Prevalence and Determinants of Multimorbidity in Older Chinese Adults: A Nationwide Cross-Sectional Study Using CLASS Data

Yi Li[1,†], Chenxi Zhao[2,3,†], Xinhao Zhuang[1,†], Songbo Jia[1], Xuelin Zhang[1], Liming Yan[4], Huaguang Zheng[5], Yiqun Pang[6], Bing Zhan[1], Xiaotian Li[1], Jinxi Zhang[1], John S. Ji[7], Yanjun Guo[8,*], Bingxiang Xu[3,*], Zhixiong Zhou[1,*] and Qingzhu Yang[1,*]

[1] Capital University of Physical Education and Sports, Beijing, 100191, China

[2] School of Sciences, Hebei University of Technology, Tianjin, 300401, China

[3] Key Laboratory of Hebei Province for Molecular Biophysics, Institute of Biophysics, School of Health Science & Biomedical Engineering, Hebei University of Technology, Tianjin, 300130, China

[4] School of Medicine, Tsinghua University, Beijing, 100084, China

[5] Beijing Tiantan Hospital, Capital Medical University, Beijing, China

[6] School of Physical Education, Southwest Petroleum University, Chengdu, Sichuan Province, 610000, China

[7] Vanke School of Public Health, Tsinghua University, Beijing, 100084, China

[8] School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei Province, 430030, China

[*]Correspondence: Qingzhu Yang (yangqingzhu@cupes.edu.cn), Zhixiong Zhou (zhouzhixiong@cupes.edu.cn), Bingxiang Xu (xubingxiang@sus.edu.cn), Yanjun Guo (yanjunguo@hust.edu.cn)

[†] These authors contributed equally: Yi Li, Chenxi Zhao, Xinhao Zhuang

Supplementary Methods

1. Variable Selection using LASSO Regression

To optimize feature selection, prevent overfitting, and mitigate potential multicollinearity among the independent variables, we applied the least absolute shrinkage and selection operator (LASSO) regression prior to constructing the Generalized Linear Mixed Model (GLMM). We conducted a 10-fold cross-validation to identify the optimal penalty parameter (λ). To prioritize a more parsimonious model with strong predictive performance, the λ value located at one standard error from the minimum cross-validation error (λ 1se) was utilized. During this procedure, all categorical variables were transformed into dummy variables. Out of 12 candidate sociodemographic and lifestyle variables, 11 were retained with non-zero coefficients. Smoking status was excluded during this penalization process. The retained variables were subsequently incorporated into the GLMM analysis (Supplementary Figure S2).

2. Random Forest Model Training and Hyperparameter Tuning

To complement the individual-level GLMM, a Random Forest (RF) algorithm was employed to evaluate the relative importance of county/community-level associated factors. The optimal hyperparameters for the RF model were determined using a comprehensive grid search strategy combined with 5-fold cross-validation, implemented via the Scikit-learn library in Python. The final optimized configuration included n_estimators = 100, max_depth = 30, min_samples_split = 2, min_samples_leaf = 1, and bootstrap = True (Supplementary Table S1). Model interpretation and feature ranking were performed using permutation importance and Shapley Additive exPlanations (SHAP) analysis.

3. Detailed Procedures for Sensitivity Analyses

To ensure the rigorousness of our individual-level findings and address potential methodological biases, we designed and executed four predefined sensitivity analyses within the GLMM framework:

1) Alternative Diagnostic Threshold: We elevated the threshold for defining multimorbidity from ≥2 to ≥3 concurrent chronic conditions to test the stability of the socio-demographic and lifestyle correlates under a stricter diagnostic criterion (Supplementary Figure S6).
2) Restriction to Core Systemic Conditions: To address concerns regarding disease heterogeneity and potential clinical diagnostic inconsistency, we restricted our disease count from the original 22 conditions to 18 core systemic conditions. Potentially non-systemic, localized, or controversial conditions—specifically deafness, reproductive system diseases, gastroenteritis, and tuberculosis—were excluded (Supplementary Figure S7).
3) Age-Stratified Analysis for Reverse Causality: To explore the potential for reverse causality—particularly regarding the positive association between receiving multiple social security benefits and higher multimorbidity odds—we stratified the cohort into relatively younger (60–68 years) and older (≥69 years) subgroups. This aimed to differentiate whether the association reflects socio-economic affluence or acts as a proxy for underlying vulnerability and disability-driven welfare eligibility (Supplementary Table S2).
4) Exclusion of Cognitive Impairment-Related Conditions: Self-reported comorbidity data can be susceptible to recall bias, especially among the oldest-old. To verify the accuracy of our primary associations, we conducted a sensitivity analysis restricting the cohort to participants without cognitive impairment-related conditions (excluding those reporting dementia, Parkinson's disease, or other severe neurological disorders) (Supplementary Figure S8).

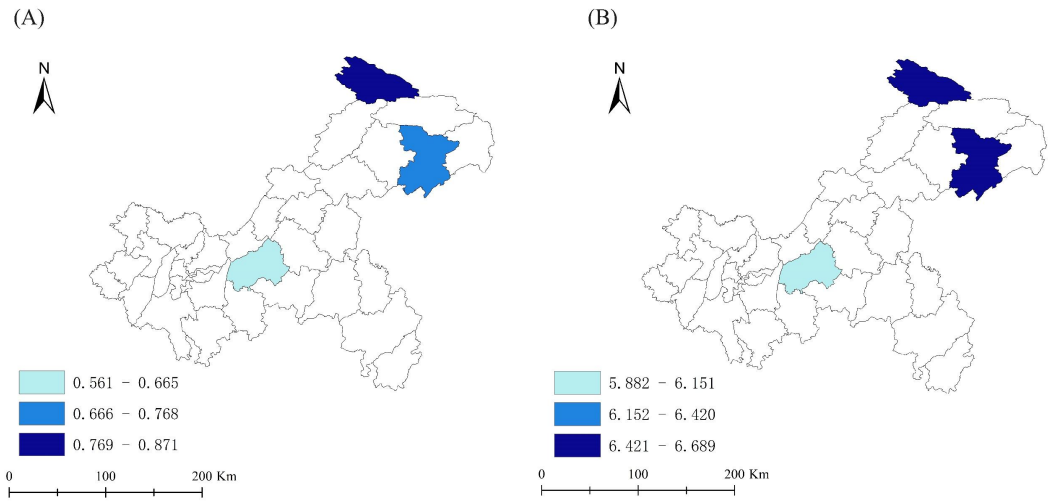
Supplemental data

Supplementary Table S1. Optimal hyperparameters for RF algorithms. The values of bootstrap, maximum depth, minimum samples per leaf, minimum samples for split, and number of estimators optimized through a 5-fold cross-validation grid search are shown.

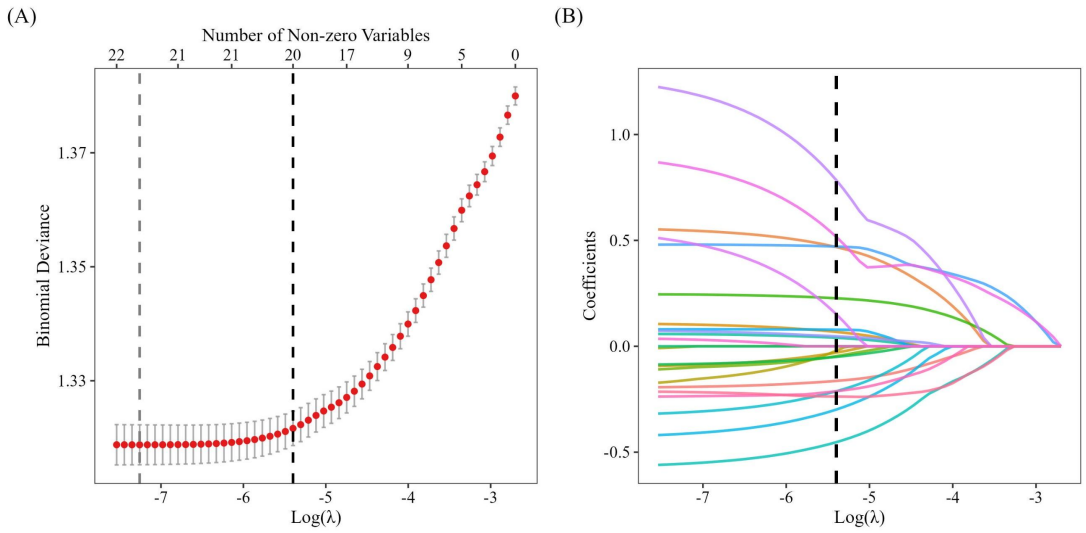
hyperparameters	value
bootstrap	True
max_depth	30
min_samples_leaf	1
min_samples_split	2
n_estimators	100

Supplementary Table S2. Age-stratified analysis of the association between receiving multiple social security benefits and multimorbidity. The Odds Ratios (OR), 95% Confidence Intervals (CI), and P-values derived from the Generalized Linear Mixed Model (GLMM) were shown for the relatively younger (60–68 years) and older (≥69 years) subgroups.

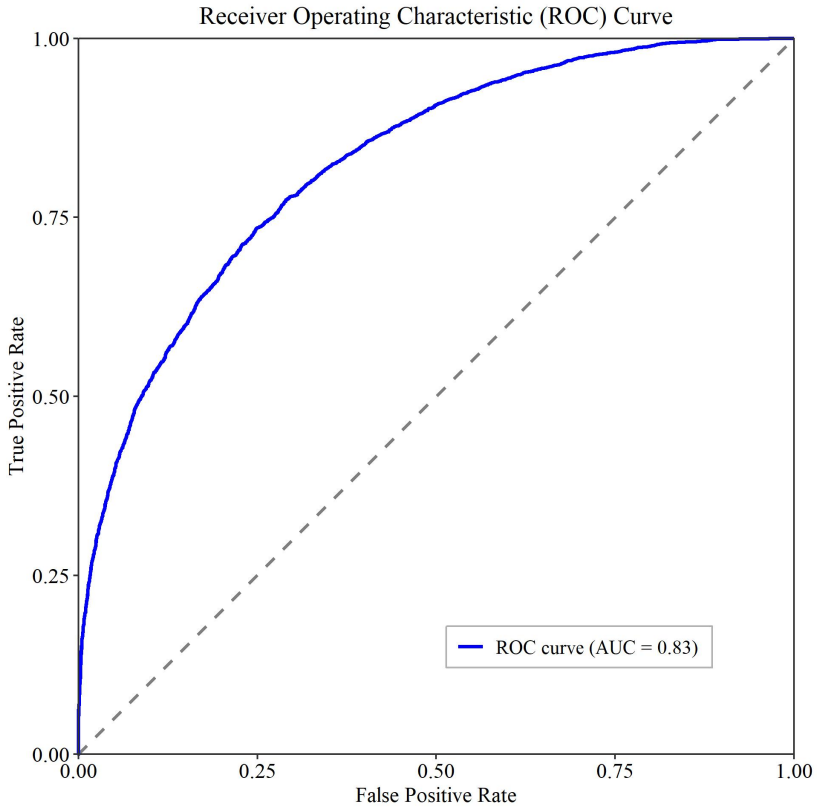
Age Subgroup	OR (95% CI)	*P*-value
60-68 years (Relatively younger/healthier)	0.17 (0.03-1.05)	0.057
⩾69 years (Older/more vulnerable)	1.86 (1.31-2.63)	<0.001



Supplementary Figure S1. Geographical distribution of the medical resources per 1,000 people in Chongqing in 2020. (A) The number of medical and health institutions per 1,000 people in Chongqing (B) The number of beds in medical and health institutions per 1,000 people in Chongqing

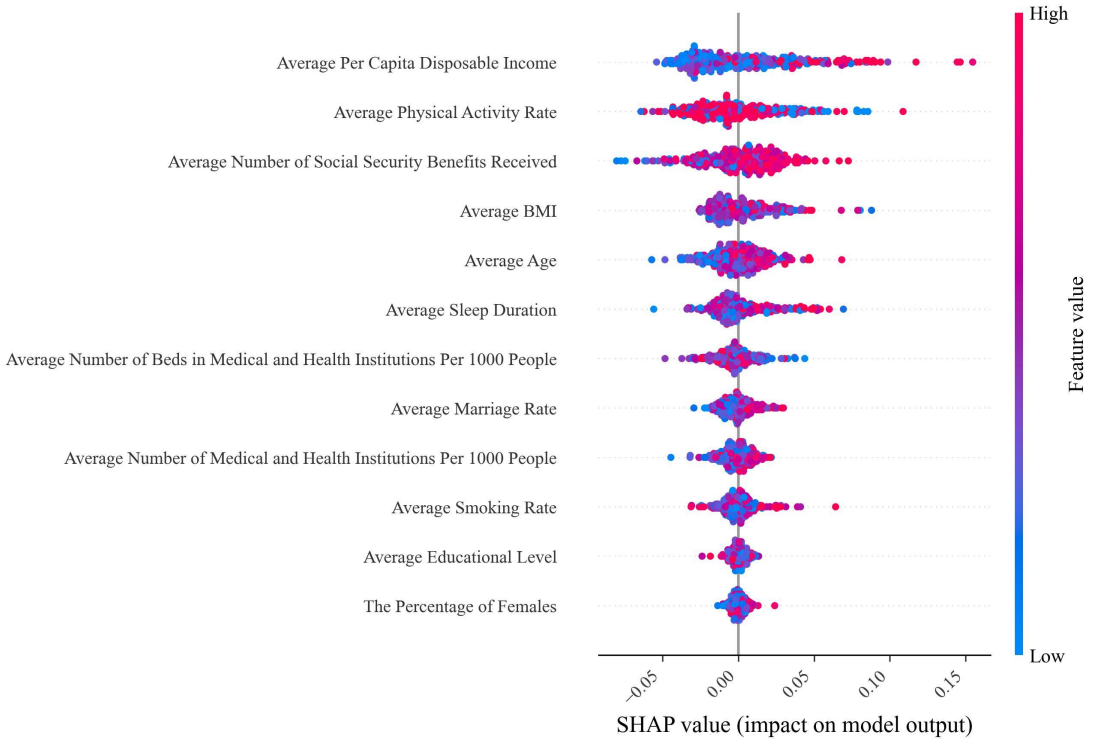


Supplementary Figure S2. LASSO regression model for variable selection. (A) The 10-fold cross-validation curve. The red dots indicate the binomial deviance values, and the grey error bars represent the standard errors across different values of the penalization parameter λ. The vertical dashed lines indicate the optimal λ values. (B) The LASSO coefficient path plot. Each colored line represents the trajectory of an individual candidate variable's coefficient as a function of log(λ).

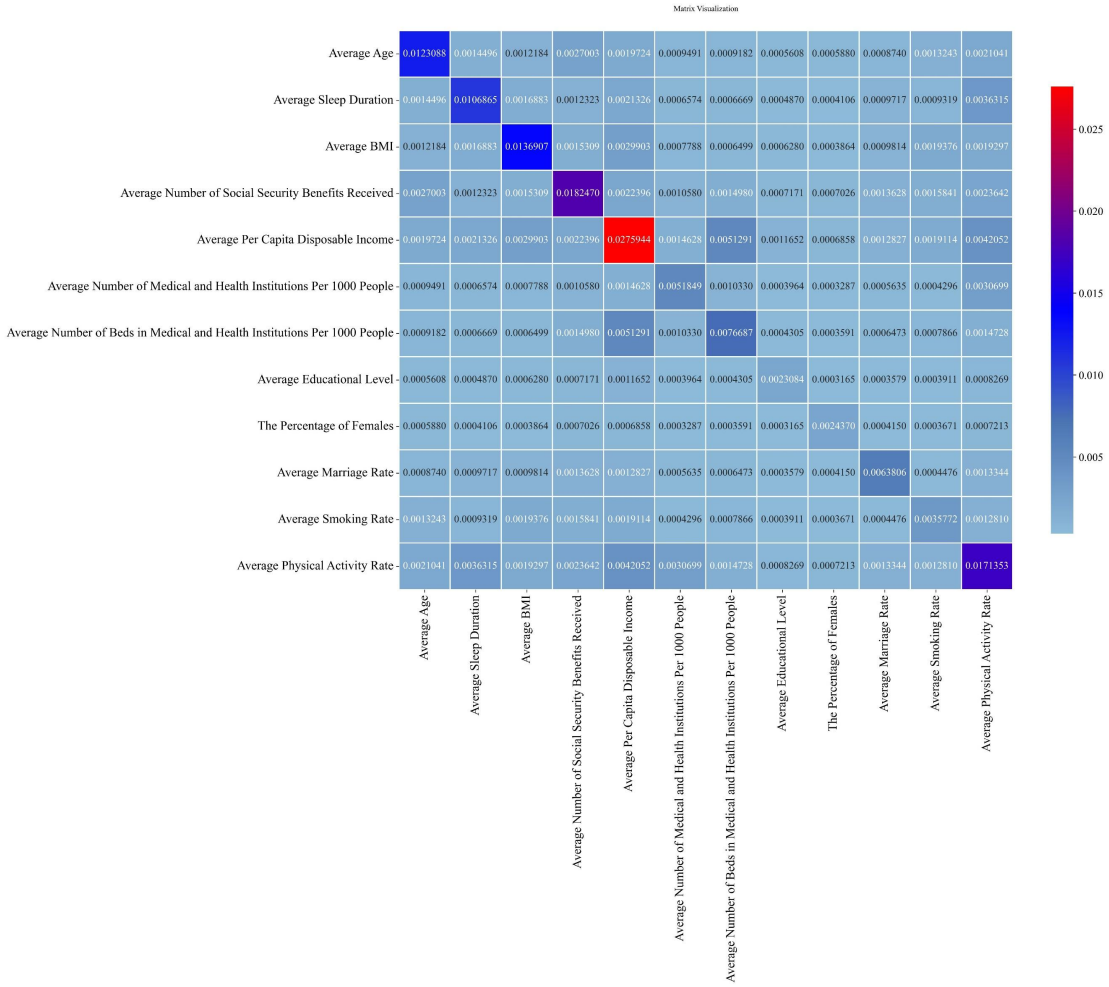


Supplementary Figure S3. ROC curves and AUC values of GLMM. The blue line represents the ROC curve for predicting individual-level multimorbidity. The dashed grey line represents the diagonal

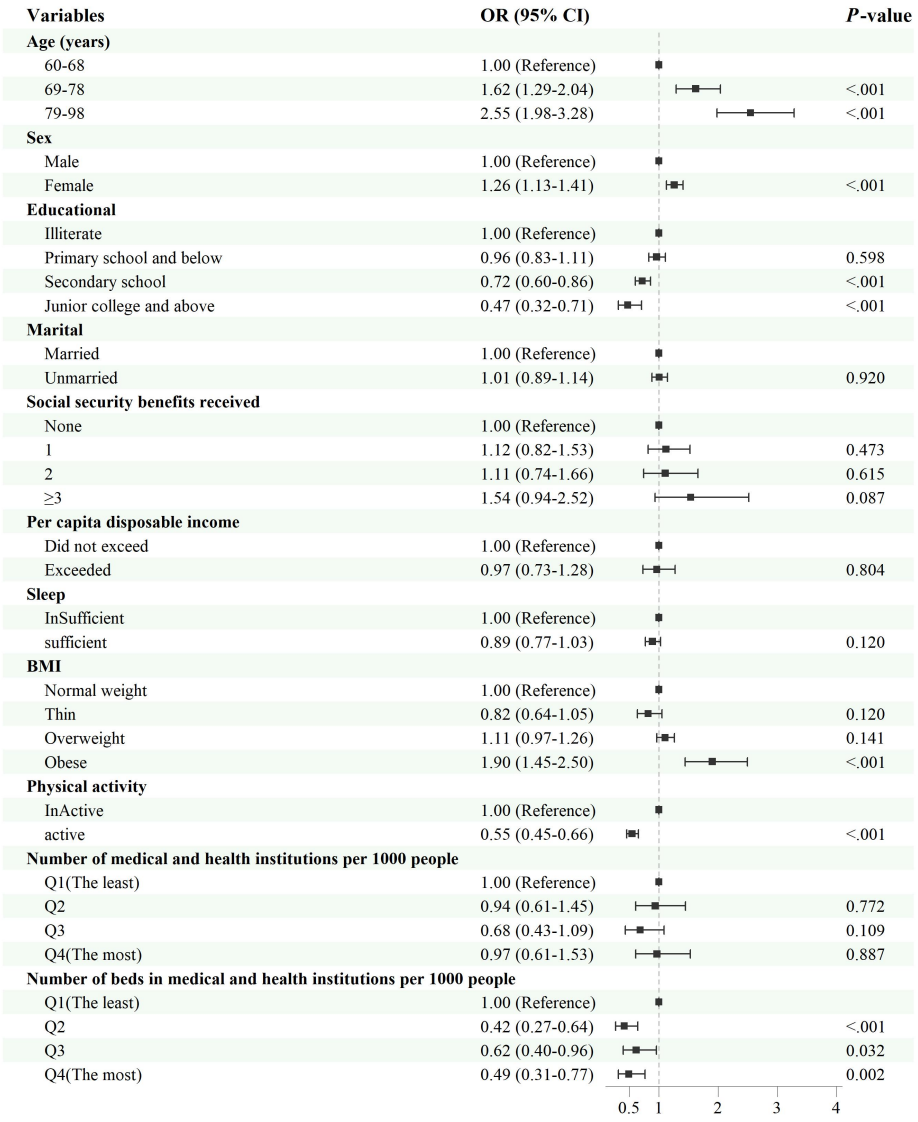
line of no discrimination. The exact Area Under the Curve (AUC) value is marked in the legend.



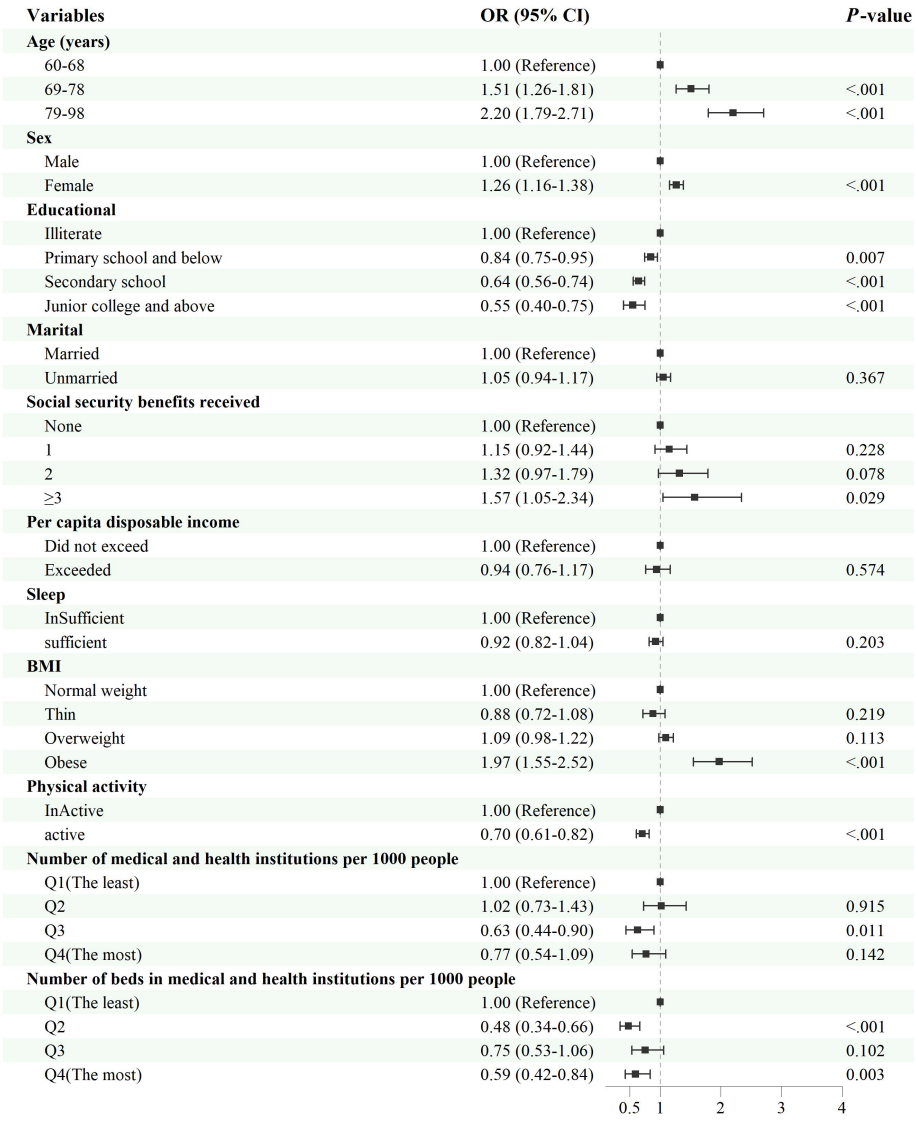
Supplementary Figure S4. SHAP value of RF. The y-axis ranks the predictive factors based on their overall importance in the Regional Random Forest model. The x-axis represents the SHAP value, indicating the impact of each feature on the model's output. Each dot represents a specific regional sample. The color gradient from blue to red represents the actual numeric value of the feature from low to high.



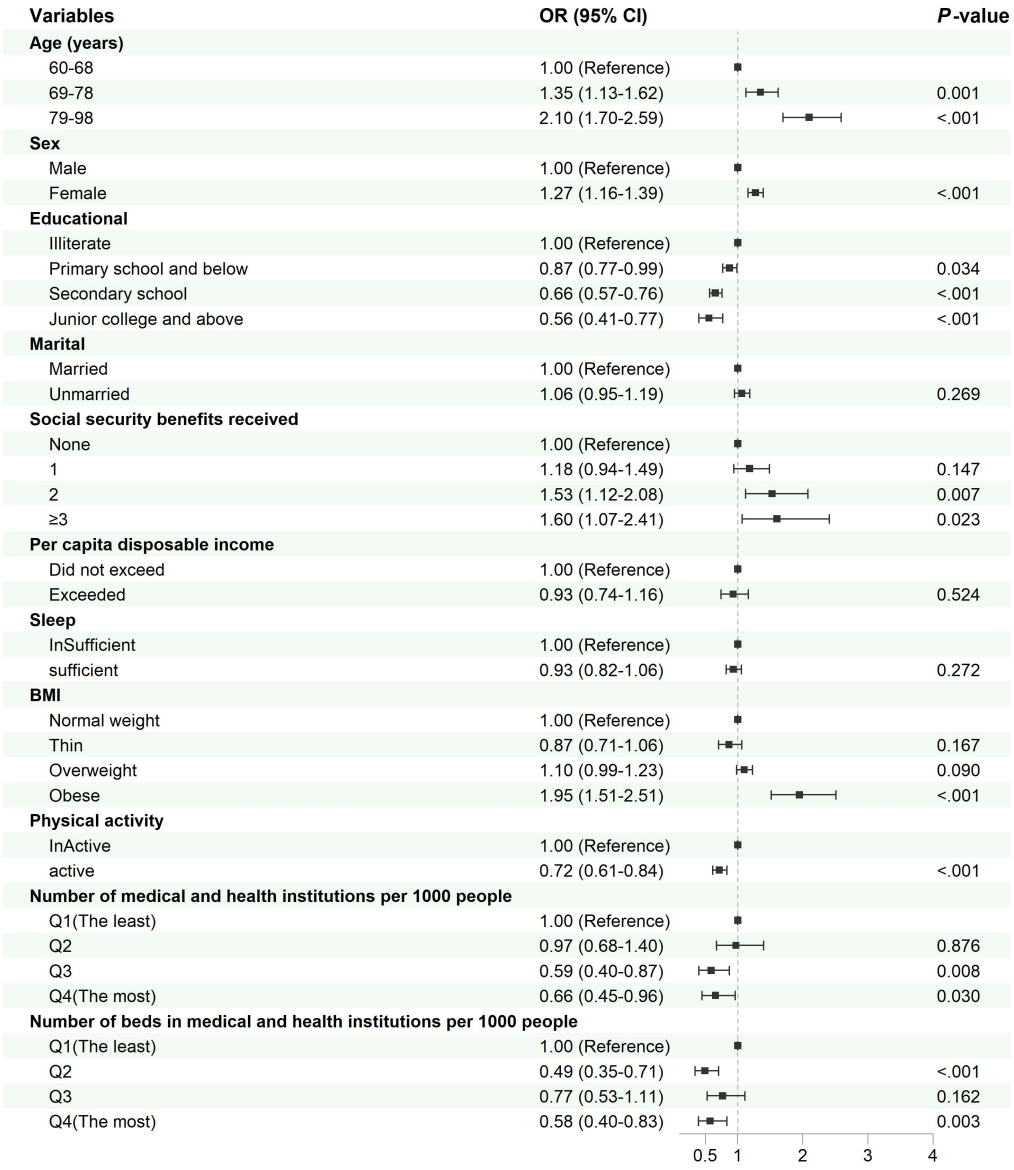
Supplementary Figure S5. Variables interaction matrix of factors influencing the prevalence of multimorbidity. The heatmap displays pairwise SHAP interaction values among factors influencing the regional prevalence of multimorbidity. The color intensity reflects the magnitude of the interaction effect. Variables are listed symmetrically on both the x-axis and y-axis.



Supplementary Figure S6. Sensitivity analysis: Associations between socio-demographic/lifestyle variables and multimorbidity using an elevated diagnostic threshold (≥3 chronic conditions). The solid squares represent the Odds Ratios (OR), and the horizontal lines represent the 95% Confidence Intervals (CI) derived from the Generalized Linear Mixed Model. The vertical dashed line indicates an OR of 1.00. The corresponding P-values for each variable category are listed on the right.



Supplementary Figure S7. Sensitivity analysis: Associations after excluding deafness and reproductive system diseases from the multimorbidity definition.



Supplementary Figure S8. Sensitivity analysis: Associations between socio-demographic, lifestyle and medical resource variables and multimorbidity after excluding participants with cognitive impairment-related conditions.