BMC Public Health

RESEARCH Open Access

Exploring factors influencing farmers' health self-assessment in China based on the LASSO method

Mingze Wu[1] and Shulin Zeng[2*]

Abstract

As the main force and practice subject of rural revitalisation, farmers' health is intricately linked to agricultural production and the rural economy. This study utilizes open data from the 2015 China Nutrition and Health Survey and employs the Least Absolute Shrinkage and Selection Operator (LASSO) method to explore the factors influencing farmers' self-assessment of health. The findings reveal that education level, proactive nutrition knowledge seeking, healthy dietary preferences and habits, and the use of clean cooking fuel positively impact farmers' health self-assessment. Conversely, age, history of illness or injury, and participation in medical insurance negatively affect their self-assessment. Furthermore, factors influencing farmers' health self-assessment exhibit heterogeneity across regions. Our findings suggest that promoting health education, disseminating nutritional dietary knowledge, and enhancing rural household infrastructure play an important role in improving farmers' self-evaluation of health. Therefore, policymakers should design more targeted health interventions and infrastructure improvement plans based on farmers' self-assessment of health and the level of regional economic development.

Keywords Farmers' health, Self-assessment of health, Dietary knowledge, LASSO method

Introduction

Health is recognized as a fundamental component of human capital, playing a crucial role in both individual comprehensive development and national advancement, with profound implications for workforce participation and labour productivity [1–3]. Naturally, the health of individuals is closely intertwined with regional development levels. According to the latest statistics from the World Bank, as of the end of 2022, the global population has surpassed 8 billion people, with approximately 4.3 billion residing in rural areas, predominantly concentrated in impoverished regions (source: World Bank, 2023). Urbanisation rates reached 56.16% worldwide in 2020, with developed countries like the United States and Europe exceeding 80%, largely completing their urbanisation process [4]. China, as the largest developing and populous country in the world, has undergone significant urbanisation progress, marked by a reduction in permanent rural residents from 790 million in 1978 to 490 million in 2022 (source: National Bureau of Statistics, 2023). However, farmers still comprise a substantial proportion (34.8%) [5]. While health issues among urban residents and the elderly have garnered widespread scholarly attention [6–8], a notable disparity persists between rural and urban areas, encompassing infrastructure, healthcare, environmental pollution, food safety, and economic conditions. Consequently, the health of farmers, as crucial component of the primary industry's human capital,

*Correspondence:
Shulin Zeng
royqingchen@gmail.com
[1] College of Economics and Management, South China Agricultural University, Guangzhou 510642, China
[2] Qidong Hospital of Traditional Chinese Medicine, Nantong 226200, Jiangsu, China

warrants increased consideration. It is of great practical significance and far-reaching social influence to study the factors influencing farmers' health.

Various disciplines emphasize different factors influencing individual health. In sociological studies, many scholars have verified the impact of factors such as education level, age, income, working environment and living conditions on people's health [9, 10]. However, economists attribute health heterogeneity to regional economic disparities, social development, institutional culture, and social welfare [11–13]. From the perspective of health medicine, factors affecting people's health include biological heredity, lifestyle, dietary preferences, interpersonal communication and emotions [14–17]. Some scholars even believe that direct behavioural and biomedical interventions can improve the health of residents in general [18]. In addition, physical education studies pay more attention to incorporating factors affecting physical health into the analysis framework from the perspective of ecological models of health behaviour [19]. A review of existing research shows that most of the literature focuses on the health status of urban residents, and health improvement among elderly and vulnerable groups.

Several crucial questions underlie our analysis. What is the health status of Chinese farmers amidst the backdrop of rapid urbanisation? Have their subjective health ratings improved compared to economically disadvantaged years? What factors significantly influence farmers' self-assessment of health? Are there important regional differences in those key influences? How can these influential factors be systematically identified through scientific means? Furthermore, what targeted recommendations should be proposed to policymakers to enhance the health status of farmers? To what extent can overall improvements in the health status of farmers, regarded as vital human capital, narrow the gap between urban and rural areas? However, we know very little about these concerns at this time, a more in-depth exploration is necessary to understand these issues.

Effective measurement and identification of individual health conditions are vital for policymakers to develop targeted interventions and implement policies to improve overall well-being. The literature has carried out much research on individual physiological health and mental health from both objective and subjective dimensions. Objective indicators, such as blood pressure, blood sugar, blood lipids, and BMI, provide insights into respondents' physical health [20]. Quality of life questionnaires such as the SF-36 and EQ-5D offer a comprehensive evaluation of the elderly people's and patients' physical health, mental health, emotion, and social participation [21, 22]. The Depressive Symptom Assessment Scale and the Anxiety Symptom Assessment Scale are widely used to measure an individual's mental health [23, 24]. Moreover, health self-assessment, a subjective measure of individuals' health perception, plays a crucial role in understanding residents' health and well-being. In this method, the simple Likert scale scoring method allows individuals to independently rate their health based on subjective feelings, and thanks to its simplicity, it has been widely adopted by scholars [25, 26]. Leveraging available data, this study employs respondent-reported health self-assessment as the dependent variable.

This study aims to accurately identify crucial factors related to farmers' health through the application of LASSO regression, a supervised machine learning method. By analysing multiple variables encompassing dimensions of farmers' individual characteristics, health literacy, dietary preferences, and cooking fuel choice, we aim to attain a comprehensive understanding of the diverse factors influencing health among individuals. The findings of this study provide valuable guidance and decision-making support for policymakers and health professionals. This includes tailored health interventions for different groups, optimized allocation of medical resources, implementation of preventive and early care measures, and enhancement of health services. This paper contributes to the literature by introducing machine learning algorithms, in contrast to traditional methods like linear regression and Poisson regression [26]. This enriches the research methodology, reduces redundant variables, and enhances the model's explanatory and generalisation capabilities.

The remainder of this paper is organized as follows. "Research methodology" section presents the research methods. "Data description" section introduces the data in this paper, including index construction, sample processing, and descriptive statistics of variables. The empirical results are presented and discussed in "Results and discussion" section. The last section comprises the conclusion and implications.

Research methodology

Logistic models are widely used for binary discrete dependent variable problems. When all the variables in the study are included in the regression model to be fitted, it is prone to the risk of multicollinearity and overfitting. To screen the factors affecting farmers' health self-assessment more scientifically and objectively, the present study was conducted with the help of LASSO method for variable screening. The LASSO algorithm implements feature selection and model parameter contraction by adding an L1 regularisation term to the loss function. This method, proposed by Tibshirani in 1996, is essentially traditional least squares estimation with penalty factors. Using the absolute value function of the model coefficients as a penalty effectively reduces the weights of the unimportant features and compresses their coefficients to zero, which leads to variable selection and parameter estimation, and results in a more refined model [27]. In contrast to conventional variable selection approaches like stepwise regression, ridge regression, principal component regression, partial least squares regression, and others, the LASSO method stands out for its robust utility in data analysis and feature selection. This method boasts distinct advantages, including heightened predictive accuracy, enhanced model interpretability, and computational simplicity [28].

LASSO is a method of logistic modelling using the LASSO method of selecting independent variables to rule out the omission of variables due to preconceived notions, as well as by scaling down the bias estimates, which in turn removes linear relationships between variables [29]. Assuming that there are independently and identically distributed observations (X^i, y_i), $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, the conditional probability expression for the Logistic model is as follows:

$$ln\frac{P(y_i = 1|X^i)}{1 - P(y_i = 1|X^i)} = \eta\beta\left(X^i\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \quad (1)$$

where X^i and y_i are the independent and dependent variables of the model, respectively. P denotes the probability of the model and β is the coefficient.

The coefficient estimates $\widehat{\beta}$ in the LASSO model can be written as Eq. (2):

$$\widehat{\beta} = argmin\sum_{i=1}^{n}\left\{y_i\eta\beta\left(X^i\right) - ln\left\{1 + exp\left[\eta\beta\left(X^i\right)\right]\right\}\right\} + \lambda\sum_{j=1}^{p}|\beta_j| \quad (2)$$

where λ is a non-negative tuning parameter that determines the degree of compression of the LASSO model coefficients. As λ increases, the coefficient estimates of each independent variable are gradually compressed, and some independent variable coefficients will be compressed to zero, resulting in a streamlined model with fewer independent variables.

In this study, LASSO analyses were carried out using the "*glmnet*" package of R software. The determination of the optimal tuning parameter λ directly affects the number of variables, the choice of type, and the results of the parameter estimation. There are several main approaches to the selection of λ in existing studies: Bayesian Optimisation, Cross-Validation, Grid Search, and Random Search [30–33]. Cross-validation was used to determine the λ in this study. The specific steps are as follows: first, make the fold assignment by randomly splitting the data into K approximately equal groups, which will be used to estimate the prediction error for each value of λ. Second, withhold one of the K folds and fit each candidate model to the remaining K − 1 folds, denoted f^k. Third, compute the prediction error of each candidate model fitted in the previous step over the withheld fold, and repeat multiple times until each fold of data has been withheld. Finally, aggregate the prediction errors obtained over the K folds, and the optimal value of λ, is chosen to be the value that corresponds to the candidate model with the smallest aggregated prediction error [34]. Since k is often taken to be 10 in practice, it can be called ten-fold cross-validation [33]. The prediction error for ten-fold cross-validation can be expressed as:

$$CV\left(f\right) = \frac{1}{N}\sum_{i=1}^{N} L[y_i, f^{k(i)}(x_i)] \quad (3)$$

where $k(i)$ denotes the indicator function of the N samples in which observation i belongs to the k th $(k = 1, 2, \ldots, K)$ data; and f^k denotes the model fitted using the exclusion of the k th data. Assuming that fitting a set of models containing tuning parameters is $f^k(x, \lambda)$, it is defined as:

$$CV\left(f, \lambda\right) = \frac{1}{N}\sum_{i=1}^{N} L[y_i, f^{k(i)}(x_i, \lambda)] \quad (4)$$

where $CV\left(f, \lambda\right)$ denotes a test error curve that varies with λ. The λ that minimises $CV\left(f, \lambda\right)$ is the tuning parameter of the LASSO model.

Data description

Indicator construction

Building upon previous literature, this study aims to explore the potential factors influencing individuals' subjective health self-assessment. To obtain quantifiable micro-level data and refer to the indicator construction of existing studies [35], the present research decomposes the perspectives of "Individual level" "Family level" "Health behaviour" and "Dietary preferences" into seven categories of influencing factors, such as "Individual conditions" and "Positive health behaviours". These factors are further refined into 24 specific indicator variables. These 24 variables constitute the indicator system of this study on the influencing factors of subjective health self-assessment among a nationally representative sample of farmers (Table 1).

Data sources and sample processing

The data used in this work are from the China Nutrition and Health Survey (CHNS). The survey was organized by The Chinese Center for Disease Control and Prevention, The National Institute of Nutrition and Food Safety,

Table 1 Indicator system of factors influencing farmers' health self-assessment

Research perspectives	Influence factors	Variable
Individual level	Individual conditions	Gender
		Age
		Education
		Marry
		Sick
		Folk doctor
Health behaviour	Positive	Dietary guideline
		Nutrition knowledge
		Insurance
	Negative	Cigarettes
		Alcohol
Dietary preferences	Unhealthy	Fast food
		Salty foods
		Soft drinks
	Healthy	Fruits
		Vegetables
Family level	Hardware	Internet
		Filter water
		Flush toilet
	Cooking fuels	Coal
		Electricity
		Liquefied petroleum gas (LPG)
		Natural gas
		Wood

and The Carolina Population Center at the University of North Carolina at Chapel Hill. The CHNS survey was initiated in 1989 and has been conducted in 10 waves. The survey adopts the multi-stage stratified cluster random sampling method with the following steps: First, the simple random sampling method was used to select a total of nine provinces (Liaoning, Heilongjiang, Jiangsu, Shandong, Henan, Hubei, Hunan, Guangxi, and Guizhou) and three municipalities directly under the Central Government (Beijing, Shanghai, and Chongqing) scattered in eastern, central, and western regions of China. Second, different counties in each province were stratified according to differences in residents' income levels, and after weighting, four counties (including one high-income, two middle-income, and one low-income) were randomly selected from each of the above provinces as representatives of the rural sample; for the urban areas. The capital of each province and a low-income city were selected as urban samples. Next, three administrative villages were randomly selected from each county. Finally, 220 communities were selected from the rural and urban samples, with approximately 20 family households interviewed in each community, for a total of 4,400 households included in the survey. The CHNS database has been widely used in research in related fields such as medicine, economics and management due to its strong rigour and representativeness [36–38]. In our study, cross-sectional data from 2015 was selected, which is the latest year of data currently available for open access and could more accurately reflect the current physical health status and dietary preferences of Chinese individuals.

The aim of this study is to explore the significant factors influencing health self-assessment among farmers. First, we focus on smallholder farmers who permanently reside outside of city areas and are engaged in agricultural production activities. Second, considering that the variables related to dietary knowledge and dietary preferences in the questionnaire are applicable to residents aged 12 and above, we further excluded individuals below the age of 12 who were minors. Finally, after excluding the samples with missing data for key variables, we obtained a total of 10,115 valid samples for 2015.

Considering that the "*glmnet*" package in R software is primarily used for regression problems involving continuous dependent variables and binary categorical variables, the "In the last twelve months, how would you define your state of health" question in the questionnaire, which implemented a 5-point Likert scale, was transformed into a binary variable in our study: scores of 1 to 3 ("very poor" "poor" and "fair") were classified as "0 = unhealthy", while scores of 4 to 5 ("good" and "very good") were classified as "1 = healthy".

Descriptive statistics

According to Table 2, among the surveyed respondents, over half of the farmers reported having good physical health. The majority of respondents were female, with an average age exceeding 50 years, and had a low level of education, with only primary school qualifications. Most of the farmers in the sample were married, accounting for approximately 83.6%. Approximately 12.6% of individuals reported experiencing illness or injury in the past month, with a relatively low proportion seeking medical care from informal practitioners, accounting for only about 5%. We found that only about 20% of farmers demonstrated proactivity in terms of their knowledge and acquisition of nutritional dietary knowledge. Almost all farmers had purchased health insurance. The proportions of individuals who smoke and consume alcohol are 27.1% and 27.4%, respectively. In terms of dietary intake, we found that the majority of farmers exhibited good dietary preferences, preferring fruits and vegetables over high-fat, high-salt "junk food." The proportion of farmers using the internet was relatively low, less than 30%.

Table 2 Variable definitions and descriptive statistics

Variable	Definition	Mean	Std. Dev
Health self-assessment	1 = good health, 0 = bad health	0.506	0.500
Gender	1 = male, 0 = female	0.474	0.499
Age	Respondents' age	50.414	16.555
Education	Respondents' education level	1.912	1.441
Marry	Currently married? 1 = yes, 0 = otherwise	0.836	0.370
Sick	1 = if respondents have been sick or injured in last 4 weeks, 0 = otherwise	0.126	0.332
Folk doctor	1 = if respondents visit a folk doctor last year, 0 = otherwise	0.054	0.227
Dietary guideline	1 = know about the Dietary Guidelines for Chinese Residents, 0 = otherwise	0.199	0.399
Nutrition knowledge	1 = proactively look for nutrition knowledge, 0 = otherwise	0.208	0.406
Insurance	1 = if respondents have medical insurance, 0 = otherwise	0.975	0.156
Cigarettes	Have you ever smoked cigarettes? 1 = yes, 0 = otherwise	0.271	0.445
Alcohol	Drank beer/alcohol last year? 1 = yes, 0 = otherwise	0.274	0.446
Fast food	Likes fast food? (KFC, pizza, hamburgers, etc.) 1 = dislike very much, 5 = like very much	2.037	0.786
Salty foods	Likes salty snack foods? (potato chips, pretzels, etc.) 1 = dislike very much, 5 = like very much	2.163	0.809
Soft drinks	Likes soft/sugared drinks? 1 = dislike very much, 5 = like very much	2.567	0.930
Fruits	Likes fruits? 1 = dislike very much, 5 = like very much	3.622	0.744
Vegetables	Likes vegetables? 1 = dislike very much, 5 = like very much	3.788	0.657
Internet	1 = if respondents use the internet, 0 = otherwise	0.282	0.450
Filter water	1 = if respondents filter water before drinking, 0 = otherwise	0.124	0.329
Flush toilet	1 = have a flush toilet at home, 0 = otherwise	0.640	0.480
Coal	1 = use coal fuel for cooking, 0 = otherwise	0.066	0.249
Electricity	1 = use electricity for cooking, 0 = otherwise	0.303	0.459
LPG	1 = use liquefied petroleum gas for cooking, 0 = otherwise	0.346	0.476
Natural gas	1 = use natural gas for cooking, 0 = otherwise	0.170	0.375
Wood	1 = use wood, sticks/straw, etc. for cooking, 0 = otherwise	0.105	0.306

Approximately 12.4% of rural households had access to filtered drinking water, while over 60% of households had flush toilets. The highest proportion of households used liquefied gas for cooking, followed by electricity, and then cleaner natural gas. However, a small portion of households still rely on wood and coal for cooking.

Results and discussion

To comprehensively analyse the factors influencing subjective health self-assessment among representative farmers in China, this study selected 24 micro-level influencing factors, including gender, age, education, marital status, and others. Logistic regression was initially employed, and subsequently, the LASSO method was applied to select variables and estimate parameters from the 24 chosen factors. The dynamic process of variable selection using the LASSO method is depicted in Fig. 1.

Figure 2 represents the graph generated by the model through cross-validation. The vertical axis represents the magnitude of model bias, while the horizontal axis represents the values of the tuning parameter λ. The two vertical lines in Fig. 2 from left to right represent the λ with the minimum average error (lambda.min) and the maximum λ with the average error within one standard deviation (lambda.1se). When lambda.min is selected, the model can be allowed to fit more features. This choice is usually suitable for paying more attention to the predictive performance of the model, but not too much about the complexity of the model; the disadvantage is that there may be a risk of overfitting. When lambda.1se is selected, the model more strongly penalizes unimportant variables, leading to the selection of fewer features and limiting the number of non-zero coefficients in the model, thereby reducing the risk of overfitting [27, 32]. Based on the results of cross-validation and the sample size, this study prioritizes the interpretability of the model. Therefore, lambda.1se is selected as the optimal λ. With this λ value, the model ultimately removes 10 variables and retains 14 variables.

The second column of Table 3 shows the regression results of the logistic model. The results showed that age, illness or injury in the last month, having medical insurance, liking salty food, and flush toilet at home had statistically significant negative effects on farmers' self-assessment of health; while education level, knowing about dietary guidelines, proactively looking for nutrition

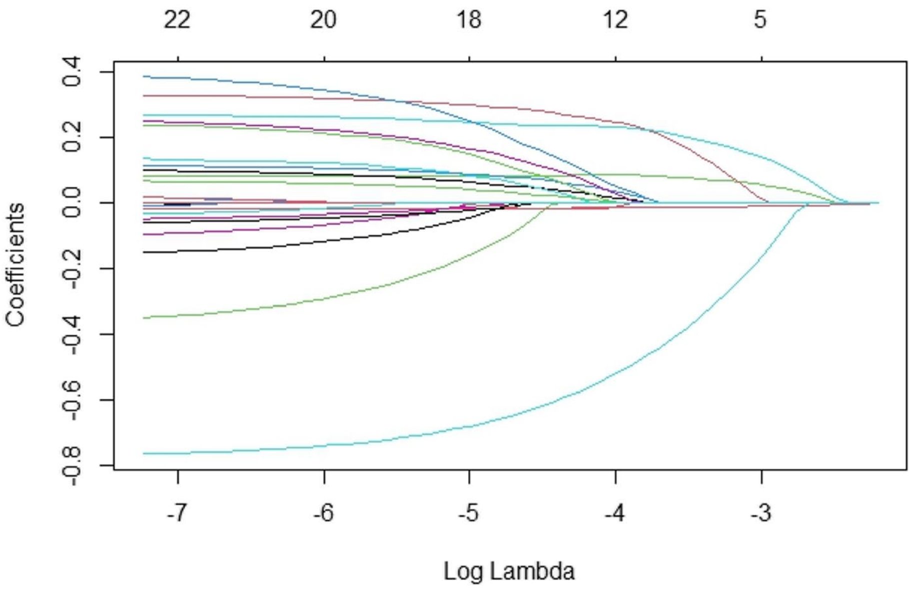


Fig. 1 Variable selection path for the LASSO model

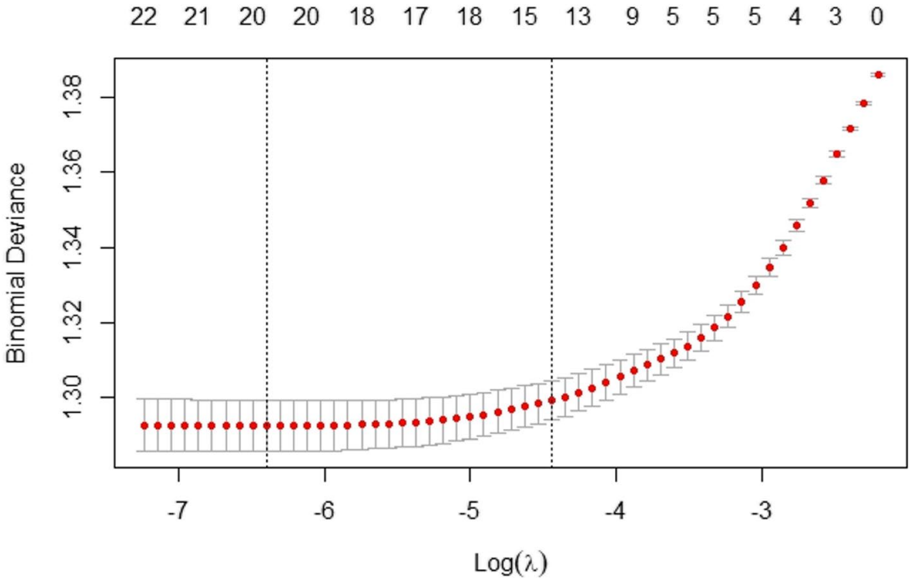


Fig. 2 Correspondence between λ and the number of independent variables

knowledge, drinking alcohol, liking fruits and vegetables, internet usage, filtering water before drinking, cooking with LPG and natural gas had statistically significant positive effects on farmers' self-assessment of health. However, there are some differences between the variables selected by the LASSO method and the statistically significant regression results in the Logistic model. The variables "salty foods" and "flush toilet" showed statistical significance in the Logistic regression analysis. However, these variables were not selected by the LASSO method, suggesting that they no longer exhibit a statistically significant relationship with farmers' health self-assessment. Interestingly, the coefficient of the variable "use electricity for cooking" has changed from positive to negative. Next, we will proceed with the analysis using the variables that were selected through the LASSO method.

We conduct a specific analysis from the following four aspects. At the individual level, as age increases, various

Table 3 The Logistic and LASSO estimation results

Variable	Logistic	LASSO
Gender	-0.029(0.056)	
Age	-0.020***(0.002)	-0.206
Education	0.085***(0.018)	0.084
Marry	-0.024(0.061)	
Sick	-0.779***(0.068)	-0.609
Folk doctor	-0.107(0.095)	
Dietary guideline	0.104*(0.063)	0.041
Nutrition knowledge	0.334***(0.062)	0.277
Insurance	-0.375***(0.138)	-0.005
Cigarettes	0.035(0.061)	
Alcohol	0.147**(0.059)	0.038
Fast food	-0.053(0.041)	
Salty foods	-0.065*(0.039)	
Soft drinks	0.024(0.026)	
Fruits	0.069**(0.035)	0.021
Vegetables	0.117***(0.040)	0.067
Internet	0.271***(0.060)	0.237
Filter water	0.261***(0.065)	0.104
Flush toilet	-0.164***(0.048)	
Coal	0.110(0.217)	
Electricity	0.099(0.204)	-0.016
LPG	0.349*(0.203)	0.072
Natural gas	0.501**(0.207)	0.148
Wood	0.060(0.212)	
Constant	0.461(0.306)	-
Chi-square	858.866***	
Observations	10,115	

Standard errors in parentheses. Significance level: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

physiological functions of the body gradually decline. This natural aging process often leads individuals to perceive their health as less robust compared to younger people [8]. However, a study has shown that compared to individuals aged 65 to 90, elderly people over 90 tend to reverse their perception of health and exhibit a better health self-assessment. This phenomenon can be attributed to the heterogeneity of the aging process and the development of adaptive mechanisms [39]. A higher level of education has a positive impact on farmers' subjective health self-assessment. This can be attributed to several factors. Farmers with higher education levels are more likely to have opportunities to access health knowledge and information, understand how to maintain health and prevent diseases, and have the ability to utilize healthcare resources effectively. Additionally, individuals with higher education levels may possess positive psychological factors such as self-confidence and an optimistic attitude towards life. These psychological factors help individuals cope with challenges and stress in life, resulting in a positive impact on their health status [40, 41]. Therefore, individuals with higher education levels are more likely to provide higher health self-assessments. Conversely, [26] found, through their research on women in developing countries like Brazil, that lower educational attainment is a predictive factor for lower self-assessments of health. It is not surprising to find that farmers who have reported experiencing illness or injury in the past month would have a lower self-assessment of their health. However, it is important to note that this does not necessarily imply that they would consistently provide an unhealthy self-assessment when asked about their health status in the future. Individuals who have experienced recent illness or injury may temporarily perceive their health as poor, but as they recover and their health improves, their self-assessment may also change.

In terms of health behaviours, farmers who possess knowledge about dietary guidelines and actively seek nutrition information tend to have higher self-assessments of their health. On the one hand, farmers' active pursuit of nutrition and health knowledge can motivate them to adopt healthier dietary behaviours. This may include increasing the consumption of fruits, vegetables, and whole grains while reducing the intake of high-sugar, high-salt, and high-fat foods [42]. Additionally, dietary knowledge empowers farmers to understand the nutritional content of different foods, promoting dietary diversity. This is crucial for preventing malnutrition, reducing the risk of chronic diseases, and improving overall health [43]. On the other hand, acquiring dietary knowledge enhances farmers' awareness of food safety and hygiene practices [44]. They can learn how to select and store food properly, avoid food poisoning, and employ appropriate cooking and processing techniques to preserve the nutritional elements of food. We found unexpected results regarding the impact of farmers' participation in medical insurance and alcohol consumption on their health self-assessments. Farmers' participation in medical insurance was negatively associated with their subjective health assessments, while drinking alcohol had a positive impact on farmers' subjective health assessments, this finding contrasts with previous research by [43]. These surprising results emphasize the complexity of the relationship between these variables and health self-assessments among farmers. We agree that the possible reason for the unexpected results may be due to the endogeneity problem caused by reverse causality. For example, individuals with poorer health may be more likely to seek medical insurance. Likewise, the group categorized as healthy demonstrated a higher likelihood of alcohol consumption compared to the group classified as frail. Alcohol consumption may be closely associated with complex social networks, and participation in social activities can lead to positive emotional and mental health, which is reflected in health self-assessment [45]. Farmers with poorer health conditions tend to practice moderation in alcohol consumption and participate less in social activities. Conversely, farmers with better health conditions tend to consume alcohol more frequently and possess greater energy to engage in social activities.

Regarding dietary preference, farmers who preferred fruits and vegetables reported better health self-assessments. There may be a substitution effect between healthy dietary preferences and unhealthy dietary preferences. Fruits and vegetables are important food sources of high-quality vitamins and dietary fibre. Supplementation of these nutrients helps the body function properly, promotes intestinal health, maintains digestive system function, and strengthens immunity, thereby improving the overall health of the individual [46].

At the family level, the subjective health self-assessments of farmers were positively influenced by Internet use. The widespread popularity of the Internet has significantly reduced the information gap between rural and urban areas, allowing farmers to access health-related knowledge more easily, and providing a platform for them to communicate and share experiences with others. At the same time, the Internet enables farmers to obtain health services remotely, such as online consultation and purchase of medicines [47, 48]. The convenient and efficient health service methods offered by Internet platforms enable farmers to effectively manage health problems and enhance their perception and evaluation of their health. Our study is consistent with [49], suggesting that unsanitary water facilities can lead to pathogen transfer and cause diseases, and the direct consumption of unclean water may have a negative impact on farmers' health. Apart from removing suspended solids from the water source, pre-drinking water filtration also has the potential to reduce the presence of harmful substances such as bacteria, viruses, and pesticide residues, which significantly affects human health [50]. Especially in most developing countries, the practice of filtering water before drinking should draw sufficient attention from relevant authorities and organisations. The use of electricity reduces farmers' self-assessment of health; while the use of clean fuels like LPG and natural gas as the primary cooking fuel has a positive effect on farmers' self-assessment of health [51]. Traditional fuels such as coal and firewood generate significant amounts of smoke and harmful gases, including PM2.5, PM10, and carbon monoxide, during the cooking process [52, 53]. These pollutants not only pose severe threats to the ecological environment, but also cause long-term irreversible damage to the respiratory and cardiovascular systems when inhaled by individuals [54]. In contrast, the use of clean fuels can reduce the health damage caused by the combustion of traditional fossil fuels, provide a safe cooking environment, and feature convenient operations. Additionally, it can also reduce greenhouse gas emissions, which has a positive impact on mitigating global warming and promoting environmental sustainability [50].

Considering the economic differences among the sample regions, we further divided the samples according to the South and the North to study the subjective evaluation of farmers' health status in different areas. The results are shown in Table 4 (The appendix shows the graphical results of cross-validation). Through comparison, we find that there are significant differences in the factors affecting the health self-assessment of farmers in the North and the South. Southern farmers who visited a folk doctor last year had lower health perceptions than their northern counterparts. Knowledge of dietary guidelines may only have a significant positive impact on the health of farmers in the South. Preferences for fruits and vegetables seem more likely to enhance the subjective health ratings of southern farmers. Filtering water sources only improved perceptions of health among southern farmers. The use of coal fuel reduced the

Table 4 Regional heterogeneity of the LASSO method

Variable	Total	South	North
Gender			
Age	-0.206	-0.016	-0.017
Edu	0.084	0.059	0.089
Marry			
Sick	-0.609	-0.610	-0.475
Folk doctor		-0.062	
Dietary guideline	0.041	0.064	
Nutrition knowledge	0.277	0.139	0.363
Insurance	-0.005		
Cigarettes			
Alcohol	0.038	0.009	0.031
Fast food			
Salty foods			
Soft drinks			
Fruits	0.021	0.031	
Vegetables	0.067	0.178	
Internet	0.237	0.169	0.383
Filter water	0.104	0.183	
Flush toilet			
Coal			-0.201
Electricity	-0.016	-0.134	
Liquified petroleum gas	0.072		0.019
Natural gas	0.148	0.065	0.079
Wood		-0.038	
Observations	10,115	6175	3940

perceived health of farmers in the North, while the burning of wood significantly diminished the perceived health of farmers in the South. The use of LPG was only able to improve the health perception of farmers in the North. The regional groupings help us identify the key factors influencing farmers' health self-assessment from a more comprehensive perspective.

Conclusions and implications

Conclusions and limitation

This study is based on micro-survey data from 10,115 representative Chinese farmers, selects 24 variables across four dimensions and employs the LASSO method to empirically study the influencing factors of farmers' health self-assessment. The results indicate the following: (1) education level has a positive impact on farmers' health perception, while age and recent experience of illness within the past month have a negative impact. (2) Actively seeking dietary and nutritional knowledge, and having a preference for healthy eating habits such as consuming fruits and vegetables, positively influence farmers' health self-assessment. Surprisingly, alcohol intake shows a positive correlation with health self-assessment, while a majority of farmers with medical insurance report an unhealthy physical condition. (3) Internet usage and filtering water sources before drinking show a positive correlation with farmers' health self-assessment. In terms of cooking fuel used in households, the adoption of clean LPG and natural gas demonstrates a positive significance in farmers' health. (4) Significant differences exist in the factors influencing the health perceptions of farmers in the North and South. Southern farmers appear to be more proactive in understanding dietary guidelines and consuming fruits and vegetables, whereas farmers in the North seem to be more inclined to adopt LPG, resulting in more positive health feedback.

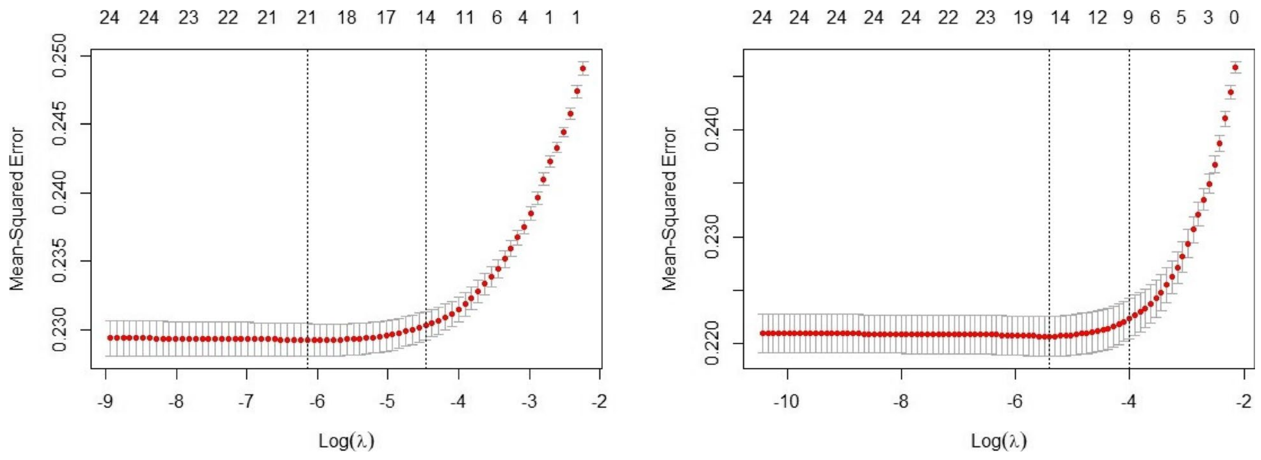
This paper employs machine learning algorithms to advance research in the field of health, enabling a more precise identification of factors influencing farmers' subjective health self-assessment. This approach makes a meaningful contribution to the existing literature and provides technical methodological support for future researchers exploring the identification of influencing factors. Certainly, it is necessary to acknowledge the limitations of this study, in particular the inclusion of only 24 factors in the model for variable screening. The full advantages of the algorithm might be realized only when a broader set of variables is considered. In future empirical analyses, the collection of more comprehensive data that incorporates a broader range of factors influencing farmers' health into the model for selection could further optimize the existing research conclusions. In addition, this study did not consider causal inferences about the mechanisms affecting the health of smallholder farmers, which could be a direction for further expansion in the future.

Implications

The results of this study provide several important implications. First, the government should strengthen health promotion and psychological counselling in rural areas, offering easily understandable and actionable health knowledge, encompassing nutrition, disease prevention, personal hygiene, and other informative aspects. Health knowledge can be disseminated and popularized through channels such as village committees and health stations, utilizing methods such as pamphlets and training lectures. Additionally, efforts should be made to provide psychological support, especially for vulnerable groups, to enhance their social support and sense of participation. This can contribute to improving their subjective health evaluation. Second, policymakers bear the responsibility of actively enhancing health awareness and promoting healthy behaviours among the farmers. Through community health promotion activities, media campaigns, and other communication channels, farmers can be motivated to adopt positive and healthy dietary habits, fostering changes in unhealthy lifestyles. For example, encouraging the diversification of diets, increasing dietary fibre intake, and limiting the consumption of high-salt and high-fat foods. These measures aim to improve the health level and overall quality of life for farmers. Third, relevant stakeholders should strengthen infrastructure according to the needs of different regions. For example, the effective operation of electricity facilities should be enhanced in the South, while water filtration equipment and flush toilets should be strengthened in the North, which is essential to improving the living environment and sanitation of farmers, thereby helping to prevent the spread of diseases. Fourth, the government and relevant organisations should actively promote the use of clean fuels, such as LPG and natural gas, especially in impoverished rural areas. Measures such as providing subsidies, tax exemptions, or offering preferential policies can be implemented to encourage farmers to use clean fuels.

Appendix

Results of cross-validation between the South and the North.



a. South

b. North

Acknowledgements

This research uses data from China Health and Nutrition Survey (CHNS). We thank the National Institute of Nutrition and Food Safety, China Center for Disease Control and Prevention, Carolina Population Center, the University of North Carolina at Chapel Hill, the NIH (R01-HD30880, DK056350, and R01-HD38700) and the Fogarty International Center, NIH for financial support for the CHNS data collection and analysis files from 1989 to 2006 and both parties plus the China-Japan Friendship Hospital, Ministry of Health for support for CHNS 2009 and future surveys.

Authors' contributions

MW: conceptualization, methodology, software, formal analysis, and writing — original draft. SZ: conceptualization and writing — review. All authors contributed to the article and approved the submitted version.

Funding

No funds will be received for this work.

Availability of data and materials

The data supporting the findings of this study are available from CHNS (https://www.cpc.unc.edu/projects/china/data/datasets).

Declarations

Ethics approval and consent to participate

The CHNS study was approved by the Institutional Review Board at the University of North Carolina and the National Institute of Nutrition and Health, Chinese Center for Disease Control and Prevention. All parents provided written informed consent for their children's participation in the survey.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest.

Received: 24 October 2023 Accepted: 18 January 2024
Published online: 31 January 2024

References

1. Montagna C, Zangelidis A. Labour Market implications of thyroid dysfunctions. Econ Hum Biol. 2023;50:101247. https://doi.org/10.1016/j.ehb.2023.101247.
2. Otsu Y, Yuen CYK, Health. Crime, and the Labor Market: theory and policy analysis. J Econ Dyn Control. 2022;144:104529. https://doi.org/10.1016/j.jedc.2022.104529.
3. Pit SW, Shrestha R, Schofield D, Passey M. Partial and complete Retirement due to Ill-Health among mature age australians. Public Health. 2013;127:561–71. https://doi.org/10.1016/j.puhe.2012.12.022.
4. Wang Q, Wang X, Li R. Does Population Aging Reduce Environmental pressures from Urbanization in 156 countries? Sci Total Environ. 2022;848:157330. https://doi.org/10.1016/j.scitotenv.2022.157330.
5. Jiang C, Li J, Liu J. Does Urbanization affect the gap between Urban and Rural areas? Evidence from China. Socioecon Plann Sci. 2022;82:101271. https://doi.org/10.1016/j.seps.2022.101271.
6. Bhan T, Patel A. Migration and Health: exploring the effect of destinations on the Health outcomes among Rural-Urban migrants in India. Soc Sci Med. 2023;116079. https://doi.org/10.1016/j.socscimed.2023.116079.
7. Bincy K, Logaraj M, Anantharaman VV. Social Network and its effect on selected dimension of Health and Quality of Life among Community Dwelling Urban and Rural Geriatric Population in India. Clin Epidemiol Glob Health. 2022;16:101083. https://doi.org/10.1016/j.cegh.2022.101083.
8. Yuan L, Yu B, Gao L, Du M, Lv Y, Liu X, Sun J. Decomposition Analysis of Health Inequalities between the Urban and Rural Oldest-Old populations in China: evidence from a National Survey. SSM - Popul Health. 2023;21:101325. https://doi.org/10.1016/j.ssmph.2022.101325.
9. Skrabski A. Social Capital in a changing society: Cross Sectional Associations with Middle aged Female and Male Mortality Rates. J Epidemiol Community Health. 2003;57:114–9. https://doi.org/10.1136/jech.57.2.114.

10. Zhang X, Wang Z, Li T. The current status of Occupational Health in China. Environ Health Prev Med. 2010;15:263–70. https://doi.org/10.1007/s12199-010-0145-2.
11. Leitão J, Pereira D, Gonçalves Â. Quality of Work Life and Organizational Performance: workers' feelings of contributing, or not, to the Organization's Productivity. Int J Environ Res Public Health. 2019;16:3803. https://doi.org/10.3390/ijerph16203803.
12. Ploetner C, Telford M, Brækkan K, Mullen K, Turnbull S, Gumley A, Allan S. Understanding and improving the experience of Claiming Social Security for Mental Health Problems in the West of Scotland: a participatory Social Welfare Study. J Community Psychol. 2020;48:675–92. https://doi.org/10.1002/jcop.22278.
13. Xu X, Cui Y, Cai Y, Asante Antwi H, Zhou L, Bangura JN. Effects of Personal Ability and Social Welfare on the Health of Rural-to-Urban Elderly migrants: an empirical analysis from Jiangsu Province, China. Healthcare. 2021;9:1760. https://doi.org/10.3390/healthcare9121760.
14. Ali T, Ali J. Factors affecting the consumers' willingness to Pay for Health and Wellness Food Products. J Agric Food Res. 2020;2:100076. https://doi.org/10.1016/j.jafr.2020.100076.
15. Ford MB. Social Distancing during the COVID-19 pandemic as a predictor of Daily Psychological, Social, and Health-related outcomes. J Gen Psychol. 2021;148:249–71. https://doi.org/10.1080/00221309.2020.1860890.
16. Kirkpatrick CF, Bolick JP, Kris-Etherton PM, Sikand G, Aspry KE, Soffer DE, Willard K-E, Maki KC. Review of current evidence and clinical recommendations on the effects of low-carbohydrate and very-low-carbohydrate (including ketogenic) diets for the management of Body Weight and other cardiometabolic risk factors: A Scientific Statement from the National Lipid Association Nutrition and Lifestyle Task Force. J Clin Lipidol. 2019;13(e1):689–711. https://doi.org/10.1016/j.jacl.2019.08.003.
17. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in Human Genetic studies. Cell. 2019;177:26–31. https://doi.org/10.1016/j.cell.2019.02.048.
18. Rodrigues CMC, Plotkin SA. Impact of vaccines; Health, Economic and Social perspectives. Front Microbiol. 2020;11:1526. https://doi.org/10.3389/fmicb.2020.01526.
19. Rhodes RE, McEwan D, Rebar AL. Theories of physical activity Behaviour Change: a history and synthesis of approaches. Psychol Sport Exerc. 2019;42:100–9. https://doi.org/10.1016/j.psychsport.2018.11.010.
20. Battista F, Ermolao A, Baak MA, Beaulieu K, Blundell JE, Busetto L, Carraça EV, Encantado J, Dicker D, Farpour-Lambert N, et al. Effect of Exercise on Cardiometabolic Health of adults with overweight or obesity: focus on blood pressure, insulin resistance, and Intrahepatic Fat—A systematic review and Meta-analysis. Obes Rev. 2021;22. https://doi.org/10.1111/obr.13269.
21. Ye Z, Sun L, Wang QA, Head-to-Head. Comparison of EQ-5D-5 L and SF-6D in Chinese patients with low back Pain. Health Qual Life Outcomes. 2019;17. https://doi.org/10.1186/s12955-019-1137-6.
22. Zare F, Ameri H, Madadizadeh F, Aghaei MR. Validity and reliability of the EQ-5D-3L (a generic preference-based instrument used for calculating Quality-Adjusted Life -Years) for patients with type 2 diabetes in Iran. Diabetes Metab Syndr Clin Res Rev. 2021;15:319–24. https://doi.org/10.1016/j.dsx.2021.01.009.
23. Chang J-J, Ji Y, Li Y-H, Pan H-F, Su P-Y. Prevalence of anxiety Symptom and Depressive Symptom among College students during COVID-19 pandemic: a Meta-analysis. J Affect Disord. 2021;292:242–54. https://doi.org/10.1016/j.jad.2021.05.109.
24. Wang W, Guo Y, Du X, Li W, Wu R, Guo L, Lu C. Associations between poor sleep quality, anxiety symptoms, and depressive symptoms among Chinese adolescents before and during COVID-19: a longitudinal study. Front Psychiatry. 2022;12:786640. https://doi.org/10.3389/fpsyt.2021.786640.
25. Cunha ACD, Santos Neto ETD, Salaroli LB. Self-Assessment of the Health Status and Leisure activities of individuals on Haemodialysis. Sci Rep. 2022;12:20344. https://doi.org/10.1038/s41598-022-23955-7.
26. Santos TADP, Guimarães RA, Pagotto V, Aredes NDA, Siqueira ISLD, Rocha SD, Carrijo CIDS, Rosso CFW. Negative Self-Assessment of Health in women: Association with Sociodemographic Characteristics, physical inactivity and Multimorbidity. Int J Environ Res Public Health. 2022;19:2666. https://doi.org/10.3390/ijerph19052666.
27. Tibshirani R. Regression shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Methodol. 1996;58:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
28. Sudheer Kumar SD, Attri KK. Singh Comparison of Lasso and Stepwise Regression technique for wheat yield prediction. J Agrometeorol. 2021;21:188–92. https://doi.org/10.54386/jam.v21i2.231.
29. McEligot AJ, Poynor V, Sharma R, Panangadan A. Logistic LASSO regression for dietary intakes and breast Cancer. Nutrients. 2020;12:2652. https://doi.org/10.3390/nu12092652.
30. Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, Hill MD, Sajobi TT. Functional outcome prediction in ischemic stroke: a comparison of machine learning algorithms and regression models. Front Neurol. 2020;11:889. https://doi.org/10.3389/fneur.2020.00889.
31. Elshewey AM, Shams MY, El-Rashidy N, Elhady AM, Shohieb SM, Tarek Z. Bayesian optimization with support Vector Machine Model for Parkinson Disease classification. Sensors. 2023;23(2085). https://doi.org/10.3390/s23042085.
32. McConville KS, Breidt FJ, Lee TCM, Moisen GG. Model-assisted Survey Regression Estimation with the Lasso. J Surv Stat Methodol. 2017;5:131–58. https://doi.org/10.1093/jssam/smw041.
33. Yokoyama T, Nakahara S, Kondo H, Miyake Y, Sakamoto T. Novel score for Predicting Early Emergency Endovascular Therapy in Trauma Care using Logistic LASSO Regression. Acute Med Surg. 2022;9. https://doi.org/10.1002/ams2.774.
34. Roberts S, Nowak G. Stabilizing the Lasso against Cross-validation Variability. Comput Stat Data Anal. 2014;70:198–211. https://doi.org/10.1016/j.csda.2013.09.008.
35. Martins CA, Do Prado CB, Ferreira JRS, Zandonade E, De Paula A, Bezerra OM, Salaroli LB. Self-rated Health Status and Associated Factors in rural workers. BMC Public Health. 2023;23:680. https://doi.org/10.1186/s12889-023-15548-4.
36. Duan R, Chen Y, Qiao T, Duan R, Chen M, Zhao L, Gong Y, Cheng G. Modern Dietary Pattern is prospectively Associated with earlier age at Menarche: data from the CHNS 1997–2015. Nutr J. 2020;19:95. https://doi.org/10.1186/s12937-020-00622-z.
37. Feng X, Liu Q, Li Y, Zhao F, Chang H, Lyu J. Longitudinal study of the relationship between Sleep Duration and Hypertension in Chinese adult residents (CHNS 2004–2011). Sleep Med. 2019;58:88–92. https://doi.org/10.1016/j.sleep.2019.01.006.
38. Zhao F, Liu Q, Li Y, Feng X, Chang H, Lyu J. Association between Alcohol Consumption and Hypertension in Chinese adults: findings from the CHNS. Alcohol. 2020;83:83–8. https://doi.org/10.1016/j.alcohol.2019.09.004.
39. Vivas I, Moreira MJG, Pinheira VH. Perceptions and Health Care Accessibility Perception of Portuguese Ageing Adults. In Longevity and Development: New perspectives on Ageing Communities; Guardado Moreira, M.J., A. Carvalho, L.S., Sequeira Gonçalves, P.J., Barreiros Pinheira, V.M., Eds.; Lecture Notes in Bioengineering; Springer International Publishing: Cham. 2023;61–73. ISBN 978-3-031-22272-6.
40. Li J, Powdthavee N. Does more education lead to Better Health Habits? Evidence from the School reforms in Australia. Soc Sci Med. 2015;127:83–91. https://doi.org/10.1016/j.socscimed.2014.07.021.
41. Taylor SE, Kemeny ME, Reed GM, Bower JE, Gruenewald TL. Psychological resources, positive illusions, and health. Am Psychol. 2000;55(1):99–109. https://doi.org/10.1037/0003-066X.55.1.99.
42. Yang Y, He D, Wei L, Wang S, Chen L, Luo M, Mao Z. Association between Diet-related knowledge, attitudes, behaviors, and Self-Rated Health in Chinese adult residents: a Population-based study. BMC Public Health. 2020;20:720. https://doi.org/10.1186/s12889-020-08896-y.
43. Jeruszka-Bielak M, Kollajtis-Dolowy A, Santoro A, Ostan R, Berendsen AAM, Jennings A, Meunier N, Marseglia A, Caumon E, Gillings R, et al. Are Nutrition-related knowledge and attitudes reflected in Lifestyle and Health among Elderly people? A study across five European countries. Front Physiol. 2018;9:994. https://doi.org/10.3389/fphys.2018.00994.
44. Ternier S. Understanding and measuring cooking skills and knowledge as factors influencing convenience food purchases and consumption. SURG J. 2010;3:69–76. https://doi.org/10.21083/surg.v3i2.1122.
45. Ma X, Piao X, Oshio T. Impact of Social Participation on Health among Middle-aged and Elderly adults: evidence from Longitudinal Survey Data in China. BMC Public Health. 2020;20:502. https://doi.org/10.1186/s12889-020-08650-4.
46. Bursać Kovačević D, Brdar D, Fabečić P, Barba FJ, Lorenzo JM, Putnik P. Strategies to Achieve a Healthy and Balanced Diet: Fruits and Vegetables as a Natural Source of Bioactive Compounds. In Agri-Food Industry

Strategies for Healthy Diets and Sustainability; Elsevier, 2020;51–88 ISBN 978-0-12-817226-1.

47. Fares N, Sherratt RS, Elhajj IH. Directing and orienting ICT Healthcare Solutions to address the needs of the Aging Population. Healthcare. 2021;9:147. https://doi.org/10.3390/healthcare9020147.
48. Siriwardhana Y, Gür G, Ylianttila M, Liyanage M. The role of 5G for Digital Healthcare against COVID-19 pandemic: opportunities and challenges. ICT Express. 2021;7:244–52. https://doi.org/10.1016/j.icte.2020.10.002.
49. Wang D, Shen Y. Sanitation and work time: evidence from the Toilet Revolution in Rural China. World Dev. 2022;158:105992. https://doi.org/10.1016/j.worlddev.2022.105992.
50. Barstow C, Bluffstone R, Silon K, Linden K, Thomas EA, Cost-Benefit. Analysis of Livelihood, Environmental and Health Benefits of a large Scale Water Filter and Cookstove distribution in Rwanda. Dev Eng. 2019;4:100043. https://doi.org/10.1016/j.deveng.2019.100043.
51. Ma W, Vatsa P, Zheng H. Cooking fuel choices and subjective well-being in Rural China: implications for a Complete Energy Transition. Energy Policy. 2022;165:112992. https://doi.org/10.1016/j.enpol.2022.112992.
52. Athira G, Bahurudeen A, Appari S. Sustainable Alternatives to Carbon Intensive Paddy Field burning in India: a Framework for Cleaner Production in Agriculture, Energy, and Construction industries. J Clean Prod. 2019;236:117598. https://doi.org/10.1016/j.jclepro.2019.07.073.
53. Xu M, Zhang S, Xie Y. Impacts of the Clean Residential Combustion Policies on Environment and Health in the Beijing–Tianjin–Hebei Area. J Clean Prod. 2023;384:135560. https://doi.org/10.1016/j.jclepro.2022.135560.
54. Basu AK, Byambasuren T, Chau NH, Khanna N. Cooking fuel choice, Indoor Air Quality and child mortality in India. SSRN Electron J. 2020. https://doi.org/10.2139/ssrn.3614243.

Publisher's Note