Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 11: 1–18
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076251314102
journals.sagepub.com/home/dhj
Sage

The impacts on population health by China's regional health data centers and the potential mechanism of influence

Jiaoli Cai[1,2], Yue Li[1] and Peter C Coyte[3]

Abstract

Background: China recently established a series of pilot regional health data centers with a mandate to acquire, consolidate, analyze, and translate data into evidence for health policy decision-making. This experiment with "big data" has the potential to influence population health and is the focus of this study.

Methods: This study used national longitudinal survey data from the China Family Panel Studies over the period 2014–2020 to empirically assess the impact of China's establishment of pilot regional health data centers on population health and health inequality. A difference-in-differences model was employed to investigate the policy effect on population health, with additional exploration of possible mechanisms of influence. The corrected concentration index was used to measure health inequality, while Wagstaff decomposition method was applied to examine the marginal influence of the policy effect on health inequality.

Results: Overall health status of local residents has improved after the establishment of the pilot regional health data centers. Using mechanism analysis, the findings demonstrated that improvements to population health were driven by promoting healthy lifestyles and innovations in medical practices. Furthermore, due to differences in individual e-health literacy, the pilot centers produced "pro-rich" health inequality where high-income groups benefited more from the establishment of the pilot centers in terms of health than low-income groups.

Conclusions: This study has highlighted the potential to improve population health, in general, with the advent of big data centers, but for these benefits be unevenly distributed among the resident population.

Keywords

Healthcare, big data, health, health inequality, e-health literacy, China

Submission date: 2 July 2024; Acceptance date: 3 January 2025

Introduction

While good health by most metrics has improved over the past century[1] and is crucial for human development,[2,3] health inequality remains a common problem.[4–6] Health inequality is often manifest in differences in health outcomes between different socioeconomic groups[7,8] with income-related health inequality shown to be greater than solely income or health inequality.[9–11] Inequalities in health among socioeconomic groups have been a focus of public health policy in many countries.[12]

To advance population health and to address income-related health inequality, China has embarked on a series of ambitious health system reforms.[13,14] The success of these reforms was critically dependent on access to and the availability of data, methods, and expertise so that health policy decision-making was evidence-based. With enhanced data collection tools, the

[1]School of Economics and Management, Beijing Jiaotong University, Beijing, China
[2]Research Center for Central and Eastern Europe, Beijing Jiaotong University, Beijing, China
[3]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

Corresponding author:
Jiaoli Cai, School of Economics and Management, Beijing Jiaotong University, No.3 Shangyuancun Haidian District Beijing 100044, P. R. China.
Email: jiaoli.cai@bjtu.edu.cn

advent of an array of valid, reliable, and linkable data sets, and the availability of sophisticated data storage, retrieval, and analysis systems, the opportunities to employ such "big data systems" to inform health policy decision-making in China is at a vital and promising stage of development.

For example, the Chinese State Council issued a circular in 2016 concerning the application and development of "big data" in the health and medical sectors.[15] This initiative aimed to acquire, consolidate, analyze, and translate data into evidence for health policy decision-making. The council acknowledged the importance of "big data" as a strategic national resource and as a tool to foster the development of health systems and to enhance population health.[16]

Medical big data represent the specialized application of "big data" in the medical domain,[17,18] and it bears significant potential in addressing contemporary medical challenges and in fostering evidence-based health policy decision-making.[16] Medical big data are becoming increasingly pivotal as a resource for crafting more effective personalized treatments and services for patients,[19] enhancing healthcare delivery and patient outcomes,[18] and minimizing resource wastage.[17]

The use of "big data" in healthcare plays an increasingly important role in medical diagnosis and management.[20–22] Existing research primarily focuses on two areas. The first provides a general overview of how big data can enhance population health management or improve health outcomes for specific diseases, such as enhancing precision health management,[23] improving cardiovascular care,[18] and advancing precision treatment for breast cancer.[24] The second focuses on the technical applications of "big data" in healthcare, such as its use in reducing medical errors,[25] supporting clinical decision-making,[26] achieving real-time health evaluations, and assisting in remote monitoring.[27] While these studies highlight the positive role of big data in improving population health and the quality of medical services, the direct impact of big data on health outcomes has rarely been empirically explored. There is also limited understanding of how big data may influence the health of individuals and populations. Specifically, research on China's regional health data center pilots is primarily limited to descriptive theoretical analyses regarding policy implications and future trends.[16] No relevant empirical studies on health outcomes have been conducted. China's regional health data center pilots serve as a testing ground for the application of big data in healthcare in China's healthcare system. Understanding the impact of these center pilots on people's health and health inequalities is crucial for the further development of big data in healthcare. Consequently, the purpose of this study is threefold: First, to evaluate the health consequences of China's establishment of a series of regional health data centers; Second, to assess their potential mechanisms of influence; and third, to estimate the impact of the pilot programs on socioeconomic-related health inequality.

This study makes three key contributions. First, this is the first study to empirically examine the impact of pilot regional health data centers on population health. Previous studies were mostly descriptive and aimed at the analysis of the practical application of "big data" in healthcare. Our study provides evidence for the health effects of pilot regional health data centers from an empirical perspective. Second, this study examines the mechanisms by which pilot programs affect population health. We suggest that healthcare big data can support healthy lifestyles and foster medical innovations, thereby promoting population health. Finally, this study focuses on e-health literacy and examines the unequal distribution of health benefits resulting from the pilot programs, which particularly favor the wealthy. This reminds policymakers to be wary of the uneven benefits of digital welfare across populations caused by the characteristics of pilot programs.

This paper is organized as follows: the second section introduces the policy background and puts forward a theoretical framework. The third section reports on the data, model, and econometric procedures employed. The empirical results are presented in the fourth section, followed by a discussion in the fifth section. The final section concludes the manuscript.

Theoretical framework

Policy background

Since October 21, 2016, a national pilot program has been initiated in stages for constructing regional health data centers across China. The program aims to diminish data barriers and enhance the application of "big data" in healthcare. A total of six regional health data centers have been established. The first batch of pilot centers was announced on October 21, 2016, and comprised two regional data center pilots in Fujian and Jiangsu Provinces. The second batch of pilot centers was announced on December 12, 2017, and comprised data centers in Shandong, Anhui, and Guizhou Provinces. The third announcement was made in July 2018 and resulted in the establishment of the Ningxia Hui Autonomous Region Regional Data Center. These centers were designed to integrate existing regional "big data" in healthcare, including health data, administrative data, and electronic medical records, to effectively address persistent medical challenges, support academic research, and provide evidence for decision-making.[16]

Theoretical hypotheses

The Chinese government's pilot policy on regional health data centers is approached from two key perspectives: the demand side, representing individual citizens with health needs, and the supply side, involving the supply and innovation level of medical services. The primary goal is to diminish data barriers for all individuals and enhance the application of "big data" in healthcare. From a supply perspective, pilot centers facilitate the sharing of "big data" resources in healthcare, which creates an accessible data foundation for medical innovation. Moreover, the government stresses the application of "big data" in healthcare, including support for the development of various internet medical enterprises and the promotion of clinical and scientific research, thereby driving innovation in medical practice. Medical innovations aid healthcare decision-making,[28,29] improve the safety and accuracy of healthcare,[30] and help to enhance the way in which healthcare is organized and delivered.[31] From the perspective of client demand, the establishment of the pilot centers helps to promote the importance of health and helps to form a variety of health applications, wearable devices, and "new technology" for people to use.[31,32] The applications can help users and patients to improve people's awareness of health, improve health planning, and encourage healthy lifestyles. The users or patients can become advocates for their own health.

Therefore, China's regional health data centers have the potential to drive innovation in medical practice and promote healthy lifestyles, thereby improving population health. Three research hypotheses have been proposed.

> Hypothesis 1: The pilot regional health data centers can enhance population health.

> Hypothesis 2: The pilot regional health data centers can enhance population health by promoting healthy lifestyles.

> Hypothesis 3: The pilot regional health data centers can enhance population health by enhancing innovation in medical practices.

E-health literacy refers to the population's subjective acceptability of accessing health information online and their ability to evaluate health information and solve health problems on electronic resources.[33,34] The pilot regional health data centers emphasize the use of "new technologies." The acceptance and use of "new technologies" are based on people's strong digital skills. People with good digital skills and high subjective acceptance can benefit more from new medical methods.[34] However, due to differences in e-health literacy among populations, people's acceptance of "new technologies" varies. For instance, higher socioeconomic groups typically exhibit greater e-health literacy,[35] are more willing to try new types of technology platform,[36] and are more prone to explore the internet for health information,[33] showing higher subjective acceptance of new technologies. Therefore, due to differences in e-health literacy, groups with higher socioeconomic status may benefit more from the pilot regional health data centers, which in turn affects health status and leads to health inequality.

> Hypothesis 4: The pilot regional health data centers have varying health impacts across socioeconomic groups, leading to "pro-rich" health inequality.

Figure 1 shows the research framework.

Methods

Data

In this empirical study, microdata were drawn from the Chinese Family Panel Studies (CFPS). The CFPS is a national longitudinal survey initiated by the Institute of Social Science Survey of Peking University in 2010 and conducted biennially. It documents changes in Chinese

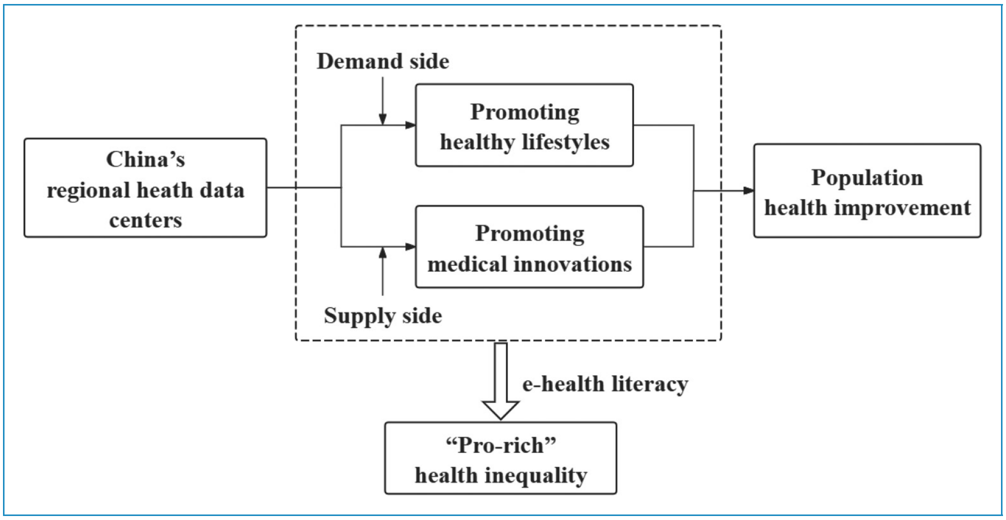
Demand side
Promoting healthy lifestyles
China's regional heath data centers
Promoting medical innovations
Supply side
Population health improvement
e-health literacy
"Pro-rich" health inequality

Figure 1. Research framework.

society, economy, demographics, education, and health to support academic research and inform public policy analysis.[37] As of 2020, six survey waves had been conducted comprising 2010, 2012, 2014, 2016, 2018, and 2020.[38] The survey spans 25 provinces, municipalities, and autonomous regions in China, representing approximately 95% of the total population.[39] The CFPS adopts proportional probability sampling with implicit stratification, multistage, multilevel, and population proportionality to enhance the validity and representativeness of its sample.[40] It is also supported by the Survey Research Center at the University of Michigan and other authoritative institutions responsible for survey design and methodology.[41] The target sample size is 16,000 households, with all household members in the selected households included as survey respondents. Prior to the launch of the CFPS national baseline survey in 2010, two pilot surveys were conducted in Beijing, Shanghai, and Guangdong in 2008 and 2009, respectively, to review the reliability and validity of the survey.[38] In 2008, while the target sample size was 2400 households, the resulting data were collected from 2375 households across 24 counties and districts. In 2009, a follow-up survey was conducted with the households sampled in 2008, with data collected from 1995 households.[42] The pilot survey sample constitutes 15% of the expected sample size of the formal survey, indicating good validity of the questionnaire. In 2010 (baseline survey), the response rates to the CFPS at the household and individual levels were 81.28% and 84.14%, respectively.[43]

The CFPS is a representative and authoritative source of microdata for academic research and public policy analysis[37] and has been used widely in academic research.[41,44,45] The CFPS is selected as the main database for microindividual analysis because of its high representativeness and authority, which provides a reliable data foundation for this study. The CFPS also encompasses multidimensional data across almost all age groups, including health, education, migration, and socioeconomic status,[46,47] thereby supporting the comprehensive analysis of various issues from multiple perspectives. It should, however, be noted that although the initial sample of the CFPS is broad in coverage, it is less representative in some remote and ethnic minority areas.[41] In addition, due to the long period of CFPS data collection, some indicators, such as exercise-related variables, may be investigated only in some survey waves. In our study, we only used four waves (2014, 2016, 2018, and 2020) of the CFPS due to the absence of key variables, such as "exercise frequency," in the 2010 and 2012 surveys.

We restricted the analysis to adult respondents aged 16 and above, as defined by CFPS. Missing values in the dependent and demographic variables were excluded. Finally, we obtained an unbalanced panel comprising 42,547 observations from 14,282 respondents.

Apart from that, patent-related data were sourced from the China National Intellectual Property Administration, as referenced in Howell's study.[48] Information on the number of medical enterprises and entrepreneurship came from QICHACHA, which provides detailed information such as address and establishment date of the enterprise.[49] Socioeconomic characteristics at the city level were obtained from China's Urban Statistical Yearbooks.

Measures

Dependent variable. This study used the "self-assessed health (SAH)" variable in the CFPS questionnaire as an indicator of individual health status. Despite the potential for subjective measurement errors, SAH can serve as a relatively accurate measure of health and is widely used in previous literature.[50,51] Self-assessed health reflects respondents' perception of their own health status, and offers a comprehensive measure of health status. It was selected from the question, "How do you assess your health?." Among the answers, 1 represents "excellent," 2 represents "very good," 3 represents "good," 4 represents "fair," and 5 represents "poor." To enhance the interpretability of health status, we inverted the health variable, assigning values 1 through 5 to denote "poor" to "excellent."

Additionally, the "health change" variable was used as an alternative dependent variable for robustness checks. Respondents were asked about the change in health status compared to one year ago. Answer 1 indicates that the health status had not changed or had improved, while answer 0 indicates that the health status had deteriorated.

Two main measures have been used in the literature to capture health inequality: single health inequality and socioeconomic-related health inequality. This study focused primarily on socioeconomic-related health inequality. The unequal distribution of wealth and power is a fundamental cause of health inequalities.[52,53] Attaining a higher social status not only ensures a more conducive and relaxed work environment but also alleviates economic constraints faced by patients during medical treatment, facilitating access to high-quality medical resources.

This study used the corrected concentration index (CCI)[54] to measure socioeconomic-related health inequalities, as the SAH is a bounded variable. We further employed the Wagstaff-type decomposition method to assess the contribution of pilot programs to health inequality.[55]

Independent variables. The core independent variable was "pilot policy," which represented the establishment of the pilot regional health data centers and was labeled as "DID." The process of implementation started on October 21, 2016, and continued until 2018. Given the lag in initiation of the data centers after implementation, this study designated 2017, 2018, and 2019 as date of intervention for the three sets of pilot data centers. Consequently, Fujian and Jiangsu provinces were designated as having operational data centers from 2017, Shandong, Anhui, and Guizhou provinces from 2018, and Ningxia Hui Autonomous Region from 2019. All other regions were assigned as control groups.

Mechanism variables. The first mechanism through which the pilot programs may improve health is by promoting healthy lifestyles. Healthy lifestyles can be reflected in both prolonged exercise duration and increased exercise frequency.[56,57] This was assessed through two variables: "exercise hours" and "exercise frequency," extracted from CFPS questionnaire queries: "How often did you exercise last week?" and "How many hours did you spend exercising last week?," respectively. The "exercise hours" indicates the time individuals allocate to physical activity, while "exercise frequency" denotes the regularity of physical activity. The classification for exercise frequency was as follows: (1) for less than once a week, (2) for once or twice a week, (3) for three to four times a week, (4) for five to seven times a week, and (5) for more than seven times per week. The term "exercise" encompassed various activities, including daily walking, running, jogging, mountain climbing, martial arts, indoor and outdoor physical exercise, ball games, and water sports. This comprehensive definition captured an individual's daily physical activity. Longer exercise time and more frequent exercise indicate a healthier lifestyle.[56]

The second mechanism through which the pilot programs may improve health is by enhancing innovations in medical practice. Drawing on the study of Fritsch and Wyrwich,[58] this study selected indicators to measure innovation in medical practice based on innovation output. Three indicators were employed: the number of medical patent applications applied for by hospitals (i.e., Medical patent applications variable), the number of medical patent inventors in hospitals (i.e., Medical patent inventors variable), and the number of medical enterprise entrepreneurship (i.e., Medical enterprises entrepreneurship variable).

Control variables. Following the Grossman model of the demand for medical care[59] and insights from prior research, this study categorized control variables into individual demographic characteristics and medical treatment status characteristics. The demographic characteristics included gender, age, age squared, retirement status, marital status, household registration status (i.e., hukou status), education, number of cigarettes smoked per day (i.e., smoking), and total income in the past year. Medical treatment status characteristics included whether the respondent had been hospitalized in the past year (i.e., hospitalized) and their out-of-pocket medical expenses in the past year. The total income and out-of-pocket medical expenses were log-transformed in 2020 RMB (1.00 USD = 6.9 RMB). All variables are defined in Table 1.

Empirical strategy

Step 1: The impacts of the pilot regional health data centers on health and its mechanisms of influence.

We employed a time-varying difference-in-differences (DID) model, as referenced by Zhou et al.,[60] to investigate the impacts of China's pilot regional health data centers on population health. The DID model is a widely used econometric method for evaluating the impact of nonrandomized interventions. This model assumes that, in the absence of treatment, the change in outcomes between the pre- and postintervention periods for the treatment group is similar to that of the control group. The implementation of the DID model involves two steps. First, in a natural experiment, a treatment group and a control group are selected: the treatment group consists of individuals affected by the intervention or policy, while the control group includes those unaffected. Second, the DID estimator (represented by the pilot policy variable in this study) is calculated.[61] This estimator captures the pure treatment effect by comparing how outcomes change before and after an intervention between individuals who have been exposed to the intervention (treated) and those who have not (untreated).[62] The DID model effectively controls for confounding factors, such as permanent differences between treatment and control groups, and time trends in outcomes unrelated to the intervention,[63] which enables us to estimate a relatively pure policy effect. Finally, the following regression model is estimated:

$$\text{Health}_{ict} = \alpha_0 + \alpha_1 \text{DID}_{ct} + \alpha_2 X_{ict} + \mu_c + \gamma_t + \varepsilon_{ict} \quad (1)$$

where i represents the individual, c represents the county, and t represents the survey year. Health_{ict} represents the SAH of individual i in county c at year t. DID_{ct} represents a policy dummy variable, indicating whether county c is a policy pilot county at year t. X_{ict} contains a series of control variables, including out-of-pocket medical expenses, whether hospitalized in the past year, total income, education, age, age squared, hukou status, gender, marital status, retirement status, and number of cigarettes smoked. μ_c represents county fixed effects, γ_t represents year fixed effects, and ε_{ict} represents random error term. The variable of interest is DID_{ct}. A statistically significant and positive DID_{ct} value suggests a positive impact of the pilot regional health data centers on population health, while a nonsignificant DID_{ct} value indicates that the policy has no effect on population health.

It is noteworthy that the Chinese government's selection process for the pilot regional health data centers may be subject to potential selection bias. The decision-making process may be guided by the evaluation of local capabilities and resources for the storage, management, and utilization of high-quality, high-capacity medical data. Therefore, the government may prioritize provinces with high levels of healthcare, abundant medical resources, and advanced

Table 1. Definition of variables.

Variable category	Variable	Definition
Dependent variable	Health	Self-accessed health (SAH) of respondents; 1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent.
	Health changes	Change in health status compared to a year ago; 1 = no change or better, 0 = worse.
Independent variable	Pilot policy (DID)	Whether the province was a pilot of China's regional heath data centers in that year; 1 = yes, 0 = no.
Mechanism variable	Exercise frequency	How often respondents exercise (frequency); 1 = less than once a week, 2 = once or twice a week, 3 = three to four times a week, 4 = five to seven times a week, 5 = more than 7 times a week.
	Exercise hours	Hours of exercise per week.
	Medical patent applications	Number of medical patent applications per 10,000 people in the province in that year, i.e., the number of patent applications in the province in the current year/the household registration population at the end of the year in the province (in 10,000 people).
	Medical patent inventors	Number of medical patent inventors per 10,000 people, i.e., the number of inventors in the province in the current year/the household registration population at the end of the year in the province (in 10,000 people); where medical patent inventors refer to a person who makes a creative contribution to the invention.
	Medical enterprises entrepreneurship	Number of start-ups of medical enterprises per 10,000 people in the province in that year, i.e., the number of start-ups in the province in the current year/the registered population at the end of the year in the province (in 10,000 people).
Control variables	Out-of-pocket medical expenses	Out-of-pocket expenses for medical expenses excluding the part that has been reimbursed or is expected to be reimbursed (10,000 yuan), in natural logarithm form.
	Total income	Total individual income from all jobs in the past year (10,000 yuan), including salary income, bonuses, cash benefits, and in-kind subsidies, in natural logarithm form. It is greater than or equal to 0.
	Hospitalized	Whether hospitalized last year ;1 = yes, 0 = no.
	Education	Level of education ;1 = illiterate/semi-illiterate, 2 = primary school, 3 = junior high school, 4 = high school/technical school/vocational high school, 5 = junior college, 6 = undergraduate, 7 = master, 8= Doctor.
	Age	The age of the respondents.
	Age squared	Respondent's age squared.
	Hukou status	Household registration status 1 = urban, 0 = rural.
	Gender	1 = male, 0 = female.
	Marital status	1 = married (with spouse), 0 = unmarried.

(continued)

Table 1. Continued.

Variable category	Variable	Definition
	Retirement status	Whether the respondent is retired ;1 = yes, 0 = no.
	Smoking	Number of cigarettes smoked per day.

developmental statuses, while excluding areas with limited medical resources. Given the potential selection bias in this empirical setting, a DID-based propensity score matching (PSM-DID) model was employed to manage potential selection bias.[64] Propensity score matching is a widely used technique that minimizes selection bias.[61] The purpose of the matching procedure is to identify a group of nonpilot provinces that exhibit characteristics similar to those of the pilot provinces, thereby minimizing the potential for bias.[65] We also incorporated county-fixed effects into our regression models to minimize the impact of regional characteristics on the selection of pilot programs. The county fixed effects controlled for any unobserved variations in the region, including the quality of care, the adequacy of medical resources, and access to healthcare.

Step 2: The mechanisms through which pilot programs influence health.

Besides examining the direct effect specified in equation (1), this study also delves into potential mechanisms of influence. Given the substantial endogenous bias inherent in traditional mediation effect tests, this study draws upon the study by Dell[66] to validate the causality channel in the mechanism test. In practice, we replaced the dependent variable in the regression with a healthy lifestyle (as captures by exercise) and medical innovation output:

$$\text{Exercise}_{ict} = \beta_0 + \beta_1 \text{DID}_{ct} + \beta_2 X_{ict} + \mu_c + \gamma_t + \varepsilon_{ict} \tag{2}$$

$$\text{MedInov}_{ct} = \gamma_0 + \gamma_1 \text{DID}_{ct} + \gamma_2 Z_{ict} + \mu_c + \gamma_t + \varepsilon_{ict} \tag{3}$$

The first potential mechanism of influence is through promoting individual's healthy lifestyles. Equation (2) assesses the influence of pilot programs on individual' lifestyle, with Exercise_{ict} denoting the healthy lifestyle of individual i in county c at year t. X_{ict} incorporates a set of control variables identical to those in equation (1).

The second mechanism of influence is through enhancing innovation in medical practices. Equation (3) is employed to examine the impact of pilot programs on local medical innovation, where MedInov_{ct} represents medical innovation in county c at year t. Consistent with prior literature, Z_{ict} encompasses a set of control variables influencing innovation output. The remaining variables in equations (2) and (3) are configured identically to those in equation (1).

Step 3: Estimate the impact of pilot regional health data centers on health inequality.

This study utilized the CCI[54] to measure socioeconomic-related health inequalities and employed the Wagstaff-type decomposition method to assess the contribution of pilot programs to health inequality.[67]

(1) *Corrected CI*

The CI, as introduced by Wagstaff et al.,[68] quantifies the degree to which inequalities in health-related variables are consistently associated with socioeconomic status. The formula used for the CI is:

$$\text{CI} = \frac{2}{n\mu}\sum_{i=1}^{n} \text{Health}_i R_i - 1 \tag{4}$$

where Health_i is the health status of individual i. μ is the mean of the health variable. R_i represents the i^{th} individual' relative rank in the income distribution.

Given that the health variable (i.e., SAH) in our study was bounded, the CCI was utilized to measure the health inequality. It is commonly used to measure socioeconomic inequality in bounded SAH.[69] The CCI is calculated as follows:

$$\text{CCI} = \frac{4\mu}{b-a}\text{CI} \tag{5}$$

where b and a represent the upper and lower bounds of the health variable, respectively. In this study, b equals to 5 and a equals to 1. The CCI ranges between −1 and 1. If CCI > 0, it indicates a "pro-rich" inequality. If CCI < 0, it indicates a "pro-poor" inequality. If CCI = 0, then there is no inequality. A larger absolute value of CCI indicates greater inequality.

(2) *Wagstaff decomposition*

In order to assess the impact of pilot programs on health inequality, a Wagstaff-type decomposition method[55] was employed as referenced by Gu et al.[69] The decomposition starts from the following equitation:

$$\frac{\text{Health}_i - a}{b-a} = \gamma + \sum_{j=1}^{q} \lambda_j x_{ji} + \delta_i \tag{6}$$

where λ_1, λ_2…, λ_q represent the coefficients for independent variables x_1, x_2…, x_q. δ_i represents the random error term. λ_j equals to $\frac{1}{b-a}\alpha_i$, where α_i is obtained using equation (1).

Substituting equation (4) and equation (5) into equation (6), we get:

$$\mathrm{CCI} = 4\left[\sum_{j=1}^{q} \lambda_j \overline{x_j}\mathrm{CI}_j + \frac{2}{n}\sum_{i=1}^{n} \delta_i R_i\right] \tag{7}$$

where x_j represents the determinants of Health$_i$, including the main explanatory variable DID$_{ct}$ and a series of control variables X_{ict} in equation (1), and $\overline{x_j}$ is the mean of x_j. CI$_j$ is the CI of determinants x_j. The contribution of determinants x_j to the total health inequality is $4\lambda_j\overline{x_j}\mathrm{CI}_j$, and the contribution rate is $\frac{4\lambda_j\overline{x_j}\mathrm{CI}_j}{\mathrm{EI}} \times 100\%$.

Results

Sample description

Table 2 provides an overview of the sample characteristics. Respondents, on average, reported a health status of 2.7, indicating generally good health. Approximately 62% of respondents perceived no change or improvement in their health compared to the previous year. 3.7% of observations were covered by the policy pilot. Nearly half (47.6%) of the respondents were male, and approximately 31% of the respondents held urban hukou. On average, respondents engaged in physical activity once or twice a week, accumulating seven hours of exercise per week. After logarithmic transformation, the mean total income was 4.440, with a median of 0. This is primarily because the sample includes individuals aged 16 and above, encompassing both teenagers attending school and retired elderly individuals, most of whom have an income of 0.

Figure 2 plots the average health status for the treatment and control groups across each survey year. As shown in Figure 2, the health status for both groups exhibited a downward trend from 2014 to 2016, with the control group consistently reporting better overall health than the treatment group. By 2016, the health status of the two groups was largely comparable. Following the announcement of the first batch of pilot regional health data centers in 2016, the health status improved significantly for both groups, with the treatment group benefiting more. By 2018, the treatment group reported better health than the control group. Consequently, Figure 2 suggests that the notable improvement in health among the treatment group may be linked to the pilot programs.

The health impact of the pilot programs

Results of the DID method. The baseline regression results are presented in Table 3. The first column showed the results from the univariate regression, followed by inclusion of all control variables in the second column. The third column provided the comprehensive results incorporating all control variables, county-fixed effects, and year-fixed effects. Regression results across the three columns showed minimal change, with the coefficient of the DID variable remaining significantly positive at the 1% statistical level. The results in the third column indicate that the implementation of the pilot programs has led to a slight improvement in the overall health status of the population in the pilot area by 0.166 units. Hypothesis 1 is confirmed.

The baseline regression passed the parallel trends test (see Supplementary Table A1), which verified noteworthy treatment effects on individuals' health conditions during and after policy implementation. Additionally, we replaced the dependent variable "health" with the "health changes" variable for additional robustness testing, as shown in column (4) of Table 3. The results remain unchanged, thus confirming hypothesis 1 once again.

For other control variables, higher total income, good education, and being married were associated with better health, while increased out-of-pocket medical expenses, hospitalization, and older age were linked to reduced health.

Robustness test and endogenous processing. There may be several endogeneity issues to be considered in this study. One pertains to the nonrandom selection of policy pilots. As we discussed in the third section, the inherent medical advantages of pilot provinces may not only elevate the chances of being chosen for the pilot but also impact the health status of local residents. To tackle this issue, this study employed the PSM-DID method, and the results closely corresponded with the baseline regression findings (see Supplementary Table A2 and Table A3). Another consideration involves the potential bias caused by omitted variables, where certain unobservable contemporaneous policies or other influencing factors could interfere with policy outcomes. A placebo test was conducted to exclude the interference of the above factors, demonstrating the inherent effectiveness of the policy (see Supplementary Figure A1). Consequently, our results remained consistent and robust after addressing endogenous issues, such as the potential selection bias, omitted variable bias. Hypothesis 1 is again confirmed.

Mechanisms analysis

Table 4 displays the outcomes related to healthy lifestyles as an influencing mechanism. Both the coefficients in the first and second columns were significantly positive, suggesting that the implementation of the pilot programs had notably increased both the hours and frequency of exercise per week. This increase is attributed to the promotion of various healthcare apps in the pilot programs and the emphasis on developing industries such as health management, health consultation, and health culture. These efforts

Table 2. Sample characteristics.

Variable category	Variable	Observations	Mean	Standard deviation	Minimum	Median	Maximum
Dependent variable	Health	42547	2.749	1.207	1.000	3.000	5.000
	Health changes	40,187	0.617	0.486	0.000	1.000	1.000
Independent variable	DID	42,547	0.037	0.190	0.000	0.000	1.000
Mechanism variable	Exercise frequency	42,547	2.189	1.457	1.000	1.000	5.000
	Exercise hours	17,105	6.962	8.827	0.017	4.700	105.000
	Medical patent applications	37,124	0.037	0.071	0.001	0.013	0.394
	Medical patent inventors	37,124	0.043	0.086	0.001	0.014	0.471
	Medical enterprises entrepreneurship	37,124	0.626	0.609	0.100	0.520	3.003
Control variables	Out-of-pocket medical expenses (log)	42,547	6.287	2.202	0.000	6.399	13.142
	Total income (log)	40,189	4.440	5.061	0.000	0.000	16.148
	Hospitalized	40,115	0.175	0.380	0.000	0.000	1.000
	Education	42,547	3.053	1.575	1.000	3.000	8.000
	Age	42,547	48.097	17.170	16.000	48.000	101.000
	Age squared	42,547	2608.156	1715.636	256.000	2304.000	10,201.000
	Hukou status	42,547	0.311	0.463	0.000	0.000	1.000
	Gender	42,547	0.476	0.499	0.000	0.000	1.000
	Marital status	42,547	0.786	0.410	0.000	1.000	1.000
	Retirement status	42,547	0.078	0.269	0.000	0.000	1.000
	Smoking	40,115	4.083	8.336	0.000	0.000	100.000

have fostered a positive atmosphere for health management and fitness in society, motivating people to invest more time in exercise and promoting healthy lifestyles. Healthy lifestyles further contribute to the improvement of individuals' health status. Hypothesis 2 is confirmed.

Table 5 shows the results related to the mechanism of enhancing innovations in medical practice. The notably positive coefficient in the first column suggests that the pilot programs had markedly enhanced the entrepreneurial activities of medical enterprises in society. Furthermore, the coefficients in the second and third columns, both significantly positive, signified a substantial increase in the number of patent applications and patent inventors in hospitals. This suggests that the pilot programs have led to the continuous emergence of professional new treatment instruments, new drugs, and new therapies. Consequently, doctors' intraoperative and postoperative medical diagnoses, as well as health monitoring, have become more precise and efficient, thereby enhancing medical capabilities and benefiting patients. Hypothesis 3 is confirmed.

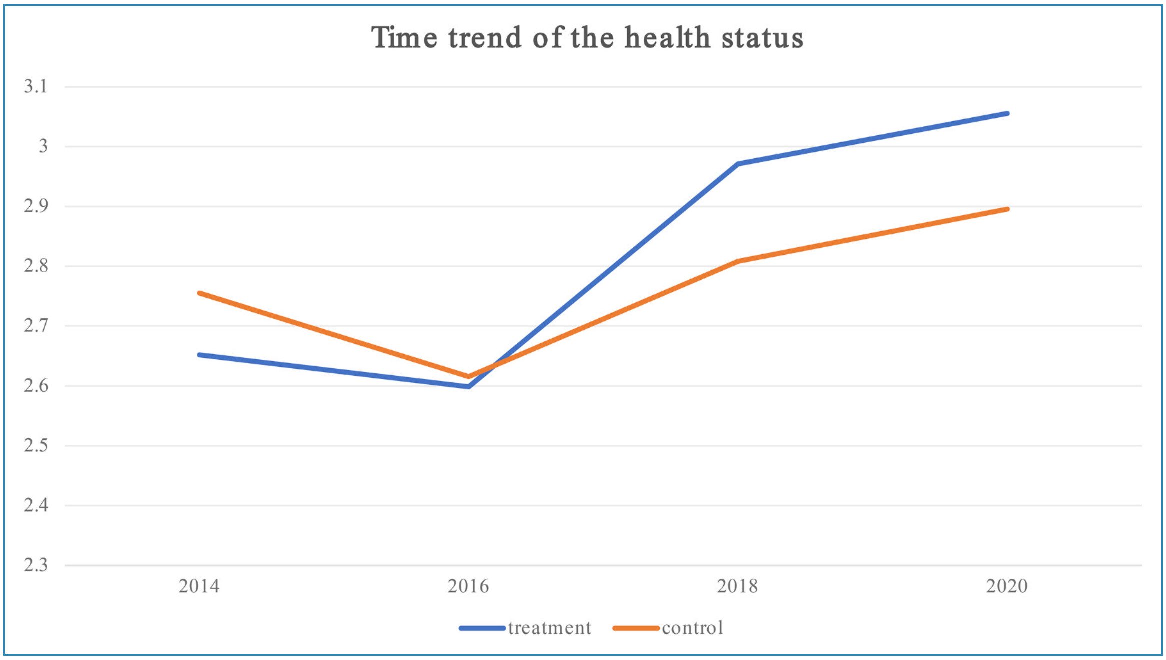

Figure 2. Time trend of health status in the treatment and control group.
Note: The figure presented the average health status of both the treatment and control groups across each survey year. The *y*-axis represented the average health status for both groups, while the *x*-axis displayed the survey years (2014, 2016, 2018, and 2020).

The impact of the pilot programs on socioeconomic-related health inequality

The CCI and Wagstaff-type decomposition method were used to investigate the impact of the pilot programs on socioeconomic-related health inequality. Table 6 revealed an overall CCI of 0.158, with a $p < 0.001$. A positive overall CCI suggested a prevalent "pro-rich" health inequality. The contribution rate of the pilot programs to the overall CCI was 1.7%, suggesting a notable positive contribution to the socioeconomic-related health inequality. Therefore, pilot regional health data centers directly enhanced the health of all residents, but potential variations in policy coverage resulted in a skew of medical resources toward high-income groups. Hypothesis 4 is confirmed.

We further validated the robustness of this result by stratifying the entire sample into "high-income" and "low-income" groups based on the mean of the income variable (see Supplementary Table A4). The regression coefficient was greater for the "high-income" group, indicating a greater impact of regional data centers on health among this demographic. This reaffirms that the pilot regional health data centers have widened the health gap between high- and low-income groups, exacerbating health inequality among these groups.

Discussion

This is the first study to explore the impact of pilot regional health data centers on population health and health inequality in China. We found that pilot programs promoted the overall health status of the residents in the pilot areas. China's "big data" in the health and medical sectors is part of the national big data strategic layout. The use of "big data" in healthcare spans various areas and presents cost-effective opportunities for global healthcare improvements.[70] For example, during the COVID-19 pandemic, "big data" in healthcare significantly aided Singapore, Taiwan, South Korea, and Hong Kong in enhancing traditional public health measures and curbing the spread of SARS-CoV-2.[71] Additionally, "big data" can be applied to public health promotion, healthcare management, drug and medical device monitoring, and routine clinical practice.[16] All these practices provide opportunities for improving people's health. Our findings are consistent with previous studies.[72,73] Dash et al.[72] highlighted the importance of "big data" in healthcare to improve people's health. Pastorino et al.[73] reported that "big data" in healthcare can benefit patients from several aspects such as increasing earlier diagnosis and the effectiveness and quality of treatments by the discovery of early signals and disease intervention, reduced probability of adverse reactions.

Establishing medical big data centers to support equitable discovery and innovation in digital healthcare is a significant global concern.[74] Health Data Research UK, for instance, is committed to "uniting the UK's health data to enable discoveries that improve people's lives."[74] Similarly, the Estonian eHealth project focuses on improving the quality and efficiency of health services and aims to digitize all patient information

Table 3. Effect of pilot program on health status.

	(1) Health	(2) Health	(3) Health	(4) Health changes
DID	0.268***	0.067**	0.166***	0.034***
	(8.706)	(2.451)	(4.054)	(2.605)
Out-of-pocket medical expenses		−0.135***	−0.134***	−0.137***
		(−46.950)	(−26.765)	(−25.147)
Total income		0.013***	0.015***	0.014***
		(9.486)	(9.345)	(8.512)
Hospitalized (yes vs no)		−0.224***	−0.221***	−0.221***
		(−14.481)	(−12.411)	(−11.703)
Education		0.024***	0.030***	0.031***
		(5.728)	(4.639)	(4.531)
Age		−0.063***	−0.060***	−0.060***
		(−30.432)	(−21.768)	(−20.693)
Age squared		0.000***	0.000***	0.000***
		(21.461)	(15.545)	(14.750)
Hukou status (urban vs rural)		0.001	0.025	0.029
		(0.079)	(1.127)	(1.225)
Gender (male vs female)		0.106***	0.106***	0.113***
		(8.649)	(7.340)	(7.327)
Marital status (married vs unmarried)		0.112***	0.095***	0.095***
		(7.586)	(5.033)	(4.597)
Retirement status (retired vs not retired)		−0.046**	−0.060**	−0.059**
		(−2.231)	(−2.436)	(−2.338)
Smoking		0.002**	0.002*	0.002*
		(2.459)	(1.833)	(1.666)
County FE	No	No	Yes	Yes
Year FE	No	No	Yes	Yes
N	42547	40,115	39,776	39,774

(continued)

Table 3. Continued.

	(1) Health	(2) Health	(3) Health	(4) Health changes
Adj.R^2	0.002	0.225	0.241	0.126

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. *T*-values were reported in parentheses, which were estimated using the robust standard errors of cluster heteroscedasticity at the county level. The first three columns displayed the results of the main regression, and the fourth column provided the result of the robustness test. Columns 3 and 4 controlled for county-fixed effects, year-fixed effects and control variables related to the individual demographic characteristics and medical treatment status characteristics.

Table 4. The potential mechanism of the effect of pilot programs on health−promoting healthy lifestyles.

	(1) Exercise hours	(2) Exercise frequency
DID	1.094**	0.099*
	(1.974)	(1.668)
Out-of-pocket medical expenses	0.027	0.014***
	(0.817)	(3.121)
Total income	−0.072***	−0.015***
	(−3.582)	(−7.059)
Hospitalized (yes vs no)	0.224	0.039*
	(1.085)	(1.803)
Education	−0.353***	0.104***
	(−5.651)	(12.461)
Age	0.106***	0.020***
	(3.515)	(5.255)
Age squared	−0.001**	−0.000
	(−2.142)	(−1.639)
Hukou status (urban vs rural)	−0.218	0.395***
	(−1.104)	(13.246)
Gender (male vs female)	0.636***	0.153***
	(3.928)	(8.381)
Marital status (married vs unmarried)	0.269	−0.095***
	(1.384)	(−4.420)

(continued)

Table 4. Continued.

	(1) Exercise hours	(2) Exercise frequency
Retirement status (retired vs not retired)	−0.409	0.025
	(−1.292)	(0.774)
Smoking	0.020	−0.010***
	(1.558)	(−10.212)
County FE	Yes	Yes
Year FE	Yes	Yes
N	16,815	39,776
Adj.R^2	0.119	0.142

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *T*-values were reported in parentheses, which were estimated using the robust standard errors of cluster heteroscedasticity at the county level. All regressions controlled for county-fixed effects, year-fixed effects and control variables related to the individual demographic characteristics and medical treatment status characteristics.

Table 5. The potential mechanism of the effect of pilot programs on health−cultivating medical innovation output.

	(1) Medical enterprises entrepreneurship	(2) Medical patent applications	(3) Medical patent inventors
DID	0.096**	0.025**	0.027**
	(2.133)	(2.289)	(2.172)
Regional GDP	−0.017	0.089**	0.086*
	(−0.111)	(2.530)	(1.923)
Added value of the tertiary industry	0.093	−0.075*	−0.074
	(0.633)	(−1.907)	(−1.491)
Local fiscal healthcare expenditure	0.210*	0.040*	0.059**
	(1.694)	(1.818)	(2.325)
Number of domestic patent applications accepted	0.010***	0.001***	0.001***
	(9.970)	(5.998)	(6.524)
Urban proportion	0.318**	−0.124***	−0.152***
	(2.185)	(−5.612)	(−5.439)

(continued)

Table 5. Continued.

	(1)	(2)	(3)
	Medical enterprises entrepreneurship	Medical patent applications	Medical patent inventors
Number of certified physician assistant in urban areas	0.011***	0.003***	0.003***
	(4.199)	(4.301)	(4.307)
Number of certified physician assistant in rural areas	−0.006	−0.002	−0.002
	(−1.120)	(−1.175)	(−1.105)
Number of general hospital	−0.000	−0.000	−0.000*
	(−0.764)	(−1.603)	(−1.737)
Number of traditional Chinese medicine hospital	−0.001*	−0.001***	−0.001***
	(−1.818)	(−3.791)	(−3.709)
Number of specialized hospitals	−0.001***	−0.000	−0.000
	(−4.375)	(−1.040)	(−1.247)
County FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	36,532	36,532	36,532
Adj. R2	0.975	0.916	0.913

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *T*-values were reported in parentheses, which were estimated using the robust standard errors of cluster heteroscedasticity at the county level. All regressions controlled for county-fixed effects, year-fixed effects. The selection of control variables was based on existing literature, including Regional GDP, added value of the tertiary industry, local fiscal healthcare expenditure, number of domestic patent applications accepted, urban proportion, number of certified physician assistant in urban areas, number of certified physician assistant in rural areas, number of general hospitals, number of traditional Chinese medicine hospitals, number of specialized hospitals. The first three variables were expressed in logarithmic form.

and prescriptions.[75] Additionally, the European Health Data Space promotes the sharing of health data between EU countries to support research and public health surveillance. Our study has revealed that the regional health data centers in China are associated with improvements in population health. It also offers valuable insights for the development of “big data” in healthcare in other countries.

We also found that the improvements to population health were driven by pilot programs promoting healthy lifestyles. “Big data” in healthcare advocates for mobile health and wellness services to implement novel and innovative ways to deliver care and coordinate health as well as wellness. Companies such as Apple and Google have developed specialized platforms, such as Apple’s Research Kit and Google Fit, for developing research applications for fitness and health statistics.[72] These applications can improve healthcare by accelerating interactive communication between patients and healthcare providers.[72] The applications can also help users and patients improve health awareness, health planning and encourage healthy lifestyles. The users or patients can become advocates for their own health. Our findings are similar to a study in Europe, which identified 10 big data priority projects to be implemented in Europe to support the sustainability of the health system by supporting healthy lifestyles.[76]

We also found that regional data center pilots can help drive innovations in medical practices, and thus improve population health. “Big data” in healthcare can promote new medical technology that can improve the cure rate of the disease.[77] Previous studies have demonstrated that the

Table 6. Decomposition of health inequality.

Variable	Coefficient (λ_j)	CI	Contributions to CCI	Contribution rate
DID	0.042	0.410	0.003	0.017
Out-of-pocket medical expenses	−0.034	0.027	0.023	0.154
Total income	0.004	0.552	0.036	0.247
Hospitalized (yes vs no)	−0.055	−0.159	0.006	0.042
Education	0.008	0.131	0.012	0.081
Age	−0.015	−0.084	0.242	1.634
Age squared	0.000	−0.156	−0.172	−1.161
Hukou status (urban vs rural)	0.006	0.026	0.000	0.001
Gender (male vs female)	0.027	0.111	0.006	0.038
Marital status (married vs unmarried)	0.024	−0.015	−0.001	−0.008
Retirement status (retired vs not retired)	−0.015	−0.301	0.001	0.010
Smoking	0.001	0.101	0.001	0.004
Total CCI	0.158			

Notes: The results showed a Wagstaff-type decomposition of health inequality, with health inequality measured using the corrected concentration index (CCI). The total CCI represented the overall health inequities within the entire sample. The contribution rate indicates the extent to which different factors contribute to overall health inequalities. A higher contribution rate reflected a greater impact of a given factor on health inequality.

application of "big data" in healthcare based on digital technology can be used to reduce medical errors,[25] support clinical decision-making,[26] and achieve real-time health evaluations and monitor remote patients.[27] Innovative technology based on big data supports precision medicine.[73]

We also found "pro-rich" health inequalities brought about by the pilot programs. One explanation for this finding was that high-income groups may benefit more extensively from the establishment of the pilot programs due to their higher e-health literacy. Health inequalities arising from medical big data constitute a significant global concern. Cruz's analysis of the American case demonstrates that to fully address health inequalities, both technical and social factors must be considered with medical big data.[78] A previous study found that individuals with good digital skills and high subjective acceptance can benefit more from new medical methods.[34] Effectively addressing health inequalities requires recognizing and confronting the disparities between groups influenced by factors such as socioeconomic status. Among these, the difference in digital literacy among different groups is an important factor that prevents different groups from enjoying digital welfare equitably.[79–81] Digital literacy must be integrated into digital health policies; otherwise, digital health policies will exacerbate health inequalities.[82]

To ensure that China's pilot regional health data centers yield more equitable health outcomes, the Chinese government should first implement a national e-health literacy promotion program designed to enhance the population's ability to use health information services, particularly among low-income and less educated groups. For example, targeted education and training programs can be developed to help these groups overcome barriers they may encounter when using health information services, such as unfamiliarity with technology, language barriers, or difficulty understanding health information. To improve the accessibility and ease of use of the platforms, digital health platforms with intuitive interfaces should be developed to lower the barriers to use for all users, especially those who are not familiar with technology. Additionally, improvements should be made to advance equity in digital access to medical services. The government should accelerate the digital transformation of medical infrastructure, facilitate seamless access to the national health information platform for primary healthcare units, and enhance the overall digital capacity of health services.

This study has some limitations. First, we only used four waves of CFPS data, so the study panel duration is shorter. Future research could use more years of panel data to explore the long-term impact of the pilot regional health data centers on people's health. Second, in addition to the two mechanisms we proposed, other possible impact mechanisms of the pilot regional health data centers on people's health level can be further explored. For example, improving the efficiency of medical matching based on data and promoting the sharing of medical resources may also significantly improve people's health. Third, the utilization of health big data faces several challenges, including heightened complexity, challenges in acquisition, large volume and variety, and privacy concerns.[19] It can pose a significant challenge to existing stakeholders within the health system, potentially disrupting established authority and power structures. Moreover, when confronted with an abundance of medical data, health big data harbor the potential to generate new discoveries and information. However, it can also deviate from established practices and understanding, resulting in heightened uncertainty until the information gains credibility. While further exploration of these aspects bears notable importance, it unfortunately lies beyond the purview of this paper's study model. All of these aspects are worth studying in the future.

Conclusion

This is the first study to explore the impact of China's pilot regional health data centers on population health and health inequality. This study reveals that while the establishment of healthcare big data centers was associated with an improvement in population health, it was also associated with an exacerbation in socioeconomic-related health inequality. China's regional health data centers can support healthy lifestyles and promote medical innovation, thereby improving health.

This study has important policy implications for the development of pilot regional health data centers. The "big data" in healthcare helps to monitor the trends of major diseases and provide evidence for policy-making in healthcare. By leveraging opportunities for "big data" development in healthcare, the Chinese government can strengthen the popularization of health care knowledge and investment in medical innovation, make full use of information technology to strengthen health care work and improve medical quality. While maintaining the strategic positioning of "big data" in health and medical care, residents' e-health literacy will be included in the scope of policy attention, and attention will be paid to the health inequality brought about by the policy. The government should implement community-based digital health literacy programs targeting vulnerable populations, such as the elderly, low-income groups, and rural residents, to enhance digital health skills. Efforts should also prioritize the development of user-friendly digital health platforms featuring intuitive interfaces and multilingual support to cater to diverse populations. Furthermore, it is essential to offer subsidies and incentives for digital health tools, which will facilitate the acquisition of necessary devices and lower the barriers to usage, thereby promoting digital health equity.

Future studies could further explore the dynamic nature of health inequalities arising from medical big data, identifying and assessing the extent and changing trends of health inequality across different populations and regions. Additionally, ethical issues associated with the use of medical big data—including personal privacy protection, data security, informed consent, and the impact on existing health system stakeholders—warrant further investigation.

Contributorship: J.C. contributed to the design of the study. Y.L. contributed to statistical analysis. J.C. and Y.L. wrote the first draft. P.C.C. critically revised the paper for important intellectual content. All authors approved the final version.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: This work was supported by National Natural Science Foundation of China (grant number 72374021).

Guarantor: JLC.

ORCID iDs: Jiaoli Cai https://orcid.org/0000-0002-8515-2577
Yue Li https://orcid.org/0009-0004-9365-9070

Supplemental material: Supplemental material for this article is available online.

References

1. World Health Organization. *World health statistics 2023: monitoring health for the SDGs, sustainable development goals*, https://www.who.int/publications-detail-redirect/9789240074323 (2023, accessed 2 August 2023).
2. Elrick H. A new definition of health. *J Natl Med Assoc* 1980; 72: 695–699.
3. World Health Organization. *Constitution of the World Health Organization*. https://www.who.int/about/governance/constitution (2006, accessed 2 August 2023).
4. Hu Y, van Lenthe FJ, Borsboom GJ, et al. Trends in socioeconomic inequalities in self-assessed health in 17 European countries between 1990 and 2010. *J Epidemiol Community Health* 2016; 70: 644–652.
5. Kachi Y, Inoue M, Nishikitani M, et al. Determinants of changes in income-related health inequalities among working-age adults in Japan, 1986–2007: time-trend study. *Soc Sci Med* 2013; 81: 94–101.
6. Mackenbach JP, Kulhánová I, Artnik B, et al. Changes in mortality inequalities over two decades: register based study of European countries. *Br Med J* 2016; 353: i1732.

7. Le DD, Leon-Gonzalez R, Giang TL, et al. Socio-economic-related health inequality in non-communicable diseases among older people in Viet Nam. *Ageing Soc* 2021; 41: 1421–1448.
8. Zhang C-Q, Chung P-K, Zhang R, et al. Socioeconomic inequalities in older adults' health: The roles of neighborhood and individual-level psychosocial and behavioral resources. *Front Public Health* 2019; 7. https://www.frontiersin.org/articles/10.3389/fpubh.2019.00318 (accessed 3 August 2023).
9. Devkota S and Upadhyay M. How do income and education affect health inequality: evidence from four developing countries. *Appl Econ* 2015; 47: 5583–5599.
10. Hurley J, Mentzakis E and Walli-Attaei M. Inequality aversion in income, health, and income-related health. *J Health Econ* 2020; 70: 102276.
11. Jiang J, Huang W, Liu Y, et al. The temporal and spatial changes of health inequality in Rural China. *Front Public Health* 2022; 10. https://www.frontiersin.org/articles/10.3389/fpubh.2022.821384 (accessed 14 October 2023).
12. Mackenbach JP, Valverde JR, Artnik B, et al. Trends in health inequalities in 27 European countries. *Proc Natl Acad Sci USA* 2018; 115: 6440–6445.
13. Cai J, Coyte PC and Zhao H. Decomposing the causes of socioeconomic-related health inequality among urban and rural populations in China: a new decomposition approach. *Int J Equity Health* 2017; 16: 128.
14. The Central Committee of the Communist Party of China and the State Council. *Healthy China 2030*. 2016.
15. The State Council. The People's Republic of China. *China to boost big data application in health and medical sectors*. http://english.www.gov.cn/policies/latest_releases/2016/06/24/content_281475379018156.htm (2016, accessed 3 August 2023).
16. Zhang L, Wang H, Li Q, et al. Big data and medical research in China. *Br Med J* 2018; 360: j5910.
17. Lee CH and Yoon H-J. Medical big data: promise and challenges. *Kidney Res Clin Pract* 2017; 36: 3–11.
18. Rumsfeld JS, Joynt KE and Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol* 2016; 13: 350–359.
19. Azzaoui AE, Sharma PK and Park JH. Blockchain-based delegated Quantum Cloud architecture for medical big data security. *J Netw Comput Appl* 2022; 198: 103304.
20. Benkner S, Arbona A, Berti G, et al. @neurIST: infrastructure for advanced disease management through integration of heterogeneous data, computing, and complex processing services. *IEEE Trans Inf Technol Biomed* 2010; 14: 1365–1377.
21. Li X, Krumholz HM, Yip W, et al. Quality of primary health care in China: challenges and recommendations. *Lancet* 2020; 395: 1802–1812.
22. Shilo S, Rossman H and Segal E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat Med* 2020; 26: 29–38.
23. Han A, Isaacson A and Muennig P. The promise of big data for precision population health management in the US. *Public Health* 2020; 185: 110–116.
24. Zhang H, Hussin H, Hoh C-C, et al. Big data in breast cancer: towards precision treatment. *Digit Health* 2024; 10: 20552076241293695.
25. Lin Y-K, Chen H, Brown RA, et al. Healthcare predictive analytics for risk profiling in chronic care: a Bayesian multitask learning approach. *MIS Q* 2017; 41: 473–496.
26. Rush B, Celi LA and Stone DJ. Applying machine learning to continuously monitored physiological data. *J Clin Monit Comput* 2019; 33: 887–893.
27. Yu C-S, Lin Y-J, Lin C-H, et al. Development of an online health care assessment for preventive medicine: a machine learning approach. *J Med Internet Res* 2020; 22: e18585.
28. Gu D, Su K and Zhao H. A case-based ensemble learning system for explainable breast cancer recurrence prediction. *Artif Intell Med* 2020; 107: 101858.
29. Pramanik MI, Lau RYK, Demirkan H, et al. Smart health: big data enabled health paradigm within smart cities. *Expert Syst Appl* 2017; 87: 370–383.
30. Pan J, Ding S, Wu D, et al. Exploring behavioural intentions toward smart healthcare services among medical practitioners: a technology transfer perspective. *Int J Prod Res* 2019; 57: 5801–5820.
31. Wang Y, Liu Y, Shi Y, et al. User perceptions of virtual hospital apps in China: systematic search. *JMIR Mhealth Uhealth* 2020; 8: e19487.
32. Wong DK-K and Cheung M-K. Online health information seeking and eHealth literacy among patients attending a primary care clinic in Hong Kong: A cross-sectional survey. *J Med Internet Res* 2019; 21; e10831.
33. Battineni G, Baldoni S, Chintalapudi N, et al. Factors affecting the quality and reliability of online health information. *Digit Health* 2020; 6: 2055207620948996.
34. Powell J and Deetjen U. Characterizing the digital health citizen: mixed-methods study deriving a new typology. *J Med Internet Res* 2019; 21: e11279.
35. Kontos E, Blake KD, Chou W-YS, et al. Predictors of eHealth usage: insights on the digital divide from the health information National Trends Survey 2012. *J Med Internet Res* 2014; 16: e3117.
36. Wang H, Liang L, Du C, et al. Implementation of online hospitals and factors influencing the adoption of Mobile medical services in China: cross-sectional survey study. *JMIR Mhealth Uhealth* 2021; 9: e25960.
37. Xie Y and Hu J. An Introduction to the China Family Panel Studies (CFPS). *Chin Sociol Rev* 2014; 47: 3–29.
38. Wang X and Wang Y. Association between digital engagement and urban-rural disparities in Chinese women's depressive symptoms: A national-level cross-sectional study. https://doi.org/10.1177/20552076241239246. Epub ahead of print 4 April 2024.
39. Li C, Khan MM and Chen Z. Public trust of physicians in China improved since the COVID-19 pandemic Began. *Soc Sci Med* 2023; 320: 115704.
40. Xie M, Huang Z and Zang W. The inequality of health-income effect in employed workers in China: a longitudinal study from China Family Panel Studies. *Int J Equity Health* 2020; 19: 96.
41. Zhang Y-J, Jin Y-L and Zhu T-T. The health effects of individual characteristics and environmental factors in China: evidence from the hierarchical linear model. *J Cleaner Prod* 2018; 194: 554–563.
42. Xie Y, Zhang X, Tu P, et al. *China family panel studies user's manual*. Beijing, China: Peking University, 2017.
43. Chen H and Meng T. Bonding, bridging, and linking social capital and self-rated health among Chinese adults: use of the anchoring vignettes technique. *PLoS ONE* 2015; 10: e0142300.
44. Chen Y, Fan Z, Gu X, et al. Arrival of young talent: the send-down movement and rural education in China. *Am Econ Rev* 2020; 110: 3393–3430.

45. Huang W and Zhang C. The power of social pensions: evidence from China's new rural pension scheme. *Am Econ J: Appl Econ* 2021; 13: 179–205.
46. Wang H, Liu H and Fu Q. Effects of common prosperity on China's education expenditure—empirical analysis based on CFPS quasi-micro panel data. *Int Rev Econ Finance* 2024; 91: 440–455.
47. Wang W and Zhang S. The impact of internet use on rural women's off-farm work participation: empirical evidence from China. *Sustainability* 2022; 14: 16972.
48. Howell A. Agglomeration, absorptive capacity and knowledge governance: implications for public–private firm innovation in China. *Reg Stud* 2020; 54: 1069–1083.
49. Chen Z, Dong B, Pei Q, et al. The impacts of urban vitality and urban density on innovation: evidence from China's Greater Bay Area. *Habitat Int* 2022; 119: 102490.
50. Alexander D and Schnell M. Just what the nurse practitioner ordered: independent prescriptive authority and population mental health. *J Health Econ* 2019; 66: 145–162.
51. Horn BP, Maclean JC and Strain MR. Do minimum wage increases influence worker health? *Econ Inq* 2017; 55: 1986–2007.
52. Mackenzie M, Skivington K and Fergie G. "The state They're in": unpicking fantasy paradigms of health improvement interventions as tools for addressing health inequalities. *Soc Sci Med* 2020; 256: 113047.
53. McCartney G, Dickie E, Escobar O, et al. Health inequalities, fundamental causes and power: towards the practice of good theory. *Sociol Health Illn* 2021; 43: 20–39.
54. Erreygers G. Correcting the concentration index. *J Health Econ* 2009; 28: 504–515.
55. Wagstaff A, van Doorslaer E and Watanabe N. On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam. *J Econom* 2003; 112: 207–223.
56. Stonerock GL and Blumenthal JA. Role of counseling to promote adherence in healthy lifestyle medicine: strategies to improve exercise adherence and enhance physical activity. *Prog Cardiovasc Dis* 2017; 59: 455–462.
57. Xiao Y, Romanelli M and Lindsey MA. A latent class analysis of health lifestyles and suicidal behaviors among US adolescents. *J Affect Disord* 2019; 255: 116–126.
58. Fritsch M and Wyrwich M. Is innovation (increasingly) concentrated in large cities? An international comparison. *Res Policy* 2021; 50: 104237.
59. Grossman M. On the concept of health capital and the demand for health. *J Polit Econ* 1972; 80: 223–255.
60. Zhou Z, Zhao Y, Shen C, et al. Evaluating the effect of hierarchical medical system on health seeking behavior: A difference-in-differences analysis in China. *Soc Sci Med* 2021; 268: 113372.
61. Hsu Y-T, Chiu Y-L, Wang J-N, et al. Impacts of physician promotion on the online healthcare community: using a difference-in-difference approach. *Digit Health* 2022; 8: 20552076221106319.
62. Karan A, Yip W and Mahal A. Extending health insurance to the poor in India: an impact evaluation of Rashtriya Swasthya Bima Yojana on out of pocket spending for healthcare. *Soc Sci Med* 2017; 181: 83–92.
63. Godard-Sebillotte C, Karunananthan S and Vedel I. Difference-in-differences analysis and the propensity score to estimate the impact of non-randomized primary care interventions. *Fam Pract* 2019; 36: 247–251.
64. Dehejia R and Wahba S. Propensity score-matching methods for nonexperimental causal studies. *Rev Econ Stat* 2002; 84: 151–161.
65. Xiao D, Yu F and Guo C. The impact of China's pilot carbon ETS on the labor income share: based on an empirical method of combining PSM with staggered DID. *Energy Economics* 2023; 124: 106770.
66. Dell M. The persistent effects of Peru's Mining *Mita*. *Econometrica* 2010; 78: 1863–1903.
67. Wagsta A, van Doorslaer E and Watanabe N. On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam. *J Econom* 2003; 112: 207–223.
68. Wagstaff A, Paci P and Van Doorslaer E. On the measurement of inequalities in health. *Soc Sci Med* 1991; 33: 545–557.
69. Gu H, Kou Y, You H, et al. Measurement and decomposition of income-related inequality in self-rated health among the elderly in China. *Int J Equity Health* 2019; 18: 4.
70. Hilbert M. Big data for development: a review of promises and challenges. *Dev Policy Rev* 2016; 34: 135–174.
71. Nageshwaran G, Harris RC and Guerche-Seblain CE. Review of the role of big data and digital technologies in controlling COVID-19 in Asia: Public health interest vs. privacy. *Digit Health*. Epub ahead of print 23 March 2021. DOI: 10.1177/20552076211002953
72. Dash S, Shakyawar SK, Sharma M, et al. Big data in healthcare: management, analysis and future prospects. *J Big Data* 2019; 6: 1–25.
73. Pastorino R, De Vito C, Migliara G, et al. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. *Eur J Public Health* 2019; 29: 23–27.
74. Ibrahim H, Liu X, Zariffa N, et al. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health* 2021; 3: e260–e265.
75. Pastorino R, De Vito C, Migliara G, et al. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. *Eur J Public Health* 2019; 29: 23–27.
76. European Commission. *Study on Big Data in public health, telemedine and healthcare*. https://ec.europa.eu/health/sites/health/files/ehealth/ docs/bigdata_report_en.pdf%0A%0A (2016).
77. Luo J, Wu M, Gopukumar D, et al. Big data application in biomedical research and health care: a literature review. *Biomed Inform Insights* 2016; 8: 1–10.
78. Cruz TM. Perils of data-driven equity: safety-net care and big data's elusive grasp on health inequality. *Big Data Soc* 2020; 7: 2053951720928097.
79. Agree EM, King AC, Castro CM, et al. "It's got to be on this page": age and cognitive style in a study of online health information seeking. *J Med Internet Res* 2015; 17: e79.
80. Crawford A and Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. *J Med Internet Res* 2020; 22: e19361.
81. Spooner KK, Salemi JL, Salihu HM, et al. Ehealth patient-provider communication in the United States: interest, inequalities, and predictors. *J Am Med Inform Assoc* 2017; 24: e18–e27.
82. Rich E, Miah A and Lewis S. Is digital health care more equitable? The framing of health inequalities within England's digital health policy 2010–2017. *Sociol Health Illn* 2019; 41: 31–49.