Postprint: A Study on Multimorbidity Patterns of Chronic Diseases Among Chinese Older Adults

Authors: Pan Ye, Liu Zhihui, Hu Qianqian, Wang Liuyi, Wang Liuyi

Date: 2023-06-08T00:00:00+00:00

Abstract

Background With population aging and extended life expectancy, multimorbidity of chronic diseases is becoming increasingly prevalent. The multiplicity of disease types and complexity of conditions pose challenges to health management for older adults. As an essential research topic, multimorbidity patterns are relatively understudied domestically.

Objective To investigate multimorbidity patterns of common chronic diseases among Chinese older adults, and to assist policymakers, researchers, and clinicians in better understanding the current status of multimorbidity.

Methods Respondents aged 60 years and above were selected from the 2018 China Health and Retirement Longitudinal Study (CHARLS) dataset, utilizing demographic characteristic data and data on 14 chronic diseases from the health status section. Four methods—association rule mining, cluster analysis, principal component analysis, and latent class analysis—were employed to explore multimorbidity patterns among Chinese older adults, and results obtained from different methods were compared.

Results Data from a total of 10,800 respondents were included. While the patterns identified by the four methods differed, consistent multimorbidity patterns emerged: (1) hypertension, diabetes or elevated blood glucose, dyslipidemia; (2) chronic lung disease and asthma; (3) arthritis or rheumatism, stomach disease or digestive system disease; (4) stroke, memory-related diseases.

Conclusion The consistent patterns identified across different methods demonstrate clear etiological relationships among the included chronic diseases; differences in multimorbidity patterns arise from the complex etiological relationships involved and the distinct principles underlying each method.

ChinaRxiv

Full Text

Patterns of Coexistence of Multiple Chronic Conditions among Chinese Elderly

PAN Ye[1], LIU Zhihui[2], HU Qianqian[3], WANG Liuyi[2]*

[1]Henan Integrative Medicine Hospital, Zhengzhou 450004, China
[2]Department of General Practice, Henan Provincial People's Hospital, Zhengzhou 450003, China
[3]Henan University People's Hospital, Zhengzhou 450003, China

Corresponding author: WANG Liuyi, Chief physician; Email: wly2000@126.com

Abstract

Background: With population aging and increased longevity, comorbid chronic diseases have become increasingly prevalent. The variety and complexity of these conditions pose significant challenges to health management for older adults. While multimorbidity patterns represent a critical research area, such studies remain relatively scarce in China. **Objective:** To investigate common multimorbidity patterns among Chinese elderly and help policymakers, researchers, and clinicians better understand the current status of multimorbidity. **Methods:** Data on demographic characteristics and 14 chronic conditions were extracted from respondents aged 60 years and above in the 2018 China Health and Retirement Longitudinal Study (CHARLS). Four analytical methods—association rules, cluster analysis, principal component analysis, and latent class analysis—were employed to explore multimorbidity patterns, with results compared across methods. **Results:** A total of 10,800 respondents were included. While patterns varied across the four methods, consistent multimorbidity patterns emerged: (i) hypertension, diabetes or elevated blood glucose, and dyslipidemia; (ii) chronic lung disease and asthma; (iii) arthritis or rheumatism, and stomach or digestive diseases; and (iv) stroke and memory-related diseases. **Conclusion:** The consistent patterns identified across different methods demonstrate clear etiological relationships among the included chronic conditions; differences in patterns arise from complex etiological relationships and varying methodological principles.

Keywords: chronic disease; multiple chronic conditions; comorbidity; association rules; clustering analysis; latent class analysis

Funding: Henan Provincial Traditional Chinese Medicine Science Research Special Project (2023ZYZD13); Henan Provincial Traditional Chinese Medicine Culture and Management Research Project (TCM2023005)

1. Materials and Methods

1.1 Data Sources

The China Health and Retirement Longitudinal Study (CHARLS) aims to collect high-quality micro-level data representative of Chinese middle-aged and elderly households and individuals to analyze population aging issues and promote interdisciplinary research on aging [10-12]. We selected the Harmonized_{CHARLS}D version, extracting 2018 data for analysis. The dataset included demographic characteristics and health status data on 14 chronic conditions. After data cleaning, we screened respondents aged 60 years and above without deleting missing values, yielding 10,800 respondents. Missing values in the 14 chronic conditions (relatively few) were imputed using the mode. Demographic characteristics were not included in pattern analysis and were presented as missing values in descriptive statistics.

1.2 Definition of Chronic Conditions

The 14 chronic conditions selected from CHARLS data were: hypertension; diabetes or elevated blood glucose (including impaired glucose tolerance and elevated fasting blood glucose); cancer or malignant tumors (excluding mild skin cancer); chronic lung diseases such as chronic bronchitis or emphysema, cor pulmonale (excluding tumors or cancer, hereafter referred to as chronic lung disease); heart disease (including myocardial infarction, coronary heart disease, angina, congestive heart failure, and other cardiac diseases); stroke; emotional or mental health problems; arthritis or rheumatism; dyslipidemia (hyperlipidemia or hypolipidemia); liver disease (excluding fatty liver, tumors, or cancer); kidney disease (excluding tumors or cancer); stomach or digestive diseases (excluding tumors or cancer); asthma; and memory-related diseases (such as dementia, brain atrophy, Parkinson's disease).

Demographic variables included: sex (male, female), age, household residence type (urban, rural), respondent's household registration type (urban, rural), education level (below primary school, primary and junior high school, high school and vocational training, college and above), and marital status (married, separated or divorced, widowed, unmarried).

1.3 Analytical Methods

Data processing and all analyses were implemented through R programming, utilizing multiple R packages to complete the four types of analysis. General descriptive statistics included respondent demographics, incidence rates of the 14 chronic conditions, multimorbidity rates, average number of chronic conditions, and variance to understand population characteristics and disease prevalence.

1.3.1 Association Rules Association rules were used to identify common two-condition and three-condition combinations. Support represented the observed incidence of disease combinations, while lift served as an independence indicator. For two-condition combinations, lift was calculated as observed incidence divided by expected incidence. Setting minimum thresholds for the number of items in antecedent and consequent sets, along with minimum support, yielded combinations with higher observed incidence. Expected incidence was calculated to determine disease independence through lift values.

1.3.2 Cluster Analysis Cluster analysis classifies observations based on quantitative characteristics, with two types: case clustering and variable clustering. The statistic reflecting similarity between cases or variables is called distance, with multiple calculation methods available. The 14 chronic conditions were processed as binary data (1 = condition present, 0 = absent), forming an $n \times p$ binary matrix where n = number of respondents (10,800) and p = number of conditions (14). Case clustering classified respondents based on an $n \times n$ distance matrix calculated from the 14 conditions. Systematic clustering of the 14 conditions used a $p \times p$ distance matrix based on co-occurrence patterns to classify diseases.

We employed k-medoids algorithm for respondent clustering and hierarchical clustering for the 14 chronic conditions to distinguish multimorbidity patterns. Distance selection for binary data depends on the relative weighting of positive matches (both 1) and negative matches (both 0). This study used Jaccard distance and Yule's Q distance—commonly used for binary data. Yule's Q incorporates information from both positive and negative matches, better reflecting all relevant information in binary data [13], while Jaccard distance emphasizes only positive matches [14]. Since multimorbidity patterns focus on positive matches (co-occurrence) and the sample included healthy individuals (all zeros), Yule's Q could not be calculated when one record contained all zeros. After comparing both methods and considering clinical significance, Yule's Q distance was selected for hierarchical clustering of chronic conditions, and Jaccard distance for k-medoids clustering of respondents. The number of clusters for hierarchical clustering was determined through dendrogram interpretation combined with clinical meaning, using the average linkage method. The optimal number of clusters for k-medoids was determined using the fviz_{nbclust} function from the factoextra package.

1.3.3 Principal Component Analysis Variables reflect information to varying degrees and exhibit correlations. Too many variables increase computational complexity and problem complexity. Principal component analysis applies appropriate mathematical transformations to create new variables as linear combinations of original variables, selecting a few principal components that account for large proportions of total variance. This study used varimax-rotated principal component analysis, determining the number of components by variance contribution and reducing 14 chronic conditions to fewer disease combinations. Conditions with absolute loading coefficients >0.36 were selected to identify component composition.

1.3.4 Latent Class Analysis (LCA) LCA classifies respondents based on their status across the 14 chronic conditions. Similar to cluster analysis, LCA groups individuals by similarity but does not define inter-class distances or select clustering methods. Classification is based on observed variable probabilities. All possible classifications were examined, with model fit assessed using indicators such as Likelihood-ratio, G^2 statistic, and Bayesian Information Criterion (BIC). Analysis was conducted using the poLCA package in R, with final class determination based on BIC values.

---

2. Results

2.1 Characteristics of the Study Population

The final sample included 10,800 respondents aged 60 and above. The population had slightly more females (51.3%), with most residing in rural areas (60.3%) and holding rural household registration (74.0%). Education levels were generally low, with 90.7% having no education beyond junior high school. Most respondents were married (77.9%), consistent with national demographic characteristics [15]. Overall, 86.2% had one or more chronic conditions, and 49.0% had 2–4 chronic conditions .

All 14 chronic conditions had multimorbidity rates above 85% . Hypertension had the highest incidence (47.14%). Emotional or mental health problems and memory-related diseases had the highest average numbers of comorbid conditions (4.16 and 4.17, respectively). While stroke, liver disease, and kidney disease had relatively low incidence rates, their average numbers of comorbidities were relatively high.

2.2 Association Rule Analysis

Table 3 presents all two-condition combinations with support >10%, and Table 4 shows three-condition combinations with support >5%. Based on support alone, association rule-derived comorbidity combinations reflect simple co-occurrence. For example, arthritis or rheumatism with hypertension had high co-occurrence in older adults but are etiologically independent, yielding an observed/expected incidence ratio close to 1. Dyslipidemia with heart disease showed high co-occurrence, and the observed/expected incidence ratio indicated an association, suggesting related etiologies.

For three-condition combinations, the ratio of observed to expected incidence and lift values were used to assess inter-disease relationships. Two groups showed high observed/expected incidence ratios and lift values: (1) heart disease, dyslipidemia, and hypertension; and (2) diabetes or elevated blood glucose, dyslipidemia, and hypertension—both showing etiological and risk factor associations.

2.3 Cluster Analysis

The dendrogram of 14 chronic conditions (Figure 1) grouped them into six clusters based on distance proximity. Yule's Q distance reflects disease occurrence information across respondents; conditions with closer distances had higher co-occurrence and shared pathogenic factors. For example, asthma and chronic lung disease are both respiratory diseases; dyslipidemia, diabetes or elevated blood glucose, stroke, and hypertension are all vascular/circulatory diseases; memory-related diseases and emotional or mental health problems are psychiatric conditions; stomach or digestive diseases and arthritis or rheumatism are metabolic system diseases. Heart disease clustered closely with memory-related diseases and emotional or mental health problems, forming a psychiatric cluster. Cancer or malignant tumors were independent of other conditions, forming a separate cluster.

For respondent clustering, the within-cluster sum of squared errors indicated the optimal number of clusters was four. Post-clustering, each respondent was assigned to one cluster. Disease prevalence and condition counts for the four clusters are shown in Table 5. Since k-medoids clustering using Jaccard distance emphasizes co-occurrence information, Cluster 1 included all healthy respondents and those with 1–2 low-comorbidity conditions (relatively healthy). Cluster 2 showed substantially higher prevalence of respiratory diseases and stomach/digestive diseases than the overall population, with slightly higher arthritis/rheumatism prevalence. In hierarchical clustering, respiratory diseases formed one independent cluster, as did stomach/digestive diseases and arthritis/rheumatism; patients with random co-occurrence of these four conditions primarily fell into Cluster 2. Cluster 3 had a high proportion of respondents with 3–5 comorbidities, with substantially higher prevalence of heart disease and dyslipidemia, and slightly higher hypertension and stroke prevalence. Common 3–5 condition comorbidities were primarily cardiovascular diseases (hypertension, diabetes/elevated blood glucose, dyslipidemia, heart disease), suggesting Cluster 3 represents this disease combination. All respondents in Cluster 4 had hypertension, with an average of 1.82 comorbidities, and lower prevalence of other conditions than the overall population, suggesting this cluster includes respondents with hypertension alone or hypertension combined with non-cardiovascular conditions.

2.4 Principal Component Analysis

The scree plot (Figure 2) indicated five principal components. Setting loading coefficient absolute values >0.36 yielded non-overlapping representative conditions for each component, with conditions in the same component showing strong associations. Component 1 included hypertension, diabetes or elevated blood glucose, heart disease, and dyslipidemia. Component 2 included chronic lung disease and asthma. Component 3 included arthritis or rheumatism and stomach or digestive diseases. Component 4 included stroke, emotional or mental health problems, and memory-related diseases. Component 5 included cancer or malignant tumors, liver disease, and kidney disease.

2.5 Latent Class Analysis

The minimum BIC criterion indicated the optimal number of classes was five. Analysis of the conditional probability distribution plot (Figure 3) and cross-tabulation of classes with chronic conditions and condition counts (Table 7) revealed: Class 5 was the relatively healthy group, including all respondents with no conditions or only one condition, most with two conditions, and with lower prevalence of all 14 conditions than the overall population. Class 1 was the metabolic system disease group, with high prevalence of arthritis or rheumatism and stomach/digestive diseases. Class 2 was the vascular/circulatory disease group (hypertension, diabetes/elevated blood glucose, heart disease, dyslipidemia). Class 3 was the high-multimorbidity group, with 5+ chronic conditions, relatively small in size, and higher prevalence of multiple conditions. Class 4 was the respiratory disease group, with high prevalence of chronic lung disease (chronic bronchitis, emphysema, cor pulmonale) and asthma.

---

3. Discussion

3.1 Differences and Consistency in Multimorbidity Patterns Across Methods

Multimorbidity concepts vary; this study specifically refers to "multimorbidity" as "the coexistence of acute or chronic conditions in one individual" [2], emphasizing co-occurrence without requiring etiological relationships or designation of an index disease. The consistent patterns identified across methods were: (i) hypertension, diabetes or elevated blood glucose, and dyslipidemia; (ii) chronic lung disease and asthma; (iii) arthritis or rheumatism and stomach or digestive diseases; and (iv) stroke and memory-related diseases. Different methods yielded varying combinations; for example, heart disease clustered closely with pattern (i) in principal component and latent class analyses but was more distantly associated in hierarchical clustering. Malignant tumors were independent of other conditions across multiple methods. Emotional or mental health problems had low prevalence and showed some associations with other conditions in certain methods but without stable comorbidity patterns. Kidney and liver diseases appeared in some classifications but showed relatively independent etiological relationships.

Methodological principles differed substantially. Association rules identify specified-size disease combinations through antecedent/consequent set control, using lift to assess independence and distinguish co-occurrence types, though support only reflects observed incidence. Hierarchical clustering and principal component analysis emphasize disease co-occurrence associations, yielding clearer etiological relationships. K-medoids and LCA focus on classifying individuals to reflect comorbidity combinations, with k-medoids using distance-based spatial partitioning (results depend on distance algorithms) and LCA providing a more objective statistical approach that classifies individuals based on different response patterns across manifest variables to identify population heterogeneity [16]. Both methods derive comorbidity combinations through population prevalence.

3.2 Comparison with Similar Studies

Our findings both differ from and align with domestic studies. Liu et al. [8] used the same data and method but deleted respondents with missing data and included ages 45+, yielding lower incidence rates. Our study retained missing data, producing results closer to the overall CHARLS respondent population and Chinese adults aged 60+. Despite data differences, conclusions were consistent regarding hypertension's numerous disease combinations and high lift values for vascular/circulatory two-condition combinations. Li et al. [17] also used 2018 CHARLS data (different version with slight data quantity differences), identifying two- and three-condition combinations through co-occurrence network analysis; our data and conclusions align with theirs. Sun et al. [7] used China Kadoorie Biobank (CKB) data and found that diabetes, coronary heart disease, stroke, and hypertension clustered together in baseline surveys, consistent with our results. However, their follow-up survey conclusions differed, as did their findings regarding malignant tumors' relationships with other conditions—likely due to different disease definitions and study populations.

Inconsistencies across similar studies stem from different disease definitions, respondent and interviewer cognition of diseases, and inherent data errors. Our finding that malignant tumors lack clear comorbidity characteristics with other conditions differs from Sun et al. [7], possibly due to different tumor definitions and populations. Association rules identified hypertension with arthritis/rheumatism as the highest-observed two-condition combination, yet this pairing did not cluster together in hierarchical clustering, respondent k-medoids clustering, or principal component analysis—reflecting etiological independence despite high random co-occurrence probability. In contrast, association rule-derived combinations of dyslipidemia-hypertension, heart disease-hypertension, and diabetes/hypertension showed observed/expected incidence ratios >1, indicating non-independent etiological relationships that other methods also grouped together. The stomach/digestive diseases and arthritis/rheumatism combination also showed observed/expected ratio >1 and appeared across other methods. While association rules are straightforward, focusing solely on incidence may miss strongly associated but low-prevalence patterns (e.g., chronic lung disease and asthma), which other methods successfully identified despite asthma's low prevalence limiting co-occurrence rates.

3.3 Strengths and Limitations

This study's strengths include a relatively large sample size compared to domestic studies, application of multiple methods for multimorbidity classification, and trial of novel algorithms. Beyond hierarchical clustering's dendrogram-clinical integration for determining cluster number, respondent k-medoids clustering used data-driven methods, principal component analysis used scree plots, and LCA used comprehensive indicator evaluation to select class numbers, reducing subjective judgment. Domestic multimorbidity studies rarely use distance algorithms like Jaccard and Yule's Q, which better capture information in binary data.

Limitations include using mode imputation for missing chronic condition data, which may affect incidence accuracy and potentially underestimate true rates. Retained missing demographic data were not processed or linked to individual classifications, precluding further analysis of risk factors and prognosis, limiting clinical relevance.

Author Contributions: WANG Liuyi conceptualized the study, provided overall research direction, ensured feasibility, and assumed final responsibility for manuscript supervision. PAN Ye performed data cleaning, processing, statistical analysis, R programming, and figure/table generation. PAN Ye, LIU Zhihui, and HU Qianqian analyzed and interpreted results and drafted the manuscript. PAN Ye and LIU Zhihui revised the manuscript.

Conflict of Interest: The authors declare no conflicts of interest.

References

[1] World Health Organization. The World Health Report 2008—primary health care (now more than ever) [M]. New York: World Health Report, 2008: 148.

[2] CHANG F, ZHANG R, LU Y, et al. Analysis and design of comorbidity concept system [J]. Chinese General Practice, 2018, 21(3): 256-260. DOI: 10.3969/j.issn.1007-9572.2017.00.219.

[3] WU F, GUO Y F, KOWAL P, et al. Prevalence of major chronic conditions among older Chinese adults: the study on global AGEing and adult health (SAGE) wave 1 [J]. PLoS One, 2013, 8(9): e74176. DOI: 10.1371/journal.pone.0074176.

[4] CORNELL J E, PUGH J A, WILLIAMS J W, et al. Multimorbidity clusters: clustering binary data from a large administrative medical database [J]. Appl Multivar Res, 2009, 12(3): 163. DOI: 10.22329/amr.v12i3.658.

[5] SCHÄFER I, VON LEITNER E C, SCHÖN G, et al. Multimorbidity patterns in the elderly: a new approach of disease clustering identifies complex interrelations between chronic conditions [J]. PLoS One, 2010, 5(12): e15941. DOI: 10.1371/journal.pone.0015941.

[6] GARCÍA-OLMOS L, SALVADOR C H, ALBERQUILLA Á, et al. Comorbidity patterns in patients with chronic diseases in general practice [J]. PLoS One, 2012, 7(2): e32141. DOI: 10.1371/journal.pone.0032141.

[7] SUN Z J, FAN J N, YU C Q, et al. Analysis of multimorbidity prevalence characteristics among adults in 10 regions of China [J]. Chinese Journal of Epidemiology, 2021, 42(5): 755-762.

[8] LIU B B, TIAN Q F, GUO J L. Analysis of chronic disease prevalence and multimorbidity patterns among middle-aged and elderly populations in China [J]. Medicine and Society, 2022, 35(8): 58-61, 66.

[9] OLAYA B, MONETA M V, CABALLERO F F, et al. Latent class analysis of multimorbidity patterns and associated outcomes in Spanish older adults: a prospective cohort study [J]. BMC Geriatr, 2017, 17(1): 186. DOI: 10.1186/s12877-017-0586-1.

[10] ZHAO Y H, STRAUSS J, YANG G H, et al. (2013). China Health and Retirement Longitudinal Study: 2011-2012 National Baseline User's Guide, National School of Development, Peking University.

[11] ZHAO Y H, STRAUSS J, CHEN X X, et al. (2020). China Health and Retirement Longitudinal Study Wave 4 User's Guide, National School of Development, Peking University.

[12] ZHAO Y, HU Y, SMITH J P, et al. Cohort profile: the China health and retirement longitudinal study (CHARLS) [J]. Int J Epidemiol, 2014, 43(1): 61-68. DOI: 10.1093/ije/dys203.

[13] LLOYD B P, KENNEDY C H, YODER P J. Quantifying contingent relations from direct observation data: transitional probability comparisons versus yule's Q [J]. J Appl Behav Anal, 2013, 46(2): 479-497. DOI: 10.1002/jaba.45.

[14] A note on the triangle inequality for the Jaccard distance [J]. Pattern Recognit Lett, 2019, 120: 36-38. DOI: 10.1016/j.patrec.2018.12.007.

[15] YANG H M. New trends in population aging and new characteristics of elderly population in China [J]. Population Research, 2022, 46(5): 104-116. DOI: 10.3969/j.issn.1000-4149.2019.01.006.

[16] WEN Z L, XIE J Y, WANG H H. Principles, steps, and procedures of latent class models [J]. Journal of East China Normal University: Educational Sciences, 2023, 41(1): 1-15. DOI: 10.16382/j.cnki.1000-5560.2023.01.001.

[17] LI Y N, WANG Y Q. Study on the status and patterns of chronic disease multimorbidity among Chinese elderly [J]. Chinese General Practice, 2021, 24(31): 3955-3962, 3978. DOI: 10.12114/j.issn.1007-9572.2021.00.295.

(Received: February 24, 2023; Revised: May 10, 2023)
(Editor: JIA Mengmeng)

Note: Figure translations are in progress. See original paper for figures.