.428)	(0.453)	(0.138)	(0.089)	(0.109)	(0.140)	
at THCs	0.843	0.834	0.149	0.154	-0.137	-0.053	397
	(0.366)	(0.374)	(0.118)	(0.062)	(0.115)	(0.094)	

Notes: THCs stands for township health centers. Column 5 gives the coefficient of *HXI×Post* in Equation 5.1. Column 6 gives the coefficent of *HXI×Post×DRG* in Equation 5.2. Stars in column 2 signify significant differences in baseline values between treatment and control counties. Standard errors in brackets are clustered at the county level. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Appendix 5.A

Table 5.A.1. Comparison between Sui and Fugou as the potential control for Xi.



Appendix 5.B

Unlike typical DRG systems that adjust the final payment rate using region-specific or hospital-type adjustment coefficients, Xi and Yiyang accounted for the variation in diseases treated and medical practices between county hospitals and THCs by developing different case grouping (different conditions or different categorization of the same condition), clinical pathways and fee schedules. Enumerating all detailed designs is beyond the scope of this paper. For reference, we provide only the list of conditions eligible for the new payment at the county general hospital in Xi. They are rotavirus gastroenteritis, esophageal cancer, stomach cancer, colon cancer, lung cancer, breast cancer, cervix cancer, benign gallbladder tumors, uterine fibroids, ovarian cyst, benign thyroid nodules, angina resulting from coronary artery disease, ST segment elevation myocardial infarction, cerebral hemorrhage, ischemic stroke, varicose veins, pneumonia, chronic rhinosinusitis, chronic tonsillitis, pneumonia, chronic obstructive pulmonary disease, stomach ulcer, appendicitis, inguinal hernia, Hirschsprung's disease, cholelithiasis, spinal disc herniation, acute pyelonephritis, ureterolithiasis, bladder stones, benign prostatic hyperplasia, hydrocele testis, adenomyosis, fallopian pregnancy, vaginal delivery, caesarean section, nasal septum deviation, clavicle fracture, humeral shaft fracture, distal radius fracture, femoral neck fracture, femoral shaft facture, patella fracture, tibial fracture, tibial and fibula fracture, tibial plateau fracture, ankle fracture, and dialysis.[27]

[27] Many conditions are categorized into multiple case groups depending on causes, procedures, severity and age of the patient.

CHAPTER 6

Conclusion

This dissertation consists of four studies divided into two parts. Part one assesses the potential determinants of the female health disadvantage and mental wellbeing among the Chinese elderly. Part two contains two chapters that aim to evaluate recent reform attempts at addressing some of the most pressing health care system challenges.

The first study focuses on the particularly problematic female disadvantage in health among the elderly Chinese. It examines three potential explanations: (i) gender-specific reporting behaviors, (ii) female disadvantage in education as a result of the culturally rooted discrimination against women, and (iii) gender differences in chronic conditions and health functioning. The three corresponding findings are as follows. First, no evidence is found for gender-specific reporting, ruling it out as an explanation of the gender health gap. Secondly, female disadvantage in education makes a gross contribution of 23.2% to the gap when only socio-demographic characteristics are examined. Thirdly, gender differences in chronic conditions and health functioning explain about two thirds of the gap that is previously left unexplained by socio-demographic variables. However, they neither fully explain the gender health gap nor wipe out the gross contribution by female disadvantage in education. In sum, the answer to the questions raised in the introduction is: female health disadvantage is not a reporting artifact, and is explained by both discrimination against women and biological differences.

It is interesting to note that previous literature suggests that education affects health through the onset of chronic conditions and health functioning (Zimmer et al. 1998; Zimmer et al. 2014). Nevertheless, after accounting for observed conditions and functioning, as much as one third of the education contribution to the gender health gap still remains. This finding suggests that educational disadvantage can negatively affect health *even without materializing into diseases and functioning limitations that are commonly seen*. Therefore, simply promoting equality in the delivery of health care services is unlikely to fully eliminate the large gender health gap. Greater investment in rural education and positive policy discrimination in favor of girls are worth considering given the widely persistent son preference and poor education resources in rural China. In fact, Ningxia province already released a regulation in 2014 demanding that girls from rural areas be given priority in high school enrollment, and in getting tuition and accommodation fee waivers.[1] Such policies seem to be promising measures that could reduce or even eliminate the large female health disadvantage in the future.

[1] Source: http://www.nxfp.gov.cn/fpxw/fpyw/12544.htm.

The second study zooms in on an important albeit much less studied health domain – mental wellbeing – among the elderly. It examines in particular the effect of having only one child on parents' mental wellbeing, given the cultural and economic motivations for multiple children and the widespread public discontent with the one-child policy. Using the instrumental variable approach, this study yields two main findings. First, having only one child does not inflict a negative net effect on the mental wellbeing of elderly parents (aged 45+). Secondly, further analysis to reveal coping strategies suggests that having only one child does not reduce parents' chances of having a child living in the same household or close by, or receiving transfers (both the instances and in net amounts) from a child. On the contrary, it significantly increases parents' chances of seeing a child at least once a month.

The finding that having more children does not imply more parent-child interactions is consistent with some of the observations from sociology studies[2] on the recent increase in suicides among the rural elderly, who typically have multiple children. In particular, one study examined suicides from a county located in the central plain area of China and therefore more susceptible to outside influences (B. Chen 2009). Four types of suicides were identified from a total of 128 cases:[3] (1) 70 cases were out of despair, typically because their giving to children was not matched with caring from children when they got old; (2) 18 cases were out of loneliness, typically because the children lived too far away to provide emotional support, (3) 23 cases were out of own will, typically because the elderly became very sick and wanted to relieve their children of the burden; and (4) 14 cases were out of rage, typically because of a lack of respect (for their authority) from children-in-law, in particular daughters-in-law.

These findings together cast doubt on the traditional expectation of economic and emotional old-age support from children, which underlies the preference for more children. Social and economic development in the past few decades has changed some of the preconditions for that expectation: e.g., a production style that was more labor-intense, low mobility of the labor force, and a cultural pressure (from local communities) for children to provide old-age support. Therefore, it is perhaps not entirely surprising that the quantity of children does not determine elderly parents' mental wellbeing nowadays. However, the generation of one-child parents is still relatively young (mostly younger than 65) and therefore active and healthy. Only children can be less capable of supporting their parents in ten years as their parents develop difficulties in daily activities or serious diseases. Therefore, the long-term effects of only children merits long-term monitoring and research attention. If follow-up research shows that parents with only one child suffer much more than those with more children in less favorable conditions, it may be worthwhile to consider health care benefits especially targeted at one-child parents.

The third and fourth studies both evaluate attempts at improving the current problematic health care system in China. The third study examines a more broad-based health reform project – the eleventh World Bank health project, Health XI. It aimed at finding effective interventions that could be replicated elsewhere in China to improve the financing and

[2] Source: http://www.cssn.cn/shx/shx_bjtj/201407/t20140730_1273075.shtml.

[3] All cases described in the paper mentioned more than one child of the deceased. Moreover, judging by their age – at least 55 years old by 2005 (i.e. 30 in 1980) – they should in general have more than one child.

delivery of medical health and public health services. Even more broadly, it also intended to develop a model that can facilitate inter-departmental coordination and cooperation, and reform design and implementation. Using household data, the study yields two main findings. First, positive project effects are found in all three domains examined – medical care, public health services provision, and self-rated health. The evidence of improvements is particularly strong for several aspects of medical care use and for all aspects of public health services provision at primary health care facilities. Secondly, lack of experience with similar reform projects does not put a county at a disadvantage in meeting project targets, and neither does poor fiscal condition.

As interventions differ across project counties, it is unlikely that the overall success of Health XI can be attributed to a specific reform. In addition, non-incentivized indicators are found to enjoy positive spillovers from incentivized ones instead of suffering from resources reallocation. These results suggest that the management of Health XI provides proper incentives to the participating counties. Specifically, annual funding disbursement is not conditional on some target values of the M&E indicators, but on the completion of activities proposed by counties and approved by the national management team. This ensures that counties do not behave opportunistically and propose only narrow-scope interventions specifically targeted at M&E indicators. The conditional funding disbursement arrangement ensures strong incentives for project counties to implement the approved activities, which often improve both M&E and non-M&E outcomes. It appears that this extra step between performance indicators and payment manages to retain the performance incentives, and at the same time avoid the side effects on non-incentivized indicators. Future reforms may consider adopting similar intermediate steps when designing how to incentivize performance.

While Health XI achieves substantial benefits in many aspects given the investment, there is no evidence of an effect on curbing the escalating inpatient expenditure. This finding reveals the difficulty in reforming the perverse incentives to overuse drugs and medical tests, which providers rely on to obtain the bulk of their income. The challenges are further illustrated in the fourth study, which examines a simplified diagnosis-related group (DRG)-based payment system piloted in two counties in the Henan province. The new system (1) set the payment rates for a selection of inpatient conditions to contain costs, and (2) imposed clinical pathways to ensure better quality of care and to encourage the use of basic services and drugs. This study yields three findings. First, the new payment system reduces the service volume for DRG-eligible conditions, does no have the intended effect on cost containment and does not affect patient satisfaction. Secondly, township health centers are more capable of responding strategically and provider response is more limited for delivery cases. Thirdly, there is some suggestive evidence of a great use of basic drugs and services that are better reimbursed by health insurance.

The finding of no effect on cost containment is discouraging given the intention of the new payment system. It is certainly challenging to design a proper incentive structure for the inpatient care delivery. However, it has been acknowledged that an effective DRG-based payment system requires competition between different providers on a playing field made level by the payment calculations (O'Dougherty et al. 2009). Unfortunately, room for competition is absent in the case evaluated. Moreover, a lack of formal gate-keeping system

makes it relatively easy for lower-level providers to refer unprofitable patients away. Future pilots of DRG-based payment systems should first address these two aspects to prevent providers from generating more inefficiency in the system and heavier burdens for the patients and the health insurance.

This dissertation has explored some of the potential contributing factors to the gender health inequalities and mental wellbeing problems among the elderly in part one. Part two provides insights on elements that are responsible for the successes and the failures of piloted interventions. However, this dissertation also raises new questions. For example, besides diseases and functioning limitations, what are other mechanisms through which disadvantage in education explains the disadvantage in self-rated health for women? Would one-child parents fare worse in mental wellbeing in less favorable situations, such as when they become widowed or need long-term care? What are the important contextual variables that contribute to the overall success of Health XI? Would competition from private clinics or other private health care providers help reduce the magnitude of undesirable provider response to financial incentives? Further research is needed to shed light on these and many other questions that remain. Just as the famous Chinese poem goes: 路漫漫其修远兮,吾将上下而求索 (the way ahead is long; I see no ending, yet high and low I will search with my will unbending).

References

Allen, J. et al., 2014. Social determinants of mental health. *International Review of Psychiatry*, 26(4), pp.392–407.

Baeten, S., Van Ourti, T. & van Doorslaer, E., 2013. Rising inequalities in income and health in China: Who is left behind? *Journal of Health Economics*, 32(6), pp.1214–1229.

Bago d'Uva, T. et al., 2008. Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, 17(3), pp.351–375.

Bago d'Uva, T. et al., 2011. Slipping Anchor?: Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. *Journal of Human Resources*, 46(October 2009), pp.875–906.

Banister, J., 1991. *China's changing population*, Stanford University Press.

Berkman, L.F., 2000. Which influences cognitive function: living alone or being alone? *The Lancet*, 355(9212), pp.1291–1292.

Bertrand, M., Duflo, E. & Mullainathan, S., 2004. How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, 119(I), pp.249–275.

Bhattacharyya, O. et al., 2011. Evolution of primary care in China 1997–2009. *Health Policy*, 100(2–3), pp.174–180.

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

Blumenthal, D. & Hsiao, W., 2005. Privatization and its discontents--the evolving Chinese health care system. *The New England Journal of Medicine*, 353(11), pp.1165–1170.

Boey, K.W., 1999. Cross - validation of a short form of the CES - D in Chinese elderly. *International Journal of Geriatric Psychiatry*, 14(8), pp.608–617.

Bongaarts, J. & Greenhalgh, S., 1985. An Alternative to the One-Child Policy in China. *Population and Development Review*, 11(4), p.585.

Busse et al., 2011. *Diagnosis-Related Groups In Europe: Moving Towards Transparency, Efficiency And Quality In Hospitals* R. Busse et al., eds.,

Cáceres-Delpiano, J. & Simonsen, M., 2012. The toll of fertility on mothers' wellbeing. *Journal of Health Economics*, 31(5), pp.752–766.

Cai, F. et al., 2012. *The Elderly and Old Age Support in Rural China*, The World Bank.

Case, A. & Paxson, C.H., 2005. Sex Differences in Morbidity and Mortality. *Demography*, 42(2), pp.189–214.

Casey, K., Glennerster, R. & Miguel, E., 2012. Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *Quarterly Journal of Economics*, 127(4), pp.1755–1812.

Centre for Health Statistics and Information of Ministry of Health, 2004. *An Analysis Report of National Health Services Survey in 2003*, Beijing: Peking Union Medical College Press.

CHARLS Research Team, 2013. Challenges of population aging in China: evidence from the national baseline survey of the China Health and Retirement Longitudinal Study (CHARLS). *Beijing: School of National Development, Peking University*.

Chen, B., 2009. Change of Inter-generational Relations and The Elderly Suicide: An empirical study in Jingshan county , Hubei province. *Sociological Studies*, 4, pp.157–176.

Chen, L. et al., 2007. The effects of Taiwan's National Health Insurance on access and health status of the elderly. *Health Economics*, 16(3), pp.223–242.

Chen, P. & Kols, A., 1982. Population and birth planning in the Peoples Republic of China. *Population Reports. Series J: Family Planning Programs*, (25), pp.577–619.

Chen, R., Copeland, J.R. & Wei, L., 1999. A meta-analysis of epidemiological studies in depression of older people in the People's Republic of China. *International Journal of Geriatric Psychiatry*, 14(10), pp.821–30.

Chen, Z., 2009. Launch of the health-care reform plan in China. *The Lancet*, 373(9672), pp.1322–1324.

Cheng, S.-T. & Chan, A.C.M., 2005. The Center for Epidemiologic Studies Depression Scale in older Chinese: thresholds for long and short forms. *International Journal of Geriatric Psychiatry*, 20(5), pp.465–470.

Clemens, J. & Gottlieb, J.D., 2014. Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health? *American Economic Review*, 104(4), pp.1320–1349.

Cutler, D. & Lleras-Muney, A., 2008. Education and Health: Evaluating Theories and Evidence. In J. House et al., eds. *Making Americans Healthier: Social and Economic Policy as HealthPolicy*. New York: Russell Sage Foundation.

Datta Gupta, N., Kristensen, N. & Pozzoli, D., 2010. External validation of the use of vignettes in cross-country health studies. *Economic Modelling*, 27(4), pp.854–865.

Deaton, A., 2010. Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2), pp.424–455.

Denizer, C., Kaufmann, D. & Kraay, A., 2013. Good countries or good projects? Macro and micro correlates of World Bank project performance. *Journal of Development Economics*, 105, pp.288–302.

Deutsch, F.M., 2006. Filial Piety, Patrilineality, and China's One-Child Policy. *Journal of Family Issues*, 27(3), pp.366–389.

Du, X., 1995. *Zhongguo nvzi jiaoyu tongshi (The General History of Female Education in China)*, Guiyang: Guizhou Education Press.

Eggleston, K., 2005. Multitasking and mixed systems for provider payment. *Journal of Health Economics*, 24(1), pp.211–223.

Elo, I.T., 2009. Social Class Differentials in Health and Mortality: Patterns and Explanations in Comparative Perspective. *Annual Review of Sociology*, 35(1), pp.553–572.

Finkelstein, A. et al., 2012. The Oregon Health Insurance Experiment: Evidence from the First Year. *The Quarterly Journal of Economics*, 127(3), pp.1057–1106.

Fratiglioni, L., Paillard-Borg, S. & Winblad, B., 2004. An active and socially integrated lifestyle in late life might protect against dementia. *The Lancet Neurology*, 3(6), pp.343–353.

Gardeazabal, J. & Ugidos, A., 2004. More on Identification in Detailed Wage Decompositions. *Review of Economics and Statistics*, 86(4), pp.1034–1036.

Geruso, M. & Layton, T., 2015. Upcoding: Evidence from Medicare on Squishy Risk Adjustment. *NBER Working Paper*, (21222).

Glass, T.A. et al., 2006. Social Engagement and Depressive Symptoms in Late Life: Longitudinal Findings. *Journal of Aging and Health*, 18(4), pp.604–628.

Goldman, D.P. & Smith, J.P., 2002. Can patient self-management help explain the SES health

gradient? *Proceedings of the National Academy of Sciences of the United States of America*, 99(16), pp.10929–10934.

Grant, D., 2009. Physician financial incentives and cesarean delivery: New conclusions from the healthcare cost and utilization project. *Journal of Health Economics*, 28(1), pp.244–250.

Greenhalgh, S., 1994. Controlling Births and Bodies in Village China. *American Ethnologist*, 21(1), pp.3–30.

Greenhalgh, S., 1985. Sexual Stratification: The Other Side of "Growth with Equity" in East Asia. *Population and Development Review*, 11(2), p.265.

Greenhalgh, S., 1986. Shifts in China's Population Poilicy, 1984-86: Views from the Central, Provincial and Local Levels. *Population and Development Review*, 12(3), pp.491–515.

Greenhalgh, S., 1990. The Evolution of the One-child Policy in Shaanxi, 1979-88. *The China Quarterly*, 122(122), pp.191–229.

Grol-Prokopczyk, H., Freese, J. & Hauser, R.M., 2011. Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior*, 52(2), pp.246–261.

Gruber, J., Hendren, N. & Townsend, R.M., 2014. The great equalizer: Health care access and infant mortality in Thailand. *American Economic Journal: Applied Economics*, 6(1), pp.91–107.

Ham, J.C., Svejnar, J. & Terrell, K., 1998. from the Czech and Slovak Republics Unemployment and the Social Safety Net During Transitions to a Market Economy : Evidence from the Czech and Slovak Republics. *American Economic Review*, pp.1117–1142.

Hardee-Cleaveland, K. & Banister, J., 1988. Fertility policy and implementation in China, 1986-88. *Population and Development Review*, 14(2), pp.245–86.

He, W., Muenchrath, M.N. & Kowal, P., 2012. *Shades of Gray : A Cross-Country Study of Health and Well-Being of the Older Populations in SAGE Countries , 2007 – 2010*, Washington, DC: U.S. Census Bureau.

Herd, P., Goesling, B. & House, J.S., 2007. Socioeconomic position and health: the differential effects of education versus income on the onset versus progression of health problems. *Journal of Health and Social Behavior*, 48(3), pp.223–238.

Hesketh, T. & Zhu, W.X., 1997. Health in china: From Mao to market reform. *British Medical Journal*, 314(7093), pp.1543–1543.

Holmstrom, B. & Milgrom, P., 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, 7, pp.24–52.

Hou, Z. et al., 2014. Effects of NCMS on access to care and financial protection in China. *Health Economics*, 23(8), pp.917–934.

Imai, K., King, G. & Stuart, E.A., 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2), pp.481–502.

Islam, A. & Smyth, R., 2015. Do Fertility Control Policies Affect Health in Old Age? Evidence from China's One-Child Experiment. *Health Economics*, 24(5), pp.601–616.

Jacobson, M. et al., 2010. How medicare's payment cuts for cancer chemotherapy drugs changed patterns of treatment. *Health Affairs*, 29(7), pp.1391–1399.

Jian, W. et al., 2015. Payment reform pilot in Beijing hospitals reduced expenditures and out-of-pocket payments per admission. *Health Affairs*, 34(10), pp.1745–1752.

Joshi, S. & Schultz, T.P., 2013. *Family Planning and Women's and Children's Health: Long-Term Consequences of an Outreach Program in Matlab, Bangladesh*,

Kapteyn, A., Smith, J.P. & Van Soest, A., 2007. Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, 97(1), pp.461–473.

King, G. et al., 2004. Enchancing the Validty and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, 98(1), pp.191–207.

King, G. et al., 2009. Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme. *The Lancet*, 373(9673), pp.1447–1454.

Knight, J., Song, L. & Gunatilaka, R., 2009. Subjective well-being and its determinants in rural China. *China Economic Review*, 20(4), pp.635–649.

Kristensen, N. & Johansson, E., 2008. New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics*, 15(1), pp.96–117.

Kruk, K.E. & Reinhold, S., 2014. The effect of children on depression in old age. *Social Science & Medicine*, 100, pp.1–11.

Langenbrunner, J., Cashin, C. & O'Dougherty, S., 2009. *Designing and Implementing Health Care Provider Payment Systems* J. C. Langenbrunner, S. O'Duagherty, & C. S. Cashin, eds., The World Bank.

Lei, L., Chen, Y. & Xiong, X., 1993. *Zhongguo nvzi jiaoyushi (The history of female education in China)*, Wuhan: Guizhou Education Press.

Lei, X. et al., 2014. Depressive symptoms and SES among the mid-aged and elderly in China: Evidence from the China Health and Retirement Longitudinal Study national baseline. *Social Science and Medicine*, 120, pp.224–232.

Li, D. et al., 2014. A meta-analysis of the prevalence of depressive symptoms in chinese older adults. *Archives of Gerontology and Geriatrics*, 58(1), pp.1–9.

Li, H., Yi, J. & Zhang, J., 2015. *Fertility, Household Structure, and Parental Labor Supply: Evidence from Rural China*,

Li, H. & Zhang, J., 2007. Do High Birth Rates Hamper Economic Growth? *Review of Economics and Statistics*, 89(1), pp.110–117.

Li, H. & Zhang, J., 2009. Testing the External Effect of Household Behavior: The Case of the Demand for Children. *Journal of Human Resources*, 44(4), pp.890–915.

Li, H., Zhang, J. & Zhu, Y., 2008. The quantity-quality trade-off of children in a developing country: identification using Chinese twins. *Demography*, 45(1), pp.223–243.

Limwattananon, S. et al., 2015. Universal coverage with supply-side reform: The impact on medical expenditure risk and utilization in Thailand. *Journal of Public Economics*, 121, pp.79–94.

Liu, G.G. et al., 2014. Chinese time trade-off values for EQ-5D health states. *Value in Health*, 17(5), pp.597–604.

Lumbiganon, P. et al., 2010. Method of delivery and pregnancy outcomes in Asia: the WHO global survey on maternal and perinatal health 2007–08. *The Lancet*, 375(9713), pp.490–499.

Ma, X. et al., 2011. Social health assistance schemes: the case of Medical Financial Assistance for the rural poor in four counties of China. *International Journal for Equity in Health*, 10(1), p.44.

MacIntyre, S., Ford, G. & Hunt, K., 1999. Do women "over-report" morbidity? Men's and women's responses to structured prompting on a standard question on long standing illness. *Social Science & Medicine*, 48(1), pp.89–98.

Malmusi, D. et al., 2012. Perception or real illness? How chronic conditions contribute to gender inequalities in self-rated health. *European Journal of Public Health*, 22(6), pp.781–786.

Mansuri, G. & Rao, V., 2012. *Localizing development: does participation work?*, World Bank Publications.

Mathauer, I. & Wittenbecher, F., 2013. Hospital payment systems based on diagnosis-related groups: experiences in low- and middle-income countries. *Bulletin of the World Health Organization*, 91(10), p.746–756A.

McGuire, T.G., 2000. Physician agency. In *Handbook of Health Economics*. pp. 461–536.

McGuire, T.G. & Pauly, M. V., 1991. Physician response to fee changes with multiple payers. *Journal of Health Economics*, 10(4), pp.385–410.

Meng, Q., 2005. *Review of health care provider payment reforms in China*, Washington, DC.

Meng, Q. et al., 2012. Trends in access to health services and financial protection in China between 2003 and 2011: A cross-sectional study. *The Lancet*, 379(9818), pp.805–814.

Miller, G. & Babiarz, K.S., 2013. Pay-for-performance incentives in low- and middle-income country health programs. In *NBER Working Paper Series*. San Diego: Elsevier, p. 18932–n/a.

Ministry of Finance, 2009. *Finance Yearbook of China*, Beijing: China Finance Press.

Ministry of Health, 2004. *China Health Yearbook 2004*, Beijing: People's Medical Publishing Company.

Ministry of Human Resources and Social Security, 2015. *China Social Security Development Annual Report 2014*, Beijing: China Labour & Social Security Publishing House.

Murray, C.J. et al., 2003. Empirical evaluation of the anchoring vignettes approach in health surveys. In C. J. L. Murray & D. B. Evans, eds. *Health systems performance assessment: Debates, methods and empiricism*. Ginebra: World Health Organization, pp. 369–399.

Nathanson, C.A., 1975. Illness and the feminine role: A theoretical review. *Social Science & Medicine*, 9(2), pp.57–62.

National Bureau of Statistics, 2011a. *China statistical yearbook 2011*, Beijing: China Statistics Press.

National Bureau of Statistics, 2011b. *China Statistical Yearbooks for Regional Economy 2011*, Beijing: China Statistics Press.

Nie, H. et al., 2011. The prevalence of mild cognitive impairment about elderly population in China: A meta-analysis. *International Journal of Geriatric Psychiatry*, 26(6), pp.558–563.

O'Dougherty, S. et al., 2009. Case-Based Hospital Payment Systems. In J. C. Langenbrunner, C. Cashin, & S. O'Dougherty, eds. *Designing and Implementing Health Care Provider Payment Systems*. Washington, DC: The World Bank, pp. 125–213.

Olken, B.A., Onishi, J. & Wong, S., 2014. Should aid reward performance? Evidence from a field experiment on health and education in Indonesia. *American Economic Journal: Applied Economics*, 6(4), pp.1–34.

Parish, W.L. & Willis, R.J., 1993. Daughters , Education , and Family Budgets Taiwan Experiences. *Journal of Human Resources*, 28(4), pp.863–898.

Peracchi, F. & Rossetti, C., 2012. Heterogeneity in health responses and anchoring vignettes. *Empirical Economics*, 42(2), pp.513–538.

Powell-Jackson, T., Yip, W.C.-M. & Han, W., 2015. Realigning demand and supply side incentives to improve primary health care seeking in rural China. *Health Economics*, 24(6), pp.755–772.

Pritchett, L. & Sandefur, J., 2015. Learning from experiments when context matters. *American Economic Review*, 105(5), pp.471–475.

Project Management Office of Henan Health Bureau, 2009. *Henan Health XI proposal*, Zhengzhou.

Pylypchuk, Y. & Selden, T.M., 2008. Discrete choice decomposition analysis of racial and ethnic differences in children's health insurance coverage. *Journal of Health Economics*, 27(4), pp.1109–1128.

Qian, N., 2009. *Quantity-Quality and the One Child Policy:The Only-Child Disadvantage in School Enrollment in Rural China*, Cambridge, MA.

Radloff, L.S., 1977. The CES-D scale a self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3), pp.385–401.

Ran, L.M. et al., 2013. An analysis of china's physician salary payment system. *Journal of Huazhong University of Science and Technology - Medical Science*, 33(2), pp.309–314.

Rao, J.K., Weinberger, M. & Kroenke, K., 2000. Visit-specific expectations and patient-centered outcomes: a literature review. *Archives of Family Medicine*, 9(10), p.1148.

Rice, N., Robone, S. & Smith, P., 2011. Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. *European Journal of Health Economics*, 12(2), pp.141–162.

Rosenzweig, M.R. & Zhang, J., 2009. Do Population Control Policies Induce More Human Capital Investment? Twins, Birth Weight and China's "One-Child" Policy. *Review of Economic Studies*, 76(3), pp.1149–1174.

Ross, C.E., Masters, R.K. & Hummer, R.A., 2012. Education and the Gender Gaps in Health and Mortality. *Demography*, 49(4), pp.1157–1183.

Ross, C.E. & Mirowsky, J., 2010. Gender and the Health Benefits. *The Sociological Quarterly*, 51(1), pp.1–19.

Schultz, T.P., 2007. Population Policies, Fertility, Women's Human Capital, and Child Quality. *Handbook of Development Economics*, 4(7), pp.3249–3303.

Shigeoka, H. & Fushimi, K., 2014. Supplier-induced demand for newborn treatment: Evidence from Japan. *Journal of Health Economics*, 35(1), pp.162–178.

Short, S.E. & Fengying, Z., 1998. Looking Locally at China's One-Child Policy. *Studies in Family Planning*, 29(4), p.373.

Van Soest, A. et al., 2011. Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 174(3), pp.575–595.

Spiers, N. et al., 2003. Are gender differences in the relationship between self-rated health and mortality enduring? Results from three birth cohorts in Melton Mowbray, United Kingdom. *The Gerontologist*, 43(3), pp.406-411–375.

Suits, D.B., 1984. Dummy variables: Mechanics v. interpretation. *The Review of Economics and*

Statistics, pp.177–180.

Sun, S. et al., 2011. Population health status in China: EQ-5D results, by age, sex and socio-economic status, from the national health services survey 2008. *Quality of Life Research*, 20(3), pp.309–320.

Sylvia, S. et al., 2015. Survey using incognito standardized patients shows poor quality care in China's rural clinics. *Health Policy and Planning*, 30(3), pp.322–333.

Terza, J. V, 1985. Ordinal probit: a generalization. *Communications in Statistics: Method and Theory*, 14(1), p.1-11-.

The EuroQol Group, 1990. A new facility for the measurement of health-related quality of life. *Health Policy*, 16(3), pp.199–208.

Tien, H.Y., 1980. Wan, xi, shao: How China meets its population problem. *International Family Planning Perspectives1*, 6(2), pp.65–70.

United Nations Population Division, 2015. *World Population Prospects: The 2015 Revision, Volume I: Comprehensive Tables*, New York.

Verbrugge, L.M., 1982. Sex differentials in health. *Public Health Reports*, 97(5), pp.417–437.

Verbrugge, L.M., 1989. The twain meet: empirical explanations of sex differences in health and mortality. *Journal of Health and Social Behavior*, 30(3), pp.282–304.

Vermeer, N., Mastrogiacomo, M. & Van Soest, A., 2016. Demanding occupations and the retirement age. *Labour Economics*, 43(9462), pp.159–170.

Vivalt, E., 2015. Heterogeneous treatment effects in impact evaluation. *American Economic Review*, 105(5), pp.467–470.

Wagstaff, A. et al., 2009. *Reforming China's rural health system*, Washington, DC: The World Bank.

Wagstaff, A. & Lindelow, M., 2008. Can insurance increase financial risk?. The curious case of health insurance in China. *Journal of Health Economics*, 27(4), pp.990–1005.

Wagstaff, A. & Yu, S., 2007. Do health sector reforms have their intended impacts?. The World Bank's Health VIII project in Gansu province, China. *Journal of Health Economics*, 26(3), pp.505–535.

Wang, F., 2012. *Family Planning Policy in China: Measurement and Impact on Fertility*,

Wang, J. et al., 2014. Use and prescription of antibiotics in primary health care settings in China. *JAMA Internal Medicine*, 174(12), p.1914.

Westfall, P. & Young, S., 1993. *Examples and Methods for p-Value Adjustment*, John Wiley & Sons.

Whyte, M.K. & Gu, S.Z., 1987. Popular Response to China's Fertility Transition. *Population and Development Review*, 13(3), p.471.

World Bank, 2015a. *China - Rural Health Project*, Washington, DC.

World Bank, 2016. *Healthy China: deepening health reform in China building high-quality and value-based service delivery*, Washington, DC.

World Bank, 2015b. *Program for results: two year review*, Washington, DC.

Yang, G. et al., 2008. Emergence of chronic non-communicable diseases in China. *The Lancet*, 372(9650), pp.1697–1705.

Yi, Z. et al., 1993. Causes and Implications of the Recent Increase in the Reported Sex Ratio at Birth in China. *Population and Development Review*, 19(2), pp.283–302.

Yip, W. & Hsiao, W.C., 2009. Non-evidence-based policy: How effective is China's new cooperative medical scheme in reducing medical impoverishment? *Social Science &*

Medicine, 68(2), pp.201–209.

Yip, W.C., 1998. Physician response to Medicare fee reductions: Changes in the volume of coronary artery bypass graft (CABG)surgeries in the Medicare and private sectors. *Journal of Health Economics*, 17(6), pp.675–699.

Yip, W.C.M. et al., 2010. Realignment of incentives for health-care providers in China. *The Lancet*, 375(9720), pp.1120–1130.

Yun, M.S., 2005. A simple solution to the identification problem in detailed wage decompositions. *Economic Inquiry*, 43(4), pp.766–772.

Yun, M.S., 2004. Decomposing differences in the first moment. *Economics Letters*, 82(2), pp.275–280.

Zeng, J., Zeng, X.X. & Tu, Q., 2013. A gloomy future for medical students in China. *The Lancet*, 382(9908), p.1878.

Zeng, Y. & Hesketh, T., 2016. The effects of China's universal two-child policy. *The Lancet*, 388(10054), pp.1930–1938.

Zhang, Z. et al., 2014. *Zhongguo nongcun weisheng fazhan xiangmu chuangxin anliji (China Rural Health Development Project case study)*, Beijing: Peking Union Medical College Press.

Zimmer, Z. et al., 1998. Educational attainment and transitions in functional status among older Taiwanese. *Demography*, 35(3), pp.361–375.

Zimmer, Z. et al., 2014. Examining late-life functional limitation trajectories and their associations with underlying onset, recovery, and mortality. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, 69(2), pp.275–286.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

645 B.M. SADABA, *Essays on the Empirics of International Financial Markets*
646 H. KOC, *Essays on Preventive Care and Health Behaviors*
647 V.V.M. MISHEVA, *The Long Run Effects of a Bad Start*
648 W. LI, *Essays on Empirical Monetary Policy*
649 J.P. HUANG, *Topics on Social and Economic Networks*
650 K.A. RYSZKA, *Resource Extraction and the Green Paradox: Accounting for Political Economy Issues and Climate Policies in a Heterogeneous World*
651 J.R. ZWEERINK, *Retirement Decisions, Job Loss and Mortality*
652 M. K. KAGAN, *Issues in Climate Change Economics: Uncertainty, Renewable Energy Innovation and Fossil Fuel Scarcity*
653 T.V. WANG, *The Rich Domain of Decision Making Explored: The Non-Triviality of the Choosing Process*
654 D.A.R. BONAM, *The Curse of Sovereign Debt and Implications for Fiscal Policy*
655 Z. SHARIF, *Essays on Strategic Communication*
656 B. RAVESTEIJN, *Measuring the Impact of Public Policies on Socioeconomic Disparities in Health*
657 M. KOUDSTAAL, *Common Wisdom versus Facts; How Entrepreneurs Differ in Their Behavioral Traits from Others*
658 N. PETER, *Essays in Empirical Microeconomics*
659 Z. WANG, *People on the Move: Barriers of Culture, Networks, and Language*
660 Z. HUANG, *Decision Making under Uncertainty-An Investigation from Economic and Psychological Perspective*
661 J. CIZEL, *Essays in Credit Risk, Banking, and Financial Regulation*
662 I. MIKOLAJUN, *Empirical Essays in International Economics*
663 J. BAKENS, *Economic Impacts of Immigrants and Ethnic Diversity on Cities*
664 I. BARRA, *Bayesian Analysis of Latent Variable Models in Finance*
665 S. OZTURK, *Price Discovery and Liquidity in the High Frequency World*
666 J. JI, *Three Essays in Empirical Finance*
667 H. SCHMITTDIEL, *Paid to Quit, Cheat, and Confess*
668 A. DIMITROPOULOS, *Low Emission Vehicles: Consumer Demand and Fiscal Policy*
669 G.H. VAN HEUVELEN, *Export Prices, Trade Dynamics and Economic Development*
670 A. RUSECKAITE, *New Flexible Models and Design Construction Algorithms for Mixtures and Binary Dependent Variables*

671 Y. LIU, *Time-varying Correlation and Common Structures in Volatility*

672 S. HE, *Cooperation, Coordination and Competition: Theory and Experiment*

673 C.G.F. VAN DER KWAAK, *The Macroeconomics of Banking*

674 D.H.J. CHEN, *Essays on Collective Funded Pension Schemes*

675 F.J.T. SNIEKERS, *On the Functioning of Markets with Frictions*

676 F. GOMEZ MARTINEZ, *Essays in Experimental Industrial Organization: How Information and Communication affect Market Outcomes*

677 J.A. ATTEY, *Causes and Macroeconomic Consequences of Time Variations in Wage Indexation*

678 T. BOOT, *Macroeconomic Forecasting under Regime Switching, Structural Breaks and High-dimensional Data*

679 I. TIKOUDIS, *Urban Second-best Road Pricing: Spatial General Equilibrium Perspectives*

680 F.A. FELSŐ, *Empirical Studies of Consumer and Government Purchase Decisions*

681 Y. GAO, *Stability and Adaptivity: Preferences over Time and under Risk*

682 M.J. ZAMOJSKI, *Panta Rhei, Measurement and Discovery of Change in Financial Markets*

683 P.R. DENDERSKI, *Essays on Information and Heterogeneity in Macroeconomics*

684 U. TURMUNKH, *Ambiguity in Social Dilemmas*

685 U. KESKIN, *Essays on Decision Making: Intertemporal Choice and Uncertainty*

686 M. LAMMERS, *Financial Incentives and Job Choice*

687 Z. ZHANG, *Topics in Forecasting Macroeconomic Time Series*

688 X. XIAO, *Options and Higher Order Risk Premiums*

689 D.C. SMERDON, *'Everybody's doing it': Essays on Trust, Norms and Integration*

690 S. SINGH, *Three Essays on the Insurance of Income Risk and Monetary Policy*

691 E. SILDE, *The Econometrics of Financial Comovement*

692 G. DE OLIVEIRA, *Coercion and Integration*

693 S. CHAN, *Wake Me up before you CoCo: Implications of Contingent Convertible Capital for Financial Regulation*

694 P. Gal, *Essays on the role of frictions for firms, sectors and the macroeconomy*

695 Z. FAN, *Essays on International Portfolio Choice and Asset Pricing under Financial Contagion*