Online ISSN: 2161-2609
Advancesin Open Access Journals

Published by advancersin@www.advancersin.org
Cancer cell research, 2026, Vol. 13, No.49

/RESEARCH ARTICLE

Title: Research on the Influencing Factors of the Risk of Dementia in the Elderly Based on Machine Learning

Han Zhou, Zhengyan Liang, Mingrui Pang, Kebuo Wang

LInstitute of Food and Nutrition, Shandong Provincial Center for Disease Control and Prevention, Jinan, Shandong 250000, China

ARTICLE HISTORY

Received: 20-Apr-2026
Revised: 13-May-2025
Accepted: 14-May-2025
DOI: *10.54762/cdpr2026-49.1-7*

Abstract: Objective: Dietary habits may be related to the development of Alzheimer's disease, but there are relatively few related studies. Methods: Data from three cohort studies conducted in the Chinese Longitudinal Healthy Longevity Survey (CLHLS) from 2008 to 2018, 2011 to 2018, and 2014 to 2018 were used. The Alzheimer's disease diagnosed in the follow-up in 2014 and 2018 was taken as the final outcome variable. Random forest chart method was used for variable selection. The multivariate COX regression model was used to predict the risk factors of Alzheimer's disease, and the random survival forest model was used to evaluate the importance of the influencing factors. Results: As of 2018, among the 11,250 elderly people included in the study, the number of Alzheimer's disease cases was 2,421, accounting for 21.5%. The multivariate COX regression model analysis showed that compared with the elderly with a long life expectancy, the young elderly (HR = 0.49, 95% CI: 0.34 - 0.70) and the general elderly (HR = 0.64, 95% CI: 0.49 - 0.83) were protective factors for Alzheimer's disease. Elderly people who consumed lard daily had a 4.70 times higher risk of Alzheimer's disease compared to those who consumed other animal oils daily (95% CI: 1.08 - 20.50). Conclusion: Elderly people who mainly consume lard have a higher risk of Alzheimer's disease.

Keywords: Elderly; Alzheimer's disease; Machine learning; COX regression

1. Introduction

Alzheimer's disease, also referred to as senile dementia, is a chronic neurodegenerative disorder that generally entails structural alterations in the hippocampal region, resulting in severe cognitive impairments as well as non - cognitive symptoms such as emotional, behavioral, and psychotic symptoms [1-2]. As a disease with an undetermined cause and no effective treatment measures, alzheimer's disease unquestionably places a substantial burden on individuals, families, and society. With the progression of the global aging population, the number of alzheimer's disease patients has also increased significantly. The world alzheimer report [3] forecasts that by 2050, the number of alzheimer's disease patients globally will reach 152 million. As one of the countries with the most rapid aging rate in the world, china had 11.5% of its total population aged 65 and above in 2019 [4]. A multitude of studies have demonstrated [5-7] that an increase in age is accompanied by an elevated risk of alzheimer's disease. Under such circumstances, how to maintain the health of the elderly and delay the onset of alzheimer's disease has become a major issue that policymakers need to address. Studies have indicated [8] that certain factors that make up a person's lifestyle can influence the occurrence and progression of alzheimer's disease. These factors encompass physical exercise and dietary habits, and controlling these factors may have a certain impact on preventing alzheimer's disease. Nevertheless, there are relatively few studies on dietary habits in relation to alzheimer's disease. Therefore, this study selected elderly individuals over 60 years old and employed machine learning and cox regression methods to investigate the factors influencing the risk of alzheimer's disease onset, with a particular emphasis on dietary and lifestyle factors, in order to propose practical approaches to delay alzheimer's disease and provide a theoretical foundation for policy formulatio.

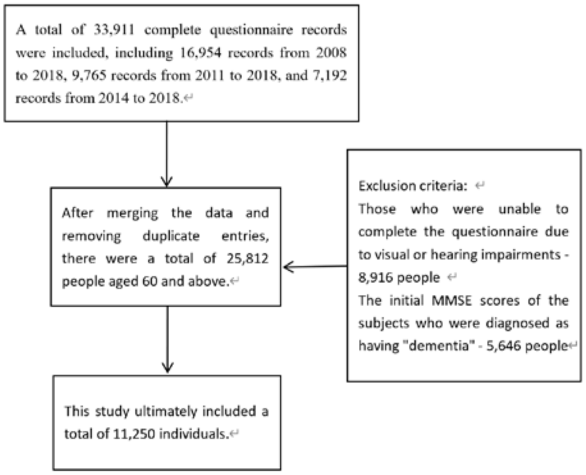
2. materials and methods

2.1 Research subjects

The data for this study were derived from the chinese longitudinal healthy longevity survey (clhls), which included 33,911 data records from three cohort studies conducted from 2008 to 2018, 2011 to 2018, and 2014 to 2018. After database merging and cleaning, a total of 1,1250 individuals were included in this study. Among them, 5,863 were male (52.12%) and 5,387 were female (47.88%), with an age range of 60 to 114 years, and the average age was (82.19 ± 11.06) years.

Figure 1 data cleaning and inclusion/exclusion process.

2.2 Variable definition and inclusion/exclusion criteria



The main outcome variable of this study is alzheimer's disease. In this study, the mini-mental state examination [9] (mmse) was used to assess the cognitive status of the elderly. The total score of the scale ranges from 0 to 30, including five categories of content: orientation, memory, attention and calculation ability, recall ability, and language ability. The normal cut-off value is defined as follows: illiterate > 17 points, primary school education > 20 points, junior high school education or above > 24 points. Those with scores below the standard within the corresponding educational level are defined as "alzheimer's disease (dementia)". The definition of the elderly age group: 60-74 years old is young-old, 75-89 years old is old, and ⩾ 90 years old is oldest-old. Living alone is defined as living alone, and having a live-in nanny is defined as living with family members. Smoking, drinking, and exercise are defined based on whether there were related behaviors in the past 12 months. Among them, active exercise for at least 3 days a week, each time for no less than 10 minutes, is defined as exercising. The cooking oil commonly used is defined as the type of oil used the most within a year.

Inclusion criteria: participants in the cohort study and those with a baseline age of 60 or above; exclusion criteria: those who cannot complete the questionnaire due to visual or hearing impairments; those whose initial mmse score is judged as "dementia"; those with other serious diseases.

2.3 Research methods

This study employed a multi-stage sampling cohort survey method. A baseline survey was conducted in 1998, followed by follow-up surveys in 2000, 2002, 2005, 2008-2009, 2011-2012, 2014, and 2017-2018. The survey covered 23 provinces, autonomous regions, and municipalities across the country, with a total of 113,000 participants. The survey was conducted through questionnaire methods, with the questionnaire covering aspects such as family basic conditions, social and economic background and family structure, economic sources and economic status, self-assessment of health and quality of life, cognitive function, personality and psychological characteristics, daily activity ability, lifestyle, life care, disease treatment, and medical expense bearing. The study received approval from the ethics committee of peking university (irb00001052-13074), and all respondents provided informed consent before participating [10].

2.4 Statistical analysis

Q-q plots and p-p plots were used to verify the normal distribution of the data. Measurement data were expressed as mean ± standard deviation, and count data were expressed as percentages. The sampling design of this study was rigorous, so it was assumed that the missing values were random. The pmm (predictive mean matching) method in multiple imputation was used to handle the missing values, thereby better preserving the natural variability of the data. The t-test was used to compare the ratios between the two groups, and the $\chi 2$ test was used to compare the ratios among multiple groups. Random forest [11] is an ensemble learning method based on decision trees. It builds multiple decision trees by randomly selecting features and samples, and combines the results of these decision trees through voting or averaging to perform classification or regression. In this study, the random forest plot method was used for variable selection, and the feature importance was evaluated to screen out the variables with a greater impact on the prediction results, in order to improve the model's prediction performance. The multivariate cox regression model was used to predict the risk factors for alzheimer's disease, and the non-parametric, nonlinear integrated machine learning random survival forest model based on survival trees was used to evaluate the importance of the influencing factors. All statistical analyses in this study were conducted using spss 25.0 and the randomforestsrc package, with the significance level set at bilateral $\alpha = 0.05$.

3. Results

3.1 Baseline characteristics of the study subjects

A total of 11,250 study subjects were included in this study. The data cleaning process is shown in figure 1. There were 3,267 young elderly individuals (aged 60-74), 5,191 general elderly individuals (aged 75-89), and 2,792 long-lived elderly individuals (aged ⩾ 90). There were 1,975 single-person dwellers, accounting for 17.56%. There were 2,405 smokers, accounting for 21.38%. There were 2,229 drinkers, accounting for 18.81%. There were 3,823 individuals who still exercised regularly, accounting for 33.98%. The proportion of cooking oil used was from high to low as other vegetable oils (84.30%), lard (15.01%), sesame oil (0.4%), and other animal oils (0.28%). 2,421 individuals were diagnosed with alzheimer's disease at the end of the follow-up, accounting for 21.5%. Among them, there were statistically significant differences in gender, age groups, whether living alone, smoking, drinking, exercising, the type of cooking oil used, years of education, bmi, and baseline mmse score between alzheimer's disease patients and elderly individuals with normal cognition ($p < 0.05$). See table 1.

Table 1 baseline characteristics of the included subjects [±sd, n(%)].

variables	dementia	normal	t/x^2	*p*
total	2421(21.5)	8829(78.5)		
gender			241.969	＜0.001
man	923(38.1)	4940(56.0)		
woman	1498(61.9)	3889(44.0)		
age			309.029	＜0.001
young-old（60-74）	357(14.7)	2910(33.0)		
old（75-89）	1311(54.2)	3880(43.9)		
oldest-old（⩾90）	753(31.1)	2039(23.1)		
live alone			16.521	＜0.001
with household member(s)	1910(78.9)	7193(81.5)		
alone	485(20.0)	1490(16.9)		
in a nursing home	26(1.1)	146(1.7)		
smoking			56.478	＜0.001
yes	383(15.8)	2022(22.9)		
no	2036(84.2)	6806(77.1)		
drinking			27.868	＜0.001
yes	388(16.0)	1841(20.9)		
no	2031(84.0)	6980(79.1)		
exercise			28.682	＜0.001
yes	712(29.5)	3111(35.5)		
no	1705(70.5)	5706(64.7)		
oil for cooking at home			9.388	0.025
other vegetable oils	2004(82.8)	7480(84.7)		
sesame oil	16(0.7)	29(0.3)		

lard	393(16.2)	1296(14.7)		
other animal oils	8(0.3)	24(0.3)		
i went to school for a few years	1.68±4.69	3.24±4.93	359.530	＜0.001
blood glucose	4.63±3.26	4.51±2.07	2.110	0.147
superoxide dismutase, sod	53.63±16.36	54.92±14.18	0.223	0.637
malondialdehyde	5.13±3.22	5.22±3.24	0.005	0.941
body mass index, bmi	20.65±3.55	21.09±3.41	30.530	＜0.001
cognitive scores at baseline	27.57±1.89	28.08±1.82	146.739	＜0.001

3.2 Variable selection based on machine learning and assessment of the importance of influencing factors

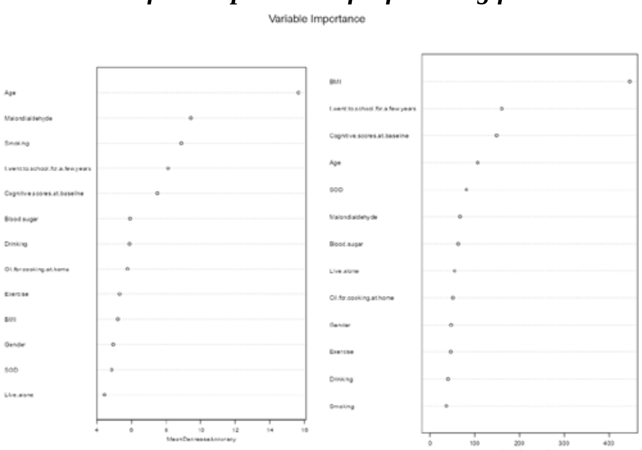


Figure 2 results of variable selection using random forest graph method

In this study, the relevant variables of alzheimer's disease were screened using the random forest graph method. Among them, the average decrease in accuracy and the average reduction in gini coefficient both represent important indicators for evaluating the contribution of features to the purity of nodes in the decision tree model. The larger the value, the more important the feature is. Age and bmi are very important for the establishment of this model. See figure 2. all relevant factors were included in the machine learning random survival forest model to obtain the importance ranking of the included factors. The average decrease in accuracy is an indicator for evaluating the impact of features on the prediction performance of the model. The results of this study found that age is still the most important factor affecting alzheimer's disease. See figure 3.

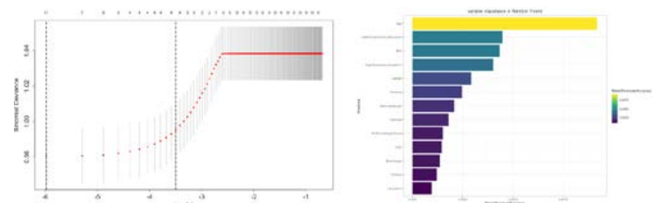
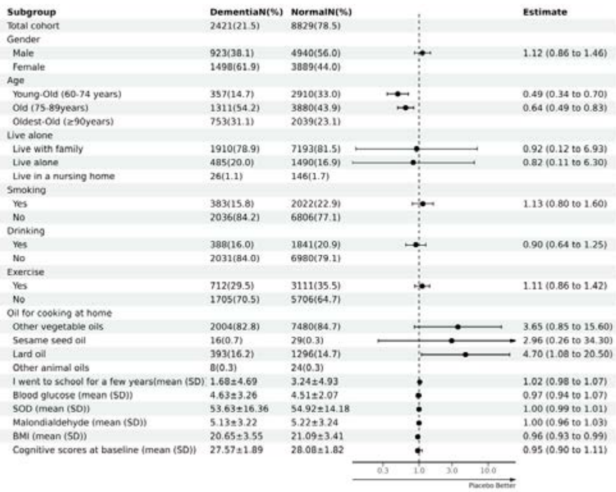


figure 3 ranking of variable importance and partial likelihood loss diagram

3.3 multivariate cox regression results

Using alzheimer's disease as the outcome variable, the survival time was the time when alzheimer's disease occurred during the follow-up. These variables were input into the multivariate cox regression model. Among them, gender, age group, living alone, smoking, drinking, exercise, and the type of oil used were all classified variables and were set the last one as the reference level. The results showed that compared with the elderly who lived a long life, young elderly people (hr = 0.49, 95% ci: 0.34, 0.70) and general elderly people (hr = 0.64, 95% ci: 0.49, 0.83) were protective factors for the occurrence of alzheimer's disease. Compared with other animal oils, using lard (hr = 4.70, 95% ci: 1.08, 20.50) was a risk factor for the occurrence of alzheimer's disease. See figures 4 and 5.



Subgroup	DementiaN(%)	NormalN(%)	Estimate
Total cohort	2421(21.5)	8829(78.5)	
Gender			
Male	923(38.1)	4940(56.0)	1.12 (0.86 to 1.46)
Female	1498(61.9)	3889(44.0)	
Age			
Young-Old (60-74 years)	357(14.7)	2910(33.0)	0.49 (0.34 to 0.70)
Old (75-89years)	1311(54.2)	3880(43.9)	0.64 (0.49 to 0.83)
Oldest-Old (≥90years)	753(31.1)	2039(23.1)	
Live alone			
Live with family	1910(78.9)	7193(81.5)	0.92 (0.12 to 6.93)
Live alone	485(20.0)	1490(16.9)	0.82 (0.11 to 6.30)
Live in a nursing home	26(1.1)	146(1.7)	
Smoking			
Yes	383(15.8)	2022(22.9)	1.13 (0.80 to 1.60)
No	2036(84.2)	6806(77.1)	
Drinking			
Yes	388(16.0)	1841(20.9)	0.90 (0.64 to 1.25)
No	2031(84.0)	6980(79.1)	
Exercise			
Yes	712(29.5)	3111(35.5)	1.11 (0.86 to 1.42)
No	1705(70.5)	5706(64.7)	
Oil for cooking at home			
Other vegetable oils	2004(82.8)	7480(84.7)	3.65 (0.85 to 15.60)
Sesame seed oil	16(0.7)	29(0.3)	2.96 (0.26 to 34.30)
Lard oil	393(16.2)	1296(14.7)	4.70 (1.08 to 20.50)
Other animal oils	8(0.3)	24(0.3)	
I went to school for a few years(mean (SD)	1.68±4.69	3.24±4.93	1.02 (0.98 to 1.07)
Blood glucose (mean (SD))	4.63±3.26	4.51±2.07	0.97 (0.94 to 1.07)
SOD (mean (SD))	53.63±16.36	54.92±14.18	1.00 (0.99 to 1.01)
Malondialdehyde (mean (SD))	5.13±3.22	5.22±3.24	1.00 (0.96 to 1.03)
BMI (mean (SD))	20.65±3.55	21.09±3.41	0.96 (0.93 to 0.99)
Cognitive scores at baseline (mean (SD))	27.57±1.89	28.08±1.82	0.95 (0.90 to 1.11)

figure 4 Cox regression analysis and forest plot of the influencing factors related to alzheimer's disease (dementia).

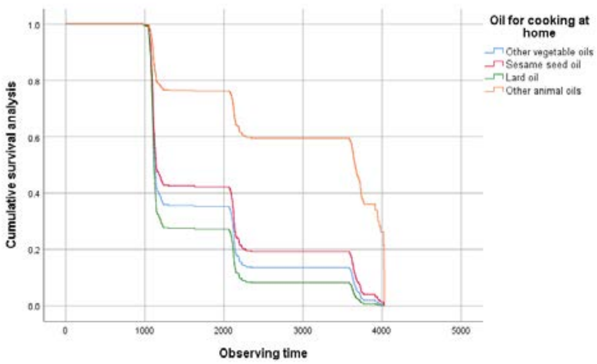


figure 5 Cox regression function graph showing the impact of different types of edible oil on alzheimer's disease (senile dementia).

4. Discussion

The causes of alzheimer's disease remain unclear. Presently, it is generally acknowledged that gender and age are significant risk factors for the onset of alzheimer's disease [10]. Our research findings did not indicate that gender was a contributing factor to the onset of alzheimer's disease. However, in comparison with long - lived elderly individuals (⩾90 years old), young elderly people (60 - 74 years old) and general elderly people (75 - 89 years old) were identified as protective factors against alzheimer's disease. The machine - learning random forest model also demonstrated that, based on the ranking of the importance of the variables involved in this study, age factors still had the most substantial impact on alzheimer's disease, which further suggests that alzheimer's disease is an inherent neurodegenerative disorder. Previous studies have proposed [11 - 12] that heart and vascular diseases can result in reduced blood flow in the brain, thereby diminishing the supply of nutrients and even causing irreversible brain damage. The resulting vascular dementia is also a crucial component of alzheimer's disease. Therefore, in addition to these unalterable factors, this study also focused on factors related to cardiovascular disease, such as exercise, smoking, and alcohol consumption, and the results were not statistically significant. Although a higher education level may postpone the onset of alzheimer's disease by enhancing cognitive reserve [13], this study revealed that there was no statistical significance in the association between the number of years of education and alzheimer's disease, which is consistent with the study by thow et al. [14]. Although education level can delay the onset of alzheimer's disease by increasing "cognitive reserve/knowledge reserve", the correlation between the number of years of education and the effect is not significant and will gradually weaken over time, eventually reaching a plateau. More research has pointed out [15] that a high level of education may delay the onset of alzheimer's disease, but once cognitive decline begins, the survival period of those with high education levels is shorter. Therefore, there may be more complex biological mechanisms between education years and alzheimer's disease.

Some scholars have already proposed that dietary factors, including the deficiency of folic acid and b vitamins, may have an impact on the development of alzheimer's disease. This study shows that compared to other animal oils (such as butter and cream), regular consumption of lard increases the risk of alzheimer's disease in the elderly by 3.7 times. The high saturated fat and high cholesterol content of lard may be related to alzheimer's disease. The results of animal experiments [16] show that long-term high-fat diet promotes the aggregation of β-amyloid protein in the brain, which is one of the main characteristics of alzheimer's disease. Population studies [17] show that a high-fat diet can cause the occurrence of alzheimer's disease by affecting lipid levels and through the microbiota-brain-gut axis. However, some studies suggest that a high-fat diet can protect mice from damage to the blood-brain barrier and brain atrophy [18]. Considering that other animal oils such as butter and cream also have high fat content, and the cholesterol content of butter is twice that of lard, we believe that the reason why lard can promote the occurrence of alzheimer's disease may be the difference in fatty acid structure. Fatty acids account for about 95% of the molecular weight in natural oils and are the main components of oils. The composition of fatty acids varies, resulting in differences in the properties of the oils, and even if the compositions are similar, physical properties can still differ due to the distribution of fatty acids at the three hydroxyl positions of glycerol. Studies [19] show that short-chain fatty acids in the diet can promote brain energy metabolism, and long-chain polyunsaturated fatty acids have a protective effect on cognitive decline [20], while saturated long-chain fatty acids are associated with an increased risk of progressing from mild cognitive impairment to alzheimer's disease [21]. Thus, different types of fatty acids have different effects on the onset and progression of alzheimer's disease.

In conclusion, as people age, long-term cooking with lard may increase the risk of alzheimer's disease to some extent. It is recommended that the elderly avoid long-term and excessive consumption of lard and choose vegetable oils or blended oils with higher contents of polyunsaturated fatty acids, which are more beneficial for the protection of cognitive function. Commonly used cooking oils, as a relatively easy-to-control and change dietary habit, should be further studied regarding the usage amount of cooking oil and the impact of different types of fatty acids on alzheimer's disease, in order to provide theoretical basis for the formulation of public nutrition policies.

Conflicts of interest

All authors declare no conflicts of interest.

Author contribution statement

zhou han: concept and design, data collection and processing, literature review, writing the first draft and revising the manuscript; liang zhengyan: data processing; pang mingrui: data processing; wang kebao: feedback comments.

acknowledgements this study received strong support from the shandong center for disease control and prevention. thanks to all participants.

References

[1] twarowski b, herbet m. inflammatory processes in alzheimer's disease-pathomechanism, diagnosis and treatment: a review[j]. international journal of molecular sciences, 2023,24(7): 6518.

[2] rostad h m, puts m t e, cvancarova småstuen m, et al. associations between pain and quality of life in severe dementia: a norwegian cross-sectional study[j]. dementia and geriatric cognitive disorders extra, 2017,7(1): 109-121.

[3] shi y, lin f, li y, et al. association of pro-inflammatory diet with increased risk of all-cause dementia and alzheimer's dementia: a prospective study of 166,377 uk biobank participants[j]. bmc medicine, 2023,21(1): 266.

[4] man w, wang s, yang h. exploring the spatial-temporal distribution and evolution of population aging and social-economic indicators in china[j]. bmc public health, 2021,21(1): 966.

[5] breijyeh z, karaman r. comprehensive review on alzheimer's disease: causes and treatment[j]. molecules (basel, switzerland), 2020,25(24): 5789.

[6] zahr n m. alcohol use disorder and dementia: a review[j]. alcohol research : current reviews, 2024,44(1): 3.

[7] fang h, li l, ai j, et al. mortality from alzheimer's disease and other dementias and its heterogeneity across states in india findings from gbd 2021 study[j]. international journal of surgery (london, england), 2025,111(7): 4830-4841.

[8] dominguez l j, veronese n, vernuccio l, et al. nutrition, physical activity, and other lifestyle factors in the prevention of cognitive decline and dementia[j]. nutrients, 2021,13(11): 4080.

[9] ma zongkang, liu xinglang, li huiui, et al. research on risk prediction model for mild cognitive impairment in the elderly [j]. preventive medicine, 2026, 38(02): 124-129.

[10] lin xueru, zheng chao, du maolin, et al. analysis of the health status and influencing factors of elderly people in china: based on the 8th round of clhls data [j]. journal of naval medical university, 2022, 43(09): 1022-1028.

[11] geng xingli, liu yingyun. hourly prediction of o3 concentration in changsha based on random forest model [j]. journal of nanhu university (natural science edition), 2025, 39(03): 16-24.

[12] graff-radford j, yong k x x, apostolova l g, et al. new insights into atypical alzheimer's disease in the era of biomarkers[j]. the lancet. neurology, 2021,20(3): 222-234.

[13] yang y, zhao x, zhu z, et al. vascular dementia: a microglia's perspective[j]. ageing research reviews, 2022,81: 101734.

[14] morgan a e, mc auley m t. vascular dementia: from pathobiology to emerging perspectives[j]. ageing research reviews, 2024,96: 102278.

[15] rosselli m, uribe i v, ahne e, et al. culture, ethnicity, and level of education in alzheimer's disease[j]. neurotherapeutics : the journal of the american society for experimental neurotherapeutics, 2022,19(1): 26-54.

[16] thow m e, summers m j, saunders n l, et al. further education improves cognitive reserve and triggers improvement in selective cognitive functions in older adults: the tasmanian healthy brain project[j]. alzheimer's & dementia (amsterdam, netherlands), 2017,10: 22-30.

[17] huang ying, huo yanli, ming yue, et al. research progress on high-fat diet and alzheimer's disease [j]. journal of neuropharmacology, 2020, 10(06): 54-59.

[18] paradise m, cooper c, livingston g. systematic review of the effect of education on survival in alzheimer's disease[j]. international psychogeriatrics, 2009,21(1): 25-32.

[19] wang m, lv j, huang x, et al. high-fat diet-induced atherosclerosis promotes neurodegeneration in the triple transgenic (3 × tg) mouse model of alzheimer's disease associated with chronic platelet activation[j]. alzheimer's research & therapy, 2021,13(1): 144.

[20] elhaik goldman s, goez d, last d, et al. high-fat diet protects the blood-brain barrier in an alzheimer's disease mouse model[j]. aging cell, 2018,17(5): e12818.

[21] lópez-ortiz s, valenzuela p l, seisdedos m m, et al. exercise interventions in alzheimer's disease: a systematic review and meta-analysis of randomized controlled trials[j]. ageing research reviews, 2021,72: 101479.

[22] barberger-gateau p, samieri c, féart c, et al. dietary omega 3 polyunsaturated fatty acids and alzheimer's disease: interaction with apolipoprotein e genotype[j]. current alzheimer research, 2011,8(5): 479-491.

[23] fan l b a w s. associations of circulating saturated long-chain fatty acids with risk of mild cognitive impairment and alzheimer's disease in the alzheimer's disease neuroimaging initiative (adni) cohort[j]. ebiomedicine, 2023: 97-104818.