Methods

Team4

1. Study Design and Overview

This study was conducted to multidimensionally analyze factors influencing the development of metabolic diseases, including hypertension, diabetes, and dyslipidemia, in non-smoking and non-drinking adults, applying the Biopsychosocial (BPS) model as the theoretical framework. The BPS model presents a conceptual structure in which Biological, Psychological, and Social/Behavioral factors interact to shape health outcomes. Accordingly, this study systematically classified variables across three domains and analyzed their contribution and directional effects on the likelihood of metabolic disease development through both traditional regression models and machine learning-based predictive models.

2. Data Source and Participants

Study data were derived by integrating health interview, examination, and nutrition survey data from the Korea National Health and Nutrition Examination Survey (KNHANES). Study participants were restricted to individuals aged 20–64 years who met the criteria for non-smoking and non-drinking status.

Inclusion Criteria

- Adults aged 20–64 years who participated in the corresponding survey year
- Participants with available responses for key items in the health interview, examination, and nutrition surveys

Exclusion Criteria

- Current smokers or individuals with a history of smoking
- Individuals with any alcohol consumption experience within the past 12 months
- Pregnant women
- Cases with excessive missing values in health and behavioral items essential for analysis
- Applying these criteria, N non-smoking and non-drinking adults were selected as the final study population.

3. Variable Construction and Data Processing

Independent variables were structured according to the BPS model into Biological, Psychological, and Social/Behavioral domains. This structured approach reflects the BPS model perspective that metabolic diseases result from complex health pathways rather than a single cause.

3.1. Dependent Variable: Metabolic Disease Status

Metabolic disease was defined as having been diagnosed by a physician with one or more of the following three conditions:

Diagnosis of hypertension, Diagnosis of dyslipidemia (including hyperlipidemia), Diagnosis of diabetes

Responses indicating "yes" were coded as 1, "no" as 0, and "don't know/no response" were treated as missing.

3.2 Biological Domain (Physiological and Physical Factors)

Sex, Age, Body mass index (BMI), Waist circumference, Systolic and diastolic blood pressure, Fasting blood glucose, Serum lipid profiles including total cholesterol, triglycerides, HDL, and LDL

All physiological indicators were used as continuous variables and standardized to adjust for scale differences.

3.3 Psychological Domain (Mental and Cognitive Factors)

Perceived stress (whether individuals perceive stress in daily life), Stress intensity (categorized as low/moderate/high based on an 8-level perceived scale)

Stress intensity was recategorized into three levels reflecting distribution characteristics and converted into dummy variables for inclusion in the analysis.

3.4 Locial/Behavioral Domain (Socioeconomic and Lifestyle Factors)

Educational attainment, Household income level, Employment status, Weekly physical activity frequency (none / 1–2 times / 3 or more times), Daily nutrient intake: total energy, protein, fat and saturated fat, sodium and potassium, vitamin D, vitamin C, folate, etc.

Nutrient intake was used as continuous values within the sample rather than absolute reference standards (normal range), standardized for use in the analysis.

3.5 Data Processing

"Don't know/no response" items were treated as missing. Missing values in continuous variables were imputed using random forest-based multiple imputation. All categorical variables were converted to dummy variables. Biological and nutritional variables were standardized (z-score) to ensure comparability across models

4. Analytical Procedures

The analysis was designed to reflect the structural order of the BPS model.

4.1 Descriptive Statistics and Group Comparisons

Descriptive statistics were calculated considering the distribution and measurement level of each variable. For continuous variables, distribution was assessed using histograms and normality tests; variables approximating normal distribution were summarized with means and standard deviations, while those with skewed distributions or outliers were presented with medians and interquartile ranges. Categorical variables were presented with absolute frequencies and relative proportions for each category.

To evaluate differences between groups with and without metabolic diseases, Student's t-test or Mann–Whitney U test was applied for continuous variables according to normality. For categorical variables, the χ^2 test was used to determine whether differences in proportions between groups were statistically significant. Through these analyses, exploratory patterns were identified regarding how biological, psychological, and social/behavioral factors differed according to metabolic disease status.

4.2 Stepwise Multivariable Logistic Regression (B→P→S Sequential Entry)

Following the structure of the Biopsychosocial (BPS) model, variables were classified by domain and entered stepwise in multivariable logistic regression.

Model 1 included only biological factors (age, sex, body mass index, waist circumference, blood pressure, fasting blood glucose, serum lipid profiles) to assess basic metabolic risk.

Model 2 added psychological factors (perceived stress, stress intensity) to Model 1 to examine whether mental factors contributed independently.

Model 3 added social and behavioral factors (educational level, income, employment, physical activity frequency, major nutrient intake) to Model 2 to evaluate the influence of living environment and behavioral factors.

Model 4 was a fully adjusted model including all variables from the three domains, serving as the final model to assess comprehensive predictive power for metabolic disease.

Odds ratios (OR) and 95% confidence intervals were calculated for each model to compare how effect sizes changed as variable groups were added stepwise, thereby evaluating the contribution of each BPS domain.

4.3 Predictive Model Development

For metabolic disease prediction, logistic regression was set as the reference model, and six machine learning algorithms were constructed for comparison. The algorithms used were AdaBoost, Random Forest, Gradient Boosting, Artificial Neural Network (ANN), XGBoost, and Support Vector Machine (SVM), all suitable for binary classification problems. All algorithms were configured to be trained with the same set of independent variables (BPS-based Biological, Psychological, and Social/Behavioral variables).

3.4 Model Evaluation Metrics

Model performance evaluation primarily used the ROC-AUC (Receiver Operating Characteristic–Area Under the Curve) as the main metric, as it most reliably reflects classification performance for distinguishing metabolic disease presence. Additionally, Accuracy, Kappa, Sensitivity, Specificity, and F1-score were calculated as supplementary metrics to compare algorithm performance from multiple perspectives. Evaluation results for each model were calculated on the validation dataset, with mean values from cross-validation also presented.

5. Methodological Limitations

First, recall bias may exist due to self-reported data. Second, the low proportion of respondents with diagnosed metabolic diseases may limit statistical power. Third, as this is cross-sectional data, interpretations should be made at the level of associations rather than causal relationships. To mitigate these limitations, data preprocessing and validation procedures were conducted rigorously.

6. Ethical Considerations

KNHANES data are publicly available de-identified data that can be analyzed without separate IRB approval. All research processes complied with the Korea Disease Control and Prevention Agency's data usage guidelines and research ethics.

8. References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. Science, 196(4286), 129–136. https://doi.org/10.1126/science.847460

Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. Journal of Statistical Software, 59(5), 1–38. https://doi.org/10.18637/jss.v059.i05