Population Segmentation and Recommendations Based on Disease Risk and Preventive Awareness

Xinzhe Zheng[1, #], Jinzhu Cai[5, #, *], Chaomin Zhou[2], Yuehan Lv[3], Chenxu Wang[4]

[1]School of Materials Science and Engineering, Jiamusi University, Jiamusi City, Heilongjiang Province 154007

[2]School of Architecture and Engineering, Jinggangshan University, Ji'an City, Jiangxi Province, 343009

[3]School of Management, Tianjin Normal University, Tianjin, 300382

[4]School of Information and Electronics Technology, Jiamusi University, Jiamusi City, Heilongjiang Province 154007

[5]School of Materials Science and Engineering, Jiamusi University, Jiamusi City, Heilongjiang Province 154007

* Corresponding Author Email: zgkjb22@sina.com

[#]These authors contributed equally.

Abstract. Chronic non-communicable diseases represent a major challenge in the global healthcare landscape. To effectively prevent and manage these diseases, understanding the disease risk and awareness of preventive measures among community residents is of paramount importance. This study collected questionnaire data and conducted comprehensive analysis. This paper's objective was to categorize residents based on multidimensional factors and provide personalized health recommendations. This paper achieved this by establishing a three-dimensional mapping model and employing the K-Medoids clustering algorithm to classify samples into four categories. Based on these categories, the paper offered diverse health advice encompassing dietary adjustments, exercise habits, regular check-ups, smoking cessation, and alcohol restriction. These recommendations aim to reduce disease risk and effectively manage chronic non-communicable diseases.

Keywords: K-Medoids clustering, Chronic disease, Three-dimensional mapping

1. Introduction

Since the beginning of the 21st century, China's rapid economic growth has garnered worldwide attention[1]. With a GDP ranking second globally, "Made in China" products have become ubiquitous, and the country is gradually advancing toward "Created in China" innovations. China's development in various aspects has improved the material and cultural standards of living for its residents, promoted advancements in technology and healthcare, and raised the average life expectancy[2]. However, alongside this economic growth, the health of the Chinese people is facing significant challenges. Chronic non-communicable diseases and injuries have emerged as primary factors affecting residents' physical well-being.Conditions like cardiovascular diseases, diabetes, and malignant tumors pose substantial threats to the health of the population. These chronic illnesses are closely linked to unhealthy lifestyles, poor dietary habits, lack of physical activity, tobacco and alcohol abuse, and shifts in dietary patterns characterized by increased fat, sugar, and salt intake, coupled with inadequate consumption of fruits and vegetables.[3-5]The physical health of urban residents is influenced by various factors, including age, [6] work environments, and lifestyle choices. Given this backdrop, it is crucial to conduct comprehensive and in-depth research on the health of Chinese residents and the factors that impact it. Society as a whole must pay greater attention to the prevention and control of chronic diseases, enhance monitoring and response to factors affecting the

physical health of urban residents, and prioritize environmental protection and ecological development[7-8].

To further safeguard the lives and health of residents, we aspire to provide personalized health guidance to different population groups, thereby enhancing the overall health and quality of life of community residents. This study serves as a vital reference for the development of more targeted health intervention measures and aids in addressing the challenges posed by chronic non-communicable diseases[9-10].

2. Methods

2.1. Data Collection

We designed a questionnaire for the surrounding residents to investigate chronic non-communicable diseases and their related epidemiological factors, and collected a large amount of data for our subsequent research. However, considering the effectiveness of the questionnaire, we need to conduct reliability and validity tests on the questionnaire results we obtained.

Firstly, to test the reliability of the questionnaire, we used Cronbach's alpha coefficient and conducted the analysis using SPSS software. In the reliability test of the survey questionnaire, there were a total of 234 items in the questionnaire, and the Cronbach's alpha coefficient (α) reached 0.601, indicating good reliability of the formal survey questionnaire. For the validity test, the team used KMO and Bartlett tests in SPSS to analyze the questionnaire data. The KMO statistic is used to test the partial correlations among variables. The overall KMO value for the questionnaire was 0.614, which is greater than 0.5, indicating that the questionnaire is suitable for factor analysis. The Bartlett test result with a p-value greater than 0.05 suggests that the questionnaire is valid.

To eliminate cases of improperly filled questionnaires in the study, we excluded certain situations in the questionnaire, including: duplicate questionnaires (i.e., the same questionnaire appearing twice with the same ID), questionnaire information deviations (such as questionnaire content not matching standards, age in the questionnaire not corresponding to reality), identical questionnaire answers (where a large number of answers in a questionnaire are the same), and entirely empty questionnaire data (i.e., rows of data with no information). In the questionnaires we obtained, there were some rows with entirely empty data. After removing these data, we ultimately processed 7,836 valid sample records.

Among the valid questionnaires, there were also some outlier data, including non-standard data or uncertain parameters set for differentiation. In this questionnaire, an age of 99 for starting smoking or drinking indicated that the specific age was forgotten. The team used the method of replacing "99" with the respective averages for these two groups of data, namely:

$$year_{ij} = \begin{cases} year_{ij} & year_{ij} \neq 99 \\ \frac{\sum_{j=1}^{n} year_{ij}}{n} & year_{ij} = 99 \end{cases} \quad (1)$$

In the context, $year_{ij}$ represents the starting age of the i-th sample for the j-th item, where j=1 corresponds to smoking, and j=2 corresponds to drinking. Here, n denotes the number of non-zero data points in the samples.

2.2. Three-dimensional mapping

First, we need to perform dimensionality reduction on the data we have obtained. Therefore, we define three-dimensional models to conduct dimensionality reduction analysis on the sample situations, mainly including: the disease status, non-human factors, and the degree coefficient of preventive measures.

Disease status: Patients can be categorized into chronic disease patients and non-chronic disease patients. Chronic diseases primarily include diabetes patients and hypertension patients, and the classification of these two types of patients is determined by the results mentioned in the above

question, through a combined assessment of physical examination results and medical history results. That is, for such patients, the following conditions are met:

$$H_i = 1 \mid D_i = 1$$
$$H_i = \begin{cases} 1 \; Ht_i = 1 \text{or } H_{si} > 139 \; or \; H_{di} > 89 \\ 0 \text{ others} \end{cases} \quad (2)$$
$$D_i = \begin{cases} 1 \; Dt_i = 1 \text{or } D_{bi} > 7 \\ 0 \text{ others} \end{cases}$$

Wherein, H_i indicates whether there is hypertension, D_i indicates whether there is diabetes. $Ht_i$and Dt_i represent the past diagnosis status of hypertension and diabetes, where 1 indicates having the disease, and 2 indicates not having the disease. Since the treatment of chronic diseases is relatively difficult, for chronic diseases, the previous diagnosis and the current physical examination data both showing the disease are consistent. Therefore, residents who have been diagnosed previously and whose current physical examination data indicates the disease are considered as the chronic disease population.

Non-human factors: For residents, whether they suffer from chronic diseases depends on genetics and their own lifestyle habits. These two factors are defined as non-human factors and human factors. In non-human factors, the main consideration is the number of people in the family who have had chronic diseases, namely:

$$\sum_{i=1}^{8} num_i \quad (3)$$

Where num_i represents the number of individuals in the family who have the i-th chronic disease.

Preventive Measures Degree Coefficient: For human factors, preventive measures can be taken through various actions. To address this, a preventive measures degree coefficient is defined. For this coefficient, first, several parameter ranges were defined based on literature and data research. Samples meeting the criteria were assigned a value of 1, otherwise, they were assigned 0. Additionally, based on literature and data, an expert scoring method was used to rank the importance of various preventive measures and provide corresponding weights and the result is show in the Table.1

Table 1. List of preventive action items

Preventive Measures	Importance Level	Weight
Daily salt intake less than 6g	3	0.15
Daily sugar intake less than 15g	3	0.15
Daily fruit consumption over 250g	2	0.10
Daily vegetable consumption over 400g	2	0.10
Dairy product consumption reaches 200g	2	0.10
Body Mass Index (BMI) less than 24	1	0.05
Waist circumference less than 90cm for men/85cm for women	1	0.05
Weekly exercise time exceeding 250 minutes	2	0.10
Non-drinker	2	0.10
Non-smoker	2	0.10

2.3. K-Medoids clustering

Similar to the K-means clustering algorithm, the K-Medoids clustering is also an efficient clustering method. It clusters data based on the central positions within clusters. It repeatedly selects sample points as medoids and continues to add them. The main algorithmic process is as follows:

Step 1: Randomly select k points as medoids.

Step 2: According to the principle of proximity to medoids, assign the remaining points to the class represented by the current best medoids.

Step 3: In each class, calculate the criterion function for each member point and select the point corresponding to the minimum criterion function as the new medoid.

Step 4: Repeat steps 2-3 until all medoid points no longer change or until the maximum number of iterations is reached, as set.

3. Results

3.1. The clustering result

Based on the data processing results, the clustering of the three-dimensional space produced the following results in the Figure 1.

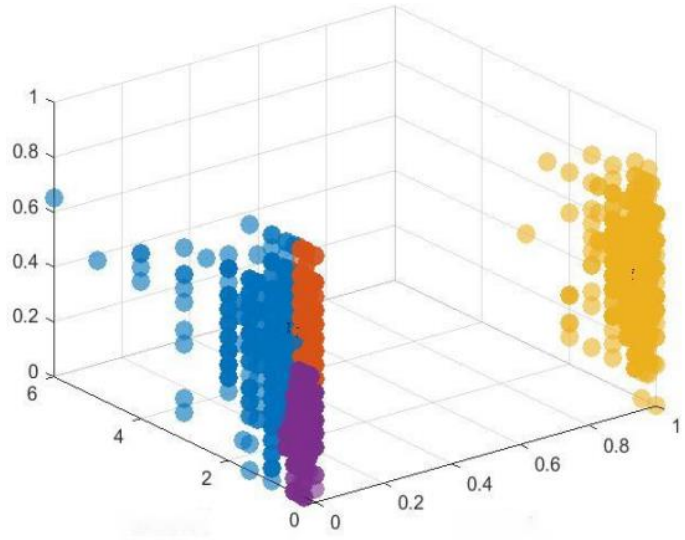


Figure 1. K-Medoids Clustering results

Further analysis of the clustering results reveals that the third category includes all data points where x equals 1, indicating the cluster of individuals with diseases, and these individuals are defined as the chronic disease population. For the first category, it represents individuals who are disease-free and have relatively well-implemented preventive measures in the z-axis dimension. These individuals have a lower risk of getting sick and have spontaneous habits for preventing chronic diseases; therefore, they are defined as individuals with awareness and a low probability of getting sick. The second category represents individuals with severe genetic factors and is defined as a high probability of getting sick. The third category represents individuals with fewer genetic factors but lack active awareness of preventive measures, and they are defined as individuals with no preventive awareness and a low probability of getting sick. The details are explained as follows:

Individuals with chronic diseases: Refers to individuals with chronic diseases, specifically, those with hypertension or diabetes.

High-risk groups: In contrast to the disease population, this group consists of individuals who do not have the disease but have a higher number of disease cases within their families. They are defined as the high probability of getting sick population. Their current disease status is 0, and the model's re-prediction result is also 0, but there are more cases of disease in their families.

Low-risk individuals without preventive awareness: Within the disease-free population, there is a group with a lower family history of disease and varying degrees of preventive awareness. Those who do not actively take preventive measures against chronic diseases are defined as the population with no preventive awareness. This group currently does not have the disease, has a low likelihood of getting sick, and has a low preventive measure index.

Individuals with preventive awareness and low disease risk: In contrast to the group with no preventive awareness, this group has a lower likelihood of getting sick and possesses some level of preventive awareness. Their families have a small number or no cases of disease, and they have good

habits and proactive preventive measures, consciously mitigating the possibility of developing chronic diseases.

Based on the clustering results, the number and proportion of each population group are calculated as Table.2 :

Table 2. The number and proportion of the population classified by residents

Type	Count	Percentage
Individuals with chronic diseases	1235	14.52%
High-risk groups	2778	35.44%
Low-risk individuals without preventive awareness	2065	26.33%
Individuals with preventive awareness and low disease risk	1858	23.70%

It can be observed that, in the overall sample, the populations with a high probability of getting sick and those with no preventive awareness are more numerous and have the largest percentage. For most residents, the risk of chronic diseases looms like a distant cloud, and they may not realize it until it's too late. Therefore, it is necessary to provide specific recommendations for residents in each group: emphasize monitoring and control for those with a high likelihood of getting sick or already suffering from diseases, encourage and educate those with no preventive awareness, and recognize and motivate those with awareness.

3.2. Health recommendations for different population groups

For individuals with chronic diseases: Dietary Structure: Monitor the diet closely, focusing on light and healthy eating.

1. Choose foods that are gentle on the digestive system, such as congee, noodles, well-cooked vegetables, lean meats, or poultry. Avoid fried, high-fat, or spicy foods to reduce gastrointestinal stress.
2. Reduce salt intake for chronic disease patients. Avoid high-sodium foods like pickles, cured meats, and processed foods. Instead, use herbs and low-sodium seasonings to enhance flavor.
3. Opt for fresh fruits and vegetables to ensure adequate nutrition. Choose a variety of colorful vegetables and fruits for diverse nutrients. Be cautious about sugar intake when selecting fruits and opt for low-sugar options like apples, oranges, and strawberries. Avoid excessive consumption of candies, desserts, and sugary beverages.
4. Select lean meats, poultry (skinless), low-fat dairy products, fish, and legumes over high-fat options like fried foods and processed meats.

Exercise: Incorporate moderate exercise to promote physical health.

1. Choose non-strenuous exercises to avoid overexertion, especially weightlifting and intense competitive sports.
2. Opt for simple and easy-to-learn exercise routines to minimize injury risks and ensure proper posture and movements. Gradually increase exercise duration and intensity. Start with shorter durations and lighter exercises, gradually progressing to longer durations and more intense workouts.
3. Aerobic exercise is crucial for controlling blood pressure and promoting cardiovascular health. Activities like walking, jogging, cycling, and swimming are excellent choices. Maintain a heart rate below 100 beats per minute during exercise to avoid excessive strain on the cardiovascular system.

For high-risk groups: Balanced Diet: Focus on a balanced diet.

1. Ensure dietary diversity, including vegetables, fruits, whole grains, healthy proteins (e.g., fish, legumes, nuts), and low-fat dairy products. Avoid excessive consumption of high-sugar, high-salt, and high-fat foods. Watch portion sizes, avoid overeating, and manage calorie intake. Choose cooking methods like steaming, boiling, baking, or stewing over frying to reduce oil consumption. Consume a daily breakfast to maintain energy and metabolism stability. Stay hydrated to maintain water balance in the body.
2. Regular Sleep: Maintain a regular sleep schedule with sufficient sleep time. Avoid staying up late for extended periods to ensure good sleep habits.

3. Exercise: Engage in at least 2-3 hours of moderate-intensity aerobic exercise per week or 75 minutes of high-intensity exercise. Develop a consistent exercise routine to improve cardiovascular health and manage weight. Incorporate more physical activity into daily life, such as taking the stairs, walking instead of driving, or using a bicycle.
4. Smoking: Smoking cessation is crucial for reducing the risk of chronic diseases. If drinking alcohol, do so in moderation: up to two standard drinks per day for men and one for women.

For low-risk individuals without preventive awareness:

1. Health Education: Disseminate information about the risk factors and preventive measures of chronic diseases. Raise awareness and provide health education through activities such as health lectures and exhibitions. Promote a healthy lifestyle through social awareness campaigns and policy support.
2. Regular Check-ups: Encourage and require regular health check-ups for this group, especially high-risk individuals and older adults. Early detection through regular check-ups helps with timely intervention and treatment.
3. Dietary Structure: Control daily salt intake to less than 6 grams, avoid high-sodium foods like pickles and instant noodles. Consume a variety of vegetables and fruits, maintain dietary diversity, and balance various nutrients, including proteins, carbohydrates, and fats.
4. Exercise: Encourage regular physical activity, such as walking, running, swimming, yoga, etc., to increase physical activity levels. Avoid high-intensity exercise to prevent unexpected reactions.
5. Smoking: Smoking cessation is crucial for reducing the risk of chronic diseases. Encourage early smoking cessation to lower disease risks.
6. Public Health Services: Provide public health services and resources, including health counseling, health assessments, and chronic disease management. Establish a health monitoring system to track disease occurrence and trends, enabling timely prevention measures and easy access to health support.

For individuals with preventive awareness and low disease risk:

1. Diet: Maintain a balanced and light diet, avoiding heavy and greasy foods. Incorporate a variety of foods like vegetables, fruits, whole grains, healthy proteins, and low-fat dairy products. Minimize high-sugar, high-salt, and high-fat food consumption. Choose healthy cooking methods like steaming, boiling, baking, or stewing to reduce oil intake.
2. Regular Check-ups: Develop the habit of regular check-ups to detect potential health issues early. Screen for specific diseases, especially for those with a family history or other risk factors, such as hypertension, hyperlipidemia, diabetes, etc. Keep personal health records for easy access and health assessments.
3. Lifestyle: Establish a healthy lifestyle, including smoking cessation and moderation in alcohol consumption (up to two standard drinks per day for men, one for women).

4. Conclusions

In this study, we collected and meticulously analyzed relevant questionnaire data, allowing us to establish a novel three-dimensional mapping model based on residents' health conditions, family medical histories, and preventive measures. This model facilitated a deeper understanding of the interplay between these factors. Utilizing the K-Medoids clustering algorithm, we classified residents into four distinct groups, representing various population types, including those with chronic diseases, high-risk individuals, those with low disease prevention awareness, and those with high awareness. This classification offered valuable insights into the health status and specific needs of these diverse segments. Consequently, we devised tailored health recommendations encompassing dietary adjustments, sensible exercise habits, regular check-ups, smoking cessation, and alcohol moderation. Emphasizing personalized guidance enhanced the precision and effectiveness of our interventions.

References

[1] Naughton B J. The Chinese economy: Adaptation and growth[M]. Mit Press, 2018.

[2] Ebenstein A, Fan M, Greenstone M, et al. Growth, pollution, and life expectancy: China from 1991–2012[J]. American Economic Review, 2015, 105(5): 226-231.

[3] Musaiger A O, Al-Khalifa F, Al-Mannai M. Obesity, unhealthy dietary habits and sedentary behaviors among university students in Sudan: growing risks for chronic diseases in a poor country[J]. Environmental health and preventive medicine, 2016, 21: 224-230.

[4] Cecchini M, Sassi F, Lauer J A, et al. Chronic Diseases: Chronic Diseases and Development 3 Tackling of unhealthy diets, physical inactivity, and obesity: health effects and cost-effectiveness[J]. Lancet, 2010, 376(9754): 1775-1784.

[5] Oyama M, Nakamura K, Tsuchiya Y, et al. Unhealthy maternal lifestyle leads to rapid infant weight gain: prevention of future chronic diseases[J]. The Tohoku Journal of Experimental Medicine, 2009, 217(1): 67-72.

[6] Prasad S, Sung B, Aggarwal B B. Age-associated chronic diseases require age-old medicine: role of chronic inflammation[J]. Preventive medicine, 2012, 54: S29-S37.

[7] Tian M, Wang H, Tong X, et al. Essential public health services' accessibility and its determinants among adults with chronic diseases in China[J]. PLoS One, 2015, 10(4): e0125262.

[8] Tian M, Chen Y, Zhao R, et al. Chronic disease knowledge and its determinants among chronically ill adults in rural areas of Shanxi Province in China: a cross-sectional study[J]. BMC public health, 2011, 11(1): 1-9.

[9] Kuhail M A, Ahmad B, Rottinghaus C. Smart resident: A personalized transportation guidance system[C]//2018 IEEE 5th International Congress on Information Science and Technology (CiSt). IEEE, 2018: 547-551.

[10] Sharon T. Self-tracking for health and the quantified self: Re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare[J]. Philosophy & Technology, 2017, 30(1): 93-121.