

Development and Analysis of a Risk Prediction Model for Rheumatoid Arthritis in Tibetan Older Adults in Tibet Based on National Health Service Survey Data

Shijie Zhang¹, Zhaxi Dawa^{1, 2, *}

¹Medical College, Xizang University, Lhasa, China

²Key Laboratory of Public Health Safety & Policy Research of Lhasa City, Xizang University, Lhasa, China

Email address:

1916797176@qq.com (Shijie Zhang), zhaxi0891@qq.com (Zhaxi Dawa)

*Corresponding author

Abstract

Background: Rheumatoid arthritis (RA) is a common chronic inflammatory joint disease that significantly affects the quality of life in older populations. However, risk prediction models specifically tailored for Tibetan older adults remain scarce. **Objective:** This study aimed to develop and validate a RA risk prediction model for Tibetan individuals aged ≥ 60 years using data from the National Health Service Survey in Tibet, thereby providing a reference for community-based early screening and health management. **Methods:** A cross-sectional design was employed, enrolling 2,144 Tibetan older adults (≥ 60 years) in Tibet. Least Absolute Shrinkage and Selection Operator (LASSO) regression was first applied for feature selection, retaining 16 variables with non-zero coefficients. These variables were then entered into multivariate logistic regression to identify independent predictors at $P < 0.05$. Five machine learning models—LASSO regression, random forest, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and naive Bayes—were constructed and compared using area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity. A nomogram was developed for model visualization. **Results:** Multivariate logistic regression identified six independent predictors: prefecture-level city, gender, employment status, insomnia symptoms, daily toothbrushing frequency, and hypertension. In the test set, LightGBM achieved the highest AUC (0.708) and the highest sensitivity (0.556), outperforming XGBoost (AUC 0.683, sensitivity 0.513) and LASSO (AUC 0.694, sensitivity 0.085). Considering the priority of maximizing true positive detection in a screening setting, LightGBM was selected as the optimal model due to its superior balanced performance (F1 score 0.371, Kappa 0.170) and acceptable trade-off in specificity. The nomogram based on logistic regression results visually illustrated the contribution of each factor to RA risk. **Conclusion:** The proposed RA risk prediction model incorporating six readily available factors demonstrates moderate predictive performance in Tibetan older adults. When the screening goal is to minimize missed diagnoses, the LightGBM-based model offers a practical risk-assessment tool for primary healthcare facilities, though its clinical utility warrants further validation through prospective studies.

Keywords

Rheumatoid Arthritis, Machine Learning, Tibetan, Older Adults