Essays on Healthcare Utilisation, Health and Labour Supply in China

Wenkun Zhang

PhD

University of York

Department of Economics and Related Studies

January, 2024

Abstract

This thesis consists of three essays on the topic of healthcare utilisation, health and labour supply in China. Chapter 1 investigates the effects of health on the labour supply of middle-aged and elderly couples in rural China by considering the interaction of own health and spousal health. The health of a couple is introduced into the general intra-household collective model to establish the theoretical framework. Identification relies on Bivariate probit model and seemingly unrelated regressions (SUR) to account for the potential correlation of the error terms in the husband's and wife's labour supply equations in the household. Findings reveal that own health improvements have a significant and positive effect on labour participation and annual work time. Better spousal health amplifies the positive own health effect on labour participation, while the positive effect of own health on annual work time is reduced by better spousal health. Spousal health impact on labour participation is negative, better own health could lessen this negative effect and even make it positive. However, spousal health has no obvious effect on annual work time.

Chapter 2 evaluates the effect of the Hierarchical Medical System (HMS) reform on the capacity and utilisation of primary healthcare institutions. Panel data (2012 – 2017) from the China Health Statistic Book is used. The HMS reform aimed to establish a two-way referral system and relieved the current pressure on second and third-tier hospitals by strengthening the capabilities of primary healthcare institutions. The reform was initiated in September 2015 and had been gradually implemented throughout the whole country by the end of 2017. Findings show that the HMS reform is effective in enhancing the proportion of nurses and practitioners in urban primary healthcare institutions but ineffective in rural areas. In addition, the HMS reform significantly increases outpatient visits, inpatient admissions and the bed occupancy rate in rural primary healthcare institutions but the impact is insignificant in urban areas.

Chapter 3 investigates the impact of the Urban-Rural Residents Basic Medical Insurance (URRBMI) integration on healthcare utilisation in rural China. The data were derived from the China Health and Retirement Longitudinal Study (CHARLS) from 2011 (wave 1) to 2018 (wave 4). Healthcare disparities continue to coexist with universal health coverage in China due to the fragmentation of social medical insurance. The Chinese government launched the URRBMI reform to establish a unified medical insurance scheme for non-working urban and rural residents in 2016. The reform was implemented gradually throughout the whole country by 2020 and is recognized as a vital step to safeguard equal healthcare and benefit to each enrollee in China. Findings reveal that the integration reform significantly increases the middle-aged and older rural population's inpatient care utilisation, including the probability and the number of nights hospitalized during the most recent hospital stay. However, the integration has limited impacts on the middle-aged and older rural population's outpatient healthcare usage.

This thesis is dedicated to my parents

Contents

List of Tables

Tables in Chapter 1

Tables in Chapter 2

Tables in Chapter 3

Tables in Appendix A

Tables in Appendix B

Tables in Appendix C

List of Figures

Figures in Chapter 1

Figures in Chapter 2

Figures in Chapter 3

Figures in Appendix A

Figures in Appendix C

Acknowledgements

Firstly, I would like to thank my supervisors Professor Andrew M. Jones and Professor Nigel Rice. I am really grateful for their kind support during my PhD journey. Many thanks for their patience in sharing their invaluable knowledge and experience, giving detailed feedback at every stage and encouraging me to think deeply. Their diligence and passion inspire me a lot.

I also wish to thank Professor Emma Tominey, my thesis advisory panel member. This thesis benefited greatly from her thoughtful comments and suggestions. I also want to thank everyone in the HEDG cluster for lots of useful discussions. A big thank you also to Dr Michael Shallcross, Maigen Savory and Vin Mc Dermott for their constant support.

A special thank you goes to my lovely friends, who made this journey such a wonderful and warm experience. I would be much closer to feel isolated during the lock-down periods without you. The great food and drinks, and numerous pleasant talks were super energy supply for me.

I delicate this thesis to my mother Liying Yan and my father Weijiu Zhang, for their unconditional love these years. I won't be where I am without their support. I haven't come back to China since September 2019 due to the pandemic and health issues. Many thanks for my parents to be my team and reminding me the precious and beautiful things I own even at the bad times. It was the strength of my family that helped me go through this PhD journey although there were many unpredictable circumstances.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author of the three chapters. This work has not previously been presented for an award at this or other University. All sources are acknowledged as References. I have orally presented earlier versions of the three papers at different seminars at the Department of Economics and Related Studies at the University of York. Chapter 1 has been presented at the 8th EuHEA PhD & Supervisor Conference, Erasmus University Rotterdam, and the 6th York Workshop on Labour and Family Economics, University of York. Chapter 2 has been presented at the 9th EuHEA PhD & Supervisor Conference, National University of Ireland. Chapter 3 has been presented at the 10th EuHEA PhD & Supervisor Conference, University of Bologna.

Wenkun Zhang

York, January 2024

Introduction

Increased longevity is one of the most remarkable success stories in human history. However, coupled with decreased fertility rates, the rapid and accelerating pace of population ageing raises challenges for high-income countries as well as in low- and middle-income countries (Mitra et al., 2020). Typical concerns range from macroeconomic slowdowns to heightened financial strains on pensions, healthcare, and other social-protection systems (Bloom et al., 2015). China, home to over one-sixth of the world's population, has witnessed an escalating trend towards an ageing society since 2000. The dramatic demographic transitions not only increase pressure on the sustainability of labour supply but also pose challenges for the healthcare system.

Population ageing poses a significant challenge for the labour market by reducing the number of workers relative to retirees. In a country dominated by a pay-as-you-go social pension scheme, a worsening support ratio forces the government to either raise taxes on workers or risk snowballing deficits, threatening macroeconomic stability (Hou et al., 2021). One effective response to population ageing in high-income countries is adjusting retirement policies, such as postponing the retirement age (Bloom et al., 2015). The Chinese government has already studied developed countries' experiences and considered gradually delaying state retirement age (Zhang et al., 2023). However, the specific plan for raising the retirement age has not been officially released. There are two main obstacles to ensuring the effectiveness of delaying the retirement age. On the one hand, whether or not retirement can be postponed critically depends on the elderly's health status (Smith et al., 2014). One the other hand, China has two distinct labour markets: the formal labour market (mainly in urban areas) and the informal labour market (mainly in rural areas). The statutory retirement age in the formal sector is 60 years old for men, 55 years old for women cadres (professionals), and 50 years old for women workers (Che and Li, 2018). Employees and retirees can expect to receive a pension upon retirement if their employers contribute to cover the basic pension as required (Mitra et al., 2020). However, retirement is an alien concept for most Chinese elderly in rural areas as the majority of the rural elderly predominantly work on family farming or raising livestock. They generally continue working as long as they are physically able to (Smith et al., 2014). Additionally, rural residents did not have a national pension scheme until 2009, when both men and women aged 60 and above became eligible to obtain limited pensions. The rural pension scheme is not employment-based, and the amount of pension income is well below the subsistence level. Therefore, it provides little disincentive to work past 60-year-old (Hou et al., 2021). Despite the massive disparities in urban and rural labour markets, it is clear that healthy ageing has the vital function of stimulating the potential capacity of the elderly to work and ensuring the labour market's sustainable operation in both formal and informal sectors.

In rural China, most residents work in the informal sector and a high proportion engage in agricultural production. Agriculture as a typical labour-intensive industry is one of the sectors substantially affected by population ageing. In the last few decades, coexisting with an ageing workforce, a growing number of young labourers in rural areas have migrated to urban areas for off-farm work, which has accelerated the shortage of labour in rural areas. To cope with the agricultural labour shortage and to protect food security, it is critical to incentivise middle-aged and elderly rural residents to provide effective labour supply in the long-term. Tapping into the older workforce can also lessen the economic burden borne by family and society. Health is regarded as a special kind of human resource (Grossman, 1972) and has been suggested as one of the most important factors affecting labour supply. For older people engaged in physically strenuous work (mostly farming activities) in rural China, their labour supply decisions might be more dependent on their health. Chapter 1 explores the effects of health on middle-aged and elderly couples' labour supply in rural China. The main contribution of this chapter is considering the interaction of own health and spousal health and conducting a comprehensive analysis of health impacts on labour supply decision-making at the household level. We make use of the general intra-household collective model to establish a theoretical framework. Health is measured by a latent stock to avoid potential reporting bias of subjective health measures, which is instead constructed by a set of objective health indicators. Bivariate probit and SUR models are employed to account for the potential correlation of error terms in couples' labour supply equations. We find interesting relationships between the impact of own health and spousal health. Improvements in own health are effective in enhancing the probability of labour participation and annual work time. Importantly, better spousal health amplifies this positive effect. In contrast, better spousal health is associated with lower labour participation probability. Better own health appears to lessen the negative effect of spousal health and even makes the spousal impact on labour participation positive. However, spousal health has a limited effect on annual work time.

The rapidly growing ageing population has posed formidable challenges for the healthcare system with soaring healthcare need. The outbreak of severe acute respiratory syndrome (SARS) in China in 2003 further challenged the healthcare system. The public complained about access to and affordability of health care as most people had no health insurance and encountered high out-of-pocket payments for health care. To respond to the increasing complaints, the government instigated a new round of systemic health reforms in 2009 to establish an equitable and effective healthcare system for all people by 2020. This complex and large-scale reform covered five targeted priorities: strengthening primary care; expanding basic health insurance; establishing an essential medicines programme; providing equitable access to public health care services; and undertaking a reform of public hospitals (Li and Fu, 2017). This systemic health reform has made admirable progress and attracted international attention (Yip et al., 2019).

Ageing is the main contributor to a broad spectrum of chronic disorders, all of which are associated with a lower quality of life in the elderly (Fang et al., 2020). Orienting a healthcare system towards primary care can enhance the continuity and coordination of care (Garrido et al., 2011) and particularly help ensure effective management of chronic non-communicable diseases. As a first reform priority, the government made efforts to strengthen the primary care system and encourage the utilisation of primary healthcare facilities. Enormous resources have been invested in primary care institutions with a particular focus on improving infrastructure. China has a three-tier health care system that consists of primary care facilities, secondary hospitals and tertiary hospitals. The primary care system can be further divided into urban and rural components, which are organised differently. There was previously no gate-keeping system in China. People had the freedom to access any kind of healthcare facility. Because of an imbalance in the allocation of medical resources among different levels of hospitals, evidence indicates that patients tended to bypass primary healthcare facilities and access higher-level hospitals when they sought medical care services. To alleviate the imbalanced distribution of medical resources and divert patient flows to primary care facilities, the government launched the Hierarchical Medical System (HMS) reform in 2015. HMS's aim is that patients should be treated at different levels of hospitals according to specific conditions. First contact in primary care facilities is encouraged, and a two-way referral system is established. Unlike the mandatory gatekeeper system in the United Kingdom and Germany, China's HMS encourages patients to access primary care by increasing subsidies for infrastructure and workforce resources to enhance primary care facilities' capacity and quality. Chapter 2 investigates the effect of the HMS reform on primary healthcare institutions' capacity and utilisation. We choose the proportion of nurses and practitioners to total staff in primary healthcare institutions as capacity indicators. The bed occupancy rate in primary care institutions and the proportion of outpatient/inpatient care provision by primary care facilities are used as utilisation indicators. Findings reveal that the HMS reform was effective in enhancing the capacity of urban primary healthcare institutions but ineffective in rural areas. Additionally, the HMS reform significantly increased the use of rural primary healthcare institutions, but its impact was not significant in urban areas.

Secondly, the move towards universal health insurance coverage ranks among one of the most impressive achievements of the 2009 reform. The share of the population covered by basic social health insurance increased from 15% in 2000 to 85% in 2008, to more than 97% in 2015 (Li and Fu, 2017), providing valuable experience of covering the population in the informal sector. Meanwhile, the generosity of health insurance schemes has been vastly improved. However, the inequality derived from the fragmentation of social insurance remains persistent (Huang and Wu, 2020). The Chinese social health insurance system mainly consists of the New Rural Cooperative Medical System (NRCMS, launched in 2003) for the rural population, the Urban Employee Basic Medical Insurance (UEBMI, launched in 1998) covering workers in the formal sector in urban areas, and the Urban Resident Basic Medical Insurance (URBMI, launched in 2007) targeting urban residents except employees. Segmentation by urban-rural and employment status involves different benefits packages. UEBMI is mandatory and has the most generous benefits, whereas NRCMS was the least generous with a relatively low reimbursement rate and limited service coverage. This led to the aggravation of disparities in healthcare utilisation and health among populations covered by different health insurance schemes (Zhou et al., 2021). The government introduced an integrated social health insurance system named "Urban-Rural Residents Basic Medical Insurance (URRBMI)" to improve equity by consolidating the two voluntary subsidised schemes: NRCMS and URBMI in 2006. After the integration, rural residents received higher levels of reimbursement for drugs and services, and enjoyed a greater choice of service items and facilities. Chapter 3 evaluates the impact of the URRBMI integration reform on healthcare utilisation in rural China. The main findings show that the reform was effective in enhancing the middle-aged and older rural population's inpatient care utilisation, including the probability and the duration of hospital stays. This is consistent with previous literature investigating the effect of the integration reform across pilot areas (Huang and Wu, 2020). However, the integration impact on middle-aged and older rural residents' outpatient healthcare utilisation was found to be insignificant. This result aligns with those of early studies, which provide evidence that the reform mainly improves inpatient benefits and has limited impacts on outpatient benefits (Su et al., 2019).

Chapters 2 and 3 focus on policy evaluations in the Chinese healthcare system. One of the most important contributions of these two chapters is providing national evidence of the HMS and URRBMI reforms. The existing literature mainly explores policy impacts in pilot areas or are restricted to specific subpopulations. The second significant contribution of the two chapters is our identification strategy for the average treatment effect on the treated (ATT). In most empirical difference-in-difference (DID) applications with multiple periods, researchers usually estimate the ATT using a two-way-fixed-effect (TWFE) linear regression, which includes dummy variables for cross-sectional units and periods with a treatment dummy. Some recent research employs different decomposition methods to show that TWFE regressions may not identify easy-to-interpret estimated coefficients when there is treatment effect heterogeneity (De Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021). The HMS reform evaluated in Chapter 2 was initiated in September 2015 and then gradually implemented throughout the country by the end of 2017. The URRBMI integration reform was launched in 2016 and was gradually implemented nationwide by 2020. The feature of staggered implementation of reforms makes treatment effect heterogeneity most likely. This means that the standard TWFEDID estimator with the underlying assumption of constant treatment effects might lead to biases. Therefore, our identification follows recent methodological extensions for staggered interventions to allow arbitrary treatment heterogeneity. We focus on the ATT for each specific cohort at each specific post-treatment period. A cohort is defined by the period when units are first treated. Specifically, we follow Wooldridge (2021) to implement the extended TWFE estimator and Callaway and Sant' Anna (2021) to implement a doubly robust estimator. In both chapters, we observe clear treatment heterogeneity across cohorts and periods, justifying our choice of approach.

Overall, as an important part of the social security system to cope with population ageing, the complex reforms in the healthcare system have had mixed success. China is on the right trajectory to build a more accessible, affordable, and efficient healthcare system. Its impressive and long-term endeavour in promoting healthcare utilisation is associated with population health improvements. Better population health is highly related to work capacity and is effective in incentivising middle-aged and elderly people to continue to participate in the job market and provide productive labour supply.

1. Chapter 1

Household Health and Labour Supply: Evidence from Rural China

Abstract

This paper uses data from the China Health and Retirement Longitudinal Study (CHARLS) (Wave1-3). We provide a comprehensive analysis of the effects of health on the labour supply of middle-aged and elderly couples in rural China by considering the interaction of own health and spousal health. In our approach, health is introduced into the general intra-household collective model to establish the theoretical framework. Next, a latent health stock index is constructed to eliminate measurement bias and we use one-period lagged latent health stock to deal with the simultaneous causality of health and labour supply. To account for the potential correlation of the error terms in the husband's and wife's labour supply equations in the household, we estimate the labour supply equations of the couples systematically by bivariate probit model and seemingly unrelated regressions(SUR). The main findings reveal that there is an obvious link between the effects of own health and spousal health on labour supply. Better own health has a positive and significant impact on labour participation and this positive impact increases by better spousal health. In contrast, better spousal health has a significant and negative effect on labour participation, and the magnitude of this negative effect decreases by better own health. Moreover, the effect of spousal health on labour participation could even become positive with relatively good own health. Conditional on engaging in the labour market, the impact of own health improvements on annual work time is positive and significant. Better spousal health decreases this positive own health impact. However, spousal health has a limited impact on annual work time.

Keywords

Intra-household labour supply; Latent health stock; Elderly couples in rural China; CHARLS; Interaction term

1.1. Introduction

China became the first country to enter an ageing society at the low-income stage in 2000 and has witnessed a rapid ageing pace in recent decades. The Chinese population constituted 18% of the world population in 2020, with 190.6 million residents aged 65 and above (Fang et al., 2020). In rural areas, the population is ageing faster than in urban areas. The percentage of people over 65 years old was 17.72% in rural areas, whereas it was 11.11% in urban areas (Shen et al., 2023). Such a massive ageing population has placed a pronounced economic burden on society and created labour shortages, especially in labour-intensive economic sectors (Bloom et al., 2015). The Chinese labour market is distinctive in the institutional segregation of urban and rural registered residents, which is further magnified in retirement (Giles et al., 2023). The formal sector, which covers the majority of urban workers, can expect to receive relatively generous social pensions payable upon reaching statutory retirement age if the employers contribute to cover the basic pension as required (Zhao and Zhao, 2018; OECD, 2017). The legal retirement ages, set in 1978, are 50 for blue-collar women, 55 for white-collar women and 60 for all men (Hou et al., 2021). The actual average retirement age in the formal sector is 54 (Zhang et al., 2023), which is much earlier than the actual retirement age in many OECD countries (Giles et al., 2023). In contrast, workers in the informal sector, which is concentrated in rural areas, working on the farm or in other agriculture-related activities only expect to receive pensions that are a small fraction of those afforded to urban residents (Lei et al., 2013). The rural pension scheme is not employment-based and rural residents are eligible to start receiving limited pensions at age 60 if they have contributed to the pension system for at least 15 years (Giles et al., 2023). The amount of pension income is well below the subsistence level, which provides little incentive for rural elderly to stop working at the age of 60 (Ning et al., 2016). Rural residents' lives after retirement (exit from work) mainly depend on family support and the depletion of their own savings (Benjamin et al, 2003), which leaves them in a more vulnerable position. Those informal retirees in rural areas may need to work as long as their health allows (Smith et al., 2014).

As a major labour-intensive agricultural producer globally, China is facing the ageing of its agricultural labour force, which is posing considerable challenges for food security and agricultural sustainability (Ren et al., 2023). Moreover, young people's off-farm employment and prolonged out-migration to urban areas have resulted in a scarcity of high-quality agricultural labour and exacerbated the effects of ageing on agricultural production (Jiang et al., 2019). It has triggered concerns throughout Chinese society regarding the question of 'who will farm in the future' (Liu et al., 2023).

To relieve the economic and social pressures caused by changes in the population structure, it is important to encourage middle-aged and elderly people in rural China to provide a long-term effective labour supply. Accordingly, the determinants of labour supply decisions among middle-aged and elderly people in rural China should be studied. As most middle-aged and elderly people live in households, other household members, especially their spouses, could affect their behaviours. Thus, it is fair to study labour supply decisions in households. What are the most important determinants of labour supply? How does the household affect an individual's labour supply decision? During the last decades, these questions have attracted renewed attention from both theoretical and empirical researchers.

The theory of labour supply is grounded on the model of a consumer making a choice between consuming goods and consuming leisure. On the one hand, this neoclassical theory model is extended by taking account of family structures' influence and production within households (Cahuc et al., 2014). Theory focusing on intra-household labour supply decisions has developed along three different lines. The first, involving the unitary model, starts from the principle that the family can be linked to a sole agent having its own proper utility function (Killingsworth and Heckman, 1986). However, this model has been criticised for arbitrarily aggregating family members' preferences (Fortin and Lacroix, 1997; Blundell and MaCurdy, 1999). The second, involving the axiomatic bargaining models (McElroy and Horney, 1981; McElroy, 1990) and non-cooperative models (Lundberg and Pollak, 1993; Chen and Woolley, 2001), has been developed to allow family members, especially partners, to have different preferences. The third, which has attracted a great deal of recent attention, involves the collective model. This model postulates that making choices is fundamentally something individuals do; that the family is no more than a particular framework enlarging (or constraining) the range of each member's choices; and that however decisions are made, the outcomes are Pareto efficient (Chiappori, 1988, 1992, 2011; Chiappori et al., 2002; Browning et al., 1994, 2006). There is a great deal of empirical evidence of the collective model from different countries (Chau et al., 2007; Oreffice and Quintana-Domeque, 2012; Giovanis and Ozdamar, 2019). On the other hand, the neoclassical labour supply theory model is extended to adapt to analyse dynamic behaviours. From a dynamic perspective, a consumer must make the choices over a 'life cycle', so the life-cycle labour supply model allows us to grasp the contrasting effects caused by a transitory change in wages or a permanent modification of the wage profile (Heckman and MaCurdy, 1980; MaCurdy, 1981). Additionally, Stephens (2002) has extended the life-cycle model (MaCurdy, 1981) into the household, assuming that the household jointly maximises utility.

Although age, gender, education and wages are important factors in labour supply decisions, there are a number of empirical studies suggesting that health is one of the most important determinants of elderly people's labour supply. The empirical works of health and labour supply mainly focus on two aspects: individual health effects on labour supply and spouse health effects on labour supply. Further, health is usually measured by subjective and objective measures and health shocks, and labour supply is always measured by labour participation, working hours, and retirement (labour exits).

Numerous studies have detailed the individual health impacts on labour supply. Although scholars have failed to reach a consensus on the magnitude of the effect and on whether it is important compared to other determinants, the majority of the literature finds that health is a crucial factor of labour supply and that the impact varies by gender, age, and sometimes education level and wealth level. The literature has traditionally used cross-sectional data (Grossman and Benham, 1974; Luft, 1975) and subjective health measures (self-reported health) (Hurd and Boskin, 1984; Haveman et al., 1994). However, cross-sectional analyses limit the ability to control for individual unobserved factors that confound the relationship between labour supply and its determinants. Additionally, subjective health measures have been criticised for measurement error (Disney et al., 2006), justification bias (Bound, 1991), and reverse causal effects (García Gómez and López Nicolás, 2006). More recently, the availability of rich longitudinal survey data allows more reliable evidence to be obtained on this topic (Jones et al., 2010, 2020; García-Gómez et al., 2010). A stream of literature attempts to deal explicitly with endogeneity and measurement error issues and instrument self-reported measures using objective measures (Stern, 1989; Cai and Kalb, 2006; Deschryvere, 2005). A large group of studies choose objective measures as additional health measures, such as body mass index (Caliendo and Gehrsitz, 2016), problems with activities of daily living (ADLs) (Kalwij and Vermeulen, 2008), instrumental ADLs (IADLs), diagnosis of chronic and acute health problems (Heinesen and Kolodziejczyk, 2013; Minor, 2013; Chatterji et al., 2017), mental health (Frijters et al., 2010), and disabilities. However, objective measures are imperfectly correlated with working capacity, making estimates subject to measurement error (Stern, 1989; Bound et al., 1999; Coile, 2004). Numerous recent studies prefer health shock measures, which are defined by acute health events, the onset of a new chronic disease and accidental injuries or falls (McClellan, 1998; Disney et al., 2006; Jones et al., 2010, 2020; Macchioni Giaquinto et al., 2022). The appeal of this approach is that it exploits the arrival of unexpected new information about health to estimate the effect of changes in health on changes in labour supply. This approach avoids the justification hypothesis concern by not using self-reported health status and more generally addressing the potential problem of (time-invariant) unobserved heterogeneity that is correlated with both health and labour supply (Coile, 2004).

Another stream of empirical research has focused on spouse health and labour supply. Traditionally, poor health or health shocks can result in a significant loss in family income if the worker reduces the labour supply, but the family can protect itself against this loss if the worker's spouse increases the labour supply, generating an 'added worker effect' (substitution effects) (Mincer, 1962; Spletzer, 1997). The literature that explores the effect of spouse health on labour supply does not present a clear consensus. Some studies find that women increase their labour supply in response to their husband's poor health (Parsons, 1977; Charles, 1999; Reis, 2007). In contrast, some studies emphasis complementarity (the complementarity of leisure time) rather than substitutionality between own and spousal labour supply. A few empirical results reveal that women reduce their labour supply to care for their sick partner (Hollenbeak et al., 2011; Jeon and Pohl, 2017). Additionally, some researchers find no significant effect on the wife's labour supply when the husband falls ill (García-Gómez et al., 2013; Braakmann, 2014). Moreover, quite a few studies find gender differences. For example, on the one hand, some previous findings conclude that decreases in potential family income have the strongest effect on wives because they tend to increase their market work when the husband dies or his health condition deteriorates (Berger, 1983; Berger and Fleisher, 1984). On the other hand, husbands react to their wife's disability or death by reducing their market work and increasing their contribution to household work. Coile (2004) finds that the added worker effect is small for men and that there is no such effect for women. Vecchio (2015) shows significant and positive responsiveness to labour participation among women when residing with a family member experiencing either a disabling cancer condition or a musculoskeletal condition. The presence of a mentally ill family member reduces the male propensity to participate in the labour market. Most recently, Acuña et al (2019) analysed the existence of the added worker effect in a life cycle, finding that women's probability of labour force entry over three years increases by 50 percentage points when their husbands between the ages of 18–44 are diagnosed with arthritis. This effect disappears in older age groups. Furthermore, a few researchers investigate health and joint retirement decisions in couples (Johnson, 2001).

Previous studies provide a rich basis for labour supply theory and reveal numerous empirical pieces of evidence of health impact on labour supply. Generally, studies investigating how own health affects labour supply always regard spousal health as an important control and vice versa. However, to the best of our knowledge, scarce research explores how the effects of own health and spousal health on labour supply decisions interact with each other in the household. This paper contributes to the existing literature by considering the interaction of own health and spousal health and conducting a more comprehensive analysis of health and labour supply in the household. We find an obvious link between the effects of own health and spousal health on labour supply. Additionally, this study also contributes to the literature on the labour supply of couples in the informal sector. The existing literature revealing the strong correlation between husbands and wives' joint labour supply decisions mainly investigates the labour supply in the formal sector (Cribb et al., 2013; Schirle, 2008; García-Miralles and Leganza, 2014; Michaud et al., 2020). In our approach, we first introduce health into the general intra-household collective model to build the theoretical framework. Our empirical models include two main parts: first, we use a set of objective health indicators, such as doctor-diagnosed health problems and limitations in daily activities, to construct a latent health stock index for eliminating possible measurement bias of self-reported health measures. We apply a random-effect ordered probit model to estimate this latent health stock. Next, in the labour supply model, we are interested in two measures of labour supply: labour participation and annual work time, and our main independent variables are own latent health, spousal latent health and their interaction terms. We use one-period lagged latent health stock to deal with the possible simultaneous causality of health and labour supply. Additionally, to account for the potential correlation of the error terms in the husband's and wife's labour supply equations in the household, we estimate the labour supply equations of the couples systematically using bivariate probit and SUR models rather than estimating them separately. Our control variables include age, education, hukou, living area, household structure and household non-labour income.

1.2. The intra-household collective model with health

The traditional approach to labour supply arises, fundamentally, out of the idea that each of us can make trade-offs between the consumption of goods and leisure. It is well-known that health can be viewed as a durable capital stock that produces an output of healthy time (Grossman, 1972). In other words, given a total time budget, health defines the time unable to be involved in the market and non-market activities due to health problems and then determines the available time for work and leisure. Thus, it seems fair to introduce health into the traditional labour supply model.

Next, we consider the individual's labour supply choice in the household. We set an intra-household collective labour supply model following Chiappori et al. (2002). There are several reasons for choosing the intra-household collective model. Firstly, the unitary model has been attacked due to both the theoretical and empirical aspects. On the one hand, in the unitary framework, the process by which individual preferences get aggregated into a household utility function is essentially a "black box", issues such as intra-household inequality and household formation/dissolution cannot be handled well. On the other hand, in the unitary model, the household's problem is equivalent to maximizing a single utility function, subject to a pooled budget constraint. A central prediction of this model is income pooling, which is the idea that the household's demands depend only on its total income and not on the sources of income. Thus, empirically, this implies that if total income is held constant, a change in the sources of income would not affect results. Another empirical implication is that a marginal increase in one source of income has the same effect on results as a marginal increase in any other source of income. However, this central prediction has consistently failed to find support in the data. Secondly, although bargaining models from cooperative game theory were among the first non-unitary models of the household, these models have disadvantages due to two aspects. On the one hand, an important feature of the cooperative bargaining model is the presence of a threat point for each household member. The threat points represent the maximal utility from some kind of a default outcome. Typically, this default outcome has been interpreted as an outside option that is external to the household (for example, divorce). However, this feature is argued to be a non-cooperative equilibrium with the household. On the other hand, the conditions derived from cooperative bargaining models turn out not restrictive, unless the agents' premarital preferences are known. Thirdly, the intra-household collective model is increasingly dominant in the literature for two main reasons. First, it rejects income pooling from the framework, which is more realistic. Additionally, the collective model only makes a very weak and general assumption-namely, that the household always reaches Pareto-efficient agreements, and the conditions deriving from this assumption are falsifiable and more testable than conditions deriving from the cooperative bargaining model. Overall, in this paper, we choose the intra-household collective model as the basic theoretical model and introduce health to the basic intra-household collective model to support the following empirical analysis.

In this framework, the household consists of two individuals with distinct utility functions, and the decision process, whatever its true nature, leads to Pareto-efficient outcomes. This assumption seems quite natural, given that spouses usually know each other's preferences pretty well (at least after a certain period) and interact very often. Therefore, they are unlikely to leave Pareto-improving decisions unexploited. Formally, let us consider a household that has two decision-makers, a husband, m(male), and a wife, f(female). Let t^m and t^f be the labour supply of the husband and wife respectively. For member i $(i = m, f)$, T^i denotes member i's total available time, and in the household, T^i is composed of L^i (member i's leisure), t^i (member i's time for work), s^i (member i's sick days) and a^i (member i's caring time for spouse). Specifically, we define s as those days unable to be involved in the market and non-market activities due to health problems, and a is defined as the time for caring spouse due to the spouse's health problems. Then, it's obvious that s is a function of an individual's own health and a is a function of a spouse's health. Let h^i be member i's health, C^i be the aggregate consumption of member i, w_i be member i's wage rates, y be non-labour income in the household, z^i be member i's K-vector of preference factors and d be the J-vector of distribution factors. Distribution factors are those variables that change the household's environment, and in particular, the members' respective bargaining positions. Distribution factors that affect opportunities of spouses outside marriage can influence the intrahousehold balance of power and ultimately the final allocation of resources (Haddad and Kanbur, 1991). Early literature has emphasized that variables indicating the situation in the marriage market are natural examples of distribution factors. Becker (1993) finds that the state of the marriage market crucially depends on the sex ratio (the relative supplies of males and females in the marriage market). When the sex ratio is favourable to the wife-that is, there is a relative scarcity of women-the distribution of gains from marriage will be shifted in her favour, and this may, in turn, affect intrahousehold decisions. There is also a variety of literature discussing distribution factors under the intrahousehold collective framework. Chiappori (2002) uses the state-level sex ratio index as a distribution factor and finds the increase in the sex ratio reduces wives' annual work time, whereas it increases husbands' labour supply. Chau Tak Wai, et al (2007) define the difference in non-labour income between spouses (the husband's non-labour income minus the wife's) and the differences in years of education between spouses as distribution factors and finds that husbands work less when the differences in non-labour income and education of husbands over wives are larger and similarly, wives work more when the difference is larger.

Here, we start from the most general version of the model, in which member i's welfare can depend on his or her spouse's consumption and leisure in a very general way, including, for instance, altruism, public consumption of leisure, positive or negative externalities, and so forth. In this general framework, member i's utility function is $U^i\ (C^m,\ L^m, C^f, L^f, z^m,\ z^f)$. Under the collective framework, intra-household decisions are Pareto-efficient. For any given $(w_m,\ w_f, y,\ z^m, z^f, d)$, hence, there exists a weighting factor $\mu(w_m,\ w_f, y,\ z^m, z^f, d)$ belonging to [0,1] such that $(C^i,\ L^i)$ solves the following program (Program A):

$$\max_{\{C^m, C^f, L^m, L^f\}} \mu U^m + (1-\mu) U^f$$

Subject to

$$w_m t^m + w_f t^f + y \geq C^m + C^f$$

$$L^m + t^m + s^m + a^m = T$$

$$L^f + t^f + s^f + a^f = T$$

$$s^m = s^m(h^m)$$

$$s^f = s^f(h^f)$$

$$a^m = a^m(h^f)$$

$$a^f = a^f(h^m)$$

where the function μ is assumed continuously differentiable in its arguments. It should thus be clear that the vector of distribution factors, d, appears only in μ, but in neither the preference nor the budget constraint.

It should, however, be emphasized that this general version of the collective model cannot be uniquely identified from the sole knowledge of labour supplies. There is a continuum of different structural models that are observationally equivalent, that is, that generate identical labour supply functions. Therefore, to help empirical analysis, we follow Chiappori et al. (2002), assuming members have egotistic preferences[1], which means: individual utilities have the form $U^i\ (C^i,\ L^i, z^i)$, where$U^i$is strictly quasi-concave, increasing, and continuously differentiable for $i = m, f$. Indeed, consider the household as a two-person economy, from the second fundamental welfare theorem, any Pareto optimum can be decentralized in an economy of this kind. Thus, we can have the following sharing rule interpretation. Under the assumption of egotistic preferences, Program A is equivalent to the existence of some function $\emptyset^i(w_m, w_f, y, z^i, d)$ such that each member $i\ (i = m, f)$ solves the following program (Program B):

[1] As shown in Chiappori(1992), the following analysis also holds in the more general cases of "caring" agents(see Becker 1991),that is, agents whose preferences are represented by a utility function that depends on both their egotistic utility and their spouses'. In fact, any decision that is Pareto-efficient under caring preferences would also be Pareto-efficient under egotistic preferences.

$$\max_{\{C^i,L^i\}} U^i\,(C^i, L^i, z^i)$$

subject to

$$w_i h^i + \emptyset^i \geq C^i$$

$$L^i + t^i + s^i + a^i = T$$

where $\emptyset^m + \emptyset^f = y$ (Proof see (Chiappori, 1992). Additionally, $\emptyset^i$ may be negative or bigger than y (for instance, if y is low and wages are very different, one member may share labour income with the other).

The interpretation is: In the labour supply decision process, household members could share income with their spouses, and then, subject to the corresponding budget constraint, each member separately chooses a labour supply (and private consumption, leisure). The function $\emptyset$ is called the sharing rule. It describes the way how non-labour income (and sometimes individuals' labour incomes) is divided up, as a function of wages, non-labour income, distribution factors, and other observable characteristics. Now, using male labour supply choice as an example, it could be expressed by the following maximization program (Program C):

$$\max_{\{C^m,L^m\}} U^m\,(C^m, L^m, z^m)$$

subject to

$$w_m t^m + \emptyset^m \geq C^m$$

$$L^m + t^m + s^m + a^m = T$$

$$s^m = s^m(h^m)$$

$$a^m = a^m(h^f)$$

The budget constraint could be also expressed in the following manner:

$$C^m + w_m t^m \leq M \equiv w_m(T - s^m(h^m) - a^m(h^f)) + \emptyset^m$$

Using λ to denote the Lagrange multiplier associated with the budget constraint, the Lagrangian Q of this program is:

$$Q^m = U^m(C^m, L^m, z^m) + \lambda(\mathrm{M} - \mathrm{C}^m - w_m L^m)$$

The first-order conditions require that the derivatives of the Lagrangian with respect to each of its arguments are zero:

$$\frac{\partial Q^m}{\partial C^m} = \frac{\partial U^m(C^m, L^m, z^m)}{\partial C^m} - \lambda = 0$$
$$\frac{\partial Q^m}{\partial L^m} = \frac{\partial U^m(C^m, L^m, z^m)}{\partial L^m} - \lambda w_m = 0$$
$$\frac{\partial Q}{\partial \lambda} = M - C^m - w_m L^m = 0$$

There are 3 unknowns - C^m, L^m and λ-and three equations, we could get the optimizing choices after solving this system of equations:

$$L^m = L^m(M, w_m, z^m)$$
$$C^m = C^m(M, w_m, z^m)$$

As we know:

$$M \equiv w_m(T - s^m(h^m) - a^m(h^f)) + \emptyset^m$$
$$t^m = T - L^m - s^m - a^m$$
$$\emptyset^m = \emptyset^m(w_m, w_f, y, z^m, d)$$

Then:

$$t^m = t^m(w_m, w_f, y, h^m, h^f, z^m, d)$$

Thus, the husband's labour supply in the household could be expressed by a function of wage rates of husband and wife, non-labour income, the health of husband and wife, preference factors and distribution factors. And for the wife's labour supply, the analysis is essentially identical.

1.3. Empirical Method

1.3.1. A model for underlying health stock

In attempting to identify the impact of health on the labour supply decision, the choice of health measures is important. Using general self-reported health measures is rejected by some researchers as these subjective measures are based on subjective judgements and may have potential bias and not be comparable across individuals.

In this section, we follow Bound et al. (1999), Disney et al. (2006) and Jones et al. (2010) in constructing an individual's underlying health stock to deal with the issues associated with the possible measurement error (reporting bias). Specifically, we assume that the i individual's health at time t is a function of a comprehensive set of objective health indicators b_{it}, a time-varying unobservable ε_{it} (uncorrelated with b_{it}), and a set of panel-level random effects μ_i. Denote this health state as γ_{it}. And then:

$$\gamma_{it} = b'_{it}\beta_t + \mu_i + \varepsilon_{it} \qquad (1)$$

Although this health state is not observed, self-reported health status is available in our data as a five-category variable. Let this categorical variable be h_{it}, and denote the latent counterpart to h_{it} as h^*_{it}, which is a simple function of γ_{it} and a term ω_{it} reflecting reporting error:

$$h^*_{it} = \gamma_{it} + \omega_{it} \qquad (2)$$

Specially, we assume that ω_{it} is uncorrelated with ε_{it}.Thus we have:

$$h^*_{it} = b'_{it}\beta_t + \mu_i + (\varepsilon_{it} + \omega_{it}) \qquad h^*_{it} = b'_{it}\beta_t + \mu_i + u_{it} \qquad (3)$$

Assuming that u_{it} is normally distributed and is independent of μ_i, Eq. (3) can be estimated as a random-effect ordered probit. Self-reported health status is used as a dependent variable and the fitted value from this regression ($\hat{h}_{it}$) are used for the latent health index. In Eq. (3): b_{it} includes doctor-diagnosed health problems and limitations in daily activities. Then, this time-varying individual latent health index ($\hat{h}_{it}$) is constructed to enter the labour supply equation.

1.3.2. A model for labour supply

In this paper, we investigate the impact of health on the labour supply by including own health, spousal health and their interaction terms into labour supply equations.

The sample for analysis is restricted to individuals who are observed for at least two points in time, labelled $t-1$ and t. These can be any consecutive waves across the waves we have observations. There could be concerns about the simultaneous correlations between health status and labour supply, and thus we introduce the one-period lagged health by exploiting the "timing of events" as the lagged health status occurs before employment status is observed. And similarly, all of the time-varying potential confounders are measured as of $t-1$. Besides, considering the potential reverse causality between wages and labour supply, we don't include wage controls in our empirical models.

Then, we employ the following functional form for the labour supply equations of the husband and wife:

$$Y_{it}^{m} = \beta_1 \gamma_{it-1}^{m} + \beta_2 \gamma_{it-1}^{f} + \beta_3 \gamma_{it-1}^{m} * \gamma_{it-1}^{f} + \beta_4 d_{it-1} + \beta_5 X_{it-1} + \varepsilon_{it}^{m} \quad (4)$$

$$Y_{it}^{f} = \beta_1 \gamma_{it-1}^{f} + \beta_2 \gamma_{it-1}^{m} + \beta_3 \gamma_{it-1}^{m} * \gamma_{it-1}^{f} + \beta_4 d_{it-1} + \beta_5 X_{it-1} + \varepsilon_{it}^{f} \quad (5)$$

Equation (4) is the labour supply regression of the husband (male), where: Y_{it}^{m} is the labour supply of husband i at time t; γ_{it-1}^{m} and γ_{it-1}^{f} are unobserved health states of the husband and his wife at time $t-1$; as γ_{it-1} is unobserved, we replace it with $\hat{h}_{it-1}$, which is the latent health index for the individual. $\gamma_{it-1}^{m} * \gamma_{it-1}^{f}$ is the interaction term of the husband's and wife's unobserved health states at time t-1, which is measured by $\hat{h}_{it-1}^{m} * \hat{h}_{it-1}^{f}$; X_{it-1} is a set of covariates at time $t-1$ including individual characteristics like husband's age, education; household characteristics (such as living area and non-labour income). Especially, we follow Chau Tak Wai, et al (2007) using the education gap as the distribution factor reflecting the respective bargaining positions in the household, d_{it-1} is the education gap between the husband and wife at time $t-1$, which is measured by comparing the education years between the husband and wife. Similarly, equation (5) is the labour supply regression of the wife (female).

Since the error terms in Eq. (4) and (5) are likely to be correlated, we estimate the system of labour supply equations of husbands and wives by bivariate probit and SUR models rather than estimating labour supply equations separately for husbands and wives. The error terms in Eq. (4)and (5) are allowed to have their own variances and be correlated with the others in the same period.

1.4. Data and Variables

1.4.1. The CHARLS dataset and sample

This study uses panel data from the China Health and Retirement Longitudinal Study (CHARLS) (Wave 1-3/Year 2011-2015). CHARLS is a nationally representative longitudinal survey of the middle-aged and elderly population of China, consisting of persons 45 years old or older, and including assessments of the social, economic, and health circumstances of community residents. The participants are followed up every two years. Before sample restrictions, the entire sample is composed of 57417 individuals: 17708 observers in 2011, 18612 observers in 2013, and 21097 observers in 2015.

In rural China, about 75% of labour participants engage in agricultural work, and more than 90% of labour participants engage in the informal labour market. It could be interesting to explore the labour supply decisions of the informal workforce. Thus, in this paper, we limit our analysis to couples who are 45 years old or older and live in rural China. This means there should be two married or partnered respondents in the household, and both of them should be 45 years old or older. The entire sample size is 26,878 (13,439 couples). In the empirical analysis part, the samples of 2011 and 2013 are used to obtain lagged variables. After removing the individuals who have missing values for the main variables in the corresponding empirical models, the respective final sample sizes are: 22,082 in the latent health model; 11,430(5715 couples) in the labour participation model; and 4848(2424 couples) in the yearly work time model (Table 1.1).

Table 1.1 Sample size

	Wave 1(2011)	Wave 2(2013)	Wave3(2015)	Total obs
Total obs	17708	18612	21097	57417
45 years old Couple in Rural China	8470	8876	9532	26878
Latent health equation	7827	7279	6976	22082
Labour participation equation	-	6366	5064	11430
Yearly work time equation	-	2682	2166	4848

1.4.2. Variables and definitions

1.4.2.1. Variable in the latent health model

The CHARLS includes a series of self-reported health variables. Of particular interest to us is the general five-point self-assessed health. To construct the index for the health stock variable, we use questions on difficulties in daily activities and specific health problems[2]. Individuals are asked about a set of functional limitations, such as ADLs and IADLs, which reflect on their ability to live independently and thrive[3]. They are also asked about a list of doctor-diagnosed health problems. We create binary dummies for the presence of each limitation in daily activities and each specific health problem. Given this broad set of health measures, it is likely that we are measuring most of the important aspects of health. Additionally, we assume that reporting bias does not influence these variables, which identify more specific health problems. Moreover, because most individuals in our sample engage in agriculture work, the doctor-diagnosed health problems and the functional limitations in the daily activities we choose can more or less affect work capacity mentally or physically. In other words, the objective health measures in our paper can very likely limit the kind, amount or efficiency of work. Table 1.2 provides detailed Variable definitions.

Table 1.2 Variable names and definitions in the latent health model

Variables	Description
Self-reported health	1.very poor/2.poor/3.fair/4.good/5.very good
1-item Activities of daily living (ADLs): Some difficulty	1 if difficulty reported, 0 otherwise. There is a dummy for difficulty with dressing
4-item Instrumental activities of daily living (IADLs): Some difficulty	1 if difficulty reported, 0 otherwise. There are individual dummies for difficulties with: 1. managing money/2. taking medications/3. preparing hot meal/4. cleaning house
7-item Other	1 if difficulty reported, 0 otherwise. There are individual dummies for

[2] See Appendix A1for the detailed process of choosing the objective variables.

[3] ADL is a term used to collectively describe the fundamental skills required to independently care for oneself, such as eating, bathing, and dressing. IADLs are things people do every day to take care of themselves and their home. We also consider some other functional limitations in various activities beyond the defined ADLs and IADLs.

functional limitations: Some difficulty	difficulties with: 1. walking 1km/2. jogging 1km/3. getting up from a chair after sitting for long periods/4. climbing several flights of stairs without resting/5. stooping kneeling, or crouching /6. lifting or carrying weights over 5kg /7. reaching arms above shoulder level
The doctor diagnosed health problems: Ever Had Condition (13 items)	1 if the problem is reported, 0 otherwise. There are individual dummies for problems with: 1. high blood pressure; 2. Diabetes; 3.caner; 4.lung disease; 5.heart problems; 6.stroke; 7.psych problem; 8. Arthritis; 9.Dyslipidaemia; 10.liver disease; 11.kidney disease; 12. Stomach digestive disease; 13. asthma

1.4.2.2. Variable in the labour supply model

The CHARLS includes some labour supply questions. We are interested in two measures of labour supply: labour participation and yearly work time. To elaborate, we summarise the main labour force status for the respondents as agricultural work, non-agricultural employed, non-agricultural self-employed, non-agricultural unpaid family business, unemployed, retired, and never worked. Further, we create a dummy for labour participation, which is assigned 1 if the respondent engaged in agricultural work, non-agricultural employed work, non-agricultural, self-employment work, or non-agricultural family business work. Yearly work time is the number of total yearly hours that the respondent works for their main job and side jobs. We drop out individuals who report over 7300 (20*365) hours because this is not realistic. Other covariates include age, education, hukou, living region, household structure, and household non-labour income. Table 1.3 displays detailed variable definitions.

Table 1.3 Variable names and definitions in the labour supply model

Variables	Description
Labour force status	1. Agricultural work/2.Non-Agri employed/3.Non-Agri self-employed/4.Non-Agri family business/5.Unemployed/6.Retired/7.Never work
Labour participation	1 if engaged in the labour market according to labour force status; 0 otherwise
Yearly work time	The number of total yearly hours that the respondent works for their main job and side jobs
Latent health index	The predicted latent health index from the Latent health model
Age	Age of the respondent
Gender	1 if female, 0 if male
Education	1. less than lower secondary/2. upper secondary & vocational training/3. tertiary

Education Gap	Husband's education years-Wife's education years
Hukou	Hukou status affects many aspects of life in China such as buying a house, buying a car, children's school enrolment and other welfare. 1. Agricultural hukou/2. Non-agricultural hukou/3. Unified residence hukou/4. Do not have hukou
Living region	The household living region defined by the National Bureau of Statistics; 1 if the household is located in a specific region and 0 if not; there are four individual dummies: East, West, Central, Northeast
Household structure	The demographic structure: the number of pre-school children (age:0-6); the number of school children (age:7-18) and the number of old people (age>75)
Household non-labour income	Summary of household government and public transfer income, other household member's total income and household rental income from non-financial assets
Category household non-labour income	1 if less than 50th percentile of household non-labour income; 2 if among 50th and 75th percentile; 3 if more than 75th percentile

1.5. Empirical Results

1.5.1. Results for latent health model

We now use objective health measures to estimate the model for latent health stock. Considering potential gender differences, we estimate latent health by gender. Table 1.4 provides variables' descriptive statistics. The total sample for the latent health model consists of 11,148 males and 10,934 females. It indicates that a considerable segment of the sample experiences challenges in Activities of Daily Living (ADLs) or has health conditions diagnosed by a physician, notably with difficulty in other functional limitations, managing money, cleaning the house, having high blood pressure, lung disease, heart problems, arthritis, dyslipidaemia or stomach/digestive disease. Females in the sample report a higher incidence of difficulties across most of these aspects compared to males.

Table 1.4 Descriptive statistics of variables in the latent health model

Variable	Obs		Mean		Std.Dev.		Min		Max	
	male	female	male	female	male	female	male	female	male	female
self-reported health	11,148	10,934	3.0606	2.8795	0.959	0.943	1	1	5	5
1-item ADLs (reference group: no difficulty)										
dressing	11,148	10,934	0.0478	0.0595	0.213	0.237	0	0	1	1

4-item IADLs (reference group: no difficulty)										
managing money	11,148	10,934	0.0843	0.154	0.278	0.361	0	0	1	1
taking medications	11,148	10,934	0.0393	0.0687	0.194	0.253	0	0	1	1
preparing hot meal	11,148	10,934	0.0849	0.0889	0.279	0.285	0	0	1	1
cleaning house	11,148	10,934	0.0910	0.121	0.288	0.326	0	0	1	1
7-item other functional limitations (reference group: no difficulty)										
walking 1km	11,148	10,934	0.122	0.200	0.327	0.400	0	0	1	1
jogging 1km	11,148	10,934	0.441	0.618	0.497	0.486	0	0	1	1
getting up from a chair	11,148	10,934	0.231	0.359	0.421	0.480	0	0	1	1
climbing stairs	11,148	10,934	0.332	0.510	0.471	0.500	0	0	1	1
stooping kneeling	11,148	10,934	0.271	0.385	0.444	0.487	0	0	1	1
lifting over 5kg	11,148	10,934	0.0764	0.173	0.266	0.379	0	0	1	1
reaching arms above shoulder	11,148	10,934	0.0967	0.129	0.296	0.335	0	0	1	1
The doctor diagnosed health problems: Ever Had Condition (13 items)										
high blood pressure	11,148	10,934	0.262	0.286	0.440	0.452	0	0	1	1
diabetes	11,148	10,934	0.0553	0.0759	0.229	0.265	0	0	1	1
caner	11,148	10,934	0.00960	0.0168	0.0975	0.129	0	0	1	1
lung disease	11,148	10,934	0.151	0.108	0.358	0.310	0	0	1	1
heart problems	11,148	10,934	0.108	0.152	0.310	0.359	0	0	1	1
stroke	11,148	10,934	0.0319	0.0252	0.176	0.157	0	0	1	1
psych problem	11,148	10,934	0.0122	0.0200	0.110	0.140	0	0	1	1
arthritis	11,148	10,934	0.363	0.445	0.481	0.497	0	0	1	1
dyslipidemia	11,148	10,934	0.101	0.117	0.302	0.321	0	0	1	1
liver disease	11,148	10,934	0.0536	0.0454	0.225	0.208	0	0	1	1
kidney disease	11,148	10,934	0.0833	0.0717	0.276	0.258	0	0	1	1
stomach/digestive disease	11,148	10,934	0.252	0.315	0.434	0.465	0	0	1	1
asthma	11,148	10,934	0.0649	0.0405	0.246	0.197	0	0	1	1

Table 1.5 presents the results of estimations by pooled oprobit and random effects oprobit. All objective health measures are nearly individually significant at the 1% level. The negative coefficient unambiguously means that an increase in the variable concerned will decrease the probability with which an individual is predicted to be in the highest health category (very good) and increase the probability with which they are predicted to be in very poor health, and vice versa. We also observe slightly different impacts for men and women, which means it is reasonable for us to estimate latent health stock by gender. The reported likelihood-ratio test shows that there is enough reason to favour a random-effect ordered probit regression over a standard pooled

ordered probit regression. Therefore, we predict the latent health index for each individual from the random-effect oprobit equation[4].

Table 1.5 Results of estimating self-assessed health

	Male		Female	
	(1)	(2)	(3)	(4)
	Pooled oprobit	Random effects	Pooled oprobit	Random effects
Dependent variable: self-reported health				
1-item ADLs (reference group: no difficulty)				
dressing	-0.16***	-0.18***	-0.15***	-0.17***
	(-2.73)	(-2.63)	(-2.91)	(-2.91)
4-item IADLs (reference group: no difficulty)				
managing money	-0.21***	-0.26***	-0.12***	-0.13***
	(-4.89)	(-5.10)	(-3.55)	(-3.50)
taking medications	-0.12*	-0.13*	-0.21***	-0.21***
	(-1.93)	(-1.91)	(-4.66)	(-4.15)
preparing hot meal	0.02	0.02	-0.29***	-0.32***
	(0.40)	(0.36)	(-5.53)	(-5.32)
cleaning house	-0.22***	-0.22***	-0.10**	-0.11**
	(-4.34)	(-3.74)	(-2.15)	(-2.07)
7-item other functional limitations (reference group: no difficulty)				
walking 1km	-0.19***	-0.26***	-0.15***	-0.19***
	(-4.83)	(-5.40)	(-4.54)	(-5.10)
jogging 1km	-0.42***	-0.47***	-0.29***	-0.30***
	(-16.46)	(-15.21)	(-11.40)	(-10.39)
getting up from a chair	-0.12***	-0.15***	-0.12***	-0.14***
	(-3.88)	(-4.11)	(-4.50)	(-4.69)
climbing stairs	-0.27***	-0.29***	-0.21***	-0.22***
	(-9.40)	(-8.57)	(-8.23)	(-7.52)
stooping kneeling	-0.15***	-0.19***	-0.16***	-0.19***
	(-5.10)	(-5.58)	(-5.85)	(-6.08)
lifting over 5kg	-0.22***	-0.24***	-0.22***	-0.23***
	(-4.60)	(-4.36)	(-6.77)	(-6.03)
reaching arms above shoulder	-0.13***	-0.15***	-0.10***	-0.09**
	(-3.21)	(-3.25)	(-2.82)	(-2.29)
The doctor diagnosed health problems: Ever Had Condition (13 items)				
high blood pressure	-0.13***	-0.15***	-0.10***	-0.12***
	(-5.06)	(-4.62)	(-3.95)	(-3.82)

[4] The random-effects oprobit model uses quadrature approximation. The accuracy depends partially on the number of integration points. We change the number of integration points to conduct the technical robustness tests.

diabetes	-0.20***	-0.25***	-0.21***	-0.24***
	(-4.30)	(-3.91)	(-5.08)	(-4.58)
caner	-0.36***	-0.43***	-0.36***	-0.43***
	(-3.35)	(-3.01)	(-4.33)	(-4.14)
lung disease	-0.19***	-0.22***	-0.16***	-0.18***
	(-5.74)	(-5.06)	(-4.23)	(-3.76)
heart problems	-0.24***	-0.25***	-0.19***	-0.22***
	(-6.79)	(-5.34)	(-6.01)	(-5.41)
stroke	-0.24***	-0.28***	-0.08	-0.13
	(-3.97)	(-3.35)	(-1.16)	(-1.53)
psych problem	-0.49***	-0.59***	-0.22***	-0.27***
	(-5.06)	(-4.63)	(-2.96)	(-2.78)
arthritis	-0.16***	-0.18***	-0.21***	-0.22***
	(-7.14)	(-6.01)	(-9.44)	(-7.71)
dyslipidemia	-0.12***	-0.14***	-0.12***	-0.12***
	(-3.15)	(-2.94)	(-3.42)	(-2.70)
liver disease	-0.23***	-0.28***	-0.13***	-0.13**
	(-4.88)	(-4.40)	(-2.58)	(-2.02)
kidney disease	-0.29***	-0.29***	-0.23***	-0.26***
	(-7.43)	(-5.67)	(-5.45)	(-4.83)
stomach/digestive disease	-0.28***	-0.31***	-0.29***	-0.31***
	(-11.18)	(-9.25)	(-12.21)	(-10.26)
asthma	-0.23***	-0.29***	-0.19***	-0.21***
	(-4.92)	(-4.59)	(-3.41)	(-2.85)
cut1	-2.85***	-3.35***	-2.73***	-3.10***
	(-86.22)	(-71.96)	(-84.38)	(-71.50)
cut2	-1.59***	-1.86***	-1.44***	-1.61***
	(-69.68)	(-58.76)	(-58.94)	(-51.49)
cut3	0.09***	0.15***	0.19***	0.28***
	(4.96)	(6.22)	(8.89)	(10.30)
cut4	0.77***	0.98***	0.86***	1.06***
	(38.26)	(34.47)	(36.06)	(33.34)
sigma2_u		0.44***		0.35***
		(15.63)		(13.82)
N	11148	11148	10934	10934
	chibar2(01) = 523.36		chibar2(01) = 367.80	
LR test vs. oprobit model	Prob >= chibar2 = 0.0000		Prob >= chibar2 = 0.0000	

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

1.5.2. Results for the labour supply model

1.5.2.1. Results for the labour participation model

In the labour participation models, we observe 5715 couples. The independent variables we are interested in are health-related: own latent health, spousal latent health, and their interaction terms. For covariates, we control age, education level, education gap, hukou, household living region, household non-labour income and household structure. Figure 1.1 and Figure 1.2 present the distributions of self-reported health and latent health stock for men and women. For self-reported health, both men and women have the highest proportion reporting their health as fair and quite a few of them reporting very poor or very good health. For latent health stock, the distribution curve for women is smoother and flatter than for men, and after the construction, more samples fall in the range of poor and fair health.

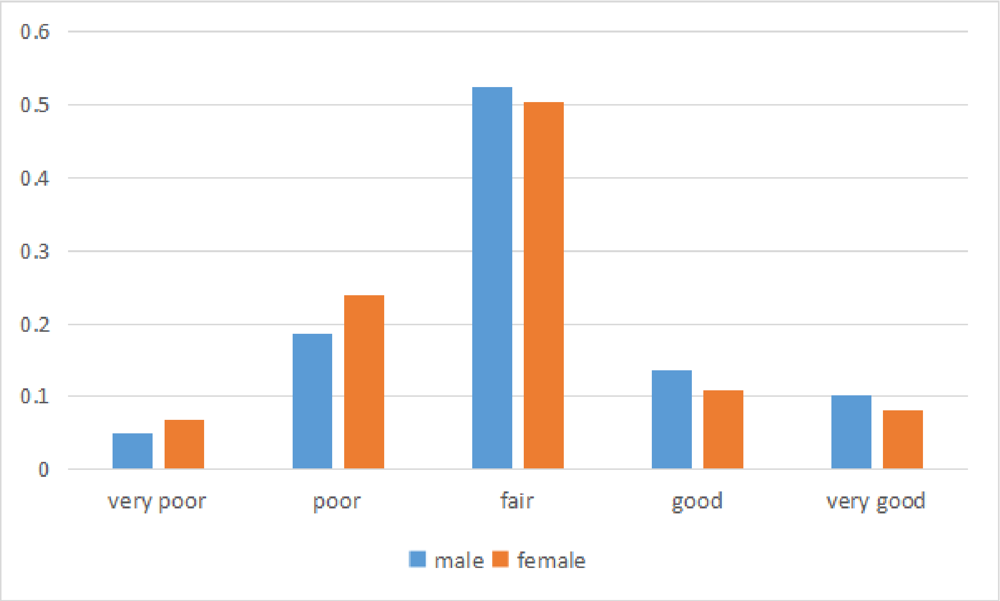


Figure 1.1 Distributions of self-reported health (labour participation model)

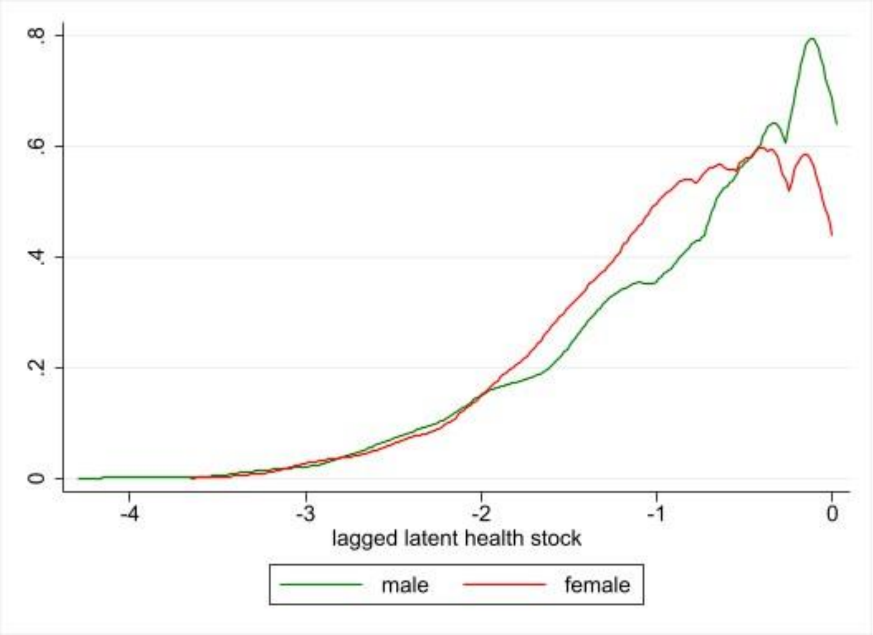


Figure 1.2 Distributions of latent health stock (labour participation model)

Table A2-A5 in the Appendix show the time variations of self-reported health across waves for males and females. From wave 1 to wave 2, 47.29% of men and 46.69% of women in the sample maintained the same self-reported health; from wave 2 to wave 3, 50.11% of men and 49.74% of women in the sample maintained the same self-reported health. Overall, 48.48% of men and 47.66% of women in the sample report the same self-reported health from wave 1 to wave 3. Figure A1 in the Appendix displays the distribution of the gap between current and lagged latent health stock for males and females, where the gap equals current latent health stock minus one-period lagged latent health stock. Both transition probabilities of self-reported health across waves and the distribution of current and one-period lagged latent health stock gaps show there are enough health variations over time. This reassures that the use of lagged health is sufficient to remove concerns over simultaneity bias.

Detailed variables' descriptive statistics are summarized in Table 1.6. The overall sample size is 5715 for both male and female. It shows that men are on average older and have a higher labour participation rate, latent health stock, and education level than women in the sample. On average, 82.2% of middle-aged and older men engage in the labour market, while 74.1% of middle-aged and older women participate in the labour market. The average one-period lagged latent stock is -0.798 for males and -0.893 for females, which both lie in the "fair" self-reported health category. The average age for men and women is 61.43 and 59.35 years old, respectively.

Table 1.6 Descriptive statistics of variables in labour participation model

Variable	Obs	Mean	Std.Dev.	Min	Max
Dependent variable: Labour participation					
male	5,715	0.822	0.382	0	1
female	5,715	0.741	0.438	0	1
Latent health index					
male	5,715	-0.798	0.738	-4.299	0.0213
female	5,715	-0.893	0.685	-3.659	0
Age					
male	5,715	61.43	8.646	45	89
female	5,715	59.35	8.166	45	91
Education					
male	5,715	1.121	0.347	1	3
female	5,715	1.031	0.177	1	3
Education gap	5,715	2.561	3.720	-11	15
Hukou					
male	5,715	1.089	0.318	1	4
female	5,715	1.030	0.196	1	4
Living region					
East	5,715	0.282	0.450	0	1
Central	5,715	0.318	0.466	0	1
West	5,715	0.335	0.472	0	1
Northeast	5,715	0.0660	0.248	0	1
Household non-labour income	5,715	10112	21807	0	600000
Category household non-labour income	5,715	1.786	0.813	1	3
Household structure					
Number of pre-school children (age:0-6)	5,715	0.325	0.639	0	5
Number of school children (age:7-18)	5,715	0.354	0.704	0	7
Number of old people (age>75)	5,715	0.0532	0.239	0	2

Table 1.7 presents the bivariate probit model results for the health-related variables we are interested in. Appendix Table A7 presents the results of covariates. Columns (1) and (3) show the coefficients for the men's labour participation equation and the women's labour participation equation, respectively. Columns (2) and (4) show the corresponding margins. Because we include an interaction term in the equations, the

sign, magnitude, and significance of the coefficients cannot be interpreted directly (Buis, 2010; Dow et al., 2019). Therefore, we show the results numerically and graphically to interpret the interaction effects more precisely (Table 1.8 and Figure 1.3) (Mize, 2019).

Table 1.7 Labour participation estimate

	Biprobit			
	Male's labour participation		Female's labour participation	
	Coefficients	Margins	Coefficients	Margins
	(1)	(2)	(3)	(4)
Own latent health	0.63***	0.11***	0.52***	0.13***
	(12.66)	(18.56)	(12.06)	(17.20)
Spousal latent health	0.07	-0.01*	0.07	0.00
	(1.23)	(-1.75)	(1.60)	(0.34)
Own latent health *Spousal latent health	0.11***	-	0.06*	-
	(2.98)	-	(1.84)	-
N	5715	5715	5715	5715
Wald test of rho=0: chi2(1) =269.75			Prob > chi2 = 0.0000	
Controls: age, education, education gap, hukou, living area, household non-labour income, household structure				

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

First, better own health significantly increases the probability of labour participation. Additionally, better spousal health enhances the positive effect of better own health on labour participation probability. Moreover, this positive own health impact on labour participation is slightly greater for women, although the gender gap narrows with the improvement of spousal health. For example, given very poor spousal health (latent health equals -3.5), a one-unit increase in own latent health improves labour participation probability by 4.76% and 9.14% for men and women, respectively. Given fair spousal health (latent health equals 0), a one-unit increase in own latent health brings a 13.07% and 14.75% increase in labour participation probability for men and women, respectively.

Table 1.8 The AME of latent health on labour participation

	Male		Female	
	AME of own latent health	AME of spousal latent health	AME of own latent health	AME of spousal latent health
Given spousal or own latent health				
-4		-0.1203***	0.0827***	
		(-3.5991)	(2.6921)	
-3.5	0.0476***	-0.1130***	0.0914***	-0.0488*
	(2.6272)	(-3.5629)	(3.5810)	(-1.7290)
-3	0.0595***	-0.0973***	0.1000***	-0.0417*
	(4.0696)	(-3.5200)	(4.8513)	(-1.6495)
-2.5	0.0714***	-0.0757***	0.1084***	-0.0314
	(6.3348)	(-3.4760)	(6.7588)	(-1.5208)
-2	0.0834***	-0.0519***	0.1166***	-0.0193
	(10.0543)	(-3.3617)	(9.7293)	(-1.2676)
-1.5	0.0954***	-0.0294***	0.1247***	-0.0072
	(15.5075)	(-2.8665)	(14.0055)	(-0.6821)
-1	0.1073***	-0.0111	0.1325***	0.0033
	(18.8102)	(-1.4402)	(17.2590)	(0.3979)
-0.5	0.1191***	0.0017	0.1401***	0.0111
	(16.7518)	(0.2246)	(15.8610)	(1.2520)
0	0.1307***	0.0091	0.1475***	0.0156
	(13.8746)	(1.2176)	(12.8447)	(1.5940)
N	5715	5715	5715	5715

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

Second, the sign and magnitude of spousal impact on labour participation change with the levels of own health. When own health is relatively poor (the latent health is below -1 and -1.5 for men and women, respectively), better spousal health significantly decreases the probability of labour participation. The magnitude of the negative effect of spousal health on labour participation is decreased by own health improvements. Additionally, this negative spousal health impact on labour participation is bigger for men. Moreover, when own health improves (the latent health is above -0.5 and -1 for men and women, respectively), the effect of spousal health on labour participation becomes positive. For example, given very poor own health (latent health equals -3.5), a one-unit increase in spousal latent health brings an 11.30% and 4.88% decrease in the probability of labour participation for men and women, respectively. With poor own health (latent health equals -3), a one-unit increase in spousal latent health decreases men's and women's labour participation probability by 9.73% and 4.17%, respectively. Given fair own health (latent health equals 0), a one-unit increase in spousal latent health brings an insignificant 0.91% and 1.56% increase in labour participation probability for men and women, respectively.

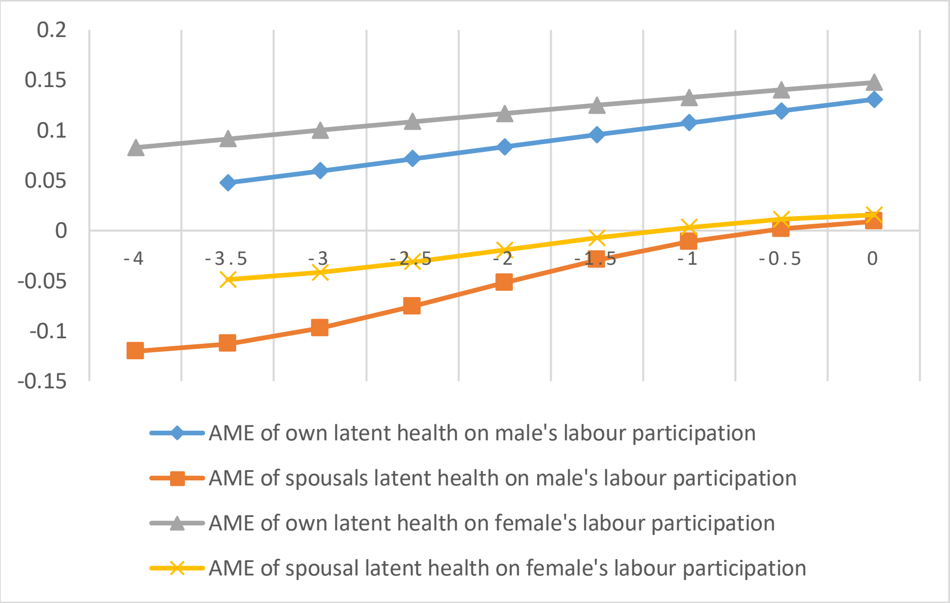


Figure 1.3 The AME of latent health on labour participation

1.5.2.2. Results for yearly work time model

In the yearly work time models, we observe 2424 couples. Appendix Table A6 summarises the variables' descriptive statistics. It shows that, on average, men have higher yearly work time than women in the sample.

Table 1.9 presents the results of health-related variables in the SUR models. Appendix Table A8 presents detailed results of controls. The correlation of residuals is approximately 0.3733. It is reasonable to estimate the equations of the husband's and the wife's labour supply systematically. Columns (1) and (2) show the respective coefficients for the men's and women's yearly work time equations without the couple's health interaction term. Columns (3) and (4) display the respective coefficients for men and women in the yearly work time equations with the couple's health interaction term.

Before adding the interaction term, better own health shows a positive and significant impact on yearly work time for both men and women. On average, a one-unit improvement of own latent health increases annual work time by 96.5 and 123.3 hours for men and women, respectively. In contrast, the impact of spousal latent health is negative but insignificant. However, when adding the interaction term, all health-related variables' effects seem to be insignificant, although we observe the negative sign for the interaction term.

Table 1.9 Yearly work time estimate

	Reduced		Interaction term	
	male	female	male	female
	(1)	(2)	(3)	(4)
own latent health	96.50**	123.30***	35.86	98.54
	(2.50)	(2.88)	(0.56)	(1.64)
spousal latent health	-36.16	-43.26	-84.67	-74.18
	(-0.88)	(-1.08)	(-1.46)	(-1.12)
own latent health *spousal latent health	-	-	-64.75	-33.10
	-	-	(-1.20)	(-0.59)
Correlation of residuals	0.3711		0.3710	
Breusch-Pagan test of independence	chi2(1) =	333.828	chi2(1) =	333.598
	Pr = 0.0000		Pr = 0.0000	
N	2424	2424	2424	2424
Controls: age, education, education gap, hukou, living area, household non-labour income, household structure				

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

Next, to investigate the influence of involving the interaction term, we calculate average marginal effects (AMEs) and draw a figure to illustrate them (Brambor et al., 2006). Overall, conditional on engaging in the labour market, better own health has significant and positive effects on increasing annual work time (Table 1.10 and Figure 1.4). This positive own health effect on annual work time is decreased by better spousal health. For example, given poor spousal health (latent health equals -2.5), a one-unit increase in own latent health brings a 197.72 and 181.28 hours increase in yearly work time for men and women, respectively. Given spousal latent health equals -1 (fair health), a one-unit improvement of own latent health increases annual work time by

100.60 and 131.64 hours for men and women, respectively. However, the effect of spousal health on annual work time seems insignificant.

Table 1.10 The AME of latent health on yearly work time

	Male		Female	
	AME of own latent health	AME of spousal latent health	AME of own latent health	AME of spousal latent health
	Given spousal or own latent health			
-3	230.10*	109.57	197.83	25.10
	(1.95)	(0.85)	(1.48)	(0.20)
-2.5	197.72**	77.20	181.28*	8.55
	(2.13)	(0.75)	(1.69)	(0.09)
-2	165.35**	44.83	164.73**	-7.99
	(2.38)	(0.57)	(2.00)	(-0.11)
-1.5	132.98***	12.45	148.18**	-24.54
	(2.70)	(0.22)	(2.46)	(-0.48)
-1	100.60***	-19.92	131.64***	-41.09
	(2.59)	(-0.46)	(2.92)	(-1.02)
-0.5	68.23	-52.29	115.09**	-57.64
	(1.51)	(-1.21)	(2.56)	(-1.23)
0	35.86	-84.67	98.54	-74.18
	(0.56)	(-1.46)	(1.64)	(-1.12)
N	2424	2424	2424	2424

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

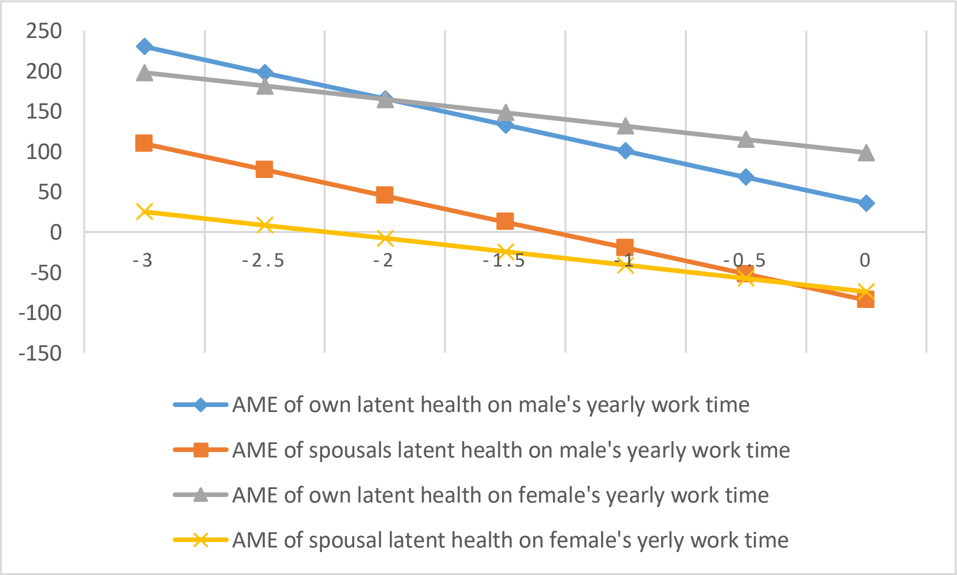


Figure 1.4 The AME of latent health on yearly work time

1.5.3. Sensitivity analysis

1.5.3.1. Different objective health measures

We further explore the sensitivity of the results to the construction of the latent health stock. We choose some different objective health measures, such as difficulty with ADLs and doctor-diagnosed health problems (Table 1.11), to estimate the latent health stock. Figure 1.5 and Figure 1.6 present the AMEs of own health/spousal health on labour participation/yearly work time (see Appendix: Table A9 and Table A10 for detailed information). The results remain robust to our main findings.

Table 1.11 Variable names and definitions in the new latent health model

Variables	Description
Self-assessed health	1.very poor; 2.poor; 3.fair; 4.good; 5.very good
6-items Activities of daily living (ADLs): Some difficulty	1 if difficulty reported, 0 otherwise. There are individual dummies for difficulties with：1.dressing; 2.bathing; 3.eating; 4.get in/out bed; 5.using the toilet; 6.controlling urination and defecation
The doctor diagnosed health problems: Ever Had Condition (13 items)	1 if the problem is reported, 0 otherwise. There are individual dummies for problems with: 1. high blood pressure; 2.diabetes; 3.cancer; 4.lungdisease; 5.heart problems; 6.stroke; 7.psych problem; 8.arthritis; 9.dyslipidemia; 10.liver disease; 11.kidney disease; 12.stomach; digestive disease; 13.asthma

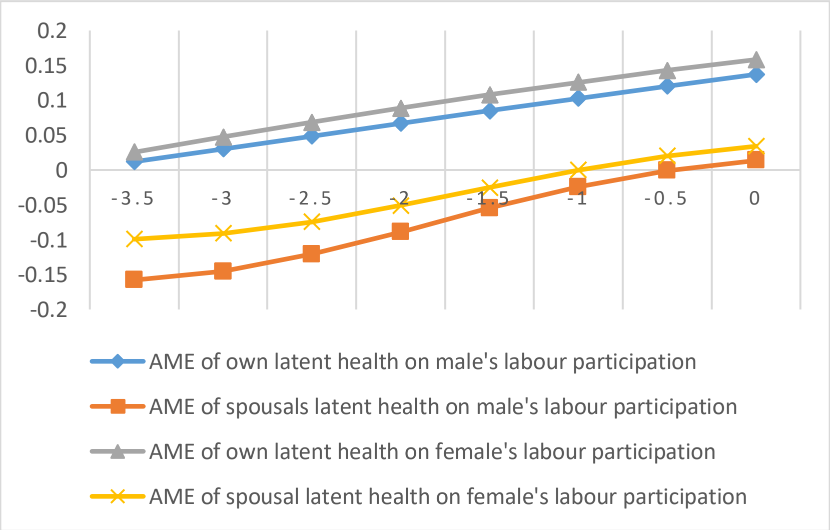


Figure 1.5 The AME of new latent health on labour participation

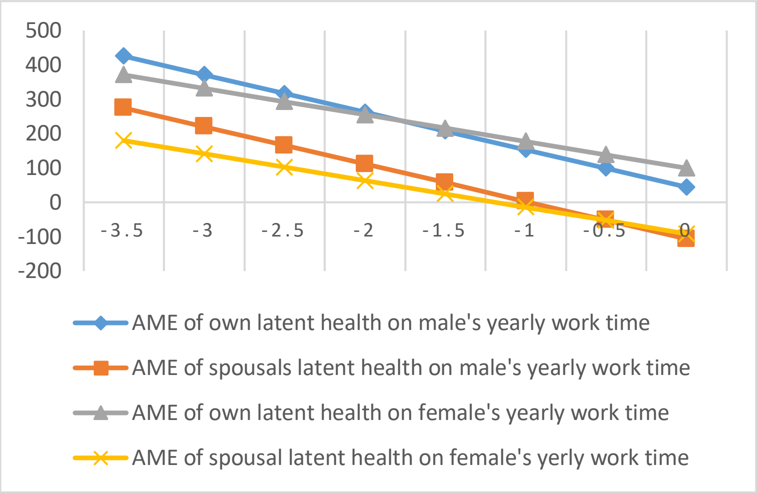


Figure 1.6 The AME of new latent health on yearly work time

1.5.3.2. Non-lagged variables

Figure 1.7 and Figure 1.8 provide the AMEs of own health/spousal health on labour participation/yearly work time when we use non-lagged variables to estimate labour supply (see Appendix: Table A11 and Table A12 for detailed information). The effects of non-lagged health-related variables on labour participation and yearly work time are essentially similar to those of one-period lagged health-related variables. However, because of potential simultaneity bias, we cannot trust the coefficient for the current health-related variables as much as the lagged ones.

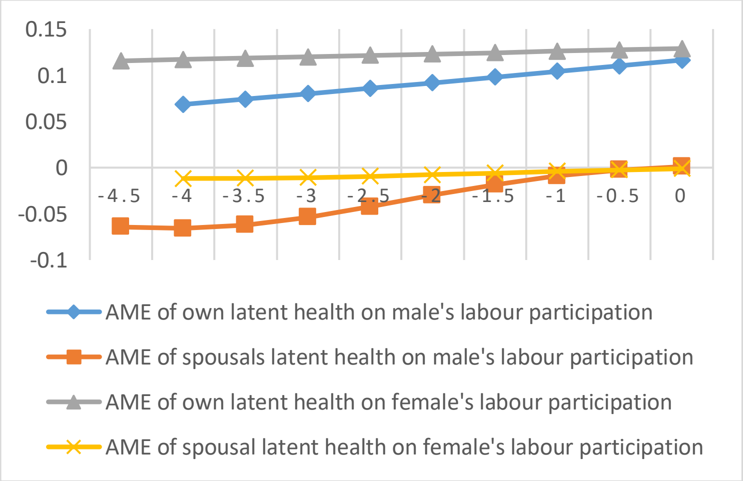


Figure 1.7 The AME of latent health on labour participation (non-lagged variables)

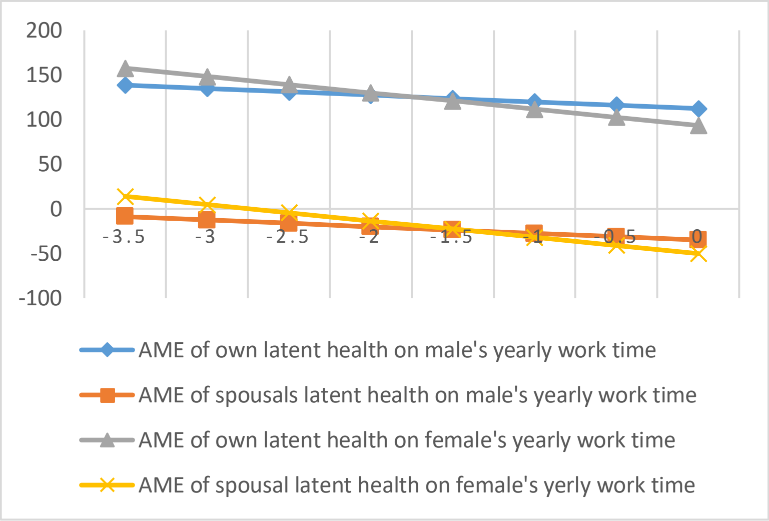


Figure 1.8 The AME of latent health on yearly work time (non-lagged variables)

1.6. Discussion and Conclusion

This study conducts a comprehensive analysis of the effect of health on the labour supply in the household. We investigate the impact of own health, spousal health and their interaction terms. The results reveal that it is worth taking all these health issues into account because of the clear link between the effects of own health and spousal health. The main findings show that improvements in own health significantly enhance the labour participation probability and that better spousal health amplifies this positive own health impact. Moreover, if own health is very poor or poor, better spousal health significantly decreases labour participation probability. The negative effect of spousal health on labour participation is reduced by own health improvements. In particular, when own health becomes relatively better, better spousal health is associated with a statistically insignificant increase in labour participation probability. Additionally, conditional on engaging in the labour market, better own health significantly increases annual work time. This positive impact of own health on annual work time is decreased by better spousal health. However, spousal health has a limited impact on annual work time.

Some studies find some degree of labour substitutability among the household members (Spletzer, 1997; Reis, 2007). Although there are significant earning losses associated with idiosyncratic health shocks at the individual level, intrahousehold labour substitution can attenuate their impact at household-level aggregates. Thus, there may be evidence for a compensating increase in the individual labour supply due to spousal poor health (the added worker effect) (Charles, 1999). In contrast, some researchers argue that there may be complementarity rather than substitution in spousal leisure, and a negative health shock could strengthen the complementarity of leisure if the affected spouse requires assistance with ADLs (and the family prefers to have the spouse provide this care—informal caregiver effect) or the affected spouse has a shortened life expectancy (Jeon and Pohl, 2017; Macchioni Giaquinto et al., 2022). The existing literature distinguishes the substitutability of labour supply between partners (the added worker effect) from the complementarity of leisure according to how to respond to poor spousal health and negative spousal health shocks. This provides insights into interpreting the results of this study.

Regarding middle-aged and elderly people in rural China, most of them engage in the informal labour market and there is no statutory retirement age. Households in rural China engage in home production, which may include domestic chores, own-farm work, raising livestock, and non-farm activities and is characterised by a high degree of specialisation or division of labour based on the age and gender of the family members. It could be interesting to explore whether there is complementarity or substitutability between the couple's labour supply in rural China. One of our findings indicates that the same one-unit of improvements in own latent health brings smaller increases in labour participation probability for individuals with very poor spousal health than those individuals with poor or fair spousal health. To some extent, this is similar to the complementarity of leisure existing on extensive margins of labour supply. Additionally, the same one-unit of improvements in own latent health brings greater increases in annual work time for individuals with poor spousal health than those individuals with fair spousal health. This reveals a similar spirit of 'the added worker effect' existing on intensive margins of labour supply. Moreover, the impact of spousal health on labour supply depends on the levels of own health. When own health is very poor or poor, better spousal health significantly decreases the labour participation probability. In other words, deterioration of spousal health brings increases in labour participation probability, which is similar to showing 'the added worker effect'. When own health becomes fair, better spousal health is associated with an increase in labour participation probability although it's not statistically significant, which is similar to indicating limited complementarity of leisure. Those findings present that there are differences between extensive and intensive margins of labour supply in terms of own and spousal health impact. In addition, whether complementarity of leisure or substitutability of labour supply plays the dominant role depends on not only the spousal health but the own health.

Our findings are consistent with the existing literature revealing own health's positive impacts on labour supply (Jiang et al., 2019). We also provide new insights into the significant moderating effects of spousal health. From a policy perspective, there is an important implication. China has witnessed a rapid demographic transition since the middle of the 20th century. Decreased fertility and an increase in life expectancy quickly led to a dramatic ageing of China's population (Ning et al., 2016). In addition, China has also seen dramatic changes in its labour market over the past few decades, with hundreds of millions of working-age farmers moving from rural to urban areas (Démurger and Li, 2013). Rural migrants working in China have been regarded as cheap labour and for a long period of time, they have been the power behind China's industrialisation and rapid economic growth (Xu, 2017). Those rural migrants contribute greatly to rural and urban development in China when they try to maximise the welfare for themselves and their families (Xu, 2017). However, many women, children and elderly parents in migrant families are left behind in rural areas. The continuous rural-urban migration forces those disadvantaged left behind family members to be the main agricultural labour force. The rapidly ageing population and the large migrant flow from rural to urban areas contribute to the constant reduction in the agricultural labour supply in rural China. Accordingly, labour inputs in agricultural production are becoming ever more dependent on the elderly. It seems urgent to alleviate the labour supply shortage in rural areas as it's related to the agricultural development and food security of the country (Jiang et al., 2019). Some researchers find that a feasible extension of working lives requires that older workers are physically and mentally capable of working (Giles et al., 2023). In response to the lack of skilled and efficient agricultural labour force, it is important to improve mid-aged and elderly rural resident's health and then incentivise them to provide productive labour supply. As healthy workers should be more productive than sick ones, mid-aged and elderly rural residents should be motivated to consistently invest in their health stock (Behncke, 2012). For instance, having regular physical examinations, investing more time in daily exercise and developing healthy behaviours may be helpful. The government could improve resident's health by providing affordable and accessible healthcare services, developing education on health literacy, and investing in sports venues or leisure facilities for exercise.

Moreover, if own health is very poor or poor, we find that better spousal health significantly decreases labour participation probability. Generally, poor health status has negative effects on labour supply (Bound, 1999; Coile, 2004; Behncke, 2012) and is associated with exit from work (Disney et al., 2006). However, rural elderly are exposed to the risks of economic vulnerability and poverty in China. A severe shortage of institutionalized risk-sharing mechanisms like sufficient public social security programs leaves them little or no choice other than to continue their work intensity even with illnesses (Cai et al., 2012; Smith et al., 2014). A recent study used hypertension as a health measure to investigate the impact of health on the labour supply of the Chinese elderly, where the results indicate that hypertension has significantly negative effects on the urban elderly but no effect on the rural elderly (Li, Lei and Zhao, 2014). One explanation provided by the study is that considerable urban-rural differences exist in the level of coverage by safety nets and the benefits received through the social welfare system. The urban elderly with their better covered through social security systems have incentives to retire early, while older rural Chinese have traditionally kept working as long as their health permits (Smith et al., 2014). The limited payment amount of pension is not large enough to cause a dramatic work disincentive for the rural elderly (Ning et al., 2016). Our findings reveal that spouses play an active role in risk sharing in rural families with limited social security support, especially when household work capacity is limited due to poor health status. From a policy perspective, although extending working lives and harnessing the human capital of the rural older population may ease some of the burden of population ageing and continuous out-migration from agriculture, it is inhumane to force the ailing elderly to continue to work. The conflict between labour supply and the welfare of the elderly requires further attention. Policymakers need to harmonize these policy interests and delineate the target population through strategic policy guidance (Jiang et al., 2019). One possible policy implication involves finding ways to incentivize mid-aged and young elderly people with health capacity to work to extend their working lives. Additionally, expanding social security support for rural elderly is also necessary, especially for the very old people or the elderly with health fragility.

This paper provides new evidence of the health impact on the labour supply in the household. The findings show how the effects of own health and spousal health interact with each other. However, after constructing latent health stock by using objective health measures, variances of health among rural middle-aged and elderly couples are found to be smaller than the variances of self-reported health because most of our samples fall in the range of poor or fair health. Therefore, our empirical results are limited to a certain range of latent health stock. To paint the full picture of health and labour supply in the household, further work can be conducted using a potentially available dataset with a wide range of health variances.

2. Chapter 2

Evaluating the Effect of the Hierarchical Medical System Reform on Primary Healthcare Institution Utilisation in China

Abstract

The Chinese healthcare system faces substantial challenges in its transformation from a profit-driven public hospital-centred system to an integrated primary care-based delivery system. The government launched a Hierarchical Medical System (HMS) reform in September 2015, and this reform was gradually implemented nationwide by the end of 2017. This study aims to evaluate the effect of the HMS reform on the capacity and utilisation of primary healthcare institutions, which is one of the reform priorities. Panel data is derived from the China Health Statistics Yearbook (2012-2017). In our approach, the Bacon-decomposition method is introduced to highlight how the standard two-way-fixed-effect (TWFE) difference-in-difference (DID) estimator would be biased in the staggered intervention set-up. Next, our empirical models make use of recent methodological extensions to allow heterogeneous treatment effects. In detail, we follow Wooldridge (2021) and Callaway and Sant' Anna (2021) to get the extended TWFE estimator and the doubly-robust estimator, respectively. The main findings reveal that the HMS reform is effective in increasing the proportion of nurses and practitioners to total workers in urban primary healthcare institutions but ineffective in rural areas. In addition, the reform significantly enhances the utilisation of rural primary healthcare institutions, including bed occupancy rate, the proportion of outpatient visits and inpatient admissions to total visits in medical institutions. Moreover, with staggered reform implementation, there is treatment effect heterogeneity across cohorts and periods.

Keywords

Hierarchical Medical System Reform; primary healthcare institution; Staggered intervention; treatment effect heterogeneity

2.1. Introduction

Healthcare systems worldwide face challenges in providing effective and efficient care to match the increasing needs arising from accelerated population ageing and the spread of chronic and infectious diseases (Hu et al., 2023; Li et al., 2020). Health services in some high-income countries are delivered in a hierarchical medical system (HMS) and through the mandatory gate-keeping mechanism that involves initial diagnoses at primary care facilities and obligatory two-way referrals among hospitals (Forrest, 2003; Brekke et al., 2007; Xiao et al., 2021). Experience from those developed countries reveals that orienting a healthcare system towards primary care can enhance the continuity and coordination of care, reduce the inappropriate use of speciality services and promote a more cost-effective and higher-quality healthcare delivery system (Hu et al., 2023; Yip and Hsiao, 2014). Unlike in those countries, there is no mandatory first contact in primary healthcare institutions in China. Additionally, before 2015, the healthcare system did not operate with a patient referral network. China has a three-tier healthcare system: primary healthcare institutions, and secondary and tertiary hospitals[5]. Primary healthcare institutions consist of primary hospitals and some unrated healthcare facilities, which directly provide essential healthcare services to all communities. The public embraced the freedom of choice at all tiers of healthcare facilities. There is an upwardly concentrated allocation of medical resources among different levels of healthcare facilities. Evidence indicated that most patients in China increasingly bypassed primary healthcare institutions and accessed the health system at higher-level hospitals when they required healthcare services, resulting in extremely overcrowded higher-level hospitals all over the country (Wang et al., 2020b; Zhou et al., 2021).

The Chinese government faces substantial challenges in guiding patients to make rational choices about different levels of healthcare institutions and improving the utilisation efficiency of medical resources. Promoting the usage of primary healthcare

[5] The Chinese Ministry of Health defines hospitals as “medical institutions having more than 20 beds”.

facilities is a priority of the systemic health reform initiated in China in 2009. Both central and local health authorities have invested enormous resources into primary care facilities. There was a dramatic increase (by 31.79%) in outpatient visits from 2010 to 2015, with varying degrees of growth in the number of patients visiting all types of facilities. However, there was no convincing evidence showing a shift in patient flow from higher-level hospitals to primary care facilities. There was even a significant decrease in the proportion of outpatient visits in primary care institutions to total outpatient visits, from 61.87% in 2010 (3.61 billion of 5.84 billion total visits) to 54.12% (4.34 billion of 7.69 billion total visits) in 2015. In contrast, the proportion of inpatient visits in primary healthcare facilities to total visits increased from 34.94% to 40.08%. Some studies have explored why the government's efforts at this national systemic reform did not seem to be leading to an ultimately successful outcome for encouraging healthcare usage in primary institutions. There are multiple possible causes and major historical and institutional factors involved. First, the public lacks trust in practitioners working at primary health facilities. This lack of trust is associated with the relatively low average educational attainment of those practitioners compared with their counterparts in high-level hospitals and with severe maldistribution in the number of licensed doctors or licensed assistant doctors among different levels of hospitals (Wu and Lam, 2016). Second, the public's lack of trust in the quality and capacity of primary care facilities is apparent and can be partly explained by the former low trust in the doctors. It is also linked to the substantial disparities in available drug varieties and infrastructure, such as buildings and medical equipment, in the tiered healthcare system (Wu et al., 2017; Liu et al., 2018b). Moreover, the absence of a gate-keeping function by primary care institutions and the lack of an effective referral system make it common for Chinese patients to bypass primary care to higher-level facilities regardless of disease type and severity (Liu et al., 2018c; Li et al., 2020). Furthermore, economic boosts and the fast development of the public transportation system in the country are contributing to patients' access to large hospitals and the patient flow across different areas (Wu and Lam, 2016).

The dramatic and lasting underutilisation of primary healthcare institutions has created a huge obstacle to building an effective healthcare system in China. This is why the government implemented the HMS reform in September 2015, which aimed to establish a two-way referral system and relieve the current pressure on second and third-tier hospitals by strengthening primary healthcare institutions' capabilities. The goal of the HMS reform is that patients should be treated at different levels of hospitals according to patients' conditions. In the HMS, patients are encouraged to go to primary healthcare institutions first when they need to visit doctors. Patients are referred to a higher-level hospital when treating their conditions is beyond the ability of the primary facilities. The HMS reform was gradually implemented throughout the country by the end of 2017. The existing literature about HMS has mainly focused on theoretical and descriptive analysis (Xu and Mills, 2017; Feng et al., 2022; Li et al., 2020; Xu et al., 2021). Those studies reveal that the absolute value of health service provision by the primary healthcare institutions has increased significantly, but the proportion of this health service provision in the whole healthcare system has continued to decline. There is scarce empirical evidence of the effects of the HMS reform. The interest in reform outcomes and the samples chosen for evaluation vary across these studies. Hu et al.(2021) investigate the impact of the HMS reform among chronic disease patients in Xiamen City using the propensity score matching and difference-in-difference (DID) methods. The findings show the effectiveness of the reform in health improvement and cost savings for chronic disease management. Zhou et al.(2021) use panel data from the China Family Panel Studies (CFPS) and employ the DID model to evaluate the effect of the HMS reform on health-seeking behaviour in China. The results indicate that the reform positively affected the probability of urban residents going to primary care facilities for contact. Basic health insurance was a significant factor in directing residents to primary care facilities.

Regarding research methods, recently, the use of the standard TWFEDID method with multiple periods has come under considerable scrutiny because of a mismatch between the model specifications and the underlying treatment homogeneity assumption. With multiple periods and staggered interventions, some recent literature builds on characterisations of the nature of the TWFE estimator and uses different decomposition methods to show what the TWFE method actually estimates. Goodman-Bacon (2021) shows that the 'static' TWFE estimator equals a weighted average of all possible two-group/two-period DID estimators that compare timing groups to each other. Some of these groups use an earlier-treated group as a control for a later-treated group. The weights on the 2x2 DIDs are proportional to timing group sizes and the variance of the treatment dummy in each pair, which is highest for units treated in the middle of the panel. Units treated in the middle of the panel are the most influential part of the summarised TWFE coefficient for no other reason than that TWFE weighs up the central treatment groups. The Bacon decomposition highlights this strange role of panel length and shows how the standard TWFE estimator is biased when effects change over time. de Chaisemartin and D'Haultfouille (2020) employ a different decomposition theorem to write the 'static' TWFE estimator as a weighted average of treatment effect parameters, some of which may have negative weights. The negative weights are an issue when the treatment effects are heterogeneous across groups or periods. The ATT may have the opposite sign than the TWFE coefficient. In addition to those decompositions focusing on the static TWFE specification, Sun and Abraham (2021) propose a decomposition of TWFE dynamic specification in the event study setting. They express relative period coefficients for ATT as a linear combination of cohort-specific effects (the ATT for a particular treatment cohort at a particular event-study relative period) from its own relative period and other relative periods. Terms that include treatment effects from other relative periods will contaminate the estimator if the treatment effect homogeneity assumption does not hold.

To avoid the relevant pitfalls of standard TWFE estimators with multiple periods, recent methodological extensions for staggered interventions allowing arbitrary treatment effect heterogeneity mainly focus on two aspects: finding alternative estimators (Sun and Abraham, 2021; Borusyak et al., 2024; Callaway and Sant'Anna, 2021) and extending the basic TWFE estimator (Wooldridge, 2021). In essence, each newly proposed estimator modifies the units that can act as effective comparison units to avoid comparing treatment units to inappropriate controls (Baker et al., 2022). However, the estimators differ in terms of which observations may serve as effective control units, how covariates are incorporated, and how flexible the covariates are. When there are no never-treated units, Sun and Abraham (2021) use the last-treated units as controls, whereas the other three scholars choose the not-yet-treated as comparisons. Additionally, Callaway and Sant' Anna (2021) only allow pretreatment covariates to be controlled; both time-invariant and time-varying covariates could be added flexibly in Borusyak et al.'s (2024) imputation estimator. Moreover, Callaway and Sant' Anna (2021) provide a flexible set of aggregations of ATTs by cohorts, periods, and relative periods, whereas Sun and Abraham (2021) and Borusyak et al. (2024) only generate aggregations by event-study relative periods. Other kinds of aggregations need to be constructed manually. Furthermore, Wooldridge (2021), different from other scholars, proposes other interesting treatment effects with causal interpretations in addition to ATT that can be identified through pooled OLS or extended TWFE regressions.

Previous studies highlighting the challenge of underutilising primary healthcare institutions in China provide an overview of the HMS reform's background. Existing literature evaluating the HMS reform reveals some interesting evidence of the reform efforts. However, it is unclear whether empirical results using data from a particular region are broadly representative of China because there are differences among the policies in each province. Additionally, pieces of evidence from a restricted sample, such as patients with chronic diseases, may be difficult to generalise among all kinds of patients because the behaviour of patients with chronic conditions may diverge from the choices of patients with other illnesses. Moreover, according to the rich body of research making methodological extensions to DID with multiple periods and variations of treatment timing, studies using the standard TWFEDID method to investigate the reform's impact could be biased. With the staggered implementation of the HMS reform across the whole country, treatment heterogeneity is most likely.

However, the conventional TWFEDID estimate relies on implying treatment homogeneity assumption. This paper aims to evaluate the effect of the 2015 HMS reform on the capacity and utilisation of primary healthcare facilities. It makes two main contributions to the existing literature: on the one hand, it evaluates the reform's impact at the national level and involves comprehensive indicators for primary healthcare facility outcomes. We choose the proportion of nurses/practitioners to total workers in primary healthcare institutions as the capacity indicators. Bed occupancy rate and the proportion of outpatient/inpatient health service provision of the primary healthcare institutions in the whole healthcare system are used as utilisation indicators. On the other hand, we employ the new staggered DID methods to allow arbitrary treatment effect heterogeneity and provide more accurate estimators. We also conduct a detailed analysis of the possible sources of treatment effect heterogeneity. In our approach, the Bacon-decomposition method is introduced to highlight how the standard TWFEDID estimator is biased. Then, we follow Wooldridge (2021) and Callaway and Sant' Anna (2021) to define and estimate the cohort-period specific ATTs. The main findings show that the HMS reform is effective in enhancing primary healthcare institutions' capacity in urban areas but ineffective in rural areas. Additionally, the impact of the HMS reform on primary healthcare institution utilisation is positive and significant in rural areas but not in urban areas. Furthermore, there is heterogeneity within cohorts and time periods: within the cohort, ATT increases based on longer exposure to policy intervention. Given the same length of exposure, ATT for the later treated cohort is greater than that for the earlier one.

2.2. Conceptual Framework

2.2.1. Institution background

2.2.1.1. The three-tiered healthcare delivery system in China

Hospitals in China are defined as 'medical institutions having more than 20 beds' and are divided into three levels based on the different tasks and functions they perform

(Figure 2.1). Primary healthcare institutions, providing basic healthcare services to all communities directly, include primary hospitals and some unrated healthcare facilities (which do not meet the definition of a hospital). Apart from primary healthcare institutions, secondary and tertiary hospitals also offer primary healthcare services. The general population can choose healthcare facilities without a mandatory gate-keeping mechanism restricting them. The primary healthcare institution system can be further divided into urban and rural parts, which are organised differently. In urban areas, primary healthcare institutions include urban community health centres and community health stations. In rural areas, primary healthcare institutions consist of township health centres and village clinics. Most community health stations and village clinics mainly provide outpatient services and rarely offer inpatient services.

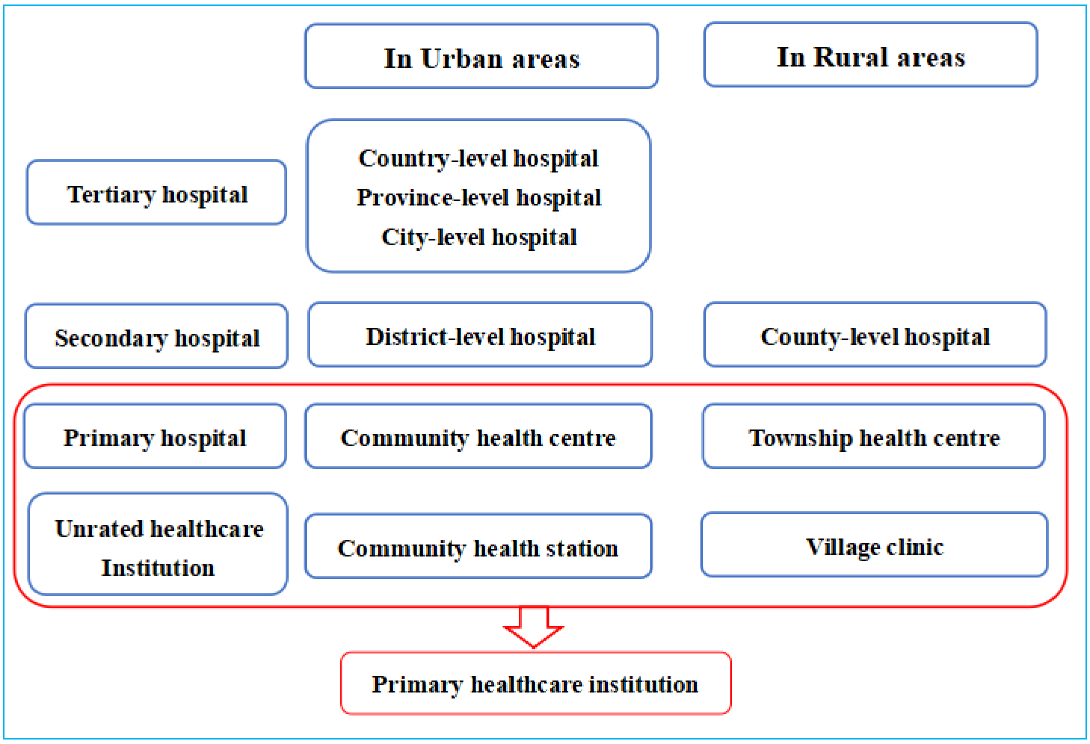


Figure 2.1 The three-level hospital system plus primary healthcare institutions in China

2.2.1.2. Provincial divisions of China

China’s provincial level (first-level) subdivisions consist of 23 provinces, five autonomous regions, four direct-administered municipalities, and two special administrative regions (Table 2.1).

Table 2.1 Provincial divisions of China

Province	A standard province.
Autonomous region	Regional autonomy for ethnic minorities in China means that, under the unified leadership of the state, regional autonomy is practised in areas where people of ethnic minorities live in compact communities. In these areas, self-government organs are established to exercise autonomy.
Direct-administered Municipality	A higher level of city directly under the Chinese government, with status equal to that of the provinces. Their political, economic and cultural status is usually higher than that of common provinces.
Special administrative region	A highly autonomous and self-governing sub-national subject of the People's Republic of China.

2.2.2. Reform background

The Chinese government launched a systemic health reform in 2009, intended to achieve universal coverage of ‘safe, effective and affordable basic healthcare services’ for all Chinese citizens by 2020. One of the main objectives of the reform was developing the primary healthcare system. During the first three years of the reform, the government invested about CNY 1409.9 billion (US$ 206 billion) in the healthcare system, of which about 44% was allocated to primary healthcare institutions (Feng et al., 2022). Enormous resources were poured into primary care infrastructure establishments, such as buildings and basic medical equipment (Wu and Lam, 2016). Compared with the 2015 HMS reform, the 2009 national reform mainly focused on establishing new primary healthcare institutions, especially in less-developed areas. This provided a sound basis for the following detailed actions to promote the development of primary care facilities.

The Chinese government issued the ‘Guiding Opinions of the General Office of the State Council on Pushing Forward the Building of the Hierarchical Medical System (HMS)’ in September 2015. HMS refers to the fact that different levels of hospitals have a clear division of labour and are responsible for undertaking different health services. Unlike the mandatory gatekeeper system established by legislation in the

United Kingdom and Germany, China’s HMS guides patients in choosing the primary healthcare institutions for first contact by imposing economic measures. There are five main aspects of detailed actions related to primary healthcare: (1) increasing subsidies for the infrastructure and workforce resources in primary healthcare institutions−the subsidies could be used to upgrade equipment, develop infrastructure, and promote job training; (2) changing incentives in primary healthcare institutions, such as introducing performance-based salaries; (3) building medical alliances, that is, cooperative associations of medical institutions in certain regions instead of strict regulations on procedures of diagnosis and treatment, flexible resource sharing is encouraged among different grades medical institutions inside each medical alliance (Sun et al., 2019); (4) introducing gradient reimbursement schemes of basic health insurance−a higher reimbursement rate has been set for primary healthcare institutions rather than large hospitals; (5) introducing a family doctor registration policy on a voluntary basis to maintain the continuity of healthcare management.

The Chinese government issued the ‘Opinions on Deepening the Hierarchical Medical System Reform’ in August 2018. They advocated several major aspects for accelerating further development of the HMS system. Compared with the 2015 HMS reform, the 2018 reform particularly highlights ‘Building medical alliances’ to promote a more balanced allocation of healthcare resources.

2.2.3. Mechanisms and hypotheses

Mechanism analysis is conducted under the guiding framework of the behavioural model of health services use, which was developed in the late 1960s and evolved over time (Andersen and Newman, 1973; Andersen, 2008). Andersen’s model is one of the most classic and comprehensive conceptual frameworks for understanding multiple dimensions of access to and utilisation of healthcare. This model suggests that healthcare utilisation is determined by three key factors, predisposing, enabling, and need, at both the individual and contextual levels (Andersen, 1995, 2008). At the individual level, predisposing factors include demographic characteristics (e.g., age, sex), socioeconomic characteristics (e.g., education, social class, and employment status), and health beliefs (e.g., attitudes, values, and knowledge of health and health services). Individual enabling factors refer to resources or means that enable individuals to obtain health services. They usually involve individual and community resources, such as health insurance, income, and availability of services. Individual need factors, conceptualised as needs perceived by the individual or needs evaluated by professionals, are the most direct and vital factors affecting health service utilisation. Contextual factors are measured at some aggregate rather than individual level. In Andersen's framework, contextual factors are also classified into three categories: contextual predisposing factors (e.g., community demographic, social, and belief factors), contextual enabling factors (e.g., supply of medical personnel and facilities), and contextual need factors (e.g., environmental and population health indices). There is considerable Chinese empirical evidence to support this behavioural model (Liu et al., 2019; Huang et al., 2019; Zeng et al., 2020). Some recent studies also reveal different health-seeking preferences for urban and rural residents in China (Qian et al., 2009; Wu et al., 2017; Liu et al., 2018b). The possible reasons for this difference are disparities in income, education, health literacy, travel distance to higher-level facilities and the relative importance attached to quality of care from a value aspect by urban and rural residents (Sun et al., 2013).

The detailed measures of the 2015 HMS reform can affect individual and contextual enabling factors and then impose effects on the usage of health services in primary health facilities (Figure 2.2). Regarding individual enabling factors, on the one hand, gradient reimbursement schemes of basic health insurance are able to reduce out-of-pocket health expenditure if patients choose to visit primary health institutions. On the other hand, voluntary family doctor contract services are assumed to alter health management from self-management to management by the professional family doctor team. This is helpful to maintain the continuity of healthcare. Regarding contextual enabling factors, that is, financial support for the infrastructure and workforce in primary health facilities, incentives' adjustment and the medical alliance are combined to improve the capacity and quality of health services in primary care facilities. Because of data availability restrictions, in this study, we only test the reform's effect on contextual enabling factors (workforce) and its overall impact on utilisation in primary care institutions. Hence, the following specific hypotheses are proposed:

Hypothesis 1: The HMS reform effectively improves primary healthcare institutions' capacity and quality.

Hypothesis 2: The implementation of HMS is effective for enhancing the utilisation of primary healthcare institutions.

Hypothesis 3: The effect of HMS may be different in primary healthcare institutions in urban and rural areas.

Furthermore, given the staggered policy implementation across the whole country, there are some possible sources for heterogeneous treatment effects. First, the composition of each treated cohort is different as some cohorts include direct-administered municipalities or autonomous regions. This creates disparities in local economic conditions and political power. Second, the implementation of the reform displays heterogeneity across provinces. Specific actions taken to achieve the HMS vary by region. Moreover, because the nationwide reform was launched by the central government in 2015 September, it is not clear whether local governments can translate its final goal into effective policies and procedures.

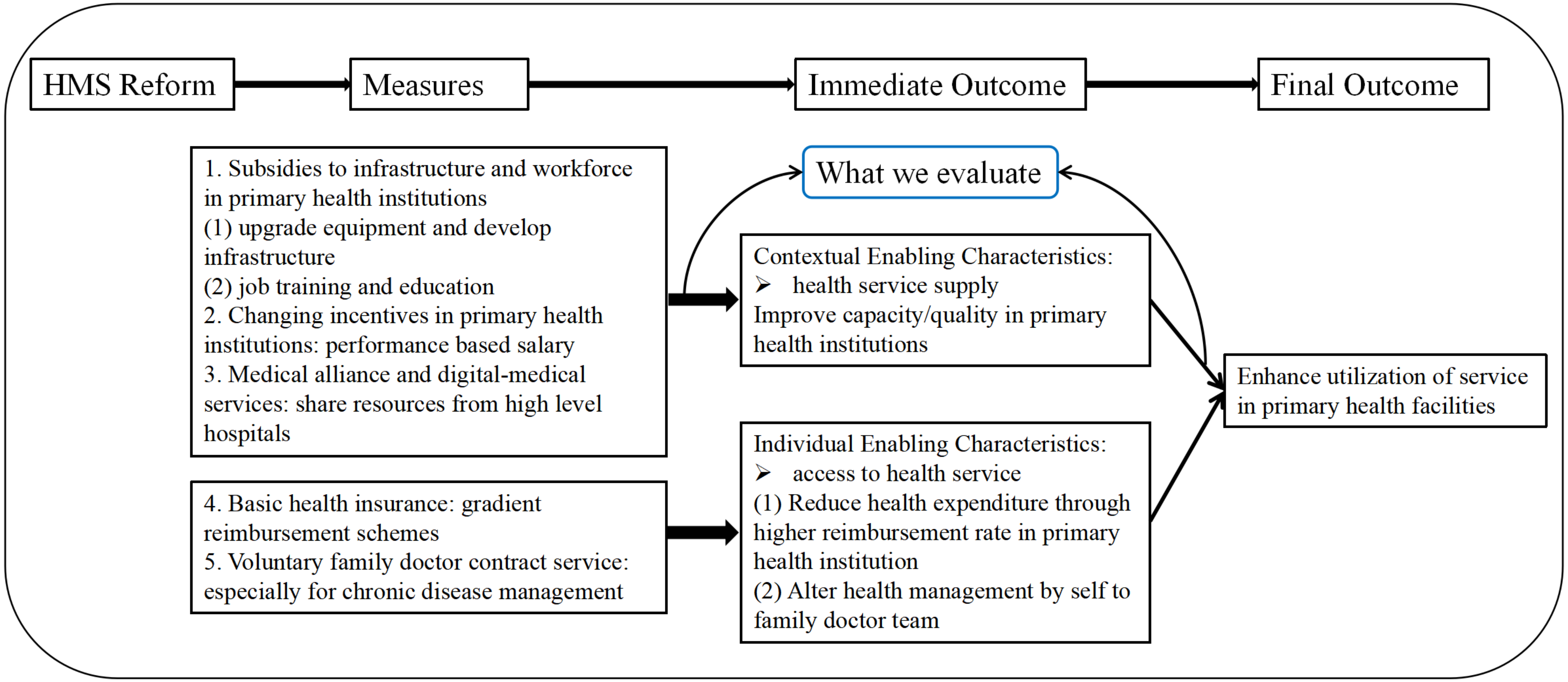

HMS Reform
Measures
Immediate Outcome
Final Outcome
1. Subsidies to infrastructure and workforce in primary health institutions
(1) upgrade equipment and develop infrastructure
(2) job training and education
2. Changing incentives in primary health institutions: performance based salary
3. Medical alliance and digital-medical services: share resources from high level hospitals
4. Basic health insurance: gradient reimbursement schemes
5. Voluntary family doctor contract service: especially for chronic disease management
What we evaluate
Contextual Enabling Characteristics:
➢ health service supply
Improve capacity/quality in primary health institutions
Individual Enabling Characteristics:
➢ access to health service
(1) Reduce health expenditure through higher reimbursement rate in primary health institution
(2) Alter health management by self to family doctor team
Enhance utilization of service in primary health facilities

Figure 2.2 The mechanism for the effect of the HMS reform on primary healthcare utilisation

2.3. Data

2.3.1. Data source and variables

Some researchers use data from resident electronic health records (Hu et al., 2021) and medical claims systems in a certain city or province in China (Xu et al.,2021) to evaluate the impact of hierarchical medical reform on health outcomes and healthcare spending among chronic disease patients. However, the Data including rich information is confidential and not readily accessible. Zhou et al (2021) use data derived from China Family Panel Studies (CFPS) 2012, 2014, 2016 and 2018. The CFPS survey data also include rich information related to healthcare and health, however, for potential outcome variables, the survey data only include the type of healthcare facilities that residents usually approach when seeking health services by the corresponding question "Where do you usually go to seek health services when you are sick". This outcome variable lacks a detailed division between outpatient and inpatient services and the self-reported variable may be a concern due to reporting bias.

Considering data availability and the richness of outcome variables, this study uses highly aggregated provincial datasets derived from the China Health Statistics Yearbook (2012-2017). The datasets contain information on health resources and health services of different kinds of medical institutions in China. We include five outcome variables for primary healthcare institutions: two capacity indicators based on human resources (nurse and practitioner proportion) and three utilisation indicators (outpatient visit proportion, inpatient admission proportion and bed occupancy rate). Table 2.2 provides detailed outcome variable definitions.

Table 2.2 Outcome variable names and definitions

Outcome variables	Definition
Nurse proportion	The proportion of nurses to total workers in primary healthcare institutions
Practitioner proportion	The proportion of practitioners to total workers in primary healthcare institutions
Outpatient visit proportion	The proportion of outpatient visits in primary healthcare institutions to total outpatient visits in health institutions
Inpatient admission proportion	The proportion of inpatient admission in primary healthcare institutions to total admission in health institutions
Bed occupancy rate	The number of beds occupied by patients/total beds

Table 2.3 displays the variables' descriptive statistics. Overall, on average, 18.6% of workers are nurses and 30.9% of workers are practitioners in primary healthcare institutions. Outpatient visits in primary healthcare institutions account for 55.4% of total outpatient visits in all kinds of healthcare facilities. For inpatient admissions, 17.4% of total admissions are in primary health facilities. Primary healthcare institutions in urban areas have higher nurse and practitioner proportions than in rural areas. Nurse and practitioner proportions are the lowest in village clinics, which are 8.03% and 23.1%, respectively. The outpatient visit proportion is higher in rural areas than in urban areas. Village clinics have the highest proportion of such visits (24%), while community health stations have the lowest proportion of such visits, which is only 2.17%. Additionally, the inpatient admission and bed occupancy rates in township centres are higher than in community centres. Notably, the inpatient admission proportion in community health centres is only 1.43%, which is much smaller than the proportion in township centres (16.9%). We have controlled variables for socioeconomic and demographic characteristics, such as per capita GDP, illiteracy ratio, sex ratio and elderly proportion.

The total sample size is 162, including 27 provinces across 6 years, and indicates that the sample size of outcome indicators for the community health station and township health centre is slightly different than that used for other outcome variables (such as outcome indicators for primary healthcare institution) The reason is the reclassification of statistical indicators in two cities, Beijing and Shanghai, which took place in 2010. In both cities, the statistics for township health centres were no longer counted separately and were merged with statistics for community health centres. Accordingly, outpatient visits in community health centres also contain outpatient visits in township health centres. Therefore, the sample size of outpatient visit proportion in community health centres remains 162 and the sample of outpatient visit proportion in township health centres reduces to 150. Similarly, statistics for community health stations were merged with statistics for village clinics in Shanghai, again leading to a sample size reduction of outcome indicators in community health stations. For example, outpatient visit proportion in community health stations. The evaluation period of this study is from 2012 to 2017. This reclassification has little impact on the evaluation.

Table 2.3 Descriptive statistics of variables

Variable	Obs	Mean	Std.Dev.	Min	Max
Nurse proportion					
Primary healthcare institution	162	0.186	0.0448	0.0846	0.304
Community health centre&station (Urban)	162	0.313	0.0379	0.214	0.416
Township health centre	150	0.227	0.0452	0.109	0.315
Village clinic	162	0.0803	0.0465	0.0140	0.275
Rural primary health institution	150	0.151	0.0461	0.0471	0.290
Practitioner proportion					
Primary healthcare institution	162	0.309	0.0523	0.210	0.430
Community health centre&station (Urban)	162	0.357	0.0295	0.290	0.425
Township health centre	150	0.347	0.0611	0.228	0.489
Village clinic	162	0.231	0.126	0.0576	0.784
Rural primary health institution	150	0.272	0.0570	0.182	0.430
Outpatient visit proportion					
Primary healthcare institution	162	0.554	0.0963	0.301	0.724
Community health centre	162	0.0649	0.0722	0.0124	0.345
Community health station	156	0.0217	0.0112	0.000283	0.0487
Township health centre	150	0.144	0.0442	0.0538	0.251
Village clinic	162	0.240	0.119	0.0153	0.497
Urban primary health institution	156	0.0761	0.0500	0.0264	0.243
Rural primary health institution	150	0.402	0.116	0.137	0.627
Inpatient admission proportion					
Primary healthcare institution	162	0.174	0.0917	0.00724	0.394
Community health centre	162	0.0143	0.00874	0.00183	0.0447
Township health centre	150	0.169	0.0806	0.0283	0.378
Bed occupancy rate					

Community health centre	162	0.511	0.147	0.189	0.977
Township health centre	150	0.558	0.130	0.283	0.811
Control variables					
Per capita GDP (CNY 10000/US $1587)	162	5.349	2.421	1.136	12.89
Illiteracy ratio (%)	162	4.658	2.350	1.230	13.01
Sex ratio (reference: female)	162	1.053	0.0375	0.974	1.204
Elderly proportion	162	0.101	0.0179	0.0637	0.143

2.3.2. Sample restrictions and evaluation set-up

This study aims to evaluate the effect of the 2015 HMS reform on primary healthcare institution outcomes, especially on the utilisation of primary healthcare institutions. To capture the explicit impact of the 2015 HMS reform without contamination by other related health reforms, the evaluation period should be restricted with care. According to the conceptual framework, one of the priorities for the first three years of the 2009 health reform was strengthening primary healthcare system's capacity, which may have affected the utilisation of primary healthcare institutions. Additionally, one priority of the 2018 HMS reform was building medical alliances, which may have also affected the utilisation of primary healthcare institutions because medical alliances provide opportunities for such institutions to share health resources from high-level hospitals. Therefore, our empirical analysis is limited to using data from 2012 to 2017.

Figure 2.3 provides the timeline of the HMS reform implementation on the Chinese mainland. It indicates the exact time when each province announced fully implementation of the HMS reform. Two provinces, Qinghai and Sichuan, implemented the HMS reform before 2015. However, because we plan to use provincial data to conduct empirical analysis, the limited sample size of the treated provinces may affect our results' validity. Therefore, dropping these two early-treated provinces is reasonable to obtain more informative inference procedures.

Finally, we consider the case with $\mathrm{T} = 6$ periods (from 2012-2017) and denote a particular time period by t where $t = 1, \cdots, 6$. The first time of intervention entry is $q = t = 4\ (year = 2015)$. At period $q + 1 = 5$, more provinces join the treated group, and so on, until period $t = q + 2 = T = 6$. In the last period $t = T = 6$, all provinces are treated. Following previous literature, we also assume the treatment is irreversible, which is quite reasonable in our case. There are three different cohorts g according to the specific first treated time, $g \in \{q, \cdots, T\}$ (see Table 2.4). In particular, cohort 5 includes three direct-administered municipalities, and cohort 6 includes three autonomous regions.

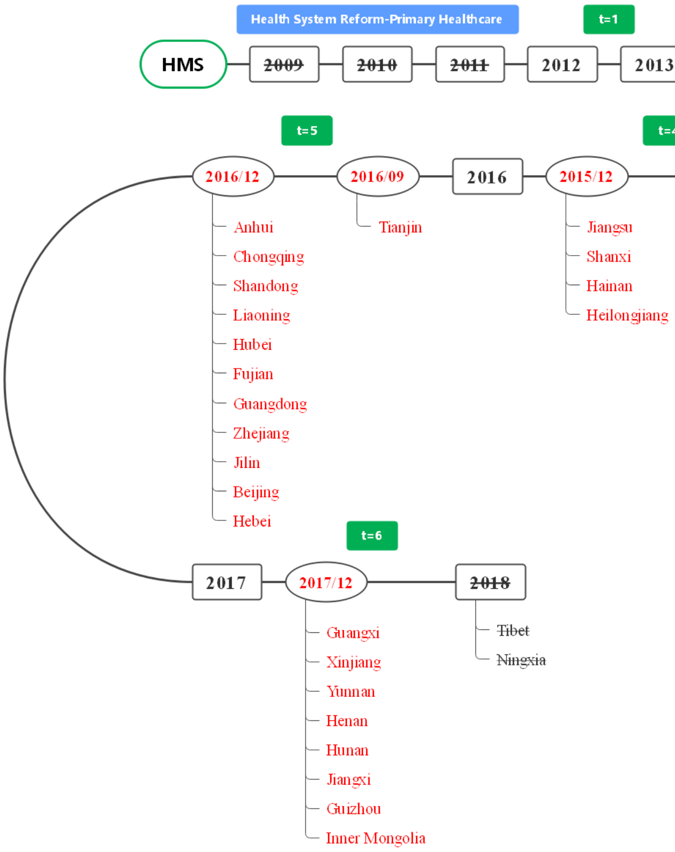


Figure 2.3 Timeline of the HMS policy implementation

Table 2.4 Provinces in different cohorts

First Entry Time	Year	Cohort	Provinces in each cohort
t=4	2015	g=4	7 provinces: Gansu, Shaanxi, Shanghai, Jiangsu, Shanxi, Hainan, Heilongjiang
t=5	2016	g=5	12 provinces: Tianjin, Anhui, Chongqing, Shandong, Liaoning, Hubei, Fujian, Guangdong, Zhejiang, Jilin, Beijing, Hebei
t=6	2017	g=6	8 provinces: Guangxi, Xinjiang, Yunnan, Henan, Hunan, Jiangxi, Guizhou, Inner Mongolia

Our identification strategy follows Wooldridge (2021) and Callaway and Sant' Anna (2021) to allow treatment effect heterogeneity. As all the provinces are finally treated, there is no never-treated group in the data, and the not-yet-treated groups are used as the comparison groups. For example, for Cohort 4, in period 4, we use all "not-yet-treated" cohorts (cohorts 5 and 6) as the comparison group. In period 5, cohort 5 gets treated, and then we use cohort 6 as the comparison group.

2.4. Empirical model

2.4.1. Extended TWFE estimator

2.4.1.1. Without covariates

Under unconditionally no anticipation and common trend assumptions, we consider estimation without covariates. Following Wooldridge (2021), Equation (1) shows how to estimate the average treatment effects on the treated in a particular cohort (cohort is defined by the first treated time) at a particular post-treatment period. Y_{it} is the primary outcome variable for province i in period t. D_{ig} is the time-invariant cohort dummy for province i, indicating when a province was first treated, $g \in \{q, \cdots, T\}$. f_{st} is the time dummy, which equals one if $s = t$ and zero otherwise, $s = q, \cdots, T$. W_{it} is the time-varying treatment indicator, which equals one if province i is eventually treated and period t is in i's post-treatment period. Wooldridge (2021) proves the equivalence between $W_{it} \cdot D_{ig} \cdot f_{st}$ and $D_{ig} \cdot f_{st}$. Interacting the cohort dummies with the time dummies corresponding to those periods where a cohort is treated allows ATTs to vary by cohort and calendar time. γ captures cohort effects, and θ captures time effects. It is obvious that τ_{gs} is the ATT for cohort g in periods in which provinces are subjected to the intervention, $s \in \{q, \cdots, T-1\}$. However, as all provinces are treated by $t = T$, the treatment effects for the last cohort are not identified. Additionally, in the last period, we have no choice but to compare the earlier treated cohorts with the final treated cohort, and so we estimate τ_{gT} for $g = q, \cdots, T-1$, which is the average effect in period T of having been treated $T - g$ period earlier rather than the cohort-period specific ATT.

$$Y_{it} = \alpha + \textstyle\sum_{g=q}^{T} \sum_{s=g}^{T} \boldsymbol{\tau_{gs}} \, (W_{it} \cdot D_{ig} \cdot f_{st}) + \gamma_q D_{iq} + \cdots + \gamma_T D_{iT} + \sum_{s=\mathrm{q}}^{T} \theta_s f_{st} \quad (1)$$

For Equation (1), pooled OLS could be used to get the parameters. We could also drop "$\gamma_q D_{iq} + \cdots + \gamma_T D_{iT}$" from Equation (1) and use extended TWFE to get the parameters. It is feasible to add covariates into Eq (1). However, it needs a large sample size and is not applicable in this study.

2.4.1.2. Simplified model

Some restrictions could be imposed on this general Eq (1) to simplify our model. Firstly, if we assume homogeneity within both cohort and calendar periods: $\tau_{gs} = \tau_{gg} = \tau_{g+1,s+1}$, then the restriction could be imposed using W_{it} rather than the triple interactions $W_{it} \cdot D_{ig} \cdot f_{st}$ in Eq (1). Secondly, if we assume treatment effects only vary by cohort, then homogeneity is imposed within the cohort by: $\tau_{gs} = \tau_{gg}$. This restriction could be imposed using the interaction term $W_{it} \cdot D_{ig}$ rather than the triple interactions $W_{it} \cdot D_{ig} \cdot f_{st}$. Thirdly, if we assume treatment effects only vary by treatment intensity, the restrictions could be written as: $\tau_{gs} = \tau_{g+1,s+1} = \tau_{s-g}$. If $s - g = 1$, then a province is in its first period of exposure. In the estimation, we need to create a set of treatment intensity indicators (according to the length of intervention exposure) to replace the triple interaction terms.

2.4.1.3. Common trend test

Generally, to test the common trend assumption, we could write the augmented equation as

$$Y_{it} = \alpha + \sum_{g=q}^{T-1} \sum_{s=2}^{q-1} \boldsymbol{\tau_{gs}} \, (W_{it} \cdot D_{ig} \cdot f_{st}) + \sum_{g=q}^{T} \sum_{s=g}^{T} \boldsymbol{\tau_{gs}} \, (W_{it} \cdot D_{ig} \cdot f_{st})$$

$$+\gamma_q D_{iq} + \cdots + \gamma_T D_{iT} + \textstyle\sum_{s=2}^{T} \theta_s f_{st} \qquad (2)$$

Then, we could jointly test the null hypothesis: $\tau_{gs} = 0, g = q, \cdots, T-1; s =$

$2, \cdots, q-1$. Alternatively, we could replace the variables $W_{it} \cdot D_{ig} \cdot f_{st}$ $(g = q, \cdots, T-1; s = 2, \cdots, q-1)$in Eq (2) with the cohort-specific trends, $D_{ig} \cdot t(t = 1, \cdots, T;\ g = q, \cdots, T-1)$. This allows for a constant trend difference between the treated and control units and is more realistic. Moreover, as $q = 4$, there are three pretreatment periods, apart form $D_{ig} \cdot t, D_{ig} \cdot t^2(t = 1, \cdots, T;\ g = q, \cdots, T-1)$ can be added to the augmented equation to allow more flexibility in the heterogeneous time trend.

2.4.2. Doubly-robust estimator

Apart from Wooldridge's regression-based estimator, Callaway and Sant' Anna (2021) propose alternative estimators deriving from a semiparametric setting.

2.4.2.1. Without covariates

When no anticipation and common trend assumptions hold unconditionally on covariates, the average treatment effect on the treated for cohort g at time period t could be obtained from Equation (3) using estimation by the analogy principle, which involves many comparisons of means. G_g signifies a cohort dummy and equals one if provinces are firstly treated at time $g, g = q, \cdots, T$. D_t is a dummy indicating the treatment status at time period t and equals one if treated.

$$ATT_{unc}^{ny}(g, t) = E[Y_t - Y_{g-1} | G_g = 1] - E[Y_t - Y_{g-1} | D_t = 0, G_g = 0] \qquad (3)$$

2.4.2.2. With covariates

Callaway and Sant' Anna (2021) provide a powerful identification strategy that extends the DID identification strategy based on the outcome regression (OR) approach of Heckman et al. (1997, 1998), the inverse probability weighting (IPW) approach of Abadie (2005), and the doubly-robust (DR) approach of Sant'Anna and Zhao (2020) to the multiple-period and multiple-group set up. The OR approach only relies on modelling the conditional expectation of the outcome evolution for the comparison groups. The IPW approach relies on modelling the conditional probability of being in group g. The DR approach exploits both OR and IPW components.

This study uses the doubly-robust (DR) estimator. DID estimators based on the DR estimands usually enjoy additional robustness against model-misspecifications compared to the IPW and OR estimands. Here, it is the detailed estimation steps: G_g signifies a cohort dummy and equals one if provinces are firstly treated at time $g, g = q, \cdots, T$. D_t is a dummy indicating the treatment status at time period t. Generalised propensity score $\hat{\varphi}(X)$ indicates the probability of being first treated at time g, conditional on pretreatment covariates X and on either being a member of group g (in this case, $G_g = 1$) or a member of the "not-yet-treated" group by time t (in this case, $(1 - D_t)(1 - G_g) = 1$) (See Eq(4)). A generalised propensity score enters the estimator as a weight (see Eq (6)). Equation (5) is the population outcome regression for the "not yet treated" by time t group.

$$\hat{\varphi}(X) = P_{g,t}(X) = \Pr\left(G_g = 1 | X, G_g + (1 - D_t)(1 - G_g) = 1\right) \qquad (4)$$

$$m_{g,t}^{ny}(X) = E[Y_t - Y_{g-1} | X, D_t = 0, G_g = 0] \qquad (5)$$

Therefore, given no treatment anticipation, conditional parallel trends and common support assumptions, the parameter we want to estimate could be expressed as Equation (6).

$$ATT_{dr}^{ny}(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{\varphi}(X)(1-D_t)(1-G_g)}{1-\hat{\varphi}(X)}}{E\left[\frac{\hat{\varphi}(X)(1-D_t)(1-G_g)}{1-\hat{\varphi}(X)}\right]}\right)\left(Y_t - Y_{g-1} - m_{g,t}^{ny}(X)\right)\right] \qquad (6)$$

Following research using the DR estimator, X is the "baseline" covariate, and we use one period before treatment as our baseline period.

To test the common trend assumption, we could estimate the chi2 statistic of the null hypothesis that all pretreatment $ATT_{dr}^{ny}(g,t), g = q, \cdots, T; t = 1, \cdots, q$, are statistically equal to zero.

2.5. Empirical Results

2.5.1. Bacon decomposition and simplified model specifications

Here, nurse proportion is used as an example to show the results of Bacon's decomposition and simplified model specifications.

We follow Goodman-Bacon (2021) to decompose our nurse proportion TWFE estimator (Table 2.5). Goodman-Bacon (2021) shows that the TWFE estimator equals a weighted average of all possible two-group/two-period DID estimators that compare timing groups to each other, some of which have no causal interpretation that involves an earlier treated group as a control for a later treated group. The decomposition results show that we use the early treated cohort g = 4 (Year 2015) as a comparison group for the later treated cohort g = 5 (Year 2016). It adjusts the path of outcomes for newly treated units by the path of outcomes for already treated units. However, this is not the path of not-yet-treated potential outcomes. It includes treatment effect dynamics, which leads to bias. This forbidden comparison group makes the total coefficient smaller than the actual value in this study. Therefore, we should interpret these standard TWFE estimates with caution.

Table 2.5 Bacon decomposition for nurse proportion TWFE estimators

treated group	control group	coefficient	Total weight	Aggregate group
		0.0013	1	
g=4(Year 2015)	g=5(Year 2016)	-0.001621	0.221053	Early vs Late
g=4/5(Year 2015/2016)	g=6(Year 2017)	0.002677	0.631579	Early vs Late
g=5(Year 2016)	g=4(Year 2015)	-0.000613	0.147368	Late vs Early

Table 2.6 provides results for nurse proportions in primary healthcare institutions using the different model specifications proposed previously. Columns (1)−(3) present the results of following Wooldridge (2021) without covariates: column (1) shows the results of assuming that treatment effects only vary by cohorts, and column (2) shows the results of assuming that treatment effects only vary by treatment intensity. Although the coefficients are insignificant, we observe heterogeneity within cohorts and different

treatment intensities. Therefore, it is worth introducing cohort-period specific ATT. Column (3) shows the results of allowing treatment effects to vary by cohort and time period. Column (4) presents the results of following Callaway and Sant' Anna (2021) without covariates. Comparing column (3) with column (4), the DR estimators are slightly smaller than the extended TWFE estimators. Additionally, the extended TWFE estimator has two more coefficients than the DR estimator, $\tau_{4,6}$ and $\tau_{5,6}$, representing the incremental effect of a one- or two-period earlier exposure to the treatment relative to the first exposure in the last period, respectively (We do not report the coefficients here because we are not interested in them).

Table 2.6 Simplified model specifications for nurse proportion

	(1)	(2)	(3)	(4)
		Cohort Effects		
g=4 (Year 2015)	0.0005			
	(0.0034)			
g=5(Year 2016)	0.0016			
	(0.0042)			
		Treatment intensity (Length of exposure)		
T0		0.0013		
		(0.0015)		
T1		0.0021		
		(0.0029)		
		Cohort-period specific effects		
ATT(4,4)			-0.0008	-0.0020
			(0.0024)	(0.0016)
ATT(4,5)			0.0029	0.0010
			(0.0038)	(0.0028)
ATT(5,5)			0.0034	0.0018
			(0.0039)	(0.0017)
Aggregation to a single			0.0021	0.0005
effect			(0.0025)	(0.0014)
2012.year	0.0000	0.0000	0.0000	
	(.)	(.)	(.)	
2013.year	0.0081***	0.0081***	0.0081***	
	(0.0013)	(0.0013)	(0.0013)	
2014.year	0.0145***	0.0145***	0.0145***	
	(0.0014)	(0.0014)	(0.0014)	
2015.year	0.0225***	0.0223***	0.0228***	

	(0.0023)	(0.0020)	(0.0022)	
2016.year	0.0298***	0.0295***	0.0284***	
	(0.0033)	(0.0033)	(0.0031)	
2017.year	0.0409***	0.0400***	0.0413***	
	(0.0041)	(0.0060)	(0.0060)	
_cons	0.1668***	0.1668***	0.1668***	
	(0.0015)	(0.0015)	(0.0015)	
N	162	162	162	135

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

Overall, because the government implemented the HMS reform gradually, and then treatment effect heterogeneity is the most likely, it is better to introduce those new methods following Wooldridge (2021) and Callaway and Sant' Anna (2021) to adapt to staggered interventions and allow heterogeneity. In the following sections, cohort−period specific ATT results for all outcome indicators are presented.

2.5.2. Results for primary healthcare institution's capacity

2.5.2.1. Results for nurse proportion

Table 2.7 displays the results for nurse proportion in primary healthcare institutions. Overall, it suggests a higher proportion of nurses to total staff in primary healthcare institutions after the reform, although the results are not statistically significant. We also investigate the impact of HMS reform on nurse proportions in different primary healthcare institutions. The common trend assumption holds unconditionally here. As discussed in Section 2.4.1.3, we employ three kinds of common trend tests. First, the event-study type test is explored, which is similar to the general approach in the standard DID setup. Second, a cohort-specific linear trend is assumed. Third, as there is more than one period before policy intervention, it is feasible to assume a cohort-specific nonlinear trend. The bottom of Table 2.7 shows the results of these tests, all three of the parallel trends tests have large p-values, providing insufficient evidence against the parallel trends assumption.

The effects of HMS reform on the nurse proportion in township health centres and village clinics are positive but insignificant. However, for the early treated cohort

(cohort 4) with one-period treatment duration (at period 5) (ATT [4,5] in Table 2.7), the HMS reform significantly increases the nurse proportion in urban primary healthcare institutions (community health centres and stations) by 0.83% (a 2.6% increase of the baseline level[6], i.e., 31.94%). DR estimators in Figure 2.4 provide similar results.

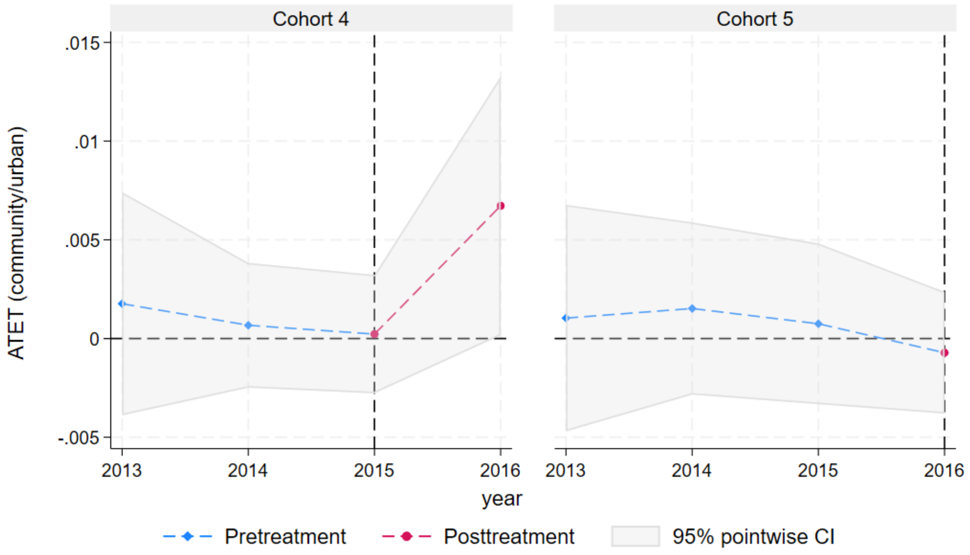


Figure 2.4 DR estimator for nurse proportion in urban primary healthcare institutions

[6] We define one period before treatment for specific cohorts as the baseline level.

Table 2.7 Results for nurse proportion in primary healthcare institutions

	(1)	(2)	(3)	(4)	(5)
	primary	urban		rural	
	overall	(community)	rural overall	township	village
Wooldridge (2021)-Extended TWFE estimator without covariates					
ATT(4,4)	-0.0008	0.0013	0.0039	0.0010	0.0113
	(0.0024)	(0.0028)	(0.0040)	(0.0039)	(0.0096)
ATT(4,5)	0.0029	0.0083*	0.0042	0.0059	0.0111
	(0.0038)	(0.0045)	(0.0053)	(0.0061)	(0.0120)
ATT(5,5)	0.0034	0.0009	0.0000	0.0054	0.0016
	(0.0039)	(0.0030)	(0.0041)	(0.0046)	(0.0081)
Aggregation to a single effect	0.0021	0.0030	0.0021	0.0044	0.0068
	(0.0025)	(0.0027)	(0.0031)	(0.0036)	(0.0079)
Common trend	√	√	√	√	√
Dig*t	F(2,26) =0.72	F(2,26) =0.53	F(2,24) =1.34	F(2,24) =0.82	F(2,26) =1.93
	Prob > F =0.4982	Prob > F =0.5972	Prob > F =0.2817	Prob > F =0.4539	Prob > F =0.1649
Dig*t^2	F(4,26) =0.92	F(4,26) =0.27	F(4,24) =0.71	F(4,24) =1.53	F(4,26) =1.04
	Prob > F =0.4668	Prob > F =0.8959	Prob > F =0.5953	Prob > F =0.2246	Prob > F =0.4065
Dig*fst,s=2,~,q-1	F(5,26) =0.76	F(5,26) =0.21	F(5,24) =0.67	F(5,24) =1.90	F(5,26) =1.27
	Prob > F =0.5872	Prob > F =0.9536	Prob > F =0.6490	Prob > F =0.1311	Prob > F =0.3052
N	162	162	150	150	162

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

2.5.2.2. Results for practitioner proportion in primary healthcare institutions

Results in Table 2.8 show the HMS reform has positive but insignificant effects on practitioner proportion in primary healthcare institutions. The impact is also investigated separately in different primary healthcare institutions. For practitioner proportion in township health centres, the common trend assumption fails to hold unconditionally. The effects of the HMS reform on practitioner proportion in township health centres and village clinics are also positive but insignificant. In contrast, for the later treated cohort (cohort 5) at period 5 (ATT (5,5) in Table 2.8), the HMS reform effectively enhances the proportion of practitioners to total staff in urban primary healthcare institutions (community health centres and stations) by 1.3% (a 3.6% increase of the baseline level, i.e., 36.16%). DR estimators provide similar results (Figure 2.5).

In summary, the HMS reform effectively enhances the nurse and practitioner proportion of primary healthcare institutions in urban areas, whereas it seems ineffective in rural areas.

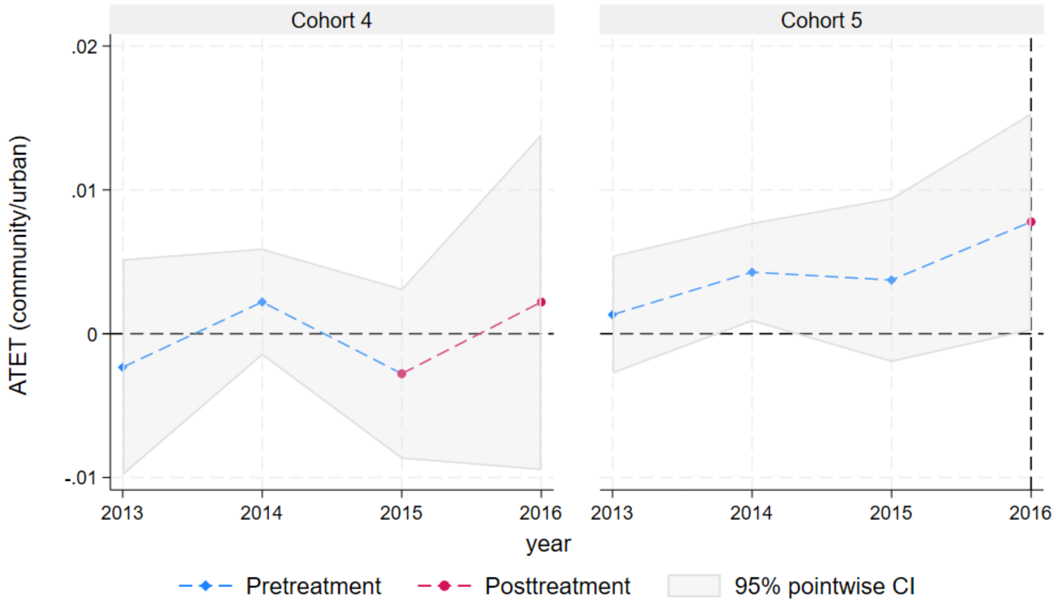


Figure 2.5 DR estimator for practitioner proportion in urban primary healthcare institutions

Table 2.8 Results for practitioner proportion in primary healthcare institutions

	(1)	(2)	(3)	(4)	(5)
	primary	urban		rural	
	overall	(community)	rural overall	township	village
Wooldridge (2021)-Extended TWFE estimator without covariates					
ATT(4,4)	0.0011	-0.0021	0.0118	0.0129	0.0066
	(0.0052)	(0.0036)	(0.0080)	(0.0082)	(0.0127)
ATT(4,5)	0.0007	0.0038	0.0111	0.0161	0.0082
	(0.0074)	(0.0069)	(0.0108)	(0.0154)	(0.0149)
ATT(5,5)	0.0033	0.0130**	0.0012	0.0088	0.0059
	(0.0052)	(0.0060)	(0.0069)	(0.0098)	(0.0087)
Aggregation to a single effect	0.0020	0.0065	0.0066	0.0118	0.0067
	(0.0046)	(0.0047)	(0.0064)	(0.0089)	(0.0093)
Common trend	√	√	√	×	√
Dig*t	F(2,26) =0.86	F(2,26) =1.62	F(2,24) =1.18	F(2,24) =2.76	F(2,26) =0.32
	Prob > F = 0.4354	Prob > F =0.2166	Prob > F =0.3236	Prob > F =0.0837	Prob > F =0.7314
Dig*t^2	F(4,26) =0.80	F(4,26) =1.10	F(4,24) =0.80	F(4,24) =2.33	F(4,26) =0.81
	Prob > F =0.5376	Prob > F =0.3757	Prob > F =0.5388	Prob > F =0.0853	Prob > F =0.5304
Dig*fst,s=2,~,q-1	F(5,26) =0.68	F(5,26) =1.35	F(5,24) =1.26	F(5,24) =2.15	F(5,26) =1.09
	Prob > F =0.6454	Prob > F =0.2761	Prob > F =0.3131	Prob > F =0.0942	Prob > F =0.3871
N	162	162	150	150	162

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

2.5.3. Results for primary healthcare institution utilisation

2.5.3.1. Results for outpatient visit proportion

The effects of HMS reform on outpatient visit proportion in primary healthcare institutions are positive but insignificant (Table 2.9). Further analysis of HMS reform's impact on different primary healthcare institutions is conducted. The common trend assumption fails to hold unconditionally for outpatient visit proportion in township health centres. The HMS reform has a limited impact on outpatient visit proportion in community health centres, community health stations and township health centres. However, the HMS reform significantly improves the proportion of outpatient visits in village clinics by 1.67% (a 7.25% increase of the baseline level, i.e., 23.02%). Additionally, there is treatment effect heterogeneity across cohorts and periods. On the one hand, ATT increases by the length of exposure within the same cohort: for cohort 4, the ATT in period 4 is about 0.72% (a 3.34% increase of the baseline level, i.e., 21.58%) and it increases to 2.57% (an 11.91% increase of the baseline level) in period 5. Given the same event study relative time period, ATT for the later treated cohort is greater than that for the early treated cohort: for cohort 5, the ATT at period 5 is about 1.7% (a 7.52% increase of the baseline level, i.e., 22.60%). DR estimators present similar results (Figure 2.6).

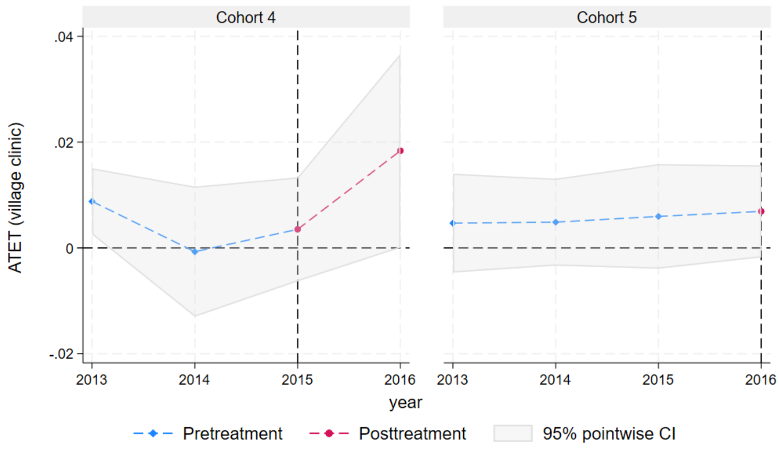


Figure 2.6 DR estimator for outpatient visit proportion

Table 2.9 Results for outpatient visit proportion in primary healthcare institutions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	primary		urban(community)			rural	
	overall	urban overall	health centre	health station	rural overall	township	village
	Wooldridge (2021)-Extended TWFE estimator without covariates						
ATT(4,4)	-0.0032	-0.0007	0.0001	-0.0015	-0.0001	-0.0061	0.0072
	(0.0078)	(0.0015)	(0.0021)	(0.0018)	(0.0090)	(0.0041)	(0.0084)
ATT(4,5)	0.0056	-0.0019	-0.0027	-0.0010	0.0070	-0.0151**	0.0257*
	(0.0122)	(0.0020)	(0.0021)	(0.0016)	(0.0140)	(0.0064)	(0.0128)
ATT(5,5)	0.0124	0.0015	-0.0019	0.0034*	0.0022	-0.0128**	0.0170*
	(0.0102)	(0.0020)	(0.0021)	(0.0019)	(0.0121)	(0.0057)	(0.0100)
Aggregation to a single effect	0.0064	0.0011	-0.0016	0.0011	0.0028	-0.0116**	0.0167*
	(0.0082)	(0.0017)	(0.0011)	(0.0012)	(0.0099)	(0.0048)	(0.0085)
Common trend	√	√	√	√	√	×	√
Dig*t	F(2,26) =0.56	F(2,25) =0.39	F(2,26) =0.04	F(2,25) =0.34	F(2,24) =0.04	F(2,24) =2.31	F(2,26) =1.83
	Prob > F	Prob > F	Prob > F	Prob > F =0.7183	Prob > F	Prob > F	Prob > F
	=0.5794	=0.6796	=0.9568		=0.9640	=0.1214	=0.1809
Dig*t^2	F(4,26) =0.75	F(4,25) =1.69	F(4,26) =0.34	F(4,25) =0.92	F(4,24) =1.85	F(4,24) =2.26	F(4,26) =1.72
	Prob > F	Prob > F	Prob > F	Prob > F =0.4680	Prob > F	Prob > F	Prob > F
	=0.5676	=0.1847	=0.8465		=0.1513	=0.0929	=0.1768
Dig*fst,s=2,~,q-1	F(5,26) =0.62	F(5,25) =1.36	F(5,26) =0.27	F(5,25) =1.62	F(5,24) =1.54	F(5,24) =2.15	F(5,26) =1.37
	Prob > F	Prob > F	Prob > F	Prob > F =0.1921	Prob > F	Prob > F	Prob > F
	=0.6885	=0.2713	=0.9230		=0.2162	=0.0937	=0.2687
N	162	156	162	156	150	150	162

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

2.5.3.2. Results for inpatient healthcare utilisation

Table 2.10 shows the results for inpatient admission proportion and bed occupancy rate in primary healthcare institutions. We also investigate the impact of HMS reform on inpatient usage indicators in different primary healthcare institutions separately. For inpatient admission proportion in community health centres, the common trend assumption fails to hold unconditionally.

Overall, the HMS reform has a limited impact on inpatient admission proportion in primary healthcare institutions. However, for early treated cohorts (cohort 4), the reform is effective in increasing the proportion of inpatient admission in primary healthcare institutions and township health centres. This positive effect increases with longer exposure (column [1] and [3] in Table 2.10).

The effects of HMS reform on the bed occupancy rate in community health centres are insignificant. In contrast, the HMS reform significantly enhances the bed occupancy rate in township health centres by 3.81% (a 7.38 percentage increase of the baseline level, i.e., 51.66%) (column [5] in Table 3.10). Moreover, apparent heterogeneity across cohorts and time periods is revealed. ATT increases by the length of exposure within the cohort: for cohort 4, the ATT in period 4 is about 2.1% (a 4.31% increase of the baseline level, i.e., 48.7%) and it increases to 3.94% (an 8.09% increase of the baseline level) in period 5. Given the same event study relative time period, ATT for the later treated cohort is greater than the early treated cohort: for cohort 5, the ATT at period 5 is about 4.67% (an 8.8 percentage increase of the baseline level, i.e., 53%). Figure 2.7 presents similar results with DR estimators.

In summary, the HMS reform effectively promotes outpatient and inpatient healthcare services usage of primary healthcare institutions in rural areas, whereas it does not work well in urban areas.

Table 2.10 Results for inpatient healthcare utilisation

	(1)	(2)	(3)	(4)	(5)
	Inpatient admission proportion			Bed occupancy rate	
	primary overall	community	township	community	township
	Wooldridge (2021)-Extended TWFE estimator without covariates				
ATT(4,4)	0.0148**	-0.0003	0.0132**	-0.0356	0.0210
	(0.0067)	(0.0014)	(0.0060)	(0.0313)	(0.0148)
ATT(4,5)	0.0253*	-0.0017	0.0241*	-0.0465	0.0394
	(0.0137)	(0.0020)	(0.0123)	(0.0452)	(0.0308)
ATT(5,5)	0.0153	-0.0014	0.0138	-0.0078	0.0467*
	(0.0130)	(0.0009)	(0.0122)	(0.0197)	(0.0272)
Aggregation to a single effect	0.0179	-0.0012	0.0163	-0.0257	0.0381*
	(0.0106)	(0.0010)	(0.0098)	(0.0233)	(0.0213)
Common trend	√	×	√	√	√
Dig*t	F(2,26) =0.95	F(2,26) =0.64	F(2,24) =0.75	F(2,26) =1.82	F(2,24) =0.36
	Prob > F =0.3985	Prob > F =0.5350	Prob > F =0.4832	Prob > F =0.1813	Prob > F =0.7014
Dig*t^2	F(4,26) =1.05	F(4,26) =4.64	F(4,24) =1.02	F(4,26) =1.30	F(4,24) =0.48
	Prob > F =0.3986	Prob > F =0.0058	Prob > F =0.4193	Prob > F =0.2950	Prob > F =0.7534
Dig*fst,s=2,~,q-1	F(5,26) =1.89	F(5,26) =4.64	F(5,24) =1.12	F(5,26) =1.74	F(5,24) =0.40
	Prob > F =0.1307	Prob > F =0.0037	Prob > F =0.3774	Prob > F =0.1610	Prob > F =0.8409
N	162	162	150	162	150

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

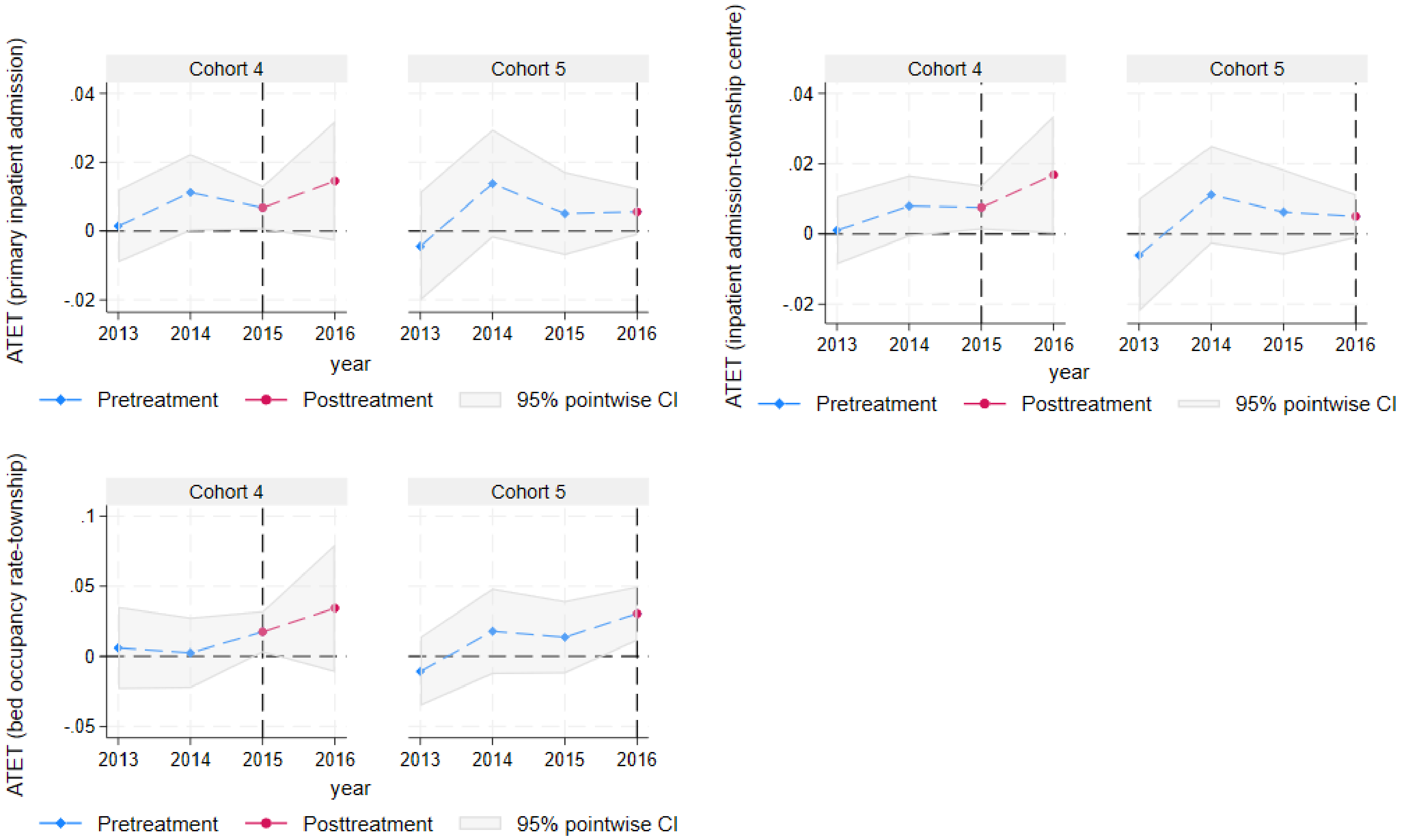


Figure 2.7 DR estimator for inpatient healthcare utilisation in urban primary healthcare institutions

2.5.4. Outcome variables fail to pass unconditional common trend assumption

Three outcome indicators fail to pass the unconditional common trend assumptions. Therefore, we assume conditional common trend assumptions and use DR estimators with covariates for those outcome indicators. Regarding inpatient admission proportion in community health centres, they remain fail to pass the test when controls are included. For practitioner proportion and outpatient visit proportion in township health centres, the common trend assumptions hold after controlling sex ratio and elderly proportion, illiteracy ratio and elderly proportion, respectively. As mentioned in Section 2.4.2.2, for each cohort, at each pretreatment period, the ATTs are estimated. The null hypothesis for the common trend assumption is all pretreatment ATTs are statistically equal to zero. Given the large p values, we couldn't reject the common trend assumption. Overall, we observe no significant effect (Table 2.11).

Table 2.11 DR estimators with covariates

	practitioner proportion (township health centre)	outpatient visit proportion (township health centre)
ATT(4,4)	0.0062	-0.0034
	(0.0057)	(0.0032)
ATT(4,5)	0.0014	-0.0057
	(0.0152)	(0.0065)
ATT(5,5)	0.0027	0.0002
	(0.0074)	(0.0017)
Aggregation to a single effect	0.0032	-0.0022
	(0.0063)	(0.0027)
Common trend	√	√
	chi2(5) =8.91	chi2(5) =8.56
	p-value=0.1127	p-value=0.1278
N	125	125
Controls	sex ratio, elderly proportion	Illiteracy ratio, elderly proportion

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

2.5.5. Sensitivity analysis

There are three kinds of robustness tests. First, we further explore the sensitivity of the results to the measurement of outcomes. Some different capacity and utilisation indicators are chosen, such as the number of beds per 10,000 population, the number of nurses/practitioners per 10,000 population, and per capita outpatient visits in primary healthcare institutions. The coefficients for the number of beds per 10,000 population are negative and significant. However, this outcome variable fails to pass the common trend assumption test even after adding covariates. Results for the other three outcome indicators remain consistent with what we find above: the HMS reform effects are positive but insignificant (Table 2.12).

Table 2.12 Results for different measures of outcome variables

	per ten-thousand population			per capita
	the number of practitioners	the number of nurses	the number of beds	outpatient visits
	Wooldridge (2021)-Extended TWFE estimator without covariates			
ATT(4,4)	-0.0487	-0.0286	-0.2691	0.0049
	(0.1660)	(0.1273)	(0.1969)	(0.0839)
ATT(4,5)	-0.0723	0.0439	-0.8202***	0.0387
	(0.2566)	(0.2076)	(0.2871)	(0.0957)
ATT(5,5)	0.0901	0.0418	-0.5733*	0.0348
	(0.1941)	(0.1745)	(0.2809)	(0.0866)
Aggregation to a single effect	0.0090 (0.1542)	0.0234 (0.1271)	-0.5579** (0.2053)	0.0278 (0.0637)
Common trend	√	√	×	√
N	162	162	162	162

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.001

Second, the robustness of the results to different estimate methods is investigated. To elaborate, we use the bed occupancy rate in township health centres as an example. Following Sun and Abraham (2021) and Borusyak et al. (2024) , we obtain the IW estimator and the imputation estimator (Table 2.13). Because they only provide the event-study aggregation of ATTs, we manually construct the cohort-period-specific ATTs without standard errors. The imputation estimators (Borusyak et al., 2024) are numerically equivalent to the extended TWFE estimators (Wooldridge, 2021). In comparison, the IW estimators (Sun and Abraham, 2021) are closer to the DW estimators (Callaway and Sant'Anna, 2021), which are slightly smaller than the former two kinds of estimators. The different choices of baseline outcome could explain this discrepancy. Both the extended TWFE estimator and imputation estimator use the average outcome from all pre-periods as baseline outcomes, whereas the IW estimates and DR estimates regard the outcome at the last period before the units are treated as baseline outcomes. Additionally, the slight discrepancy between IW and DR estimates is related to different comparison groups. The IW estimates use the last-treated units as comparisons, whereas the DR estimates use the not-yet-treated units as controls.

Table 2.13 Comparisons of different estimators for bed occupancy rate in township health centres

	extended TWFE estimator	DW estimator	IW estimator	Imputation estimator
ATT(4,4)	0.0210	0.0174**	0.0253	0.0210
	(0.0148)	(0.0076)		
ATT(4,5)	0.0394	0.0343	0.0343	0.0394
	(0.0308)	(0.0233)		
ATT(5,5)	0.0467*	0.0304***	0.0304	0.0467
	(0.0272)	(0.0099)		
Common trend	√	√	√	√
Aggregation by event-study relative period				
T0	0.0376*	0.0258***	0.0286***	0.0376**
	(0.0195)	(0.0076)	(0.0102)	(0.0184)
T1	0.0394	0.0343	0.0343	0.0394
	(0.0308)	(0.0233)	(0.0251)	(0.0291)

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

Third, for those outcome variables failing to hold the common trend assumption, we attempt different combinations of covariates and summarise the results passing the common trend test in Appendix B (Table B1). The coefficients remain positive and insignificant with slight changes in their magnitude.

2.6. Discussion and Conclusion

This study explores the effect of HMS reform on primary healthcare utilisation at the national level. Our findings show that the HMS reform effectively enhances primary healthcare institutions' capacity in urban areas. However, urban residents do not immediately respond to these improvements. The effect of the HMS reform on healthcare usage in urban primary healthcare institutions is not significant. This result may be related to China's long history of having a hospital-centred system, especially in cities. Urban residents' preferences seem difficult to change. The gap in healthcare service quality between primary healthcare institutions and higher-level hospitals remains huge despite slight quality improvements in primary healthcare institutions. This is consistent with previous studies showing the declining trend in the proportion of inpatient care provision by primary care facilities after the reform (Feng et al., 2022).

Additionally, the HMS reform has a positive but insignificant impact on improving primary healthcare institutions' capacity in rural areas. However, the reform is effective in promoting the usage of primary healthcare institutions in rural areas. The possible reasons for these results are as follows: first, the quality indicators chosen in the paper are insufficient. We only focus on the workforce perspective due to data restrictions and overlook equipment or other aspects. Second, medical insurance payments play an important role in the HMS reform (Zhou et al., 2021). Gradient reimbursement schemes could incentivise primary healthcare institution utilisation according to our theoretical analysis. The positive reform impact on rural primary healthcare institution utilisation may be associated with this economic measure rather than healthcare capacity improvement.

Moreover, our results are different from those of the previous literature (Zhou et al., 2021). Their findings show that implementing HMS has a significantly positive effect on the probability of urban residents going to primary care facilities for contact and that the impact of HMS is not significant for rural residents. One possible explanation for this might be that they use the standard DID method, which overlooks the staggered intervention of the reform and may lead to bias. Another explanation is that using a self-reported health facility choice as the outcome variable may create measurement bias. The provincial level aggregated dataset we choose might be less concerned about measurement bias. The richness of the information on health resources and health services in the dataset enables us to explore the impact of the HMS reform on the capacity and the healthcare utilisation in primary health facilities. However, as this is a highly aggregated provincial dataset, the sample size is relatively small.

Furthermore, as we use the newly proposed heterogeneous DID methods, a detailed analysis of treatment effect heterogeneity is available. On the one hand, we find that the impact of the HMS reform is greater in the later treated provinces, which could be partly explained from a policy diffusion perspective. The later-treated cohorts could learn from the earlier-treated cohorts and have more time for detailed policy design or preparations. The heterogeneity across the cohorts may also be related to the unique features of each cohort because their components are different. For example, cohort 5 includes three municipalities that have relatively high levels of economic development and political value. Thus, cohort 5 might be more capable of carrying out effective measures to achieve the goal of the reform. On the other hand, the impact of the HMS reform grows stronger as more time passes after the implementation, and the immediate effect seems to be limited for some outcome indicators. This is not surprising because it usually takes time for reforms to work. Additionally, as we define the time of the reform implementation according to the provinces' self-announced times, there may be a gap between the self-announced time and the actual implementation time.

From the perspective of policy, our findings have some important implications. For rural primary healthcare institutions, it appears challenging to attract highly educated professional health workers. The lack of competence of medical staff in rural primary facilities is often associated with the lack of trust in primary facilities, which induces rural residents to choose higher-level facilities (Liu et al., 2018b). Further policies for promoting high-quality human resources to be allocated to rural primary healthcare facilities are necessary. For instance, offering generous salaries or benefits and providing more career advancement opportunities would be helpful. Additionally, nurses and practitioners in rural primary facilities usually have lower average medical educational attainment than health workers in higher-level hospitals. It is also important to improve the competence of current medical staff. For example, it would be useful to provide more job training opportunities that are tailored to their needs and to keep pace with modern technology and practices. Moreover, encouraging current medical staff to seek on-the-job education would help improve the average education attainment and increase professional skills.

For urban healthcare institutions, the reform helps to attract professional healthcare workers. However, in itself, this appears not to effectively improve primary healthcare visits. Previous studies reveal that urban residents attach less weight to cost factors (such as medical expenses, waiting time, travelling distance and opportunity costs) when they seek healthcare services (Wu et al., 2017). In contrast, urban residents are more concerned about organizational factors, such as the reputation of the institution, advanced equipment, drug variety and average education level of doctors (Liu et al., 2018b). The unavailability of certain drugs or advanced equipment in primary facilities usually pushes urban residents to higher-level hospitals (Liu et al., 2018b; Wu et al., 2017). Therefore, in addition to human resources, policies focusing on other aspects of healthcare provision, such as extending drug availability and improving access to advanced equipment could be considered to encourage the urban population to visit primary healthcare institutions. Besides, improving service attitude to enhance the reputation of urban primary healthcare institutions may also be helpful.

3. Chapter 3

Evaluating the Impact of the Urban-Rural Health Insurance Integration Reform on Healthcare Utilisation in Rural China

Abstract

China has been making efforts to establish a universal healthcare coverage system through multiple social health insurance schemes. However, healthcare disparities coexist with universal health coverage due to the fragmentation of medical insurance. The Chinese government launched the Urban-Rural Residents Basic Medical Insurance (URRBMI) reform to establish a unified medical insurance scheme for urban and rural residents except employees in 2016. The integration reform was implemented nationwide gradually by 2020 and is recognized as a vital step to reduce inequality. This study adopts a newly proposed heterogeneous difference-in-difference approach to evaluate the effect of integration on healthcare utilisation in rural China. The data were derived from the China Health and Retirement Longitudinal Study (CHARLS) from 2011 (wave 1) to 2018 (wave 4). Findings indicate that the integration reform significantly increases the middle-aged and older rural residents' inpatient care utilisation, including the probability of a hospitalization and the number of nights hospitalized during the most recent hospital stay. The reform has limited impacts on outpatient healthcare usage. Moreover, with the staggered policy implementation, significant treatment effect heterogeneity across cohorts and periods is observed.

Keywords

urban-rural health insurance integration; healthcare utilisation; staggered intervention; heterogeneous difference-in-difference; CHARLS; rural China

3.1. Introduction

In 2005, the World Health Organization (WHO) established a resolution committing to the pursuit of universal health coverage (UHC), encompassing three key dimensions: population coverage (broadness), service coverage (depth), and cost coverage (height). UHC means that all people can access the health services they need without suffering financial hardship (WHO, 2005). UHC has been gradually identified as a priority for the global health agenda (WHO, 2013). Governments around the world are actively taking action to respond to this goal. Some low and middle-income countries, especially emerging economies such as China, Thailand, South Africa and Mexico, have implemented social health insurance reforms as the first step to move towards UHC (Basu et al., 2012; Giedion et al., 2013; Su et al., 2019).

In China, the establishment and dramatic expansion of social health insurance in recent decades have initiated significant improvements in UHC. By 2018, over 95% of the Chinese population (more than 1.34 billion people) were insured, up from less than 50% in 2005 (Ren et al., 2022). However, accompanying the impressive coverage expansion was a sharp fragmentation of social health insurance among geographic units and social groups (Huang & Wu, 2020). Before 2016, social health insurance in China was characterized by a clear divide between rural and urban areas, with the New Rural Cooperative Medical System (NRCMS, launched in 2003) for rural residents, the Urban Employee Basic Medical Insurance (UEBMI, launched in 1998) for urban employees and the Urban Resident Basic Medical Insurance (URBMI, launched in 2007) for the remainder of urban residents. Segmentation by employment and residential status involves significant differences in terms of benefit packages, fundraising and operation. This has been an important factor for inequitable access to health care and financial protection for people covered by different schemes. For example, NRCMS was generally the least generous in terms of a relatively high patient cost-sharing rate for limited-service coverage although it covered over 65% of the total population (Zhou et al., 2022). Therefore, rural populations have more restricted access to health care than urban residents and also have a larger financial burden. The general demand for a more equitable health insurance system has become more vocal: discontent due to the unfair distribution of benefits is likely to lead to social unrest and instability (Munro and Duckett, 2016; Yip et al., 2019). To address the inefficiency and inequality derived from the fragmentation of health insurance schemes, the Chinese government explored integrating URBMI and NRCMS in some pilot areas since 2009, as these two schemes had more similarities in their target population and scheme design (Meng et al., 2015). The national integration reform was launched to build a new scheme named Urban-Rural Resident Basic Medical Insurance (URRBMI) for all residents except employees in 2016.

The new URRBMI scheme was implemented gradually throughout the whole country by 2020. The goal of the consolidation was to ensure equitable benefits for urban and rural residents and improve the well-being of the population. Health insurance integration and expansion is not unique to China. Other countries and regions, such as South Korea, Japan, Thailand and Taiwan, underwent similar experiences of building universal and unified health insurance systems (Kwon, 2003; Lee, 2003; Kondo and Shigeoka, 2013; Panpiemras et al., 2011; Chen et al., 2007). Evidence from these countries suggests that integration or expansion of health insurance coverage significantly increases health care utilisation. However, most such experiences also reveal nontrivial implementation challenges to cover the informal sector, low-income, and other vulnerable populations. It has been a common difficulty to ameliorate access and health disparities between sub-populations with different coverage pre-UHC (Ikegami et al., 2011; Zhou et al., 2022). In this context, less is known about whether the integration reform covering the informal sector in China can be implemented smoothly and meet its goal.

Since the pilot consolidation of URBMI and NRCMS in 2009, a number of studies have investigated the impact of the reform on healthcare expenditure, healthcare utilisation and health. The findings indicate that the integration is effective in reducing out-of-pocket expenditures, increasing the actual reimbursement rate and improving health-care-related financial risk protection (Zhou et al., 2022; Huang and Wu, 2020; Ren et al., 2022; Liu et al., 2018a). However, there is no consensus on whether it has been effective in promoting healthcare utilisation and health outcomes. Huang and Wu (2020) find the integration facilitates middle-aged and older rural residents' inpatient healthcare usage but has limited impact on health. Zhou et al.(2022) claim the policy significantly improves the self-assessed health of the rural population but has no short-term effect on inpatient healthcare usage. A recent paper confirms the positive impact of the reform on health outcomes; it is associated with reduced functional limitations of middle-aged and elderly rural residents (Hao and Yeo, 2023). Outpatient and urban residents' healthcare usage seems to be unaffected by the policy (Zhou et al., 2022; Su et al., 2019).

Reducing urban-rural healthcare inequality has been a priority of the consolidation reform. Some researchers have explored the impact of the reform on inequality in terms of healthcare expenditure or utilisation. The evidence is mixed: Some findings show the policy improves equity in utilisation (Li et al., 2019) and reduce inequality of incidence of catastrophic health expenditures (Wang et al., 2020a). In contrast, Yang, Acharya and Liu (Yang et al., 2022) argue that the urban-rural gap in medical expenditure hasn't narrowed following the reform. There are also some empirical studies focusing on the impact of the reform on a broader range of outcome variables rather than health-related outcomes. Evidence suggests the policy increases total non-medical household consumption (Chen et al., 2022) and encourages families' reasonable allocation to risk assets[7] and risk-free assets[8] to a certain extent (Li and Yang, 2021).

Although previous studies provide a rich basis for exploring the effects of URRBMI reform, two obvious limitations exist. First, some studies are based on local data (Liu et al., 2018a), or are limited to specific groups of people. For those studies using the national dataset, the findings are still restricted to the impact in pilot areas. There is scarce empirical evidence for the evaluation of the 2016 national integration reform. Second, from a methodology perspective, studies using a standard difference-

[7] Risky assets are composed of stocks, loans, funds, derivatives, Internet financial management, financial management, non-RMB assets, gold, and other risky assets.

[8] Risk-free assets are composed of cash, bonds, demand deposits, and time deposits.

in-difference (DID) approach are likely to be biased as treatment heterogeneity is most likely with staggered implementation (De Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021). Conventional DID estimates rely on the assumption of treatment homogeneity. Under heterogeneous treatment effects, standard DID estimates in general do not converge to a convex combination of the individual treatment effects for units under the treatment condition, even when the common trend assumption is valid (Chiu et al., 2023). Recently, researchers have proposed several new estimators to produce causally interpretable estimates under heterogeneous treatment effects and parallel trends (Sun and Abraham, 2021; Baker et al., 2022; Callaway and Sant'Anna, 2021; De Chaisemartin and d'Haultfoeuille, 2018; Imai et al., 2023; Borusyak et al., 2024; Wooldridge, 2021, 2023).

This study aims to evaluate the effect of the integration reform on healthcare utilisation for the rural population. This work addresses the limitations of previous studies and contributes to the growing literature evaluating Chinese basic medical insurance programs. Our research also provides new evidence on the impact of health insurance expansion on the informal sector in low- and middle-income countries. Firstly, we use the newly proposed heterogeneous DID methods to allow arbitrary treatment effect heterogeneity and provide more robust estimation (Wooldridge, 2021, 2023; Callaway and Sant'Anna, 2021). In addition, the richness of the data derived from the China Health and Retirement Longitudinal Study (CHARLS) (2011, 2013, 2015 and 2018) allow us to evaluate the 2016 national policy effect rather than only the pilot impact. To the best of our knowledge, few studies have provided national evidence of the impacts of the integration. Moreover, a comprehensive set of healthcare usage indicators are used. The main findings reveal that the reform is effective in facilitating middle-aged and older rural population's inpatient care utilisation, including the probability of a hospitalization and the number of nights hospitalized during the most recent hospital stay. However, the integration had limited impacts on outpatient healthcare usage. Obvious treatment heterogeneity across both groups and periods is observed, which confirms the underlying constant treatment effect assumption in standard DID estimates is implausible in this context.

3.2. Conceptual Framework

3.2.1. Institutional background

Prior to 2016, the Chinese social health insurance system mainly consisted of UEBMI, NRCMS and URBMI. These three schemes are separately managed and operated at local levels, with different premiums, financing, and benefits packages. The UEBMI system was first introduced as a pilot program in the Two Rivers region of China in 1994, before its official national implementation in 1998 (Su et al., 2019). There was a marked gap between UEBMI and the other two schemes as UEBMI was mandatory for all urban employees and retirees in China. It included comprehensive benefits and was financed by high contributions from both employers and employees.

The NRCMS targeting the rural population was first proposed by the Chinese government in 2003 and was rolled out nationwide in 2007. The URBMI was piloted in 88 cities in China in 2007 and was fully implemented in 2010. It was a healthcare insurance system designed for urban residents who were not covered by UEBMI (Yip et al., 2019). The NRCMS and the URBMI programs are two major voluntary and subsidized schemes which offer a lower level of healthcare protection than UEBMI. These two schemes primarily cover inpatient and outpatient major medical needs for the insured. It is important to note that these two programs operate independently and are strictly divided based on the Hukou system (Chen et al., 2022). Hukou, which roughly translates as “Household Registration System” or “Residence System”, divides people into rural and urban residents.

By 2010, the three social health insurance schemes covered 95% of the total population (more than 1.27 billion people), among which NRCMS, URBMI, and UEBMI accounted for 66%, 15%, and 19%, respectively (Zhou et al., 2022). This achievement means China has taken the first key step towards UHC. However, the long-lasting systematic division in the health insurance system has become the main obstacle to meeting a growing demand for equity and social protection.

3.2.2. A brief overview of URRBMI integration reform

To eliminate access and health disparities, some provinces and municipalities have made a series of attempts to merge NRCMS and URBMI since 2009. In January 2016, the State Council issued Opinions on the Integration of the Basic Medical Insurance System for Urban and Rural Residents. It announced the combination of URBMI and NRCMS into a unified basic medical insurance named URRBMI, covering all urban and rural residents except those who should be covered by UEBMI.

The goal of the policy is to provide equitable, affordable, and efficient health care for all citizens. The Opinions guided six key areas of consolidation, which included integrating the coverage of the medical insurance system, the fundraising policies, the benefits, the health insurance directory, the management of funds and selected agencies of the URBMI and NRCMS. All provincial governments were required to follow the policy advisory from the State Council and modifications were allowed to be made based on their local conditions. Prefecture- or county-level governments made more concrete plans and carried them out under the guidance of provincial governments accordingly. Under top-down pressure from the central government, the reform has been implemented nationwide gradually by 2020.

3.2.3. Possible channels affecting healthcare usage

Table 3.1 gives a brief comparison of the basic health insurance schemes before and after URRBMI. The integration is likely to stimulate rural residents' usage of medical services through several channels. Firstly, the fund pooling and management of NRCMS were moved up from the county level to the municipal level. The expanded and upgraded funding pool provides more stable and sustainable funding which facilitates greater healthcare utilisation. In addition, the treatment levels have markedly improved alongside the expanded scope of hospitals designated for health insurance. This enlarges people's healthcare options and enables rural residents to seek care in higher-level facilities. Moreover, more generous benefit packages help decrease medical prices when enrollees utilize medical services and then promote healthcare usage after the integration. Specifically, on the one hand, unified coverage brings an expanded list of medical insurance drugs and medical service items. On the other hand, most reimbursement benefits have increased after the reform, such as higher reimbursement rates and larger reimbursement caps. Besides, the reimbursement mechanism has changed from "pay first and claim reimbursement later" to immediate reimbursement in some areas. Those benefit changes reduce the financial burden of medical care and enable individuals to obtain health services at affordable prices. However, as reimbursement deductibles have also increased, whether the actual reimbursement ratio has improved is unknown. Furthermore, the new URRBMI scheme covers more inpatient benefits and relatively limited outpatient benefits. Therefore, the impact of the integration on inpatient and outpatient care usage may vary.

3.2.4. Possible source for treatment effect heterogeneity

Given the staggered policy implementation, there are some possible sources for heterogeneous treatment effects. For pilot areas, their first exploration of integration before the issue of the Opinions was based on local conditions, such as local fiscal capacity and their ability to narrow the gaps between the subsystems. Those areas further refined and improved the reform under the central government's guidance after 2016. This may bring treatment effect heterogeneity across periods. Additionally, the implementation of integration displays heterogeneity across provinces and even smaller geographic units, such as municipalities and counties. Benefit levels vary by region. Local governments with higher fiscal power and social risk are more capable of extending generous health insurance benefits (Ratigan, 2017; Meng and Su, 2021). Moreover, as the nationwide unification reform was endorsed by the central government in 2016, it is not clear whether this transition goal has been translated into effective policies and procedures in lower levels of government.

Table 3.1 Summary of basic health insurance schemes before and after URRBMI

Insurance	Before integration		After integration
	URBMI	NRCMS	URRBMI
Enrolment unit	Voluntary	Voluntary	Voluntary
Enrolment (n, %)	221.16 million (82.90%)	832.00 million (97.50%)	897.36 million (98.00%)
Risk-pooling	Municipal level	County level	Municipal level
Source of financing	Per capita premium was 360 RMB, government subsided 282 RMB (78.3%), individual paid 78 RMB (21.7%)	Per capita premium was 370 RMB, government subsided 303 RMB (81.9%), individual paid 67 RMB (18.1%)	Per capita premium was 723 RMB, government subsided 497 RMB (68.7%), individual paid 226 RMB (31.3%)
Number of drugs covered (insurance coverage, %)	2208 (48.35%)	1746 (30.30%)	The drug directories of URBMI and NRCMS were merged, the merged catalogue was expanded and not less than any of them
Insurance benefits	Mainly covered inpatient care, supplemented with outpatient care for catastrophic diseases	Mainly covered inpatient care, supplemented with outpatient care for catastrophic diseases and relatively expensive outpatient care	Mainly covered inpatient care, supplemented with outpatient care with serious illnesses and relatively expensive outpatient care
Inpatient reimbursement ratio	62%	66%	75%

Note: Data were from the 2011 - 2018 National Health Statistical Yearbook; The National Statistical Bulletin on the development of basic medical insurance in 2018. The data of enrolment, source of financing, number of drugs covered and inpatient reimbursement ratio for URBMI and NRCMS were obtained around 2011, the data for URRBMI was obtained around 2018. 1 RMB = 0.16 USD in 2018

3.3. Data and Study Design

3.3.1. Data and variables

3.3.1.1. Data source

We compile our data using the city-year level datasets on the staggered implementation of the integration policy and an individual-level panel survey. The policy dataset is constructed from a comprehensive collection of local health insurance regulations and government documents. All related documents are hand-collected from official websites of prefectural cities' governments and human resource and social security bureaus.

The individual-level data are obtained from the China Health and Retirement Longitudinal Study (CHARLS), including a baseline survey in 2011 (wave 1) and the follow-up survey in 2013 (wave 2), 2015 (wave 3) and 2018 (wave 4). CHARLS is a nationally representative dataset that aims to collect high-quality microdata representing families and individuals aged 45 years and older in China. The CHARLS questionnaire includes a comprehensive set of information on demographic characteristics, socio-economic factors, health, healthcare and insurance. To ensure the adoption of best practices and international comparability or results, CHARLS shares the same basic guidelines as the Health and Retirement Study (HRS) and related ageing surveys such as the English Longitudinal Study of Aging and the Survey of Health, Aging and Retirement in Europe (Zhao et al., 2020). The CHARLS is the only large-scale national representative data in China with questions specific to health insurance including URBMI, NRCMS and URRBMI before and after the reform. Other national longitudinal survey datasets used in the literature have limitations to some extent. The China Household Finance Survey (CHFS) doesn't collect information on outpatient care utilisation but only focuses on inpatient care utilisation (Zhou et al., 2022). The China Family Panel Studies (CFPS) doesn't include questions related to insurance type in its early waves (Yang et al., 2022).

3.3.1.2. Sample restriction

Before restrictions, the overall sample size is 77,233 across four waves. The first wave of CHARLS was conducted between June 2011 and March 2012. This initial sample included 17,708 respondents in 10,257 households in 450 villages/urban communities in 150 counties/districts in 28 provinces (Zhao et al., 2020). In general, the CHARLS baseline is a good representation of the middle-aged and elderly population of China. The overall response rate was 80.51%, of which the rural response rate was as high as 94.15% (Su et al.,2019). The second wave was conducted between July 2013 and January 2014 and included a refreshment sample consisting of individuals aged between 43 and 44 at Wave 1 and their partners. The third wave was conducted between July 2015 and January 2016 and included a refreshment sample consisting of individuals aged between 41 and 42 at Wave 1 and their partners. The fourth wave was conducted between July and November 2018 and included a refreshment sample consisting of individuals who were 40 years old at Wave 1 and their partners.

The sample is constructed by applying the following screening procedures. First, the sample is restricted to the rural population (defined by hukou status: with agricultural hukou) who have NRCMS in wave 1 and only have one (public) health insurance in the following waves. This is related to our evaluation objective, which is investigating the integration reform impact on rural residents who were eligible for NRCMS before the reform. We also exclude migrants whose health insurance account was set up in a county different from their residential county. After the first step of restrictions, the overall sample size is 44,334 across four waves. Second, the balanced panel data is constructed by dropping individuals who do not have continuous observations in waves 1-4. The sample size has been reduced to 20,112. Third, we drop those individuals in jurisdictions where the timing of the reform implementation was unclear or where the reform was implemented before wave 1. After this step, the sample size is 17,280. Fourth, we restrict the interview month of wave2-4 to July and August. The survey is mainly conducted from around July to August every year, however, there are few respondents interviewed after August. The restriction of interview month helps us to observe a more accurate length of policy exposure, which is useful to investigate the dynamic treatment effect more precisely in the following section. At this step, the sample size decreases to 14,592. Fifth, as extreme values might lead to unreliable results, we winsorize continuous outcome variables at the 99.9th percentile. Respondents with missing values in outcome variables and covariates are also excluded. In addition, CHARLS is a panel survey of people aged 45 and over and their partners regardless of age in China. Given the elderly nature of the CHARLS survey and the low numbers of young people in the sample, we delete observations who are younger than 40 years old. We also delete observations who are older than 90 years old in case their self-reported information lacks credibility. Overall, the final sample for our analysis contains 14320 observations in 83 prefecture-level cities in 21 provinces (including municipalities and autonomous regions) for 4 waves. Table 3.2 gives an overview of sample loss in each restriction step.

Table 3.2 Sample restrictions

Restriction steps	Wave 1	Wave 2	Wave 3	Wave 4	Total observations
No restrictions	17708	18612	21097	19816	77233
1.Health insurance	11255	11618	10719	10742	44334
2.Panel data	5028	5028	5028	5028	20112
3. Policy implementation time	3853	3853	3853	3853	17280
4. Interview time	3648	3648	3648	3648	14592
5. Outliers and missing values	3580	3580	3580	3580	14320

3.3.1.3. Evaluation set-up

Figure 3.1 provides the timeline of the urban-rural health insurance integration reform implementation with the corresponding CHARLS survey time in our sample. It indicates the exact time when each prefectural-level city fully implemented the reform. As mentioned in the sample restriction section, cities with an unclear implementation time or those that implemented the reform prior to 2011 were not included in the study.

We consider the case with $\mathrm{T} = 4$ periods (from Wave1-Wave4) and denote a particular time period by t where $t = 1, \cdots, 4$. The first time of intervention entry is

$q = t = 2$ (Wave 2). At period $q + 1 = 3$, more cities join the treated group, and so on, until period $t = q + 2 = T = 4$. In the final period $t = T = 4$, there remain 9 cities untreated. There are three different cohorts g according to the specific first treated period, $g \in \{q, \cdots, T\}$ and a never-treated group (see Table 3.3). The evaluation period is limited from Wave 1 to Wave 4, and the not-yet-treated group is used as the comparison groups. Our identification strategy follows Wooldridge (2021) to allow treatment effect heterogeneity.

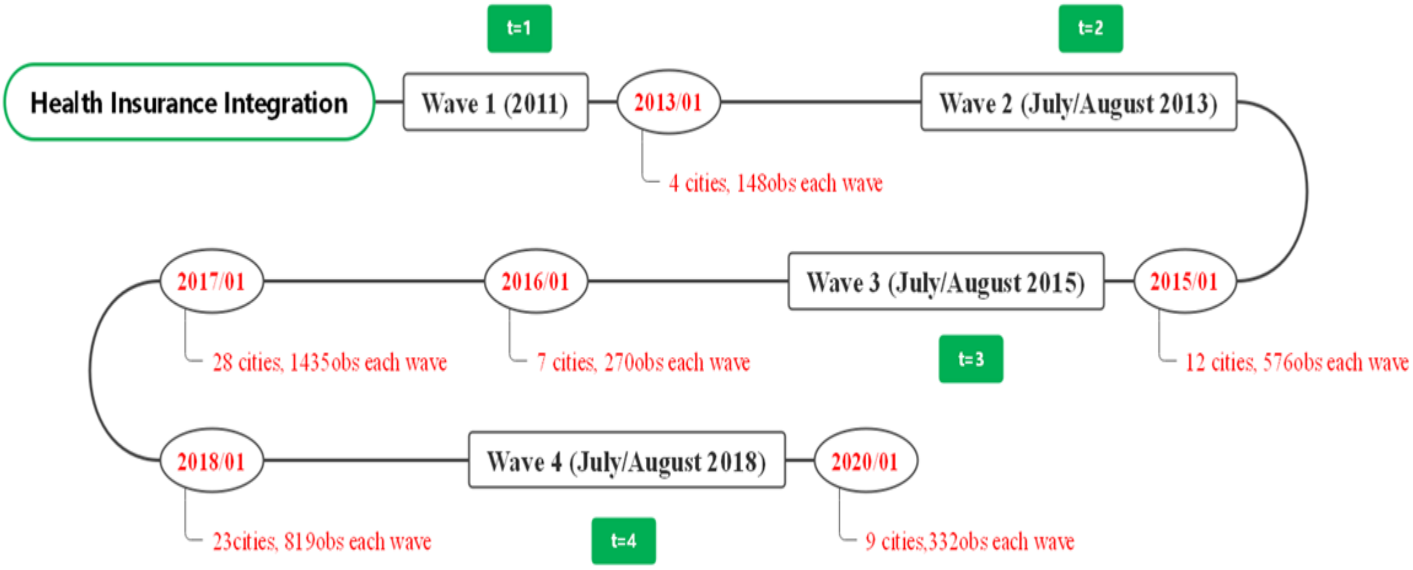

Health Insurance Integration
t=1
Wave 1 (2011)
2013/01
4 cities, 148obs each wave
t=2
Wave 2 (July/August 2013)
2015/01
12 cities, 576obs each wave
Wave 3 (July/August 2015)
t=3
2016/01
7 cities, 270obs each wave
2017/01
28 cities, 1435obs each wave
2018/01
23cities, 819obs each wave
Wave 4 (July/August 2018)
t=4
2020/01
9 cities,332obs each wave

Figure 3.1 Timeline of the staggered reform implementation

Table 3.3 Cities in Different Cohorts

First Entry Time	Wave	Cohort	Provinces in each cohort
t=2	2	g=2	4 cities, 148 observations each wave
t=3	3	g=3	12 cities, 576 observations each wave
t=4	4	g=4	58 cities, 2524 observations each wave
Never treated	-	-	9 cities, 332 observations each wave

3.3.1.4. Variables

The key independent variable in this study is the health insurance integration treatment variable. We define two different treatment variables. One is the time-invariant treatment cohort dummy, which indicates when a unit was first subjected to the intervention. Another is the time-varying treatment dummy, which equals one when a unit is eventually treated and at its post-treatment periods. Both treatment variables are based on the city's policy implementation time. The main outcome variables include the probability of using inpatient and outpatient healthcare services, duration of hospital stay and the number of outpatient visits. To help isolate the impact of integration from other underlying differences, we control for individual demographic characteristics, including age, gender, marital status and education years. Detailed variable definitions are shown in Table 3.4.

Table 3.4 Variable names and definitions

variables	Definitions
Inpatient probability	Binary variable-whether receive inpatient care in the past year: 1, if yes; 0, if no
Hospital nights	hospital nights for the most recent hospitalization in the past year
Outpatient probability	Binary variable-whether receive outpatient care in the last month: 1, if yes; 0, if no
Outpatient visits	the number of outpatient care visits in the last month
Gender	Binary variable: 1, if female; 0, if male
Education years	Recoding education years according to categorical education variable: 0 years, if no formal education (illiterate); 3 years if did not finish primary school but can read; 5 years if Sishu (Private tutoring) or elementary school; 8 years if middle school; 11 years if high school or vocational school; 13 years if two/three-year college; 15 years if bachelor's degree; 18 years if post-graduated
Marriage	Binary variable: 1, if married with spouse present, married but not living with spouse temporarily for reasons such as work or cohabitated; 0, if separated, divorced, widowed or never married
Age	Age at interview

Table 3.5 reports the summary statistics of variables, broken down by treatment cohort and wave. As mentioned in Figure 3.1, most cities implemented the integration policy between wave 3 and wave 4. Therefore, in our evaluation period, cohort 4 is the last treated group and is also the biggest group, which contains 2524 observations in each wave. Cohort 2 is the earliest treated group, which is also the smallest group including 148 observations in each wave. Overall, the average age of the sample is around 57

years old and the average education years is relatively low, at around 4 years. Over 90% of the sample is married. The disparities in healthcare utilisation among different cohorts are clear. Cohort 4 has the highest probability of using inpatient and outpatient services at wave 1, which are 8.2% and 19,5%, respectively. Cohort 4 also has the largest duration of hospital stays at wave 1, which is 0.809 nights and increases to 1.833 nights at wave 4. The number of outpatient visits in cohort 2 is the biggest at wave 1, which is 0.439.

Table 3.5 Descriptive statistics

Variable	Obs each	Wave1		Wave2		Wave3		Wave4	
	eave	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cohort2									
Inpatient probability	148	0.0473	0.213	0.0676	0.252	0.101	0.303	0.169	0.376
Hospital nights	148	0.318	1.641	0.520	2.302	0.858	3.385	1.318	3.186
Outpatient probability	148	0.176	0.382	0.304	0.462	0.223	0.418	0.216	0.413
Outpatient visit times	148	0.439	1.263	0.527	0.907	0.574	1.429	0.507	1.417
Gender	148	0.574	0.496	0.574	0.496	0.574	0.496	0.574	0.496
Education years	148	4.277	3.559	4.277	3.559	4.277	3.559	4.277	3.559
Marriage	148	0.899	0.303	0.878	0.328	0.885	0.320	0.872	0.336
Age	148	57.91	8.867	59.87	8.832	61.87	8.783	64.85	8.766
Cohort3									
Inpatient probability	576	0.0556	0.229	0.0868	0.282	0.0938	0.292	0.106	0.308
Hospital nights	576	0.578	2.806	0.984	4.052	0.950	3.416	1.134	4.011
Outpatient probability	576	0.137	0.344	0.123	0.329	0.125	0.331	0.120	0.325
Outpatient visit times	576	0.330	1.248	0.243	0.834	0.319	1.232	0.281	1.067
Gender	576	0.538	0.499	0.538	0.499	0.538	0.499	0.538	0.499
Education years	576	3.953	3.512	3.953	3.512	3.953	3.512	3.953	3.512
Marriage	576	0.908	0.289	0.892	0.310	0.870	0.337	0.837	0.370
Age	576	57.42	8.533	59.44	8.545	61.43	8.534	64.42	8.545
Cohort4									
Inpatient probability	2,524	0.0820	0.274	0.122	0.327	0.132	0.339	0.177	0.382
Hospital nights	2,524	0.809	3.742	1.346	4.783	1.411	4.900	1.833	5.044
Outpatient probability	2,524	0.195	0.396	0.221	0.415	0.200	0.400	0.177	0.381
Outpatient visit times	2,524	0.403	1.119	0.503	1.331	0.443	1.205	0.362	1.048
Gender	2,524	0.549	0.498	0.549	0.498	0.549	0.498	0.549	0.498
Education years	2,524	3.930	3.362	3.930	3.362	3.930	3.362	3.930	3.362
Marriage	2,524	0.916	0.278	0.901	0.298	0.889	0.314	0.853	0.355
Age	2,524	57.41	8.556	59.40	8.559	61.39	8.558	64.39	8.559
Never treated									

Inpatient probability	332	0.0572	0.233	0.105	0.308	0.142	0.349	0.123	0.330
Hospital nights	332	0.645	3.114	1.515	5.411	1.340	4.059	1.292	4.022
Outpatient probability	332	0.133	0.340	0.190	0.393	0.160	0.367	0.120	0.326
Outpatient visit times	332	0.187	0.578	0.404	1.124	0.289	0.782	0.208	0.657
Gender	332	0.530	0.500	0.530	0.500	0.530	0.500	0.530	0.500
Education years	332	4.587	3.177	4.587	3.177	4.587	3.177	4.587	3.177
Marriage	332	0.940	0.238	0.934	0.249	0.916	0.278	0.880	0.326
Age	332	56.52	7.687	58.52	7.696	60.50	7.705	63.51	7.694

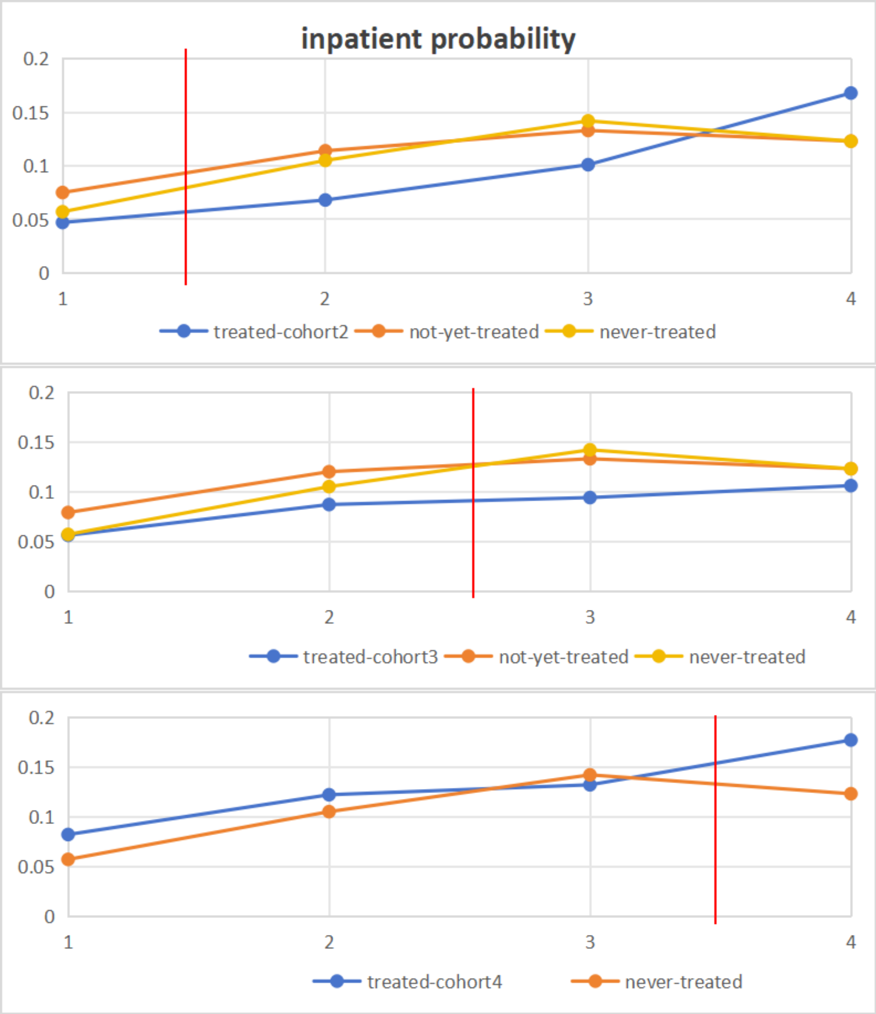


Figure 3.2 Inpatient probability comparisons between treated and control groups

Figure 3.2 compares the trend of inpatient probability between different treated cohorts and their corresponding control groups, which provides insight into the parallel trend assumption (See Appendix Figure C1 for other outcome variables). For cohort 2, the policy implementation time is between wave 1 and wave 2, and as there is only one period before treatment, it is hard to capture the pre-trend. For cohort 3, which implemented the policy between wave 2 and wave 3, the pre-trend appears parallel to the not-yet-treated group. For cohort 4, the comparison group becomes the never-treated group. Before implementing the policy, there was an upward trend of inpatient probability from wave 1 to wave 3 for cohort 4, which is similar to the trend of inpatient probability for the never-treated group.

3.3.2. Empirical models

3.3.2.1. Wooldridge (2021) Extended TWFEDID estimator

Wooldridge (2021) proposes an extended TWFEDID estimator to allow considerable heterogeneity and to avoid the corresponding pitfalls of using the standard TWFEDID estimator in the design of staggered interventions. The paper also establishes the equivalence between the extended TWFEDID estimator and an estimator obtained from a pooled ordinary least squares regression that includes unit-specific time averages and time-period specific cross-sectional averages.

Following Wooldridge (2021), Equation (1) shows how to estimate the average treatment effects on the treated in a particular cohort (cohort is defined by the first treated time) at a particular post-treatment period.

$$\begin{aligned} Y_{it} &= \delta + X_{it}\,K + \sum\nolimits_{g=q}^{T} \gamma_g D_{ig} + \sum\nolimits_{g=q}^{T} \sigma_g (D_{ig} \cdot X_{it}) \\ &+ \sum_{s=2}^{T} \theta_s f_{st} + \sum\nolimits_{s=2}^{T} \mu_s (f_{st} \cdot X_{it}) + \sum_{g=q}^{T} \sum_{s=g}^{T} \boldsymbol{\tau}_{\boldsymbol{gs}} \left(W_{it} \cdot D_{ig} \cdot f_{st}\right) \\ &+ \sum_{g=q}^{T} \sum_{s=g}^{T} \rho_{gs} \left(W_{it} \cdot D_{ig} \cdot f_{st} \cdot \dot{X}_{igt}\right) \qquad (1) \end{aligned}$$

$$\dot{X}_{igt} = X_{it} - \bar{X}_{gt} \qquad (2)$$

Y_{it} is the outcome variable for individual i in period t. D_{ig} is the time-invariant cohort dummy for i, indicating when i was first treated, $g \in \{q, \cdots, T\}$. f_{st}

is the time dummy, which equals one if $s = t$. W_{it} is the time-varying treatment indicator, which equals one if i is eventually treated and period t is in i's post-treatment period. Interacting the cohort dummies with the time dummies corresponding to those periods where a cohort is treated allows ATTs to vary by cohort and calendar time. γ captures cohort effects, and θ captures time effects. X_{it} is a set of covariates. Interaction terms $D_{ig} \cdot X_{it}$ and $f_{st} \cdot X_{it}$ indicate that the effects of covariates can change with treatment cohort and calendar time; and the interaction term $W_{it} \cdot D_{ig} \cdot f_{st} \cdot \dot{X}_{igt}$ implies treatment effects are allowed to change with cohort, calendar time and controls. Here, as Equation (2) shows, the controls have been centred around the mean $\bar{X}_{gt}$ for cohort g in period t. This ensures τ_{gs} to be the ATT for cohort g in periods in which cities are subjected to the intervention, $s \in \{q, \cdots, T\}$.

While choosing appropriate covariates, there are two aspects requiring attention. For time-invariant controls, Wooldridge (2021) proves that including time-constant controls and interacting them with time-constant cohort indicators D_{ig} does not change the estimate of ATTs. Therefore, time-constant controls are only added when interaction terms $f_{st} \cdot X_i$ and $W_{it} \cdot D_{ig} \cdot f_{st} \cdot \dot{X}_{ig}$ need to be included, which allows the effects of the covariates in the untreated state to change over time and allows ATTs to change with time-constant controls. For time-varying controls, researchers are cautious to include them unless they are strictly exogenous. At a minimum, time-varying covariates should not be influenced by the policy intervention. Based on these restrictions, covariates in our paper include two time-invariant controls “gender”, “education years” and two time-varying controls “marriage status”, “age”.

3.3.2.2. Special cases and model extensions

There are some special cases based on the general Eq (1). The general equation allows time-varying covariates and the special cases with different assumptions are “X is null” (which implies the unconditional common trend assumptions) or “only allow time-constant covariates”. We include two special cases as robust tests.

We further extended the model to allow multiple treatment levels. In principle, we could define more detailed cohorts based on the initial length of exposure as well as the first treated period. In this study, as the fact that the exact gaps between two adjacent periods (waves) are 2 or 3 years, even in the same cohort (defined by the first treated period as before), the exact length of exposure for different cities at the following post periods varies. Therefore, we define more detailed cohort indicators D_{iga} where g is the initial treatment period and a is the initial length of exposure (see Table 3.6). Specifically, we define a according to the length of exposure at the first post-period. For example, for the previous cohort 4, 58 cities' first intervention period is period 4 (see Figure 3.1). However, 23 cities were implemented in 2018 Jan, therefore, a=1, which means the length of exposure in period 4 is about 0.5 years. 28 cities implemented in 2017 Jan, therefore, a=2, which means the length of exposure at period 4 is about 1.5 years. 7 cities implemented in 2016 Jan, therefore, a=3, which means the length of exposure at period 4 is about 2.5 years. By replacing the cohort indicator D_{ig} in Eq (1) with D_{iga}, then we get the parameters τ_{gas}, $s \in \{q, \cdots, T\}$, which is the ATT for cohort g with a initial treatment duration at post-period s.

Table 3.6 Cities in more detailed cohorts

First Entry Time	Wave	Cohort	a	Provinces in each cohort
t=2	2	g=2	1	4 cities, 148 observations each wave
t=3	3	g=3	1	12 cities, 576 observations each wave
t=4	4	g=4	1	23 cities, 819 observations each wave
t=4	4	g=4	2	28 cities, 1435 observations each wave
t=4	4	g=4	3	7 cities, 270 observations each wave
Never treated	-	-	-	9 cities, 332 observations each wave

3.3.2.3. Common trend test

Generally, to test common trends assumption, we could write the augmented equation as

$$
\begin{aligned}
Y_{it} &= \delta + X_{it}K + \sum\nolimits_{g=q}^{T} \gamma_g D_{ig} + \sum\nolimits_{g=q}^{T} \sigma_g (D_{ig} \cdot X_{it}) \\
&+ \sum_{s=2}^{T} \theta_s f_{st} + \sum\nolimits_{s=2}^{T} \mu_s (f_{st} \cdot X_{it}) + \sum_{g=q}^{T} \sum_{s=g}^{T} \boldsymbol{\tau_{gs}} \left(W_{it} \cdot D_{ig} \cdot f_{st} \right) \\
&+ \sum_{g=q}^{T} \sum_{s=g}^{T} \rho_{gs} \left(W_{it} \cdot D_{ig} \cdot f_{st} \cdot \dot{X}_{igt} \right) \\
&+ \sum_{g=q}^{T} \sum_{s=2}^{g-1} \boldsymbol{\tau_{gs}} \left(W_{it} \cdot D_{ig} \cdot f_{st} \right) \quad (3)
\end{aligned}
$$

Then, we could jointly test the null hypothesis: $\tau_{gs} = 0, g = q, \cdots, T; s = 2, \cdots, g-1$. However, this typical event-study-type approach can result in many restrictions to test. The test is also hard to check for alternatives where violations of common trends may depend on covariates as it would be costly in terms of degrees of freedom. Alternatively, we could replace the variables $W_{it} \cdot D_{ig} \cdot f_{st}$ $(g = q, \cdots, T; s = 2, \cdots, g-1)$ in Eq(3) with the cohort-specific linear trends, $D_{ig} \cdot t(\, t = 1, \cdots, T - 1;\ g = q, \cdots, T)$. Although with many pre-treatment periods, one could add more functions of time, such as $D_{ig} \cdot t^2$ for each cohort, it seems that if important differences in trends are present, a linear trend will identify these in most cases (Wooldridge, 2023). This test conserves degrees of freedom. Generally, there is a trade-off between the event-study-type test and the heterogeneous linear trend test because the latter has fewer degrees of freedom but does not search in all directions where a common trend might be violated.

The tests only provide information about whether the common trends assumption holds, however, we might also be curious about what to do when there is a violation of parallel trends. As discussed in Wooldridge (2021), the event-study approach is generally inappropriate as a correction for pre-trends, as it would require that violation of parallel trends disappear immediately when treatment occurs. By contrast, the assumption that each cohort has a separate linear trend in the absence of the intervention is a reasonable- albeit not completely general-model of heterogeneous trends. In other words, the linear trend specification allows for a constant difference in trends between the treated units and the control units. Therefore, we further discuss the model allowing cohort-specific linear trends by interpreting the results with $D_{ig} \cdot t$ when the common trends assumption doesn't hold. With enough data, we might also include interactions $D_{ig} \cdot t \cdot X_i$ to allow pre-trends to depend on the observed covariates.

3.4. Empirical Results

3.4.1. Results for inpatient healthcare utilisation

3.4.1.1. Basic models for inpatient healthcare utilisation

Table 3.7 displays the results for inpatient healthcare utilisation, including the probability of having hospitalisation and hospital nights. Standard errors are clustered at the city level. Columns (1) and (3) show results without covariates, and columns (2) and (4) show results allowing both time-invariant and time-varying controls. Including controls has little impact on the ATT estimates or their standard errors. However, coefficients are uniformly larger in magnitude. The parallel trends test has large p-values, providing no evidence against the parallel trend assumption in the linear means. As discussed in Section 3.3.2.3, two kinds of common trend tests are employed in this study. First, the event-study type test is explored, which is similar to the general approach in the standard DID setup. Second, we follow Wooldridge (2021) to assume a cohort-specific linear trend. In either case, we are testing for pre-trends prior to the first intervention period in the staggered set-up. The evaluation period here is relatively short. For the earliest treated group, as there is only one period before policy intervention, it is not feasible to assume a cohort-specific nonlinear trend. There would be concern about whether a cohort-specific linear trend is enough to capture the pre-trends. Wooldridge (2021) notes that a linear trend will pick up important differences in trends in most cases. The bottom of Table 3.7 shows the results of these tests.

Table 3.7 ATTs for inpatient healthcare utilisation

	Probability		Hospital nights	
	No covariates	With covariates	No covariates	With covariates
Cohort-period ATTs	(1)	(2)	(3)	(4)
ATT（2, 2）	-0.0188	-0.0149	-0.3448	-0.3287
	(0.0408)	(0.0385)	(0.2878)	(0.2626)
ATT（2, 3）	0.0006	0.0050	-0.0587	-0.0741
	(0.0115)	(0.0137)	(0.1486)	(0.1429)
ATT（2, 4）	0.0687***	0.0817***	0.4923*	0.5357**
	(0.0209)	(0.0205)	(0.2717)	(0.2302)
ATT（3, 3）	-0.0114	-0.0142	-0.1571	-0.2114
	(0.0180)	(0.0182)	(0.1755)	(0.1801)
ATT（3, 4）	0.0013	-0.0034	0.1184	0.0680
	(0.0225)	(0.0274)	(0.2944)	(0.3469)
ATT（4, 4）	0.0430**	0.0440**	0.5187**	0.5681***
	(0.0166)	(0.0177)	(0.2050)	(0.1826)
Aggregation to a single effect	0.0268**	0.0271*	0.3156*	0.3328**
	(0.0133)	(0.0144)	(0.1609)	(0.1482)
Pre-trend test	√	√	√	√
Event Study p-value (3 df)	0.5279	0.4857	0.6192	0.5411
Heterogeneous Trend Test (2 df)	0.4201	0.3843	0.6811	0.5707
N	14320	14320	14320	14320

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

Column (1) provides pooled OLS estimators without covariates for the probability of an inpatient stay. Overall, the reform significantly increases the rural population’s probability of using inpatient healthcare services by 2.68%, a 35.40% increase on the baseline level in treated cities. We define one period before the integration reform as the baseline level. Here, the baseline level of inpatient care utilisation is 7.57%. However, the canonical TWFEDID estimator is 0.77% (a 10.17% increase of the baseline level), which is smaller than the aggregation of heterogeneous ATTs and insignificant. We follow Goodman-Bacon (2021) to decompose the TWFE estimator (see Appendix Table A1). Goodman-Bacon (2021) shows that the TWFE estimator equals a weighted average of all possible two-group/two-period DID estimators that compare timing groups to each other, some of which have no causal interpretation as using an earlier treated group as a control for a later treated group. The decomposition results show that we indeed use the early-treated cohort as a comparison group for the later-treated cohort. It adjusts the path of outcomes for newly treated units by the path of outcomes for already treated units, which includes treatment effect dynamics and leads to bias. Additionally, weights for each two-group/two-period parameter are sensitive to the size of each group, the timing of treatment, and the total number of periods (Callaway and Sant'Anna, 2021). It gives larger weights for those groups whose treatment occurs closer to the middle of the time window. The forbidden comparison group and unsuitable weights make the total coefficient smaller than the actual value in this study.

Column (2) in Table 5.1 provides pooled OLS estimators with covariates for inpatient probability. Overall, the reform significantly increases the rural population's probability of using inpatient healthcare services by 2.71% (a 35.80% increase of the baseline level, i.e., 7.57 percentage points). There is treatment effect heterogeneity across both cohorts and periods. On the one hand, ATTs increase by the length of exposure within the cohort: for cohort 2, the ATT in period 2 is about -1.49% (a 31.51% decrease of the baseline level, i.e., 4.73 percentage point) and slightly increases to 0.5% (a 10.57% increase of the baseline level) in period 3, then rapidly increases to 8.17% (a 172.76% increase of the baseline level) in period 4. On the other hand, given the same event study relative period, ATT for the later treated cohort is bigger than the earlier treated cohort: for cohort 4, the ATT at period 4 is about 4.40% (a 33.26% increase of the baseline level, i.e., 13.23 percentage point), while the ATT for cohort 3 at period 4 is about -1.42% (a 16.36% decrease of the baseline level, i.e., 8.68 percentage point).

Column (4) shows pooled OLS estimators with covariates for hospital nights. Overall, the reform significantly increases the rural population's hospital stays by 0.33 nights (a 44% increase of the baseline level, i.e., 0.75 nights). Treatment effect heterogeneity across both cohorts and periods is also observed.

Tere are a few negative coefficients in Table 3.7. There are two possible reasons for negative coefficients in the first or second post-period for some cohorts: Firstly, the indicators for inpatient healthcare utilisation are related to "in the past year", which refers to approximately one year ago until interview time. Secondly, it takes time for the reform to work and a limited short-term impact would appear reasonable.

3.4.1.2. More detailed cohorts for treatment heterogeneity analysis

The analysis of treatment heterogeneity requires attention. For example, although we find heterogeneity across different cohorts with the same event study relative period by comparing ATT(2, 2) and ATT(4, 4), the length of exposure for cohort 2 and cohort 4 at period 4 is different. For cohort 2, the approximate length of exposure at period 2 is 0.5 years. However, for cohort 4, it aggregates cities with 0.5/1.5/2.5-year exposure in period 4. Therefore, we introduce the model allowing for multiple treatment levels. Table 3.8 provides the results for inpatient healthcare utilisation with more detailed cohort information. It illustrates treatment effect heterogeneity more precisely. Similar to what is revealed in Table 3.7, including controls has little impact on the ATT estimates or their standard errors. Column (1) shows pooled OLS estimators with controls for the probability of an inpatient stay. There is heterogeneity across cohorts and periods. ATTs increase by the length of exposure within the same cohort (see comparisons among ATT(2,1,2), ATT(2,1,3) and ATT(2,1,4)). Given the same length of exposure, ATTs for the later treated cohort is larger than the earlier treated cohort: Given 0.5-year exposure, ATT for cohort 3 at period 3 (ATT(3,1,3)) is about -1.41% (a 12.24% decrease of the baseline level, i.e., 8.68 percentage point), while the ATT for cohort 4 at period 4 (ATT(4,1,4)) is about 4.66% (a 38.17% increase of the baseline level, i.e., 12.21 percentage point). Given a 2.5-year exposure, ATT for cohort 2 at period 3 (ATT(2,1,3)) is about 0.51% (a 10.78% increase of the baseline level), while ATT for cohort 4 at period 4 (ATT(4,3,4)) is about 1.69% (a 12.77% increase of the baseline level). Column (2) displays a similar pattern of treatment effect heterogeneity for the number of hospital nights stayed.

Table 3.8 ATTs for inpatient healthcare utilisation (more detailed cohorts)

	Length of exposure (year)		Probability	Hospital nights
	Initial	Total	(1)	(2)
ATT（2, 1, 2）	0.5	0.5	-0.0149	-0.3272
			(0.0385)	(0.2628)
ATT（2, 1, 3）	0.5	2.5	0.0051	-0.0711
			(0.0136)	(0.1426)
ATT（2, 1, 4）	0.5	5.5	0.0818***	0.5377**
			(0.0205)	(0.2303)
ATT（3, 1, 3）	0.5	0.5	-0.0141	-0.2098
			(0.0182)	(0.1800)
ATT（3, 1, 4）	0.5	3.5	-0.0034	0.0688
			(0.0275)	(0.3471)
ATT（4, 1, 4）	0.5	0.5	0.0466**	0.5350**
			(0.0217)	(0.2362)
ATT（4, 2, 4）	1.5	1.5	0.0474**	0.6177***
			(0.0201)	(0.2238)
ATT（4, 3, 4）	2.5	2.5	0.0169	0.3761
			(0.0203)	(0.2571)
Pre-trend test	-	-	√	√
Event Study p-value (7 df)	-	-	0.2319	0.1581
Heterogeneous Trend Test (4 df)	-	-	0.1266	0.0635
N	-	-	14320	14320

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

3.4.2. Results for outpatient healthcare utilisation

3.4.2.1. Basic models for outpatient healthcare utilisation

Table 3.9 shows the results for outpatient healthcare utilisation. Coefficients and standard errors are uniformly larger in magnitude when including covariates. Overall, the integration reform has positive but insignificant effects on the probability of outpatient care and the number of visits. Analysis of separate cohort-period ATTs displays significant effects only for cohort 2. In detail, the reform improves the probability of using outpatient services by 10.86% (a 61.81% increase of the baseline level, i.e., 17.57%). However, there are concerns about the credibility of the common trends assumption for outpatient visit times. We plot the observed means to assess the evolution of outpatient visit times for each cohort respectively (See Appendix Figure

C1). It is not surprising that opposing pre-treatment trends between cohort 3 (treated) and the control group are observed.

Table 3.9 ATTs for outpatient healthcare utilisation

	Probability		Visit times	
	No covariates	With covariates	No covariates	With covariates
Cohort-period ATTs	(1)	(2)	(3)	(4)
ATT（2, 2）	0.1059***	0.1204***	0.0077	0.0543
	(0.0335)	(0.0303)	(0.1109)	(0.0946)
ATT（2, 3）	0.0429*	0.0756***	0.1043	0.2113**
	(0.0246)	(0.0257)	(0.1183)	(0.1016)
ATT（2, 4）	0.0718*	0.1299***	0.1159	0.2980***
	(0.0389)	(0.0410)	(0.0811)	(0.1075)
ATT（3, 3）	0.0016	-0.0038	0.0423	0.0491
	(0.0145)	(0.0136)	(0.0582)	(0.0525)
ATT（3, 4）	0.0320	0.0176	0.0832	0.0839
	(0.0319)	(0.0362)	(0.0772)	(0.0865)
ATT（4, 4）	0.0114	0.0093	-0.0024	-0.0097
	(0.0269)	(0.0325)	(0.0632)	(0.0800)
Aggregation to a single effect	0.0196	0.0193	0.0243	0.0329
	(0.0209)	(0.0253)	(0.0497)	(0.0625)
Pre-trend test	√	√	×	×
Event Study p-value (3 df)	0.1673	0.1064	0.0117	0.0078
Heterogeneous Trend Test (2 df)	0.2055	0.1405	0.0379	0.0310
N	14320	14320	14320	14320

Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

3.4.2.2. Relaxing the common trend assumption

As discussed in Section 3.3.2.3, we can relax the common trend assumption. Aggregated ATTs for outpatient visit times are reported in Table 3.10 and separate cohort-period ATTs are plotted in Figure 3.3, where controls are included. Column (1) of Table 3.10 gives results from the basic model in Section 3.4.2.1, the ATT for the number of outpatient visits is 3.29% (an 8.40% increase of the baseline level, i.e., 39.16 percentage points). Column (2) provides results by assuming each cohort has a separate linear trend in the absence of the intervention, the ATT for outpatient visit time is 18.62%

(a 47.55% increase from the baseline level). Column (3) displays results allowing pre-trends to depend on cohorts and covariates. The ATT for outpatient visit time is 18.36% (a 46.88% increase of the baseline level), which is quite close to the estimate in column (2). Column (4) reports results allowing event-study type trends, which as discussed in Wooldridge (2021), is not realistic and gives strange estimates. Allowing cohort-specific linear trends or allowing pre-trends to depend on cohorts and controls has limited impact on ATTs for cohorts 2 and 4, but enlarges the ATT for cohort 3 (See Figure 3.3).

Table 3.10 Outpatient visit times-relax common trend assumption

	(1)	(2)	(3)	(4)
Outpatient Visit times	0.0329	0.1862**	0.1836**	-0.0697
	(0.0625)	(0.0886)	(0.0812)	(0.0825)

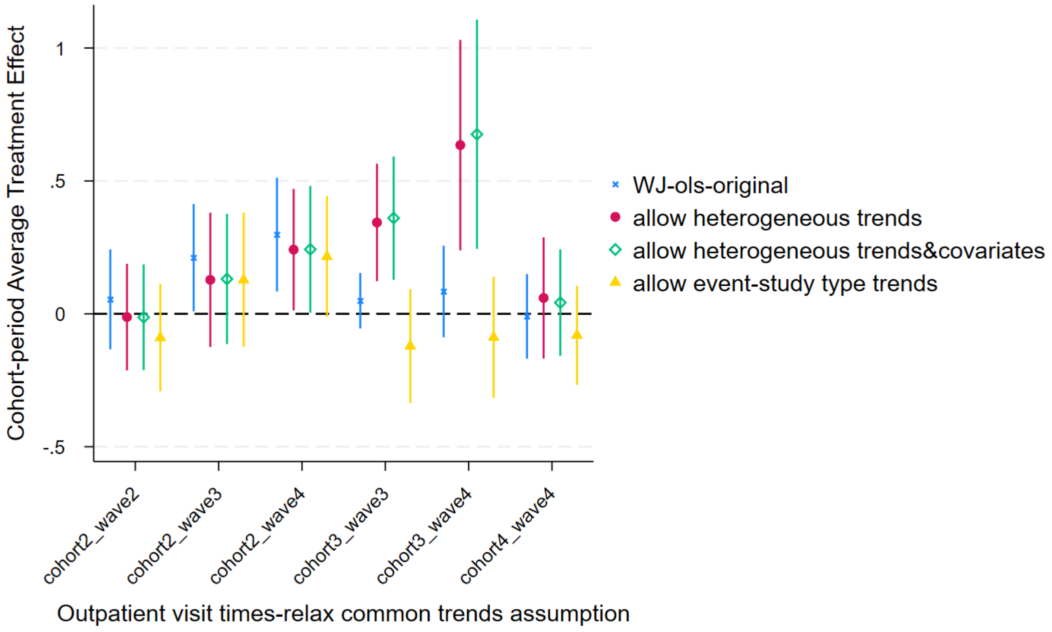


Figure 3.3 Outpatient visit times-relax common trend assumption (95% CI)

3.4.3. Sensitivity analysis

3.4.3.1. Different estimators

We explore further the sensitivity of the results to different estimation methods. First, Wooldridge (2023) extends the pooled OLS estimation provided by Wooldridge (2021) to adapt to the setting where the nature of the response variable may warrant a nonlinear model by using the pooled quasi-maximum likelihood estimator (QMLE) in the linear exponential family (LEF). Therefore, a pooled probit estimator is estimated for inpatient probability and a pooled poisson estimator is used for the number of hospital nights stayed. Additionally, we also follow Callaway and Sant' Anna (2021) to get a doubly robust estimator (CSDID) using a "not-yet-treated" control group. Figure 3.4 displays results for the probability of inpatient care from different estimators (See Appendix Figure C2 for the number of hospital nights stayed).

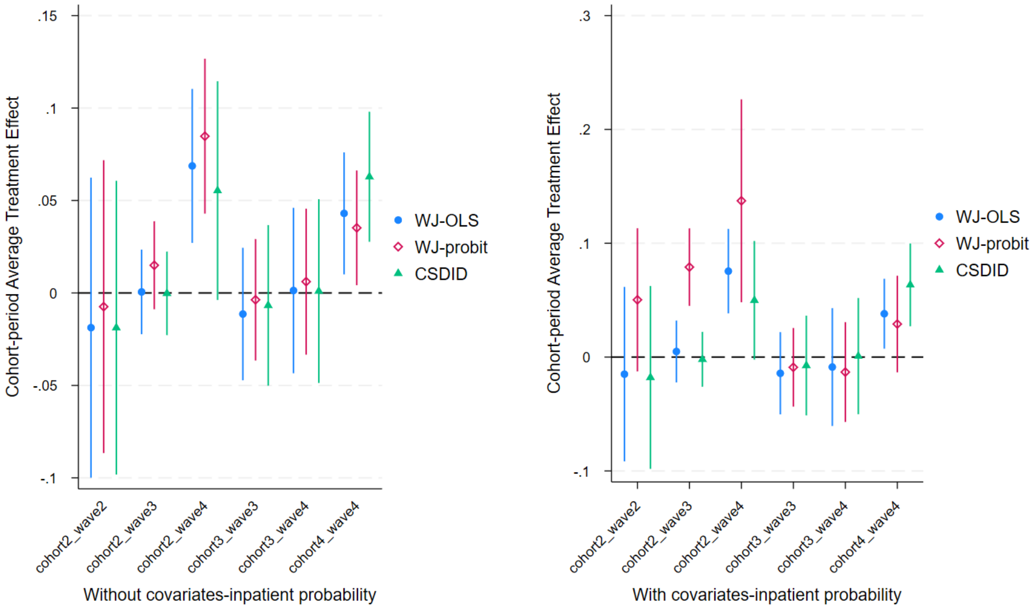


Figure 3.4 Different estimators for inpatient probability (95% CI)

Adding controls brings little changes to the magnitude of the parameters, but makes standard errors a little larger. With covariates, the aggregated ATT, using pooled

OLS, pooled probit and CSDID estimators are 2.71%, 2.67% and 3.90%, respectively. The precision of the pooled OLS estimates is notably better than either pooled probit or CSDID. Both pooled probit and CSDID estimators have slightly larger standard errors. Pooled OLS and probit estimators are very close in magnitude, while CSDID estimators are relatively bigger. The discrepancy between CSDID estimators and pooled regression estimators could be explained by the different choices of baseline outcome and different ways of incorporating covariates. Both pooled OLS and pooled probit estimates use the average outcome from all pre-periods as the baseline outcome, while the CSDID estimates regard the outcome at the last period before treatment as the baseline outcome. Additionally, CSDID estimators only use baseline (one period before treatment) covariates. This leads to a difference when including time-varying controls.

3.4.3.2. Different set of controls and relaxed sample restriction

Robustness checks are also made using different set of covariates and sample definitions (See Figure 3.5 for the probability of inpatient care and Figure C3 in the Appendix for the durations of hospital stays).

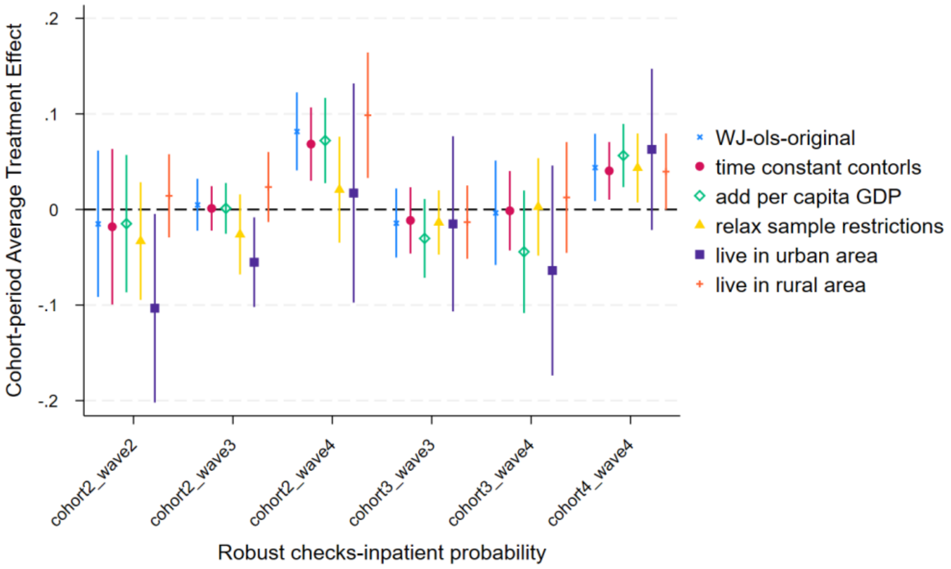


Figure 3.5 Robust checks for inpatient probability

Firstly, we only allow time-constant covariates. Secondly, we add time-varying city-level per-capita GDP. Thirdly, we relax sample restrictions on interview time and outliers. Furthermore, as residential status in this study is defined by Hukou, we run regression separately by living area. Apart from the results using a sample with the urban living area (the parallel trend assumption fails to hold here), all other results remain robust to our basic model.

3.5. Discussion and Conclusion

A possible concern for this work is the difficulty of arguing random policy adoption. Previous studies reveal local adoption of integration reform is associated with the political will of local governments, the governance ability to narrow the gaps between the subsystems, or other external environmental conditions, such as local fiscal capacity and urbanization level (Huang and Kim, 2020). The potential endogenous implementation of the integration reform may prevent the common trend assumption from holding (Huang and Wu, 2020). From this perspective, we test the parallel trends assumption unconditionally and conditionally (conditional on individual-level covariates in main results and city-level covariates in the robustness checks). It is reassuring to see the assumption holds for most of our outcome indicators both unconditionally and conditionally. For outpatient visits which fail to pass the test, further analysis of relaxing the assumption is explored.

Our findings suggest that the URRBMI integration reform has a significant and positive effect on middle-aged and elderly rural residents' inpatient healthcare usage. On average, it enhances the probability of using inpatient services by 2.71% (a 35.80% increase of the baseline level, i.e., 7.57 %) and increases the duration of hospital stays by 0.33 nights (a 44% increase of the baseline level, i.e., 0.75 night). The impressive policy impact may be associated with the largely expanded medical options and enlarged benefit packages. The systematic segmentation before the reform severely restricted the flexibility of access to healthcare services. The integration enables rural residents to seek affordable care in more hospitals or higher-level facilities. The increasing benefit brought by the integration policy lowers the “price” of medical services and facilitates more utilisation. This is consistent with earlier research investigating pilot impacts (Huang and Wu, 2020). We provide new evidence about the positive impact of the national integration reform on inpatient healthcare usage. It remains significant as what found in the pilot areas.

The consolidation has a limited impact on the probability and the frequency of using outpatient services. This is not surprising as the reform focuses more on covering expenditures for inpatient care rather than outpatient services. The new URRBMI scheme still has high deductibles for outpatient care and limited or insufficient funding for outpatient reimbursement. It aligns with previous studies showing outpatient usage seems to be unaffected by the policy (Zhou et al., 2022; Su et al., 2019). However, if we allow for a constant difference in trends between the treated group and the control group, the reform significantly increases the number of outpatient visits by 18.62% (a 47.55% increase from the baseline level, i.e., 39.16 %).

One important contribution of this study is providing a detailed analysis of potential treatment effect heterogeneity. The Bacon-decomposition method is introduced to highlight how the standard TWFEDID estimator with underlying constant treatment effect assumption would be biased in the staggered intervention set-up. This reassures our choice of using the newly proposed extended TWFEDID estimator to allow arbitrary treatment effect heterogeneity. For the reform’s impact on inpatient healthcare utilisation, we observe heterogeneous treatment effects across both cohorts and periods. On the one hand, the ATT increases by the length of exposure within the same cohort. Particularly, for those groups who implemented the policy before 2016 (early treated groups), the integration has no discernible impact on their inpatient care utilisation with a short length of exposure. In contrast, a significant and relatively larger effect is observed with longer exposure (after 2016). There is a delayed reaction to the policy change in medical behaviours in the short term, which is consistent with previous studies (Zhou et al., 2022). Additionally, local governments further refined and improved the reform after the 2016 Opinions were issued, which might also explain the increased treatment effect by periods. On the other hand, given the same length of exposure, ATTs for the later-treated cohort are bigger than the earlier-treated cohorts. This kind of heterogeneity can be explained from three aspects. First, it is common for those who implemented the integration reform after 2016 to take advantage of the experiences in pilot areas. More preparation time before implementation could induce more detailed policy designs. Besides, it is natural for later treated cohorts to avoid some potential difficulties which early treated areas experienced. Second, as the 2016 national reform was compulsory, due to top-down pressure, provinces and cities are more likely to deliver efficient policy design and procedures following the central government's guidance. This may induce further heterogeneity between polit areas and later treated areas. Third, in line with customary patterns observed in major reforms in China, the implementation of URRBMI reform displays heterogeneity across provinces and even sub-national geographic units, such as municipalities and counties. Local governments with higher fiscal strength and social risk are more capable of extending generous health insurance benefits (Ratigan, 2017; Meng and Su, 2021). Especially, for outpatient healthcare utilisation, the positive and significant effects only appear in early treated cohorts. This is because outpatient reimbursement policies drastically differed across regions. In either NRCMS (before the integration reform) or URRBMI (after the integration reform) schemes, the establishment of pooling funds mainly compensates for hospitalisation and only a few areas have attempted to compensate for outpatient costs (Su et al., 2019). Those cities starting the integration autonomously before 2016 seem to be associated with better economic development. For instance, the per capita GDP of early treated cohorts is larger than later treated cohorts in wave 1 (pre-intervention period). Therefore, those pilot areas are more capable of producing a portfolio of generous policies related to outpatient services. Apart from the local economic development, local politicians with strong career incentives are also more likely to prompt localities to be pilot areas. And such career incentives could be associated with a greater degree of effort during a trial (when local efforts are showcased) than during national implementation (Wang and Yang, 2021). Compared with the improvement of inpatient service coverage, policies related to outpatient services are easier and more likely to get high returns with relatively low costs in terms of local politicians' career incentives. This could partly explain the positive reform impact on outpatient healthcare utilisation in pilot areas and the insignificant impact in later treated areas.

Our findings may provide some meaningful policy implications. The main findings in this study provide national evidence of the reform's positive impacts on rural residents' inpatient utilisation and limited effects on outpatient utilisation. To improve residents' welfare, further reform should cover outpatient costs. The corresponding increased demand is likely to generate more pressure on the financing side. Therefore, the premiums of URRBMI could increase gradually over time for both individuals and the public sector, serving as an additional source of financing. Additionally, there are ongoing reform efforts to reduce disparities in healthcare access among rural and urban residents due to the fragmentation of the health insurance system (Zhou et al., 2022). However, the treatment effect heterogeneity across cohorts analysed in the paper reveals considerable inter-regional inequality throughout China, which also requires attention. It is worth considering how to narrow the gap among different regions in future policies.

Conclusion

This thesis presents a detailed analysis of recent reforms in the Chinese healthcare system and investigates the relationship between health and labour supply in rural China. Due to the rural-urban dual system in China, persistent gaps have existed between rural and urban populations in terms of healthcare utilisation, health and labour supply. Our three essays focus on the rural population as they're more vulnerable and disadvantaged than urban citizens.

China maintains relatively young statutory retirement ages in the formal employment sector, ranging from 45 to 60 years old. The retirement ages differ based on gender, individual health status, job occupations, and sectors. Within the informal employment sector, and particularly in rural areas, formal retirement is less common (Smith et al., 2014). Health deterioration and changes in family situations are believed to affect changes in work decisions more. Chapter 1 explores how health plays a significant role in determining labour supply decisions in middle-aged and elderly rural households. The contribution of the paper is introducing an interaction between own health and spousal health into the labour supply framework. The results add new evidence in the literature that highlights the added worker effect and the complementarity of leisure. We argue that the question of whether complementarity or substitutability plays the dominant role depends on both own health and spousal health. Additionally, the findings indicate that better health has positive effects on stimulating older people's labour supply, which can reduce financial burden related to population ageing (Hou et al., 2021). This reveals the importance of achieving 'healthy ageing'. However, as there are few alternative sources of income (limited pension support) for the rural elderly; some older people participate in the job market based on necessity rather than choice (Ning et al., 2016). From this perspective, further improvement of the social security system to protect the rural elderly from the risks of vulnerable economic conditions and poverty is necessary. Moreover, this study contributes to the literature investigating ageing farmers worldwide. The ageing of farmers and potential intergenerational changes are not unique topics in China. They pose common difficulties for newly industrialising countries across Asia (Liu et al., 2023). In those countries, smallholder farming remains the prevalent organizational form. This type of farming faces challenges in implementing mechanisation and remain labour-intensive. The younger rural generation is generally not interested in low-income small-scale farming and instead pursue non-agricultural employment in urban areas, which creates large intergenerational differences with their parents' generation. Some scholars have found that the agricultural labour force in many rural areas is mainly composed of left-behind elderly individuals (especially women) over 60 years old(Jiang et al., 2019). Our findings provide insights into dealing with ageing farmers' concerns by improving population health and encouraging longer effective labour supply of near-old and older people. Furthermore, we only investigate the relationship between health and labour supply in middle-aged and older rural households in the informal sector. Further studies might examine whether a similar relationship exists in younger households and amongst those employed in the formal sector.

China experienced both economic and demographic transitions within the past few decades, greatly increasing the demand for accessible and affordable healthcare. The reforms in the healthcare system have made laudable achievements, such as the expansion of social health insurance and the strengthening of the primary care system. However, there are still challenges in establishing a more efficient healthcare delivery system and in eliminating inequality in access to healthcare and health. Chapter 2 finds that the implementation of the HMS reform has improved the capacity of urban primary healthcare institutions in terms of human resources. However, this effect has not yet manifested in rural areas. This might be due to the difficulties of attracting skilled workers to participate in rural primary healthcare facilities, and implies a greater need to strengthen the workforce in rural areas (Xu and Mills, 2017). We also find that the HMS reform significantly increases utilisation of rural primary healthcare facilities, while the positive effect of the reform is more limited in urban areas. This disparity of response to the policy between urban and rural areas is worthy of further analysis and scrutiny. One limitation of Chapter 2 is the relatively small sample size derived from a highly aggregated (provincial-level) dataset, which may affect the precision of the estimators. A high-quality longitudinal dataset at an individual level that includes detailed health and healthcare information across the full reform period would improve the analysis.

Findings from Chapter 3 show a significant and positive effect of the URRBMI integration reform on inpatient healthcare use, and a limited impact of the reform on outpatient healthcare utilisation. This is consistent with previous studies and highlights the limited benefits of expansions in outpatient services, which requires further attention in future reforms. The potential limitation of Chapter 3 is the lack of exploration of mechanisms underpinning the findings. We assume several possible channels through which the integration reform affects healthcare use. However, we have not yet investigated those channels empirically due to a lack of suitable data.

Given staggered policy implementations, Chapters 2 and 3 employ the newly proposed heterogeneous DID methods. We find clear treatment effect heterogeneity across cohorts and periods; this is a major contribution to the existing literature. We also explore the potential source of treatment effect heterogeneity from several aspects. Firstly, the central government only provides rough guidance on the reforms and local governments have the freedom to formulate detailed policy designs. This led to geographical disparities in implementation. Secondly, there are pilot areas and the rollout of the reforms is staggered, such that later-treated cohorts were able to take advantage of the experiences of early-treated cohorts and had greater time to plan and prepare. Thirdly, it takes time to allow policies to work. Therefore, there might be no significant short-term impact but significant impact in the long run. All of these features of the reform will contribute to treatment effect heterogeneity. It is highly likely that other substantive social reforms implemented in China will also display treatment effect heterogeneity. Accordingly, it is useful to consider the new heterogenous DID approach in future Chinese policy evaluations.

In the early 1980s, the family planning policy, which centres on the one-child policy, became a basic national policy of China (OECD, 2017). Over the past several decades since then, this policy has changed the course of China’s population transition from high fertility rates (5.5% in 1970) to low fertility rates (1.15% in 2021). Meanwhile, because of improved nutrition, sanitation, and healthcare services, life expectancy has been significantly prolonged in China, increasing from 67.9 years in 1981 to 78.2 years in 2021 (Che and Li, 2018). China, as a developing country, has been undergoing an unprecedented demographic transition and rapid population ageing as a result of the decline in infertility rates and rising life expectancy (Ning et al., 2016). The main concerns of population ageing can be summarised as providing income and health security at older ages and doing so with affordable budgets (Smith, 2012; Lee and Mason, 2010).

The changing population demographics have led to a shrinking labour force contributing to the pension system but an increasing aged population eligible to receive a retirement pension. This poses a possible threat to the stability and sustainability of the current social pension system in China. Evidence from developed countries reveals that delaying retirement can not only reduce human capital waste but also ease the payment pressure of pension funds. Therefore, gradually postponing the outdated legal retirement age (promulgated in 1978) has been considered by the Chinese government. However, the statutory retirement age only works in the formal sector in urban China. While retirement hazard at statutory retirement ages displays sharp spikes for urban workers, the age pattern of actual retirement is very smooth among rural residents. Without the legal retirement age, it is generally accepted that the retirement of older farmers is a relatively gradual process, which is related to both economic factors (including social security and subsidies) and physical and psychological health factors (Farrell et al., 2020; Chiswell, 2018).

From a policy perspective, extending working lives critically depends on the health status of the elderly, especially for rural residents engaged in farming activities. One of the main policy implications from Chapter 1 is that encouraging the rural elderly to invest in their health stock is effective in improving their health capacity to work and then promoting long-term productive labour supply. Policy actions such as providing more accessible and affordable healthcare services and offering free access to physical exercise facilities would be helpful. Moreover, it is also important to ensure a good life for older people by maintaining income security and lowering health risks at older ages (Smith et al., 2014). It is extensively acknowledged that traditional systems of family support to rural elderly have collapsed due to the decline in birth rate, shrinking of family size and massive labour rural-urban migration. Furthermore, a lack of sufficient social security systems puts the rural elderly at risk of vulnerable economic conditions and poverty. Therefore, in addition to incentivising the labour supply of old farmers, expanding access to social insurance and pensions for rural elderly is necessary.

On the other hand, population ageing is also likely to place stress on the Chinese healthcare system, which has focused on diseases at younger ages and infectious rather than chronic diseases (Mitra et al., 2020). Efficient primary care is helpful to provide effective chronic disease management (Garrido et al., 2011). The hierarchical medical system reform attempts to achieve orderly treatment of 'first treatment in primary medical facilities, two-way referral, separate treatment for acute and chronic diseases, and linkage between upper and lower level healthcare institutions' (Zhang and Wang, 2024). Our findings in Chapter 2 reveal the mixed success of the HMS reform as it effectively enhances the visit ratio to rural primary healthcare institutions but has no significant impact on healthcare utilization in urban health facilities. Therefore, further policy actions to increase health resources (such as better equipment and easier-to-operate IT systems) are needed to improve the capacity of urban primary health institutions and then guide urban residents to seek healthcare services in primary health facilities. Although the HMS reform significantly encourages rural residents to go to primary healthcare institutions, the reform does not enhance the health technicians' proportion in rural primary health facilities. It is necessary to employ more powerful actions (such as providing economic incentives and offering job training opportunities) to attract educated and experienced health technicians to go to rural primary healthcare institutions.

As mentioned above, encouraging elderly people to extend their working life highly depends on their health status. Expanding access to equitable and affordable healthcare services is helpful to improve residents' health. The rural-urban health insurance integration evaluated in Chapter 3 attempts to narrow the disparities among different health insurance schemes in fund level and benefits package and then provide equitable access to healthcare and financial protection for residents. Our main findings show the positive impact of the integration reform on healthcare utilization. However, treatment effect heterogeneity analysis reveals inter-regional inequality throughout China. Future reform should pay attention to finding ways to reduce the gaps among different regions.

The thesis involves both administrative statistical datasets and national survey datasets. However, there are some limitations in terms of these datasets. On the one hand, relying on self-reported data (Chapters 1 and 3) might introduce bias through misclassification errors. The low educational levels among middle-aged and elderly rural residents may heighten this concern on the assumption that misclassification is a function of cognitive ability. However, we have restricted the sample to exclude observations that appear to lack credibility. For example, in Chapter 1, individuals who report over 7300 (20*365) annual working hours have been dropped. Although farmers in rural China usually have high work intensity, such an extreme annual working time is not realistic. Additionally, in Chapter 3, to ensure the credibility of data, observations who are older than 90 years old have been removed and all continuous outcome variables have been winsorized at the 99.9th percentile. On the other hand, although the highly aggregated administrative statistical dataset in Chapter 2 can mitigate measurement errors to some extent, the limited sample size and information loss due to aggregation also require further attention. Some studies use data derived from health record systems or medical claim systems for a certain area to investigate the impact of the Hierarchical Medical System reform. However, those datasets while including richer information are confidential and not readily accessible. In the future, available access to such health record systems or medical claim systems in hospitals would greatly improve research in this area. Despite these data limitations, for each chapter, we’ve explored possible datasets used in the existing literature and chose those most suitable to the study design and which are accessible.

There are some possible extensions following from the thesis which could guide the future research. Firstly, previous studies reveal the correlations in the retirement of spouses in the formal sector. Evidence from different countries shows positive correlations between preferences for joint leisure in a couple (García-Miralles and Leganza, 2014; Michaud et al., 2020). Given the sharp urban-rural differences in terms of retirement systems in China (Giles et al., 2023), it’s unknown whether similar joint retirement could be revealed within couples of informal retirees. Chapter 1 contributes to the literature investigating the labour supply of couples in the informal sector, while we do not say much about the direct impact of spousal retirement decisions on individuals’ retirement decisions in farmer couples in China. Future research could further explore rural couples who retire jointly and how health impacts the interdependent joint retirement decisions within rural couples. In addition, our study design in Chapter 1 focuses on the labour supply decisions of mid-aged and elderly rural residents who are close to retirement age. To understand the factors behind the retirement decision and investigate the feasibility of increasing the average retirement age, it is worth investigating the retirement and labour supply patterns across different age cohorts in both urban and rural areas. Therefore, with access to available datasets, a similar research question could be examined in younger households and amongst those employed in the formal sector in the future. Secondly, due to data limitations, we only test the capacity improvement of primary healthcare facilities as one of the channels affecting the healthcare utilization of primary health institutions in Chapter 2. If we have access to health record systems or medical claim systems in hospitals in the future, we could investigate more fully mechanisms by considering the proposed channels such as changes in economic incentives and improvement of other health resources apart from human capital. Moreover, with richer information, the heterogeneity analysis among different population groups and the further impact of the HMS reform on population health (such as mortality) is worth exploring. Thirdly, we have proposed several possible channels for how the health integration reform impacts healthcare utilization in Chapter 3. However, there is a lack of detailed mechanism tests due to data limitations. If health record systems or medical claim systems in hospitals are available, more analysis of possible channels such as changes in reimbursement benefits could be undertaken. Additionally, we could investigate more outcome indicators apart from healthcare utilization in the future.

Appendix A: Appendix to Chapter 1

A1. The process of choosing appropriate objective health measures

CHARLS includes a brunch of objective health measures, such as ADLs, IADLs, other functional limitations and doctor-diagnosed health problems (see Table A1). We're interested in whether the functional difficulties or health problems limit the kind or amount of work from both theoretical and practical perspectives. The ADLs and other functional limitations present the upper or lower body mobility, IADLs provide information about cognitive abilities, and the doctor diagnosed health problems indicate the specific chronic or critical diseases. Thus, we could argue that they are all work-related health measures. However, some objective health measures, like walking 100 meters, are fundamentally basic, people reporting difficulties with these activities are reasonable to be regarded as lack of daily life freedom as well as work ability. To be more precise, we use all these objective health measures to estimate the latent health stock first, and the results show bathing, eating, getting in/out of bed, shopping, walking 100 meters and picking up a coin from the table seem to have an insignificant impact on self-reported health. Therefore, in our main results, we drop these six health measures.

Table A1 Variable names and definitions in the latent health model

Variables	Description
6-item Activities of daily living (ADLs): Some difficulty	1 if difficulty reported, 0 otherwise. There are individual dummies for difficulties with：1.dressing/2.bathing/3.eating/4.get in/out bed/5.using the toilet/6.controlling urination and defecation
5-item Instrumental activities of daily living (IADLs): Some difficulty	1 if difficulty reported, 0 otherwise. There are individual dummies for difficulties with: 1. managing money/2. taking medications/3. shopping/4. preparing hot meal/5. cleaning house
9-item Other functional limitations: Some difficulty	1 if difficulty reported, 0 otherwise. There are individual dummies for difficulties with: 1. walking 100 metre/2. walking 1km/3. jogging 1km/4. getting up from a chair after sitting for long periods/5. climbing several flights of stairs without resting/6. stooping kneeling, or crouching /7. lifting or carrying weights over 5kg /8. reaching arms above shoulder level/9. picking up a coin from the table
The doctor diagnosed health problems: Ever Had Condition (13 items)	1 if the problem is reported, 0 otherwise. There are individual dummies for problems with: 1. high blood pressure/2. diabetes/3. caner4. lung disease/5. heart problems/6. stroke/7. psych problem/8. arthritis/9. dyslipidaemia/10. liver disease/11. kidney disease/12. stomach/digestive disease/13. asthma

A2. Descriptive statistics of variables

Table A2 Transition probability of male's self-reported health from wave 1 to wave 2

	Male's self-reported health (wave 2)					
Wave 1	very poor	poor	fair	good	very good	Total
very poor	22	55	17	6	2	102
	21.57	53.92	16.67	5.880	1.960	100
poor	52	197	247	31	19	546
	9.520	36.08	45.24	5.680	3.480	100
fair	39	167	766	164	85	1,221
	3.190	13.68	62.74	13.43	6.960	100
good	7	28	219	129	57	440
	1.590	6.360	49.77	29.32	12.95	100
very good	0	10	65	51	71	197
	0	5.080	32.99	25.89	36.04	100
Total	120	457	1,314	381	234	2,506
	4.790	18.24	52.43	15.20	9.340	100

Table A3 Transition probability of male's self-reported health from wave 2 to wave 3

	Male's self-reported health (wave 3)					
Wave 2	very poor	poor	fair	good	very good	Total
very poor	17	34	16	4	0	71
	23.94	47.89	22.54	5.630	0	100
poor	38	130	121	15	7	311
	12.22	41.80	38.91	4.820	2.250	100
fair	33	142	659	104	86	1,024
	3.220	13.87	64.36	10.16	8.400	100
good	1	24	144	74	60	303
	0.330	7.920	47.52	24.42	19.80	100
very good	1	10	72	33	69	185
	0.540	5.410	38.92	17.84	37.30	100
Total	90	340	1,012	230	222	1,894
	4.750	17.95	53.43	12.14	11.72	100

Table A4 Transition probability of female's self-reported health from wave 1 to wave 2

	Female's self-reported health (wave 2)					
Wave 1	very poor	poor	fair	good	very good	Total
very poor	35	64	36	5	2	142
	24.65	45.07	25.35	3.520	1.410	100
poor	74	302	275	43	14	708
	10.45	42.66	38.84	6.070	1.980	100
fair	39	226	717	133	67	1,182
	3.300	19.12	60.66	11.25	5.670	100
good	8	36	174	84	52	354
	2.260	10.17	49.15	23.73	14.69	100
very good	0	9	45	34	32	120
	0	7.500	37.50	28.33	26.67	100
Total	156	637	1,247	299	167	2,506
	6.230	25.42	49.76	11.93	6.660	100

Table A5 Transition probability of female's self-reported health from wave 2 to wave 3

	Female's self-reported health (wave 3)					
Wave 2	very poor	poor	fair	good	very good	Total
very poor	32	46	29	1	3	111
	28.83	41.44	26.13	0.900	2.700	100
poor	59	178	175	25	8	445
	13.26	40	39.33	5.620	1.800	100
fair	35	141	628	84	82	970
	3.610	14.54	64.74	8.660	8.450	100
good	1	20	117	52	48	238
	0.420	8.400	49.16	21.85	20.17	100
very good	4	8	54	12	52	130
	3.080	6.150	41.54	9.230	40	100
Total	131	393	1,003	174	193	1,894
	6.920	20.75	52.96	9.190	10.19	100

Table A6 Descriptive statistics of variables in yearly work time model

Variable	Obs	Mean	Std.Dev.	Min	Max
Dependent variable: yearly work time					
male	2,424	1637	1115	8.660	7274
female	2,424	1476	1165	4.330	6547
Latent health index					
male	2,424	-0.659	0.639	-3.431	0.0213
female	2,424	-0.751	0.589	-3.356	0
Age					
male	2,424	58.83	7.380	45	87
female	2,424	56.99	6.927	45	88
Education					
male	2,424	1.124	0.347	1	3
female	2,424	1.037	0.193	1	3
Education gap	2,424	2.552	3.724	-11	11
Hukou					
male	2,424	1.073	0.307	1	4
female	2,424	1.038	0.225	1	4
Household living region					
East	2,424	0.252	0.435	0	1
Central	2,424	0.315	0.465	0	1
West	2,424	0.363	0.481	0	1
Northeast	2,424	0.0693	0.254	0	1
Household non-labour income	2,424	10658	22492	0	600000
Category household non-labour income	2,424	1.847	0.824	1	3
Household structure					
Number of pre-school children (age:0-6)	2,424	0.310	0.620	0	5
Number of school children (age:7-18)	2,424	0.346	0.699	0	7
Number of old people (age>75)	2,424	0.0672	0.270	0	2

A3. Detailed empirical results

Table A7 Labour participation estimate (results for covariates)

	Biprobit			
	Male’s labour participation		Female’s labourparticipation	
	Coefficients	Margins	Coefficients	Margins
Age	-0.06***	-0.01***	-0.04***	-0.01***
	(-19.85)	(-21.33)	(-15.11)	(-15.92)
Education (reference group: Less than lower secondary)				
upper secondary & vocational	-0.04	-0.01	-0.05	-0.01
training	(-0.53)	(-0.52)	(-0.40)	(-0.40)
tertiary	-0.29	-0.07	-1.07	-0.36
	(-1.11)	(-1.02)	(-1.55)	(-1.53)
Education Gap	-0.01	-0.00	0.00	0.00
	(-1.55)	(-1.55)	(0.11)	(0.11)
Hukou (reference group: Agricultural hukou)				
Non-agricultural hukou	-0.28***	-0.06***	-0.15	-0.04
	(-3.67)	(-3.40)	(-1.21)	(-1.17)
Unified residence hukou	-0.08	-0.02	0.10	0.03
	(-0.26)	(-0.25)	(0.27)	(0.28)
Do not have hukou	0.24	0.04	4.90	0.26***
	(0.39)	(0.43)	(0.01)	(45.17)
Household Structure				
Number of preschool children	-0.04	-0.01	-0.10***	-0.03***
(age:0-6)	(-1.03)	(-1.03)	(-3.26)	(-3.27)
Number of school children(age:7-18)	-0.03	-0.01	0.04	0.01
	(-1.02)	(-1.02)	(1.29)	(1.29)
Number of old people (age>75)	0.11	0.02	0.12	0.03
	(0.99)	(0.99)	(1.35)	(1.35)
Household non-labour income (reference group: Less than p (50))				
P (50)-P (75)	-0.08	-0.02	-0.10**	-0.03**
	(-1.57)	(-1.56)	(-2.19)	(-2.18)
More than P (75)	-0.02	-0.00	-0.01	-0.00
	(-0.27)	(-0.27)	(-0.10)	(-0.10)
Living region				
East	0.12	0.02	-0.20**	-0.06**
	(1.27)	(1.27)	(-2.44)	(-2.44)
Central	0.21**	0.04**	0.04	0.01
	(2.29)	(2.30)	(0.49)	(0.49)
West	0.30***	0.06***	0.28***	0.08***
	(3.31)	(3.32)	(3.42)	(3.43)
_cons	4.92***	-	3.45***	-

	(24.49)	-	(20.14)	-
N	5715	5715	5715	5715

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

Table A8 Yearly work time estimate (results for covariates)

	Reduced		Interaction term	
	male	female	male	female
Age	-6.32*	-9.87***	-6.44**	-9.95***
	(-1.94)	(-2.80)	(-1.98)	(-2.82)
Education (reference group: Less than lower secondary)				
upper secondary & vocational training	150.31**	212.37*	153.91**	213.09*
	(2.06)	(1.72)	(2.11)	(1.73)
tertiary	16.76	254.54	22.96	259.30
	(0.06)	(0.34)	(0.08)	(0.34)
Education Gap	-21.97***	-9.56	-22.11***	-9.56
	(-3.44)	(-1.46)	(-3.46)	(-1.46)
Hukou (reference group: Agricultural hukou)				
Non-agricultural hukou	-266.62***	-184.82	-265.60***	-184.15
	(-2.64)	(-1.32)	(-2.63)	(-1.31)
Unified residence hukou	-259.63	-205.52	-264.95	-207.93
	(-1.10)	(-0.70)	(-1.12)	(-0.71)
Do not have hukou	-133.11	829.83	-145.42	823.31
	(-0.26)	(0.77)	(-0.28)	(0.76)
Household non-labour income (reference group: Less than p (50))				
P (50)-P (75)	-54.43	-70.65	-54.38	-70.63
	(-0.99)	(-1.24)	(-0.99)	(-1.24)
More than P (75)	-50.33	-128.81**	-50.39	-128.84**
	(-0.89)	(-2.19)	(-0.89)	(-2.19)
Household Structure				
Number of pre-school children (age:0-6)	15.51	-26.17	15.27	-26.30
	(0.42)	(-0.68)	(0.41)	(-0.68)
Number of school children(age:7-18)	27.06	80.41**	26.17	79.95**
	(0.83)	(2.37)	(0.80)	(2.36)
Number of old people (age>75)	-22.06	122.00	-19.01	123.58
	(0.00)	(1.40)	(-0.23)	(1.42)
Living Region				
East	273.94***	201.36**	277.04***	203.01**
	(2.80)	(1.98)	(2.83)	(2.00)
Central	12.85	-36.61	13.74	-36.18
	(0.13)	(-0.37)	(0.14)	(-0.36)
West	267.85***	380.13***	271.36***	381.81***
	(2.84)	(3.88)	(2.87)	(3.90)

_cons	1946.96*** (9.55)	1976.07*** (9.17)	1916.25*** (9.33)	1960.98*** (9.04)
N	2424	2424	2424	2424

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

Table A9 The AME of new latent health on labour participation

	Male		Female	
	AME of own latent health	AME of spousal latent health	AME of own latent health	AME of spousal latent health
Given spousal or own latent health				
-3.5	0.0119	-0.1578***	0.0255	-0.0992***
	(0.3756)	(-3.7675)	(0.5469)	(-2.6187)
-3	0.0304	-0.1453***	0.0474	-0.0910**
	(1.1794)	(-3.6567)	(1.2545)	(-2.4809)
-2.5	0.0488**	-0.1209***	0.0685**	-0.0742**
	(2.4262)	(-3.5362)	(2.3248)	(-2.2835)
-2	0.0670***	-0.0888***	0.0886***	-0.0511**
	(4.5110)	(-3.4045)	(4.0656)	(-1.9780)
-1.5	0.0851***	-0.0548***	0.1077***	-0.0250
	(8.2381)	(-3.1388)	(7.0937)	(-1.3843)
-1	0.1028***	-0.0243**	0.1257***	-0.0001
	(13.9372)	(-2.2126)	(11.8801)	(-0.0043)
-0.5	0.1202***	-0.0011	0.1426***	0.0203**
	(16.0213)	(-0.1263)	(14.5597)	(1.9835)
0	0.1371***	0.0136	0.1583***	0.0341***
	(13.4442)	(1.4717)	(12.5267)	(2.8358)
N	6110	6110	6110	6110

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

Table A10 The AME of new latent health on yearly work time

	Male		Female	
	AME of own latent health	AME of spousal latent health	AME of own latent health	AME of spousal latent health
Given spousal or own latent health				
-3.5	425.01*	274.29	370.52	179.30
	(1.79)	(1.11)	(1.45)	(0.73)
-3	370.61*	219.90	331.78	140.56
	(1.88)	(1.06)	(1.55)	(0.69)
-2.5	316.22**	165.50	293.05*	101.83
	(1.99)	(0.99)	(1.68)	(0.62)
-2	261.82**	111.11	254.31*	63.09
	(2.17)	(0.86)	(1.89)	(0.50)
-1.5	207.43**	56.71	215.57**	24.35
	(2.43)	(0.61)	(2.22)	(0.28)
-1	153.03***	2.32	176.84***	-14.38
	(2.68)	(0.04)	(2.69)	(-0.24)
-0.5	98.64**	-52.08	138.10***	-53.12
	(1.96)	(-1.03)	(2.63)	(-1.02)
0	44.24	-106.47	99.36	-91.85
	(0.62)	(-1.60)	(1.44)	(-1.24)
N	2577	2577	2577	2577

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

Table A11 The AME of latent health on labour participation (non-lagged variables)

	Male		Female	
	AME of own latent health	AME of spousal latent health	AME of own latent health	AME of spousal latent health
	Given spousal or own latent health			
-4.5		-0.0641**	0.1155***	
		(-2.4683)	(5.4418)	
-4	0.0684***	-0.0657**	0.1170***	-0.0117
	(5.2294)	(-2.5092)	(6.3650)	(-0.5009)
-3.5	0.0741***	-0.0621**	0.1184***	-0.0116
	(6.7329)	(-2.5623)	(7.6117)	(-0.5346)
-3	0.0800***	-0.0538***	0.1199***	-0.0109
	(8.9288)	(-2.6390)	(9.3644)	(-0.5816)
-2.5	0.0859***	-0.0423***	0.1214***	-0.0095
	(12.3016)	(-2.7406)	(11.9324)	(-0.6483)
-2	0.0919***	-0.0299***	0.1229***	-0.0077
	(17.5504)	(-2.8172)	(15.7465)	(-0.7374)
-1.5	0.0980***	-0.0183***	0.1244***	-0.0058
	(24.0484)	(-2.6081)	(20.6244)	(-0.7991)
-1	0.1041***	-0.0090*	0.1259***	-0.0040
	(25.6451)	(-1.6720)	(22.9474)	(-0.6554)
-0.5	0.1102***	-0.0026	0.1274***	-0.0024
	(21.2015)	(-0.5188)	(19.5987)	(-0.3717)
0	0.1164***	0.0011	0.1289***	-0.0012
	(16.7913)	(0.2339)	(15.0863)	(-0.1689)
N	9356	9356	9356	9356

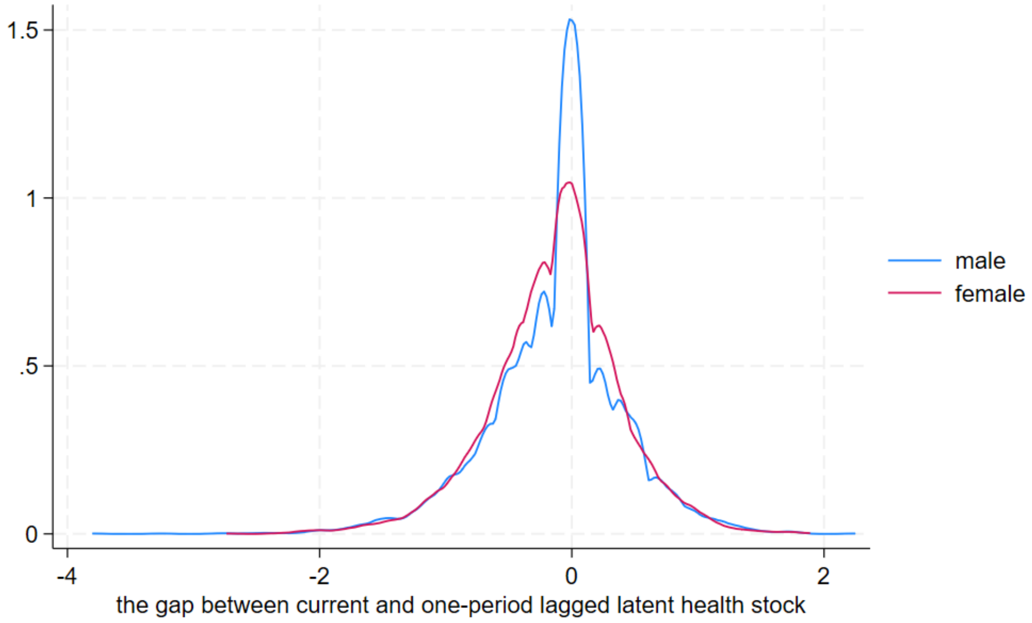
t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

Table A12 The AME of latent health on yearly work time (non-lagged variables)

	Male		Female	
	AME of own latent health	AME of spousal latent health	AME of own latent health	AME of spousal latent health
	Given spousal or own latent health			
-3.5	138.52	-8.81	157.52*	13.95
	(1.61)	(-0.10)	(1.69)	(0.16)
-3	134.78*	-12.55	148.35*	4.79
	(1.91)	(-0.16)	(1.91)	(0.07)
-2.5	131.04**	-16.29	139.19**	-4.37
	(2.36)	(-0.27)	(2.24)	(-0.08)
-2	127.30***	-20.03	130.03***	-13.53
	(3.06)	(-0.43)	(2.74)	(-0.32)
-1.5	123.56***	-23.77	120.87***	-22.69
	(4.12)	(-0.70)	(3.47)	(-0.75)
-1	119.82***	-27.51	111.71***	-31.86
	(4.86)	(-1.04)	(4.14)	(-1.28)
-0.5	116.08***	-31.25	102.54***	-41.02
	(3.97)	(-1.12)	(3.62)	(-1.38)
0	112.34***	-34.99	93.38**	-50.18
	(2.77)	(-0.94)	(2.46)	(-1.22)
N	5431	5431	5431	5431

t statistics in parentheses;* p<0.1, ** p<0.05, *** p<0.01

Figure A1 Distributions of the gap between current and lagged latent health stock



Appendix B: Appendix to Chapter 2

Table B1 DR estimators with different covariates' combinations

	practitioner proportion in township health centres					
	(1)	(2)	(3)	(4)	(5)	(6)
ATT(4,4)	0.0058	0.0062	0.0059	0.0063	0.0059	0.0055
	(0.0058)	(0.0057)	(0.0059)	(0.0064)	(0.0058)	(0.0069)
ATT(4,5)	0.0055	0.0014	0.0045	0.0007	0.0055	0.0077
	(0.0138)	(0.0152)	(0.0135)	(0.0189)	(0.0151)	(0.0235)
ATT(5,5)	0.0013	0.0026	0.0061	0.0056	0.0148	0.0162
	(0.0062)	(0.0074)	(0.0065)	(0.0096)	(0.0115)	(0.0136)
N	125	125	125	125	125	125
Common trend	√	√	√	√	√	√
	chi2(5) =8.9215	chi2(5) =8.9106	chi2(5) =6.3514	chi2(5) =8.9863	chi2(5) =5.5299	chi2(5) =7.1236
	p-value=0.1122	p-value=0.1127	p-value=0.2735	p-value=0.1096	p-value=0.3547	p-value=0.2116
Control	1. sex ratio, 2. illiteracy ratio, 3. per-capita GDP, 4. elderly proportion					
	1	1,4	3,4	1,2,4	1,3,4	1,2,3,4

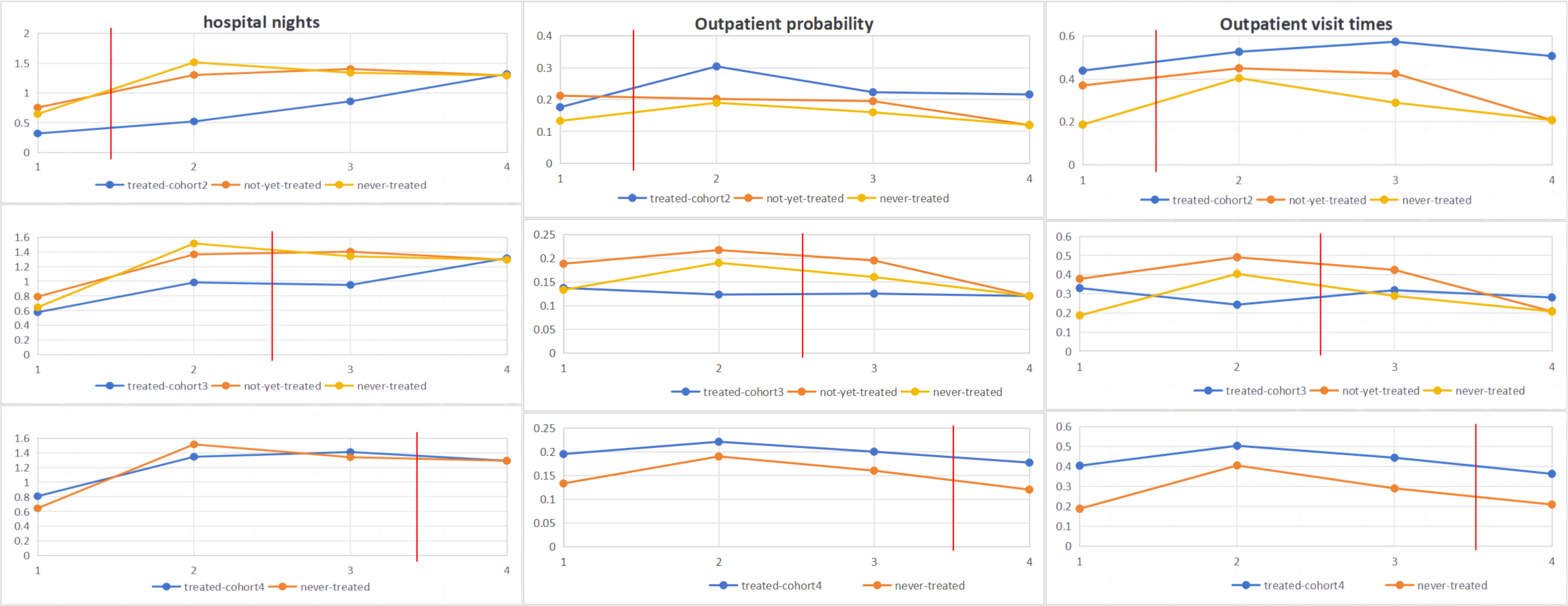
Standard errors in parentheses; * p<0.1, ** p<0.05, *** p<0.01

Appendix C: Appendix to Chapter 3

Table C1 Bacon decomposition for inpatient probability TWFE estimators

treated group	control group	coefficient	Total weight	Aggregate group
		0.007692	1	
g=2(Year 2013)	g=3(Year 2015)	-0.010980	0.008938	Early vs Late
g=2(Year 2013)	g=4(Year 2018)	-0.007806	0.078332	Early vs Late
g=3(Year 2015)	g=4(Year 2018)	-0.007938	0.304860	Early vs Late
g=2/3/4(Year 2013/2015/2018)	Never treated	0.026487	0.359233	Timing vs. never-treated
g=3(Year 2015)	g=2(Year 2013)	-0.054547	0.017876	Late vs Early
g=4(Year 2018)	g=2(Year 2013)	-0.034341	0.078332	Late vs Early
g=4(Year 2018)	g=3(Year 2015)	0.032617	0.152430	Late vs Early

Figure C1 Outcome variables comparisons between treated and control groups



Note: The vertical solid red line represents the treated time.

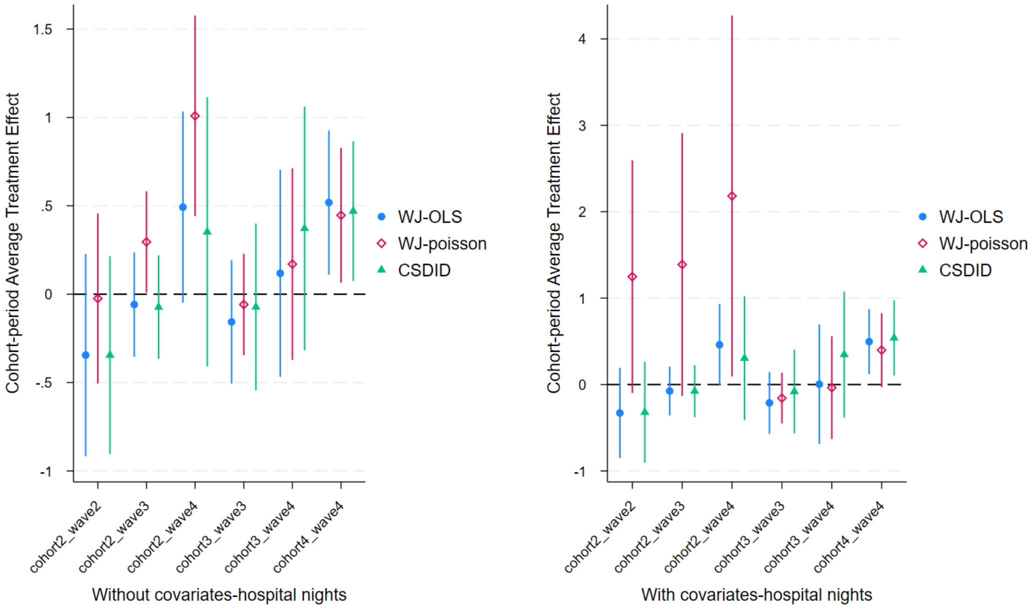


Figure C2 Different estimators for hospital nights (95% CI)

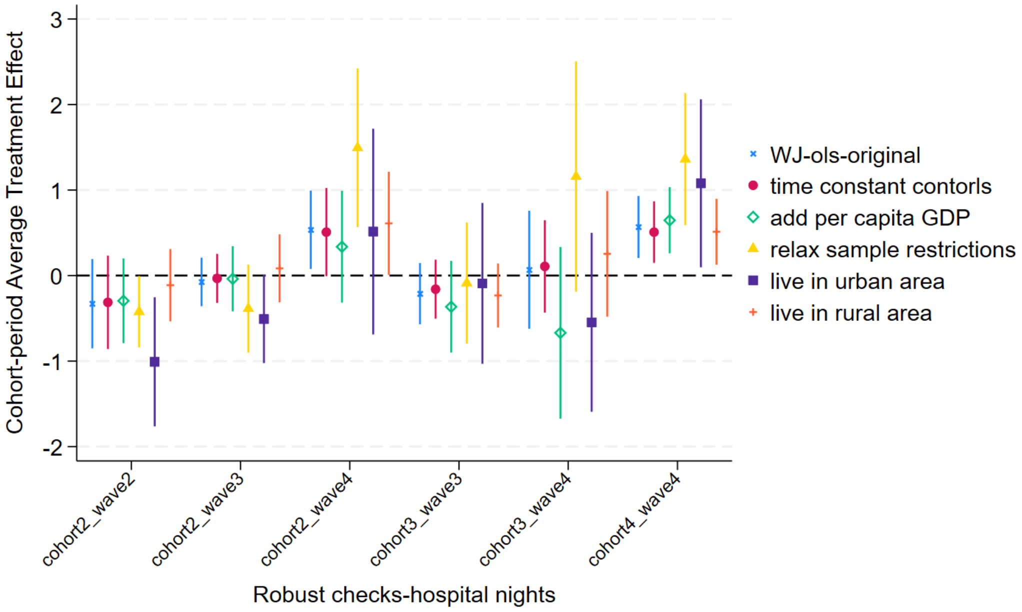


Figure C3 Robust checks for hospital nights (95% CI)

Reference

Abadie, A. (2005) Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72 (1): 1–19.

Acuña, C., Acuña, H. and Carrasco, D. (2019) Health shocks and the added worker effect: a life cycle approach. *Journal of Applied Economics*, 22 (1): 273–286.

Andersen, R. and Newman, J.F. (1973) Societal and individual determinants of medical care utilization in the United States. *The Milbank Memorial Fund Quarterly. Health and Society*, pp. 95–124.

Andersen, R.M. (1995) Revisiting the behavioral model and access to medical care: does it matter? *Journal of Health and Social Behavior*, pp. 1–10.

Andersen, R.M. (2008) National health surveys and the behavioral model of health services use. *Medical Care*, pp. 647–653.

Assembly, WHO(2005) Sustainable Health Financing, Universal Coverage and Social Health Insurance. *WHA58*, 33.

Baker, A.C., Larcker, D.F. and Wang, C.C. (2022) How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics*, 144 (2): 370–395.

Basu, S., Andrews, J., Kishore, S., et al. (2012) Comparative performance of private and public healthcare systems in low-and middle-income countries: a systematic review. *PLoS Medicine*, 9 (6): e1001244.

Becker, G.S. (1993) *A Treatise on the Family: Enlarged Edition*. Harvard University Press.

Behncke, S. (2012) Does retirement trigger ill health?. *Health economics*, 21(3), pp.282-300.

Benjamin, D., Brandt, L. and Fan, J. (2003) Health and labor supply of the elderly in rural China. Unpublished manuscript, Department of Economics, University of Toronto.

Berger, M.C. (1983) Labor supply and spouse's health: The effects of illness, disability, and mortality. *Social Science Quarterly*, 64 (3): 494.

Berger, M.C. and Fleisher, B.M. (1984) Husband's health and wife's labor supply. *Journal of Health Economics*, 3 (1): 63–75.

Bloom, D.E., Chatterji, S., Kowal, P., et al. (2015) Macroeconomic implications of population ageing and selected policy responses. *The Lancet*, 385(9968), pp.649-657.

Blundell, R. and MaCurdy, T. (1999) Labor supply: A review of alternative approaches. *Handbook of Labor Economics*, 3: 1559–1695.

Borusyak, K., Jaravel, X. and Spiess, J. (2024) Revisiting Event Study Designs: Robust and Efficient Estimation. Available at: http://arxiv.org/abs/2108.12419 (Accessed: 18 January 2024).

Bound, J. (1991) Self-Reported Versus Objective Measures of Health in Retirement Models. *The Journal of Human Resources*, 26 (1): 106–138.

Bound, J., Schoenbaum, M., Stinebrickner, T.R., et al. (1999) The dynamic effects of health on the labor force transitions of older workers. *Labour Economics*, 6 (2): 179–202.

Braakmann, N. (2014) The consequences of own and spousal disability on labor market outcomes and subjective well-being: evidence from Germany. *Review of Economics of the Household*, 12: 717–736.

Brambor, T., Clark, W.R. and Golder, M. (2006) Understanding interaction models: Improving empirical analyses. *Political Analysis*, 14(1), pp.63-82.

Brekke, K.R., Nuscheler, R. and Straume, O.R. (2007) Gatekeeping in health care. *Journal of Health Economics*, 26 (1): 149–170.

Browning, M., Bourguignon, F., Chiappori, P.-A., et al. (1994) Income and outcomes: A structural model of intrahousehold allocation. *Journal of Political Economy*, 102 (6): 1067–1096.

Browning, M., Chiappori, P.-A. and Lechene, V. (2006) Collective and unitary models: A clarification. *Review of Economics of the Household*, 4: 5–14.

Buis, M.L. (2010) Stata tip 87: Interpretation of interactions in nonlinear models. *The Stata Journal*, 10(2), pp.305-308.

Cahuc, P., Carcillo, S. and Zylberberg, A. (2014) *Labor Economics*. MIT Press.

Cai, F., Giles, J., O'Keefe, P. and Wang, D. (2012) Old-Age support in rural China: Challenges and prospects. Washington, DC: World Bank.

Cai, L. and Kalb, G. (2006) Health status and labour force participation: evidence from Australia. *Health Economics*, 15 (3): 241–261.

Caliendo, M. and Gehrsitz, M. (2016) Obesity and the labor market: A fresh look at the weight penalty. *Economics & Human Biology*, 23: 209–225.

Callaway, B. and Sant'Anna, P.H. (2021) Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225 (2): 200–230.

Charles, K.K. (1999) Sickness in the family: Health shocks and spousal labor supply. Working paper.

Chatterji, P., Joo, H. and Lahiri, K. (2017) Diabetes and labor market exits: evidence from the health & retirement study (hrs). *The Journal of the Economics of Ageing*, 9: 100–110.

Chau, T.W., Hongbin, L., Liu, P.W., et al. (2007) Testing the collective model of household labor supply: Evidence from China. *China Economic Review*, 18 (4): 389–402.

Che, Y. and Li, X. (2018) Retirement and health: evidence from China. *China Economic Review*, 49, pp.84-95.

Chen, H., Ding, Y., Tang, L., et al. (2022) Impact of urban–rural medical insurance integration on consumption: Evidence from rural China. *Economic Analysis and Policy*, 76: 837–851.

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

Chen, X., Giles, J., Yao, Y., Yip, W., Meng, Q., Berkman, L., Chen, H., Chen, X., Feng, J., Feng, Z. and Glinskaya, E. (2022) The path to healthy ageing in China: a Peking University–Lancet Commission. *The Lancet*, 400(10367), pp.1967-2006.

Chen, Z. and Woolley, F. (2001) A Cournot–Nash model of family decision making.

The Economic Journal, 111 (474): 722–748.

Chiappori, P.-A. (1988) Rational household labor supply. *Econometrica: Journal of the Econometric Society*, pp. 63–90.

Chiappori, P.-A. (1992) Collective labor supply and welfare. *Journal of Political Economy*, 100 (3): 437–467.

Chiappori, P.-A. (2011) Collective labor supply with many consumption goods. *Review of Economics of the Household*, 9: 207–220.

Chiappori, P.-A., Fortin, B. and Lacroix, G. (2002) Marriage market, divorce legislation, and household labor supply. *Journal of Political Economy*, 110 (1): 37–72.

Chiswell, H.M. (2018) From generation to generation: changing dimensions of intergenerational farm transfer. *Sociologia Ruralis*, 58(1), pp.104-125.

Chiu, A., Lan, X., Liu, Z., et al. (2023) What to do (and not to do) with causal panel analysis under parallel trends: lessons from a large reanalysis study. Working paper. Available at: http://arxiv.org/abs/2309.15983 (Accessed: 18 January 2024).

Coile, C. (2004) Health shocks and couples' labor supply decisions.Working paper

Cribb, J., Emmerson, C. and Tetlow, G. (2013) Incentives, shocks or signals: labour supply effects of increasing the female state pension age in the UK. IFS working papers (No. W13/03).

De Chaisemartin, C. and d'Haultfoeuille, X. (2018) Fuzzy differences-in-differences. *The Review of Economic Studies*, 85 (2): 999–1028.

De Chaisemartin, C. and D'Haultfœuille, X. (2020) Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110 (9): 2964–2996.

Démurger, S. and Li, S. (2013) Migration, remittances, and rural employment patterns: evidence from China. In *Labor Market Issues in China*, pp.31-63. Emerald Group Publishing Limited.

Deschryvere, M.K. (2005) Health and retirement decisions: an update of the literature. ENEPRI Research Report No. 6.

Disney, R., Emmerson, C. and Wakefield, M. (2006) Ill health and retirement in Britain:

A panel data-based analysis. *Journal of Health Economics*, 25 (4): 621–649.

Dow, W.H., Norton, E.C. and Donahoe, J.T. (2019) Stata tip 134: multiplicative and marginal interaction effects in nonlinear models. *The Stata Journal*, 19(4), pp.1015-1020.

Fang, E.F., Xie, C., Schenkel, J.A., et al. (2020) A research agenda for ageing in China in the 21st century: Focusing on basic and translational research, long-term care, policy and social networks. *Ageing Research Reviews*, 64, p.101174.

Farrell, M., Kinsella, A., O'Donoghue, C., Mahon, M. and Leonard, B. (2020) Risky (farm) business: Perceptions of economic risk in farm succession and inheritance. *Journal of Rural Studies*, 75, pp.57-69.

Feng, J., Gong, Y., Li, H., et al. (2022) Development trend of primary healthcare after health reform in China: a longitudinal observational study. *BMJ Open*, 12 (6): e052239.

Forrest, C.B. (2003) Primary care gatekeeping and referrals: effective filter or failed experiment? *BMJ*, 326 (7391): 692–695.

Fortin, B. and Lacroix, G. (1997) A test of the unitary and collective models of household labour supply. *The Economic Journal*, 107 (443): 933–955.

Frijters, P., Johnston, D.W. and Shields, M.A. (2010) Mental health and labour market participation: Evidence from IV panel data models. IZA Discussion Paper No. 4883

García Gómez, P. and López Nicolás, A. (2006) Health shocks, employment and income in the Spanish labour market. *Health Economics*, 15 (9): 997–1009.

García-Gómez, P., Jones, A.M. and Rice, N. (2010) Health effects on labour market exits and entries. *Labour Economics*, 17 (1): 62–76.

García-Gómez, P., Van Kippersluis, H., O'Donnell, O., et al. (2013) Long-term and spillover effects of health shocks on employment and income. *Journal of Human Resources*, 48 (4): 873–909.

García-Miralles, E. and Leganza, J.M. (2024) Joint retirement of couples: Evidence from discontinuities in Denmark. *Journal of Public Economics*, 230, p.105036.

Garrido, M.V., Zentner, A. and Busse, R. (2011) The effects of gatekeeping: a systematic review of the literature. *Scandinavian Journal of Primary Health Care*, 29(1), pp.28-38.

Giedion, U., Andrés Alfonso, E. and Díaz, Y. (2013) The impact of universal coverage schemes in the developing world: a review of the existing evidence. World Bank Publications.

Giles, J., Lei, X., Wang, G., Wang, Y. and Zhao, Y. (2023) One country, two systems: Evidence on retirement patterns in China. *Journal of Pension Economics & Finance*, 22(2), pp.188-210.

Giovanis, E. and Ozdamar, O. (2019) A collective household labour supply model with disability: Evidence from Iraq. *Journal of Family and Economic Issues*, 40 (2): 209–225.

Goodman-Bacon, A. (2021) Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225 (2): 254–277.

Grossman, M. (1972) On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, 80 (2): 223–255.

Grossman, M. and Benham, L. (1974) “Health, hours and wages.” In The Economics of Health and Medical Care: Proceedings of a conference held by the International Economic Association at Tokyo. 1974. Springer. pp. 205–233.

Haddad, L.J. and Kanbur, S.R. (1991) Intrahousehold inequality and the theory of targeting. World Bank Publications.

Hao, H. and Yeo, Y.H. (2023) Does the integration of urban and rural health insurance influence the functional limitations of the middle-aged and elderly in rural China? *SSM-Population Health*, 23: 101439.

Haveman, R., Wolfe, B., Kreider, B., et al. (1994) Market work, wages, and men’s health. *Journal of Health Economics*, 13 (2): 163–182.

Heckman, J.J. and MaCurdy, T.E. (1980) A life cycle model of female labour supply. *The Review of Economic Studies*, 47 (1): 47–74.

Heinesen, E. and Kolodziejczyk, C. (2013) Effects of breast and colorectal cancer on labour market outcomes-average effects and educational gradients. *Journal of Health Economics*, 32 (6): 1028–1042.

Hollenbeak, C.S., Farley Short, P. and Moran, J. (2011) The implications of cancer survivorship for spousal employment. Journal of Cancer Survivorship, 5: 226–234.

Hou, B., Wang, G., Wang, Y., et al. (2021) The health capacity to work at older ages in urban China. *China Economic Review*, 66, p.101581.

Hu, H., Liang, H. and Wang, H. (2021) Longitudinal study of the earliest pilot of tiered healthcare system reforms in China: Will the new type of chronic disease management be effective? *Social Science & Medicine*, 285: 114284.

Hu, H., Wang, R., Li, H., et al. (2023) Effectiveness of hierarchical medical system policy: an interrupted time series analysis of a pilot scheme in China. *Health Policy and Planning*, 38 (5): 609–619.

Huang, M., Zhang, H., Gu, Y., et al. (2019) Outpatient health-seeking behavior of residents in Zhejiang and Qinghai Province, China. *BMC Public Health*, 19 (1): 967.

Huang, X. and Kim, S.E. (2020) When top‐down meets bottom‐up: Local adoption of social policy reform in China. *Governance*, 33 (2): 343–364.

Huang, X. and Wu, B. (2020) Impact of urban-rural health insurance integration on health care: evidence from rural China. *China Economic Review*, 64: 101543.

Hurd, M.D. and Boskin, M.J. (1984) The effect of social security on retirement in the early 1970s. *The Quarterly Journal of Economics*, 99 (4): 767–790.

Ikegami, N., Yoo, B.-K., Hashimoto, H., et al. (2011) Japanese universal health coverage: evolution, achievements, and challenges. *The Lancet*, 378 (9796): 1106–1115.

Imai, K., Kim, I.S. and Wang, E.H. (2023) Matching Methods for Causal Inference with Time‐Series Cross‐Sectional Data. *American Journal of Political Science*, 67 (3): 587–605.

Jeon, S.-H. and Pohl, R.V. (2017) Health and work in the family: Evidence from spouses' cancer diagnoses. *Journal of Health Economics*, 52: 1–18.

Jiang, J., Huang, W., Wang, Z., et al. (2019) The effect of health on labour supply of rural elderly People in China-An empirical analysis using CHARLS data. *International Journal of Environmental Research and Public Health*, 16(7), p.1195.

Johnson, R.W. and others (2001) Retiring together or working alone: The impact of spousal employment and disability on retirement decisions. CRR Working Paper No. 2001-01

Jones, A.M., Rice, N. and Roberts, J. (2010) Sick of work or too sick to work? Evidence on self-reported health shocks and early retirement from the BHPS. *Economic Modelling*, 27 (4): 866–880.

Jones, A.M., Rice, N. and Zantomio, F. (2020) Acute health shocks and labour market outcomes: Evidence from the post crash era. *Economics & Human Biology*, 36: 100811.

Kalantaryan, S., Scipioni, M., Natale, F., et al. (2021) Immigration and integration in rural areas and the agricultural sector: An EU perspective. *Journal of Rural Studies*, 88, pp.462-472.

Kalwij, A. and Vermeulen, F. (2008) Health and labour force participation of older people in Europe: What do objective health indicators add to the analysis? *Health Economics*, 17 (5): 619–638.

Killingsworth, M.R. and Heckman, J.J. (1986) Female labor supply: A survey. *Handbook of Labor Economics*, 1: 103–204.

Kondo, A. and Shigeoka, H. (2013) Effects of universal health insurance on health care utilization, and supply-side responses: evidence from Japan. *Journal of Public Economics*, 99: 1–23.

Kwon, S. (2003) Healthcare financing reform and the new single payer system in the Republic of Korea: Social solidarity or efficiency? *International Social Security Review*, 56 (1): 75–94.

Lee, J.-C. (2003) Health Care Reform in South Korea: Success or Failure? American *Journal of Public Health*, 93 (1): 48–51.

Lee, R. and Mason, A. (2010) Fertility, Human Capital, and Economic Growth over the Demographic Transition. *European Journal of Population*, 26(2), pp.159-182.

Lei, X., Zhang, C. and Zhao, Y. (2013) Incentive problems in China's new rural pension program. In *Labor Market Issues in China*, 37, pp.181-201. Emerald Group Publishing Limited.

Li, C., Tang, C. and Wang, H. (2019) Effects of health insurance integration on health care utilization and its equity among the mid-aged and elderly: evidence from China. *International Journal for Equity in Health*, 18 (1): 166.

Li, L. and Fu, H. (2017) China's health care system reform: Progress and prospects. *The International Journal of Health Planning and Management*, 32 (3): 240–253.

Li, Q., Lei, X. and Zhao, Y. (2014) The effect of health on the labor supply of mid-aged and older Chinese. *China Economic Quarterly*, 13(3), pp.917-938.

Li, S. and Yang, Y. (2021) An empirical study on the influence of the basic medical insurance for urban and rural residents on family financial asset allocation. *Frontiers in Public Health*, 9: 725608.

Li, X., Krumholz, H.M., Yip, W., et al. (2020) Quality of primary health care in China: challenges and recommendations. *The Lancet*, 395 (10239): 1802–1812.

Liu, J., Fang, Y., Wang, G., et al. (2023) The aging of farmers and its challenges for labor-intensive agriculture in China: A perspective on farmland transfer plans for farmers' retirement. *Journal of Rural Studies*, 100, p.103013.

Liu, P., Guo, W., Liu, H., et al. (2018a) The integration of urban and rural medical insurance to reduce the rural medical burden in China: a case study of a county in Baoji City. *BMC Health Services Research*, 18 (1): 796.

Liu, Y., Kong, Q. and de Bekker-Grob, E.W. (2019) Public preferences for health care facilities in rural China: a discrete choice experiment. *Social Science & Medicine*, 237: 112396.

Liu, Y., Kong, Q., Yuan, S., et al. (2018b) Factors influencing the choice of health system access level in China: a systematic review. *The Lancet*, 392: S39.

Liu, Y., Zhong, L., Yuan, S., et al. (2018c) Why patients prefer high-level healthcare facilities: a qualitative study using focus groups in rural and urban China. *BMJ Global Health*, 3 (5).

Luft, H.S. (1975) The impact of poor health on earnings. *The Review of Economics and Statistics*, pp. 43–57.

Lundberg, S. and Pollak, R.A. (1993) Separate spheres bargaining and the marriage market. *Journal of Political Economy*, 101 (6): 988–1010.

Macchioni Giaquinto, A., Jones, A.M., Rice, N., et al. (2022) Labor supply and informal care responses to health shocks within couples: Evidence from the UK. *Health Economics*, 31(12), pp.2700-2720.

MaCurdy, T.E. (1981) An Empirical Model of Labor Supply in a Life-Cycle Setting. *Journal of Political Economy*, 89 (6): 1059–1085.

McClellan, M.B. (1998) "Health events, health insurance, and labor supply: Evidence from the health and retirement survey." In *Frontiers in the Economics of Aging*. University of Chicago Press. pp. 301–350.

McElroy, M.B. (1990) The empirical content of Nash-bargained household behavior. *Journal of Human Resources*, pp. 559–583.

McElroy, M.B. and Horney, M.J. (1981) Nash-bargained household decisions: Toward a generalization of the theory of demand. *International Economic Review*, pp. 333–349.

Meng, Q., Fang, H., Liu, X., et al. (2015) Consolidating the social health insurance schemes in China: towards an equitable and efficient health system. *The Lancet*, 386 (10002): 1484–1492.

Meng, T. and Su, Z. (2021) When top-down meets bottom-up: Local officials and selective responsiveness within fiscal policymaking in China. *World Development*, 142: 105443.

Michaud, P.C., Van Soest, A. and Bissonnette, L. (2020) Understanding joint retirement. *Journal of Economic Behavior & Organization*, 173, pp.386-401.

Mincer, J. (1962) "Labor force participation of married women: A study of labor supply." In Aspects of labor economics. Princeton University Press. pp. 63–105.

Minor, T. (2013) An investigation into the effect of type I and type II diabetes duration on employment and wages. *Economics & Human Biology*, 11 (4): 534–544.

Mitra, S., Gao, Q., Chen, W., et al. (2020) Health, work, and income among middle-aged and older adults: A panel analysis for China. *The Journal of the Economics of Ageing*, 17, p.100255.

Mize, T.D. (2019) Best practices for estimating, interpreting, and presenting nonlinear interaction effects. *Sociological Science*, 6, pp.81-117.

Munro, N. and Duckett, J. (2016) Explaining public satisfaction with health‐care systems: findings from a nationwide survey in China. *Health Expectations*, 19 (3): 654–666.

Ning, M., Gong, J., Zheng, X., et al.(2016) Does new rural pension scheme decrease elderly labor supply? Evidence from CHARLS. *China Economic Review*, 41, pp.315-330.

OECD. (2017) The silver and white economy: the Chinese demographic challenge. OECD Publishing, Paris.

Oreffice, S. and Quintana-Domeque, C. (2012) Fat spouses and hours of work: are body and Pareto weights correlated? IZA Journal of Labor Economics, 1 (1): 6.

Organization, W.H. (2013) World health report 2013: Research for universal health coverage. World Health Organization.

Panpiemras, J., Puttitanun, T., Samphantharak, K., et al. (2011) Impact of Universal Health Care Coverage on patient demand for health care services in Thailand. *Health Policy*, 103 (2–3): 228–235.

Parsons, D.O. (1977) Health, Family Structure, and Labor Supply. *The American Economic Review*, 67 (4): 703–712.

Qian, D., Pong, R.W., Yin, A., et al. (2009) Determinants of health care demand in poor, rural China: the case of Gansu Province. *Health Policy and Planning*, 24 (5): 324–334.

Ratigan, K. (2017) Disaggregating the developing welfare state: Provincial social policy regimes in China. *World Development*, 98: 467–484.

Reis, M.C. (2007) Added worker effect: Evidence from health shocks in the Brazilian informal labor market. Unpublished paper, IPEA.

Ren, C., Zhou, X., Wang, C., Guo, Y., et al. (2023) Ageing threatens sustainability of smallholder farming in China. *Nature*, 616(7955), pp.96-103.

Ren, Y., Zhou, Z., Cao, D., et al. (2022) Did the integrated urban and rural resident basic medical insurance improve benefit equity in China? *Value in Health*, 25 (9): 1548–1558.

Rigg, J., Phongsiri, M., Promphakping, B., et al. (2020) Who will tend the farm? Interrogating the ageing Asian farmer. *The Journal of Peasant Studies*, 47(2), pp.306-325.

Sant'Anna, P.H. and Zhao, J. (2020) Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219 (1): 101–122.

Schirle, T. (2008) Why have the labor force participation rates of older men increased since the mid-1990s?. *Journal of Labor Economics*, 26(4), pp.549-594.

Shen, D., Liang, H. and Shi, W.(2023) Rural Population Aging, Capital Deepening, and Agricultural Labor Productivity. *Sustainability*, 15(10), p.8331.

Smith, J.P. (2012) Preparing for population aging in Asia: Strengthening the infrastructure for science and policy. In *Aging in Asia: Findings from New and Emerging Data Initiatives*. National Academies Press (US).

Smith, J.P., Strauss, J. and Zhao, Y. (2014) Healthy aging in China. *The Journal of the Economics of Ageing*, 4, pp.37-43.

Spletzer, J.R. (1997) Reexamining the added worker effect. *Economic Inquiry*, 35 (2): 417–427.

Stephens, Jr., M. (2002) Worker Displacement and the Added Worker Effect. *Journal of Labor Economics*, 20 (3): 504–537.

Stern, S. (1989) Measuring the effect of disability on labor force participation. *Journal of Human Resources*, pp. 361–395.

Su, D., Chen, Y., Gao, H., et al. (2019) Effect of integrated urban and rural residents medical insurance on the utilisation of medical services by residents in China: a propensity score matching with difference-in-differences regression approach. *BMJ Open*, 9 (2).

Sun, L. and Abraham, S. (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225 (2): 175–199.

Sun, X., Shi, Y., Zeng, Q., et al. (2013) Determinants of health literacy and health behavior regarding infectious respiratory diseases: a pathway model. *BMC Public Health*, 13 (1): 261.

Vecchio, N. (2015) Labour force participation of families coping with a disabling condition. *Economic Analysis and Policy*, 45: 1–10.

Wang, J., Zhu, H., Liu, H., et al. (2020a) Can the reform of integrating health insurance reduce inequity in catastrophic health expenditure? Evidence from China. *International Journal for Equity in Health*, 19 (1): 49.

Wang, S. and Yang, D.Y. (2021) Policy experimentation in china: The political economy of policy learning. National Bureau of Economic Research. Woking paper (No. w29402)

Wang, Y., Castelli, A., Cao, Q., et al. (2020b) Assessing the design of China's complex health system–Concerns on equity and efficiency. *Health Policy Open*, 1: 100021.

Wooldridge, J.M. (2021) Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. Working Paper. Available at SSRN 3906345.

Wooldridge, J.M. (2023) Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*, 26 (3): C31–C66.

Wu, D. and Lam, T.P. (2016) Underuse of primary care in China: the scale, causes, and solutions. *The Journal of the American Board of Family Medicine*, 29 (2): 240–247.

Wu, D., Lam, T.P., Lam, K.F., et al. (2017) Health reforms in china: the public's choices for first-contact care in urban areas. *Family Practice*, 34 (2): 194–200.

Xiao, Y., Chen, X., Li, Q., et al. (2021) Towards healthy China 2030: Modeling health care accessibility with patient referral. *Social Science & Medicine*, 276: 113834.

Xu, H. (2017) The time use pattern and labour supply of the left behind spouse and children in rural China. *China Economic Review*, 46, pp.S77-S101.

Xu, J. and Mills, A. (2017) Challenges for gatekeeping: a qualitative systems analysis of a pilot in rural China. *International Journal for Equity in Health*, 16 (1): 106.

Xu, J., Wang, X., Hao, H., et al. (2021) Impact of hierarchical hospital reform on patients with diabetes in China: a retrospective observational analysis. *BMJ Open*, 11 (4): e041731.

Yang, D., Acharya, Y. and Liu, X. (2022) Social health insurance consolidation and urban-rural inequality in utilization and financial risk protection in China. *Social Science & Medicine*, 308: 115200.

Yip, W. and Hsiao, W. (2014) Harnessing the privatisation of China's fragmented health-care delivery. *The Lancet*, 384 (9945): 805–818.

Yip, W., Fu, H., Chen, A.T., et al. (2019) 10 years of health-care reform in China: progress and gaps in universal health coverage. *The Lancet*, 394 (10204): 1192–1204.

Zeng, Y., Xu, W., Chen, L., et al. (2020) The Influencing Factors of Health-Seeking Preference and Community Health Service Utilization Among Patients in Primary Care Reform in Xiamen, China. *Patient Preference and Adherence*, Volume 14: 653–662.

Zhang, L., Gu, J. and An, Y. (2023) The optimal delayed retirement age in aging China: Determination and impact analysis. *China Economic Review*, 79, p.101972.

Zhang, S. and Wang, Z. (2024) The effect of hierarchical medical treatment reform——research based on 205 prefecture-level cities. *Applied Economics*, pp.1-15.

Zhao, Y., Strauss, J., Chen, X., et al. (2020) China health and retirement longitudinal study wave 4 user's guide. National School of Development, Peking University, pp. 5–6.

Zhou, Q., He, Q., Eggleston, K., et al. (2022) Urban-rural health insurance integration in china: impact on health care utilization, financial risk protection, and health status. *Applied Economics*, 54 (22): 2491–2509.

Zhou, Z., Zhao, Y., Shen, C., et al. (2021) Evaluating the effect of hierarchical medical system on health seeking behavior: a difference-in-differences analysis in China. *Social Science & Medicine*, 268: 113372.