Can Education Improve Group Health Awareness? Dynamic Pseudo-panel Evidence from China's Immigration

Shixi Kang[1*], Jingwen Tan[2]

[1]School of economics, Henan University

[2]University of Copenhagen

keywords: Migrant, Health equity, Pseudo-panel, China

[*] Correspondence: Shixi Kang, School of economics, Henan University, Kaifeng, Henan, 475000, China. Email: ksx@henu.edu.cn

Can Education Improve Group Health Awareness? Dynamic Pseudo-panel Evidence from China's Immigration

Abstract

Improving residents' willingness to participate in basic health care services is a key initiative to optimize the allocation of health care resources and promote equitable improvements in group health. In this paper, we use a systematic GMM model based on pseudo-panel synthetic five-year data of migrants repeated cross-sectional data to explore the effect of education level on the completion rate of residents' health records. The results show that: (1) Under a static perspective, higher cohort education would generate a significant positive return to resident file completion rates, and this return would be underestimated when ignoring pseudo-panel cohort heterogeneity; (2) Under a dynamic perspective, there is a significant cumulative effect and long-term inertia of the resident health record completion rate, and there is a significant effect of the health record completion rate in previous years on the current year's completion. (3) The positive relationship between education level and the rate of completing health records of the population was also characterized by heterogeneity in terms of gender and education category. Among them, men groups were more likely to improve their willingness to make decisions due to higher education levels, and this improvement effect was more pronounced among those with only basic education.

Keywords: Migrant, Health equity, Pseudo-panel, SYS-GMM; China

1. Background

In the past decades, health economics related to health equity has always been a popular research area in welfare states. In recent years, how to effectively utilize health resources, improve national health awareness, and thus promote social health benefits has also gradually become a pressing concern for developing countries. In many large Chinese cities, the migrants is an important part of the labor pool. As the size of the migrant population expands (Figure 1), the potential role of this group for economic development becomes more and more negligible. However, the health status of the migrant population remains worrisome compared to that of residents. As an important human resource driving economic growth, the health and equality of the migrants need urgent care.

Before exploring what policymakers, who wish to improve health equity need to consider, we first clarify the definition of health equity. Objectively, health equity can be expressed as a balance in the provision of health resources between regions. For example, Aday, etal (1984) and Whitehead (1991) argue that health equity is the equal opportunity for everyone to improve their health status without the need for and ability to access health resources disparate by socioeconomic class, gender, or race. What is often overlooked is that the true health needs of some groups, such as the migrant population, are far underestimated. For example, in many large cities in China, the participation rate of the migrant population in public health activities such as filling out resident health records is generally low (Figure 2), and this willingness to participate is largely related to many objective factors such as the ability to access information and the efforts of the health authorities to publicize such activities. Since health awareness is directly related to the amount of health resources they are willing to actively access, this low willingness to participate in public health activities is actually a manifestation of health inequality.

To ameliorate this inequality, policymakers will be concerned with these two questions: First, where can investments be made to best ameliorate this health inequality? Since differences in income, working hours, and living area will have a significant impact on people's health awareness, education, as a process of improving individual knowledge, values, and cognitive abilities, is conducive to motivating people's need for health, and this motivation will accumulate and grow throughout their life course (Mirowsky, et al., 2005). Therefore, it is necessary to analyze in depth whether the government can implement effective health construction and

interventions by improving the educational level of this special group of migrants, to break the cycle of poverty of the disadvantaged group and thus achieve health equity for this group.

In addition, policymakers are also more concerned about which groups within the migrants would be more likely to benefit from such policies. Knowing the beneficiary groups helps policymakers develop more targeted policies, but distinguishing between these groups is not an empirical exercise. For example, Ross, et al. (2010) found that the mitigation of health impairment by higher levels of education was more pronounced in the female group. However, in the less educated sample, women's health status was conversely weaker than men's. There is also a common view that health inequalities are universal across generations and across the life course (West, 1997), and Ross, et al. (1996) suggest that health disparities among people with different levels of education will vary with age. According to cumulative advantage theory (CET), the positive effect of educational attainment on health increases with age, leading to greater heterogeneity and inequality in health among older adults.

This paper explains these questions by examining the causal relationship between the education level of the Chinese migrants and their health awareness: first, does the level of education directly affect their willingness to participate actively in public health activities? It is important to examine this factor for the improvement of health equity because differences in income, working hours, and living area will have a significant impact on people's health awareness, and education, as a process of improving individuals' knowledge, values, and cognitive abilities, may improve health equity by motivating people's need and concern for health, and this motivation will accumulate and grow throughout their life course (Mirowsky, 2005). This incentive will accumulate and grow throughout their lives (Mirowsky, et al., 2005). However, the direction of the effect of this factor is not obvious. For example, migrants may increase their concern for their own health due to increased education or expand their ability to access information about health activities through this increase. However, migrants with increased education may also be less likely to engage in health activities due to increased hours and intensity of work.

Our analysis is based on a pseudo-panel data consisting of five years of cross-sectional data, and there are many advantages to this type of data processing. First, when estimating returns to education, the consequences it entails are prone to contain more possibilities due to sample differences because of the broad nature of the education concept itself, and the pseudo-panel form of the data can attenuate individual heterogeneity by constructing cohorts. Second, it improves the sample comparability that is reduced by tracking sample attrition in our long-term surveys and facilitates the extension of the time horizon of our study. To ensure that we obtain unbiased estimates of the effect of education level on health awareness among migrants, we also consider other possible errors. For example, there may be a bidirectional causal relationship between education level and health awareness, i.e., those who are able to achieve higher levels of education will themselves need to develop additional concerns about their health (Berniell, et al., 2020). To weaken the endogeneity due to dynamic bias, we introduced a systematic GMM model to investigate the health payoffs in the long term. In addition, we demonstrate the robustness of health returns by enumerating regression results from multiple models and adjustments for instrumental variables in the systematic GMM model.

The contributions of this paper are as follows: first, the pseudo-panel model constructed in this paper extends the use of the non-tracking repeated cross-sectional survey database. As Deaton (1989) suggests, true panel data for longer time spans do not exist for the vast majority of countries due to the sample rotation and non-random attrition problems of statistical surveys. Instead, the for-panel model allows for different individuals to be observed in each observation period and focuses more on the statistical characteristics of the cohort composed of individuals. The pseudo-panel data constructed in this paper on the one hand eliminates individual measurement error and on the other hand demonstrates that investigators do not need to continuously track fixed individuals to obtain long time span panel data, further increasing the potential for the use of non-tracking repeated cross-sectional databases.

Moreover, this paper provides an innovative definition of health equity. Existing studies have focused more on the geographical distribution of health resources on the supply side, and on the demand side on the unequal share of health resources among residents due to differences in income and social status. For example, Wu, et al.(2019) found that there are eight times more hospitals per square kilometer in the East than in the West, and that health resources are more concentrated in developed areas with wealthy populations. Whitehead, et al.(2001) found that the poor are more likely to be trapped in a vicious cycle of disease and poverty due to lower income and lack of health insurance compared to the rich. However, both alleviating the inequitable distribution of resources among the population and promoting a comprehensive and balanced expansion of health resources in the country are high-investment, difficult-to-promote approaches that do not guarantee the efficient use of invested resources. Health equity is not only about the objective equal distribution of health care resources, nor is it only about income differences among residents, but it should also include the subjective willingness of residents to use health care resources. This approach also maximizes the use of excess health resources in some areas and reveals the areas and groups that are most in need of resource investment. This paper explores the relationship between educational attainment and local health record completion rates, which have a significant impact on individual growth and development, in order to provide a new perspective on the impact of educational attainment on population health awareness.

The remainder of this paper is organized as follows: Part II describes the research methodology of the pseudo-panel construction; Part III presents the data sources and variable selection; Part IV reports the results of the empirical analysis; and Part V gives conclusions and recommendations.

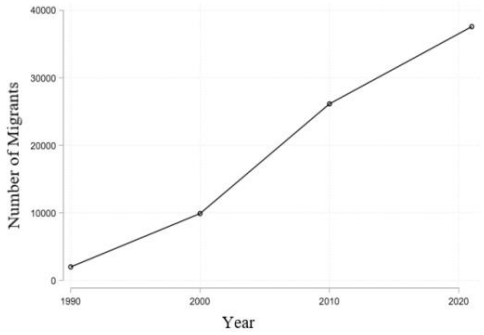

Figure 1 Changes in the size of the migrants
Note: Date from National Bureau of Statistics of the People's Republic of China

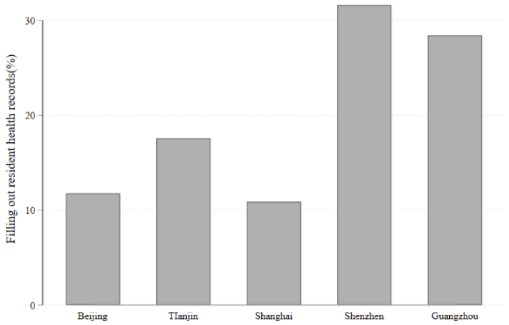

Figure 2 Proportion of migrants filling out resident health records in Chinese cities in 2018
Note: Date from National Health Commission of the People's Republic of China

2. Method

Panel data play a crucial role in inferring long-term causal relationships. However, an inevitable problem in collecting panel data is the loss of tracking samples. Deaton (1985) introduced the concept of Pseudo-Panel Data for this purpose. The cross-sectional data of the sample are randomly selected at different times, and the total sample is divided into cohorts according to certain characteristics, and the mean value of each variable in each cohort is taken separately for each year of data, and these means become the variable values of these cohorts with common characteristics, and the cohorts will replace the individuals in the original sample to form the new panel data. Since the pseudo-panel has the advantage of filling the missing data of the real panel, this paper will do the pseudo-panel treatment on the five-year cross-sectional data used based on this method.

We first simply combine the five-period cross-sectional data into pool data and assume that there is a linear function of health equality of the following form:

$$Health_{it} = \alpha + \beta_1 edu_{it} + \beta_1 X' + \lambda_i + \varepsilon_{it} \quad i = 1,...,N \quad t = 1,...,T \qquad (1)$$

$Health_{it}$ is the health equity of the sample i at period t. The repeated question "Do you have a local health record" in the five-year data of Migrants Dynamic Monitoring Survey (CMDS) is used as a proxy variable

for $Health_{it}$ which is transformed into a proportional form in the pseudo-panel. Thus $Health_{it}$ is transformed into the percentage of people in the group who fill out a health record. edu_{it} is the number of years of education in period t for the ith sample, which is the core explanatory variable in this paper, and the direction and significance of its coefficient β_1 is our focus. X' contains a series of individual characteristics and urban control variables, λ_i is an unobserved individual effect, and u_t is then used to control for annual trends and any unobservable linear variation over time in the independent variables.

There are still many influences that cannot be quantified in the completion of residents' health records, for example, the degree of secrecy of residents' health records by the government or health institutions affects migrants' willingness to participate in the completion, but this indicator is difficult to measure in practice. Therefore, we first tested Eq. (1) for omitted variables (Oster, 2019) and found that unobservable variables had a large effect on the regression results and there was omitted variable bias (Table.1).

To reduce the effect of omitted variables on the regression structure, we construct a pseudo panel model. First, we define a set of cohorts $c(1,...,C)$, where each sample i will be uniquely included in a certain cohort. These cohorts are often identified by some common characteristics of individuals, such as year of birth, race, gender, and other non-time-varying variables (Russell, et al., 2005). Taking the mean value of each variable in each cohort over time, a pseudo panel model is obtained as follows:

$$\overline{Health_{ct}} = \bar{\alpha} + \beta_1 \overline{edu_{ct}} + \beta_2 \overline{X'} + \overline{\lambda_c} + \bar{u}_t + \bar{\varepsilon}_{ct} \quad c = 1,...,C \;\; t = 1,...,T \qquad (2)$$

At this moment, all error components associated with the original sample i are removed, and the substitution of the cohort effect $\overline{\lambda_c}$ for the individual effect λ_i not only controls the effect of individual heterogeneity on the regression results, but its mean implementation form also reduces the measurement error of individuals (Antman, et al., 2007). Also, Deaton's (1985) study showed that cohort fixed effects estimates corrected for measurement error were consistent. Since the sample size tends to vary for each year of cross-sectional data, the resulting calculated cohort mean effects may also differ across years, but the reason for not retaining a t subscript for $\overline{\lambda_c}$ is that $\overline{\lambda_{ct}}$ can be approximated as $\overline{\lambda_c}$ when the number of individuals included in the cohort is sufficiently large.

However, although we attenuate endogeneity due to measurement error, there may still be a bidirectional causal relationship between education and health equity: those who are able to achieve higher levels of education inherently need good health status. In addition, considering that the question "Have you established a resident record in your local area" includes the possibility of establishing a new record in the previous year or in the current year, in order to remove the effect of the previous year on the current year and to accurately capture the causal relationship between the education level and the record completion rate of the current year's migrants, this paper includes in the additional dynamic analysis a first-order lag of the health The first-order lag term of file filling rate is added to the additional dynamic analysis in this paper to obtain a pseudo-panel model:

$$\overline{Health_{ct}} = \bar{\alpha} + \beta_1 \overline{Health_{ct-1}} + \beta_2 \overline{edu_{ct}} + \beta_3 \overline{X'} + \overline{\lambda_c} + \bar{u}_t + \overline{\varepsilon_{ct}} \quad c = 1,...,C \;\; t = 1,...,\mathrm{T} \qquad (3)$$

To alleviate the possible endogeneity in dynamic panel models, Arellano, et al.(1991) proposed a differential GMM approach using instrumental variables to derive moment conditions. However, due to certain weak instrumental variable problems with the differential GMM approach, Blundell, et al.(1998) proposed a more perfect systematic GMM approach. Systematic GMM is able to correct for unobserved individual heterogeneity, omitted variable bias, measurement error, and another potential endogeneity. And it can reduce the potential bias and error rate due to the use of first-order difference GMM estimation methods. Meanwhile, Roodman(2009) et al. argue that the standard error of the two-step GMM estimates will be significantly lower

with a limited sample. Therefore, in this paper, the two-step systematic GMM is used as the main model under the dynamic perspective and is supplemented by the difference GMM to demonstrate the robustness of the main regression results. Meanwhile, we tested for omitted variables again for Eq.(2) and Eq.(3) and found that unobservable variables had to be at least 5.17-6.05 to have a significant effect on the regression results, indicating a low probability of the existence of omitted variables.

A point of concern is that there is a non-negligible problem in determining the number of cohorts: more cohorts can increase the heterogeneity of the pseudo-panel by increasing the common characteristics of individuals, but it is also accompanied by a decrease in the number of individuals in the cohort. Therefore, identifying cohorts with a larger number of individuals is necessary for the pseudo-panel to accurately estimate subgroup means (Moffitt, 1993; Verbeek, et al., 2005). For the selection of cohorts, birth year, gender, and region were chosen as criteria for delineation in order to observe the returns of education to health equity in cohorts with generational, gender, and regional differences. Regarding the classification of birth year in cohorts, Blundell, et al. (1998) suggested that samples from every decade should be grouped into one cohort, while Browning, et al. (1985) adopted the criterion of defining cohorts every five years. Due to the large amount of data in this paper and the large number of individuals in the cohort, every five years was chosen as the basis for dividing the cohort, and a total of nine cohorts were obtained in the birth year section. Similarly, we divide the gender (male and female) and region (eastern, central and western, by economic geography concept) into 2 and 3 cohorts respectively, and then each year of cross-sectional data is divided into 54 cohorts in total. The final total number of cohorts is ($9 \times 2 \times 3 \times 5 = 270$). In the study by Verbeek (2008), the pseudo-panel estimates of subgroup means were more accurate when the number of individuals in each category was greater than 100. The demonstration of the number of individuals in each cohort in Table 2 allows us to conclude that our design is reliable.

Table 1 Oster bounds for OLS(Pool Data and Pseudo-panel Data)

	$R_{max} = 1.3\tilde{R}$	
	$\delta(Estimated\ \beta = 0)$	$\beta^*(\delta = 1)$
Pool Data		
edu_{it}	-0.14884	[0.57293, 0.00962]
Pseudo-panel		
$\overline{edu_{ct}}$ (Eq.2)	5.17120	[0.66907, 0.00138]
$\overline{edu_{ct}}$ (Eq.3)	6.05593	[0.20614, 0.01843]

Table 2 Number of individuals in the cohort

Year of birth	Sex	Region		
		East	Central	West
1955-1959	Male	3758	1142	3190
1960-1964	Male	7785	2996	6740
1965-1969	Male	16542	6388	13178
1970-1974	Male	25078	9528	19237
1975-1979	Male	26690	9969	18216
1980-1984	Male	33097	11347	20525
1985-1989	Male	36606	12821	23366
1990-1994	Male	21071	7146	13457
1995-1999	Male	7128	2241	5003
1955-1959	Female	2508	555	1854
1960-1964	Female	4930	1709	4015

1965-1969	Female	11366	4600	8753
1970-1974	Female	18892	7114	13366
1975-1979	Female	20842	7802	13449
1980-1984	Female	27461	9463	16328
1985-1989	Female	36750	13206	22745
1990-1994	Female	26893	10040	17875
1995-1999	Female	9626	3185	6541

Note: Data are from the sum of the number of individuals in each year of CMDS 2014, 2015, 2016, 2017, and 2018.

3. Data

3.1. Sources

The five-year repeated cross-sectional data of China Migrants Dynamic Survey (CMDS2014-2018) organized by the National Health Care Commission of China were selected as the samples in this paper. Among them, the original sample size was 200937 in 2014, 206000 in 2015, 169000 in 2016, 169,989 in 2017, and 152,000 in 2018.Since the CMDS data before and after 2014 had questions on the health status of the migrants' large discrepancies, therefore, only the post-2014 period was selected as the study year in this paper.

The survey covers 31 provinces (municipalities and autonomous regions) in mainland China, and uses a stratified, multi-stage, large-scale PPS sampling method to conduct a dynamic monitoring survey covering various aspects such as individual characteristics, employment status and health status for the migrant population who have stayed in the local area for more than one month, which is an important data to measure various characteristics of the migrant population. In addition, because of the small sample of some prefectural cities in CMDS data for each year, the control variables of cities at the prefectural level cannot meet the need of estimating a large number of overall, and there are many missing data for some prefectural cities and corps in the statistical yearbook of CMDS, so this paper selects some variables from the China Urban Statistical Yearbook at the provincial level to control the characteristics of cities. Table 3 gives both the expected negative signs and data sources for the main explanatory and explanatory variables.

Table 3 Description of Data

Variable	Expected sign	Definition	Source
Health	-	Percentage of migrants filling out health records	CMDS
edu	Positive	Years of education	
Income	Negative	Logarithm of average monthly household income .	
flowt	Negative	The time of the inflow of immigrants to the place	
living	Positive	Number of family members living together	
hos	Positive	Total number of hospitals (per province)	CCSY
doc	Positive	Total number of doctors (per 1,000 population and per province)	
bed	Positive	Total number of beds (per 1,000 population and per province)	
gender	-	Male=1; Female=0	CMDS
education level	-	Higher education level>9; Basic education level<=9	

***Note*:** CMDS from National Health Commission of the People's Republic of China. CCSY comes from National Bureau of Statistics of the People's Republic of China.

3.2. Variables

The main object of this paper is the return of education on health equity. Since the resident health record is a systematic information resource for all residents in the jurisdiction, recording their various age stages and covering various health factors to provide them with medical and health services. It is an important prerequisite for residents to fully enjoy medical resources, and it can also represent the health decision-making intention of the migrants. Therefore, in this paper, the questionnaire "Have you established a local health record" was selected as the explanatory variable to measure the degree of health equality in each year. The pseudo-panel

mean is the percentage of people who have established a health record in each year, which reflects the health decision willingness of different groups with the same characteristics. The transformed years of education was also selected as the core explanatory variable.

The control variables in this paper include individual advantage and urban characteristics. Since there is a sample with age below 25 years, there is a possibility of updating individual education level, so a first-order lag term of years of education is added to control its change. The final individual characteristics include first-order lags of education level, number of cohabitants, duration of mobility, and logarithm of total household income, while the urban characteristics are total number of hospitals, number of beds per 1,000 people, and number of practicing physicians per 1,000 people. Meanwhile, this paper takes 2014 as the base period and adjusts the total household income of the sample in 2015-2018 according to the consumer price index (CPI), and excludes samples with income higher than the first 2.5% and lower than the second 7.5%, in order to reduce the influence of outliers on the regression results. Table 4 gives the results of descriptive statistics for the main variables.

Table 4 Descriptive statistics of main variables

Variables	Obs.	Mean	Std.	Min	Max
Percentage of people with a health record（health）	686,113	0.354570	0.093728	0.181856	0.649154
Number of years of education(edu）	686,113	10.10032	1.261168	5.872881	12.32104
Number of people living together(living)	686,113	2.905024	0.455329	1.701707	3.662313
Mobility time(flowt)	686,113	5.599776	2.023126	2.251090	11.05050
Total household income is taken as logarithm(income)	686,113	8.569698	0.159939	8.061189	8.921130
Total number of hospitals(hos)	686,113	33747.15	6950.848	24773.16	56054.24
Total number of beds per 1,000 people(bed)	686,113	5.343850	0.477565	4.595852	6.614391
Total number of doctors per 1,000 people(doc)	686,113	2.026219	0.278158	1.531212	2.631216

4. Results

4.1. Baseline results

The main and control regressions for the benchmark analysis are given in Table 5. Columns 1 and 2 show the regression results for Pooled OLS and Pseudo-panel fixed effects OLS without the dependent variable lag term, respectively, and both regressions use robust standard errors clustered by cohort. The regression results suggest that a pseudo-panel controlling for cohort effects and individual differences exists with higher estimation accuracy, and model 2 has significantly higher regression coefficients on returns to education compared to model 1 based on Pool Data. This finding suggests that the neglect of cohort differences can lead to significant utility underestimation. One possible explanation for this phenomenon is that people with higher levels of education tend to be engaged in complex tasks that require a lot of time, and the time occupation makes them inclined to delay plans to fill out health records. This negative correlation between education level and the unobservable error term could lead to an underestimation of the regression coefficients, and the introduction of cohort fixed effects in model 2 avoids this endogeneity bias. Again, this is like Warunsiri, et al. (2010) and Juodis (2018) regarding the emphasis on the importance of cohort fixed effects. Therefore, we will default to model 3 to model 6 analysis in the form of controlling for cohort effects.

Model 3 to Model 6 introduces a lagged term for the dependent variable, demonstrating the process of providing robust estimates from the main regression6 by showing regression results based on different STEPS and IVs. Column 3 reports the regression results for the fixed effects OLS, the two-step differenced GMM estimates based on the first IV are reported in column 4, while columns 5 and 6 report the one-step versus two-step systematic GMM estimates based on the second IV, respectively. The reason for choosing to use the two-step method instead of the one-step method in the main regression is that the first-order difference can lead to a large number of missing observations due to the loss of recent values of the variables (Roodman, 2019), while the two-step estimation can effectively avoid data loss due to the internal transformation problem (Arellano, et al.,1995). From Table 5, the standard deviation of column 6 estimated based on the two-step method is much smaller than that of column 5 estimated by the one-step method only, which also further demonstrates the

reliability of using the two-step method.

Also, we put explanatory variables that may be potentially related to the nuisance term, such as education level (edu), into endogenous variables, and macro-level urban control variables, such as total number of hospitals (hos), into exogenous variables, and follow Roodman (2009) that when N is small and IV is too much it may be necessary to use collapse to reduce the bias for the corresponding The table also reports the Hansen J test value, Diff-in-Hansen value, AR(1) and AR(2) values. Table 5 shows that all instrumental variables passed the overidentification test and that there was no serial correlation in the model residual terms. The models in Table 5 are convergent in their estimates of the education coefficients, which further demonstrates the robustness of the results.

In column 6, the coefficient of the first order lagged term of the dependent variable is 0.806, indicating that there is long-term inertia in the resident health record completion rate, with the effect of previous years' record completion rate reaching 80.6% in the current year, making the inclusion of dynamic considerations necessary to accurately determine the return to education on health equity. And after controlling for this effect, the effect of education on resident health record completion rates falls back to a more precise 4.7% from the uncontrolled 20.4%.

In addition, the expected signs of the explanatory variables were largely consistent with the direction of the actual regression coefficients. For example, the number of family members is significantly and positively correlated with the rate of completing residents' health records, with each additional family member increasing the proportion of migrant's groups completing health records by 28%. It relies on a simple logic that improved health awareness largely stems from social support and supervision by groups such as family members. The higher-income migrant groups tend to have longer working hours and have difficulty following up on the completion of residents' health records over time. Other macro control variables such as the total number of hospitals, the number of physicians per 1,000 population, and the number of beds per 1,000 population also fit the pattern of positive correlation with the proportion of resident file completion.

Table 5 Baseline regression results

Health	Pool OLS(1)	OLS(2)	OLS(3)	DIFF-GMM(4)	SYS-GMM(5)	SYS-GMM(6)
edu	0.010***	0.204**	0.184***	0.228***	0.040**	0.047***
	(0.000)	(0.091)	(0.069)	(0.041)	(0.019)	(0.011)
L.health			0.024	-0.153**	0.644***	0.806***
			(0.062)	(0.073)	(0.104)	(0.059)
lninco1	-0.011***	-0.008	-0.046*	-0.052***	0.007	-0.011
	(0.001)	(0.040)	(0.027)	(0.016)	(0.055)	(0.031)
flowt	0.002***	-0.013	-0.049	-0.106***	-0.054***	-0.036***
	(0.000)	(0.052)	(0.041)	(0.018)	(0.019)	(0.011)
living	0.021***	0.305*	0.101	0.259***	0.393***	0.280***
	(0.001)	(0.168)	(0.152)	(0.101)	(0.071)	(0.049)
hos	-0.000***	-0.000	-0.000	-0.000	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
doc	-0.133***	-2.249***	0.109	-0.991***	1.064***	1.146***
	(0.001)	(0.697)	(0.671)	(0.301)	(0.300)	(0.116)
bed	0.064***	-0.386	-0.669**	-1.098***	1.500***	1.258***
	(0.001)	(0.393)	(0.304)	(0.173)	(0.232)	(0.125)
Constant	0.173***	2.724	0.925		-14.766***	-13.031***
	(0.011)	(3.365)	(2.984)		(2.033)	(0.978)
Number of Cohorts		270	216	162	216	216
Time Dummies	√	√	√	√	√	√
Step				twostep	onestep	twostep
Number of IV				42	63	63
IV（group）				1	2	2
Hansen J test				0.402	0.926	0.108
Diff-in-Hansen					0.899	0.899
AR(1)(p-value)				0.001	0.000	0.000
AR(2)(p-value)				0.472	0.128	0.108

Note: The first group is gmm(edu income flowt living), IV(i.cohort i.year hos doc bed); The second group is gmm(edu hos income ,lag(5 .)) gmm(flowt living, lag(4 .) collapse), iv(i.cohort i.year doc bed). The first group is gmm(edu income flowt living), IV(i.cohort i.year hos doc bed); The second group is gmm(edu income ,lag(5 .)) gmm(flowt living, lag(4 .) collapse), iv(i.cohort i.year doc bed). We follow the idea of using collapse to limit the bias

through over-instrumentation in Roodman (2009) when there is less N and too much IV. The Hansen test is used to determine whether the instrumental variables used by the system GMM are valid overall. Diff-in-Hansen was used to determine the validity of the necessary additional moment limits in the system GMM.

4.2 Robustness analysis

The robustness analysis of the benchmark model is given in Table 6, where column 4 shows the results of the benchmark regression. In the actual measurement, a study using only one proxy variable may imply only one situation-specific outcome. Therefore, we changed the proxy variable measuring immigrants' health awareness in column 1 from the proportion of residents' health records established to the proportion participating in community-based public health education activities. This is a non-compulsory and universal health care activity, and the proportion of participation in this activity can reflect the health awareness of the immigrant group to some extent. From column 1, the regression coefficient of education level is significantly positive after replacing the dependent variable, and the coefficient size is less different from the baseline results in column 4, which proves the robustness of our choice of proxy variables.

Column 2 introduces a differential GMM approach for additional estimation while ensuring that other things remain the same. Although the differential GMM provides slightly higher estimates of the education coefficient than the baseline regression, the reason for this phenomenon is more likely based on the removal of the effect of non-time-varying variables by the differential GMM. For example, immigrants with older birth years tend to be more receptive to and understand health records, and when this factor is excluded, its incentive effect on the proportion of records created is shifted to other independent variables (e.g., education level), which in turn generates an overestimate of the coefficients on these independent variables. Overall, however, column 2 exhibits the same positive significance as the baseline regression, indicating that our use of the systematic GMM as a regression model is robust.

Column 3 demonstrates the robustness of the choice of instrumental variables in the benchmark regression by varying the assumptions of some of the instrumental variables. We exclude the number of doctors and beds per 1,000 population from the strictly exogenous variables, assuming that there may still be an endogenous effect on the health record establishment rate at the macro level and find that the estimated coefficients on education level differ less from the benchmark regression in direction and magnitude, which initially indicates the robustness of the latter results. In addition, the estimated standard deviations of both column 3 education level and last period health record establishment rate were larger than those of the benchmark regression, further indicating the validity of the benchmark regression results.

Table 6 Robustness analysis

Health	Change of Dependent (1)	Change of Model(2)	Change of IV(3)	Baseline(4)
edu	0.050***	0.144***	0.039*	0.047***
	(0.012)	(0.031)	(0.023)	(0.011)
L.health	0.155***	-0.131***	0.900***	0.806***
	(0.052)	(0.022)	(0.191)	(0.059)
Control	√	√	√	√
Constant	0.538		-5.984	-13.031***
	(1.144)		(5.210)	(0.978)
Number of Cohorts	216	162	216	216
Time Dummies	√	√	√	√
Step	twostep	twostep	twostep	twostep
Number of IV	63	42	65	63
IV（group）	2	2	1	2
Hansen J test	0.421	0.203	0.749	0.108
Diff-in-Hansen	0.742		0.707	0.899
AR(1)(p-value)	0.003	0.000	0.001	0.000
AR(2)(p-value)	0.553	0.346	0.121	0.108

Note: The first IV group is 'gmm(edu income flowt living, lag(4 .)collapse) iv(i.cohort i.year)'; The second IV group is 'gmm(edu hos income ,lag(5 .)) gmm(flowt living, lag(4 .) collapse), iv(i.cohort i.year doc bed)'. The Hansen J test is used to determine whether the instrumental variables used by the system GMM are valid overall. Diff-in-Hansen was used to determine the validity of the necessary additional moment limits in the system GMM. The presentation of regression results for the control variables has been omitted.

4.2. Disaggregation by gender and education level

In addition to the overall estimate of the return to health equity from education level, we still need to consider which groups would have greater utility from additional investment in health records by government or health agencies. Immigrant groups are divided into two dimensions in this section: gender and education stratum. Because China has a nine-year compulsory education, in the education subgroup samples, we record those with greater than nine years as those with higher education and those with less than or equal to nine years as those with basic education.

Table 7 shows the results of the subgroup regressions. Overall, the four subgroup regressions demonstrated a positive effect of education level in the cohort on the rate of health record establishment. Columns 1 and 2 show that the coefficient of education level is 0.075 for male immigrants and insignificant for female immigrants. According to the results of the gender decomposition, the return to education in health awareness was more pronounced for males, indicating that each year of increase in education level was associated with a 7.5% increase in record establishment rate. This is like the findings of Beckfield, et al. (2018) in their study of the gender divergence of social investment on health equity in European countries. And Ross, et al. (2010) showed that women with low levels of education had less access to good health than men with the same level of education.

Columns 3 and 4 show the health record establishment rates among the cohorts under basic and higher education, respectively. For immigrants with less than 9 years of education, education level brings a 14.7% additive effect on health record creation, while among immigrants with higher levels of education, education level does not have a significant positive effect on health record creation. One possible explanation is that those with higher levels of education tend to work longer hours (Zhang, 2008), and this form of work apparently reduces their likelihood and willingness to establish a record. Although, in general, the level of education motivates people to increase their willingness to participate in health activities, this motivational effect is rather significantly weakened in groups with higher levels of education. In other words, even if higher income earners have access to more health resources, their health awareness and needs are still not promising. Even if the basic education audience is more likely to have the time to fill out a health record, access to adequate and quality health resources is still a pressing concern.

Table 7 Heterogeneity Analysis

health	Male(1)	Female(2)	Basic Education(3)	Higher Education(4)
meanedu	0.075*	0.045	0.147*	-0.003
	(0.041)	(0.053)	(0.087)	(0.022)
L.meandangan	0.858***	0.990***	0.852***	0.593***
	(0.163)	(0.172)	(0.180)	(0.074)
Control	√	√	√	√
Constant	-6.438*	-8.110**	-1.387**	-0.882
	(3.883)	(3.526)	(0.539)	(0.784)
Number of Cohorts	108	108	80	136
Number of IV	36	36	31	63
Hansen J test	0.915	0.974	0.920	0.832
Diff-in-Hansen	0.821	0.946-	0.855	0.685
AR(1)(p-value)	0.004	0.033	0.067	0.001
AR(2)(p-value)	0.306	0.525	0.147	0.700

Note: The presentation of regression results for the control variables has been omitted.

5. Conclusion

This study used a systematic GMM model to estimate the health returns to increased education in the Chinese migrant population group for this group based on a pseudo-panel constructed from five-year cross-sectional data from the CMDS. The pseudo-panel data format weakened the estimation bias due to individual heterogeneity,

and the GMM model reduced the dynamic error of education level in influencing the health record completion rate. The results show that education can give transient and positive returns to health decision-making intentions among the migrants, and such returns are underestimated when cohort heterogeneity is ignored; at the same time, there is a significant cumulative effect of file completion rate, and file completion in previous years will have a positive effect on the current year. The positive relationship between education and willingness to make health decisions is also characterized by heterogeneity by gender and education level itself. Among them, education is more likely to increase the willingness of male migrants to participate in public health activities, and this incentive effect is more significant in the group receiving basic education.

The high return to health from education found in the study and the different returns found in the disaggregated probes provide answers to the questions of policy makers. First, this paper affirms the finding that increasing the education level of the migrants can significantly improve this group's willingness to participate in public health activities, with the overall estimate of a 4.7% increase in the rate of completing health records for the migrants from rising education levels, and that increased financial spending on education may be another national measure to improve national health awareness. In addition, health care inequities cannot be fundamentally addressed through government investment in health care alone. The distribution of health care resources is one-way; access to resources is two-way. Some low-education groups can be covered by healthcare resources, but there are few proactive initiatives to utilize the resources, instead resulting in a waste of resources. Improving education can improve the awareness of the public to use medical resources and increase the efficiency of medical resources. At the same time, for the higher level of health returns shown by the male group, the state should pay attention to the health inequality of the female group through the gender education gap while strengthening their education in order to improve the overall health level, so as to achieve true health equity in a group sense. In addition, while we should develop basic education, which has a more significant incentive effect on the willingness of migrants to participate in public health activities, this does not mean that we should invest less in higher education. Exploring the mechanisms that influence group demand for health at higher levels of education and, in turn, addressing this mechanism remains a key initiative to achieve health equity. Finally, with the arrival of the new crown epidemic, a large-scale exogenous shock factor, a timely extension of the time horizon of existing studies would provide policy makers with a more detailed and nuanced empirical evidence base.

References

[1] Aday, Fleming, Anderson. An overview of current access issues [J]. Access to Medical Care in the US: Who Have It, Who Don, 1984, 1: 1-18.

[2] Whitehead. The concepts and principles of equity and health [J]. Health promotion international, 1991, 6(3): 217-28.

[3] Mirowsky, Ross. Education, learned effectiveness and health [J]. London Review of Education, 2005, 3(3): 205-20.

[4] Ross, Mirowsky. Gender and the health benefits of education [J]. The Sociological Quarterly, 2010, 51(1): 1-19.

[5] West. Health inequalities in the early years: is there equalisation in youth? [J]. Social science & medicine, 1997, 44(6): 833-58.

[6] Ross, Wu. Education, age, and the cumulative advantage in health [J]. Journal of health and social behavior, 1996: 104-20.

[7] Berniell, Bietenbeck. The effect of working hours on health [J]. Economics & Human Biology, 2020, 39: 100901.

[8] Deaton. Panel data from time series of cross-sections [J]. Journal of econometrics, 1985, 30(1-2): 109-26.

[9] Wu, Yang. Inequality trends in the demographic and geographic distribution of health care professionals in China: Data from 2002 to 2016 [J]. The International journal of health planning and management, 2019, 34(1): e487-e508.

[10] Whitehead, Dahlgren, Evans. Equity and health sector reforms: can low-income countries escape the medical poverty trap? [J]. The Lancet, 2001, 358(9284): 833-6.

[11] Oster. Unobservable selection and coefficient stability: Theory and evidence [J]. Journal of Business & Economic Statistics, 2019, 37(2): 187-204.
[12] Russell, Fraas. An application of panel regression to pseudo panel data [J]. Multiple linear regression viewpoints, 2005, 31(1): 1-15.
[13] Antman, McKenzie. Poverty traps and nonlinear income dynamics with measurement error and individual heterogeneity [J]. The journal of development studies, 2007, 43(6): 1057-83.
[14] Arellano, Bond. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations [J]. The review of economic studies, 1991, 58(2): 277-97.
[15] Blundell, Duncan, Meghir. Estimating labor supply responses using tax reforms [J]. Econometrica, 1998: 827-61.
[16] Roodman. How to do xtabond2: An introduction to difference and system GMM in Stata [J]. The stata journal, 2009, 9(1): 86-136.
[17] Moffitt. Identification and estimation of dynamic models with a time series of repeated cross-sections [J]. Journal of Econometrics, 1993, 59(1-2): 99-123.
[18] Verbeek, Vella. Estimating dynamic models from repeated cross-sections [J]. Journal of econometrics, 2005, 127(1): 83-102.
[19] Browning, Deaton, Irish. A profitable approach to labor supply and commodity demands over the life-cycle [J]. Econometrica: journal of the econometric society, 1985: 503-43.
[20] Verbeek. Pseudo-panels and repeated cross-sections [M]. The econometrics of panel data. Springer. 2008: 369-83.
[21] Warunsiri, McNown. The returns to education in Thailand: A pseudo-panel approach [J]. World Development, 2010, 38(11): 1616-25.
[22] Juodis. Pseudo panel data models with cohort interactive effects [J]. Journal of Business & Economic Statistics, 2018, 36(1): 47-61.
[23] Roodman D, Nielsen M Ø, MacKinnon J G, et al. Fast and wild: Bootstrap inference in Stata using boottest[J]. The Stata Journal, 2019, 19(1): 4-60.
[24] Beckfield, Morris, Bambra. How social policy contributes to the distribution of population health: the case of gender health equity [J]. Scandinavian journal of public health, 2018, 46(1): 6-17.
[25] Arellano, Bover. Another look at the instrumental variable estimation of error-components models [J]. Journal of econometrics, 1995, 68(1): 29-51.
[26] Beckfield, Morris, Bambra. How social policy contributes to the distribution of population health: the case of gender health equity [J]. Scandinavian journal of public health, 2018, 46(1): 6-17.
[27] Muennig, Robertson, Johnson, et al. The effect of an early education program on adult health: the Carolina Abecedarian Project randomized controlled trial [J]. American journal of public health, 2011, 101(3): 512-6.
[28] Zhang. The way to wealth and the way to leisure: The impact of college education on graduates' earnings and hours of work [J]. Research in Higher Education, 2008, 49(3): 199-213.