Statistical Analysis of Korean National Health Insurance Service Dataset: Relationships Between Physiological Parameters and Lifestyle Factors

Zhen Tang, Haoyang Wang, Hongan Zhu, Shuyang Chen, Jierui Li

April 13, 2025

Abstract

This study conducts a comprehensive statistical analysis of the Korean National Health Insurance Service (NHIS) dataset to investigate relationships between physiological health indicators and lifestyle factors. Specifically, we focus on three key objectives: (1) providing descriptive statistics of physiological parameters including blood pressure, cholesterol, hemoglobin, and glucose levels; (2) testing hypotheses regarding the relationship between smoking and drinking status and various health indicators using appropriate statistical methods; and (3) performing regression analyses to evaluate the associations between physiological parameters and lifestyle behaviors. Our findings reveal significant differences in several health markers across smoking and drinking categories, with particularly pronounced effects observed for hemoglobin levels and Gamma-GTP.

Keywords: physiological parameters, smoking, alcohol consumption, statistical analysis, public health

1 Introduction

1.1 Background

Lifestyle factors such as smoking and alcohol consumption have been consistently associated with adverse health outcomes and increased risk of chronic diseases including cardiovascular disease, liver dysfunction, and metabolic disorders. However, the specific physiological mechanisms through which these behaviors influence health remain an area of active research. Comprehensive health datasets, such as the one maintained by the Korean National Health Insurance Service (NHIS), provide valuable opportunities to investigate these relationships at a population level and identify potential biomarkers of health risk associated with these behaviors. Understanding the precise relationships between modifiable lifestyle factors and physiological parameters can inform public health interventions and clinical practice. This research contributes to the growing body of evidence regarding the health impacts of smoking and alcohol consumption in a large East Asian population.

1.2 Dataset

The dataset analyzed in this study is a CSV file containing records of 10,000 individuals from the Korean National Health Insurance Service. The dataset includes six physiological indicators: Total Cholesterol, Systolic Blood Pressure, Diastolic Blood Pressure, Hemoglobin, Fasting Blood Glucose, and Gamma-Glutamyl Transferase (Gamma-GTP). Additionally, the dataset contains information about participants' smoking status (categorized as 1 = never smoked, 2 = former smoker, 3 = current smoker) and drinking status (binary classification of drinker/non-drinker).

1.3 Research Objectives

This study aims to:

- Provide comprehensive descriptive statistics for all physiological parameters stratified by smoking and drinking status
- Test specific hypotheses regarding the relationship between lifestyle factors and health indicators
- Quantify the magnitude of associations between smoking, alcohol consumption, and physiological parameters through appropriate statistical modeling

2 Visualization of Data

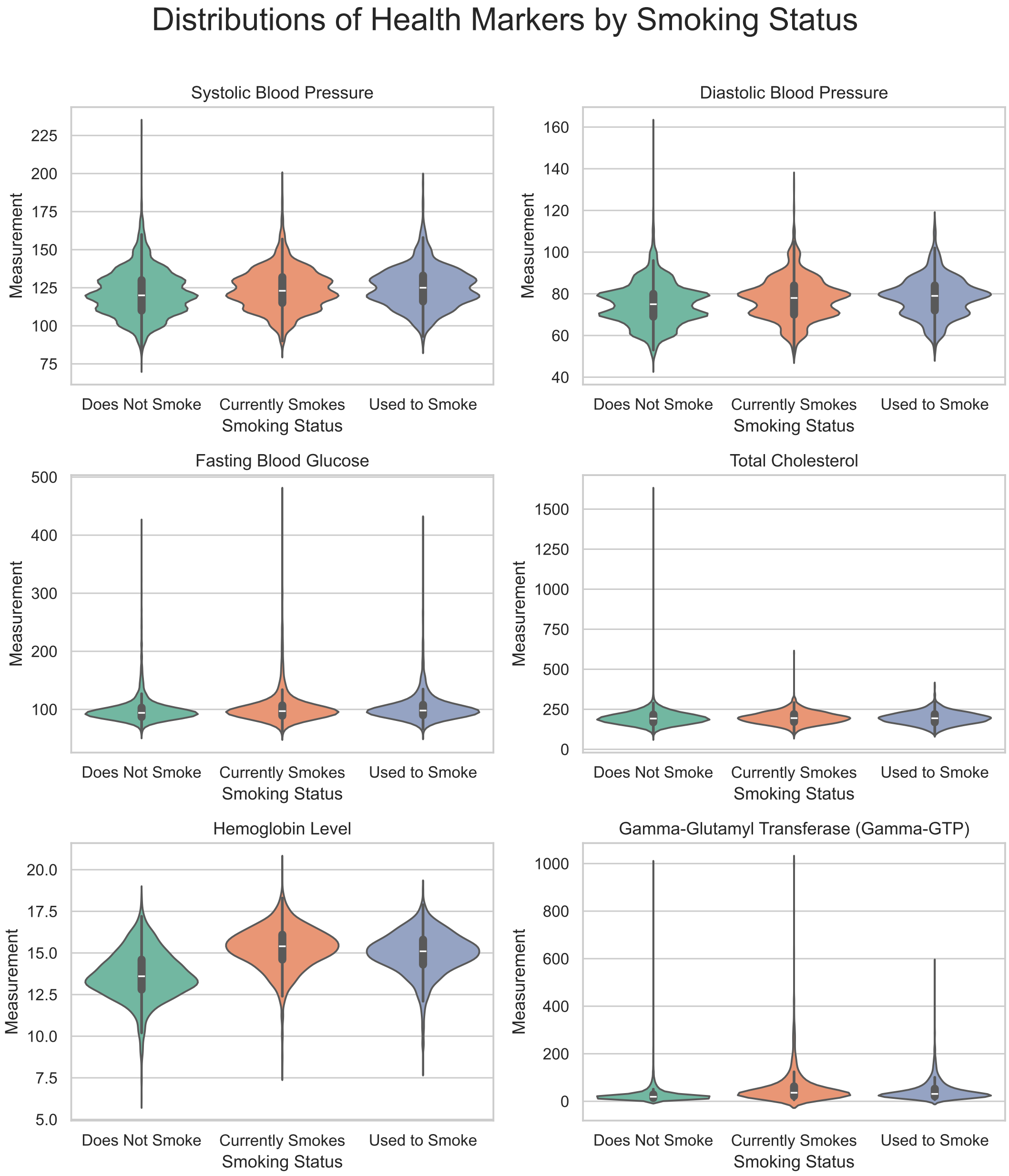


Figure 1

Figure 1 presents a series of violin plots displaying the distributions of various health markers categorized by smoking status. Each violin plot includes the probability density of values along with internal box plots indicating the median and inter-quartile ranges. The most visually apparent differences between groups are observed in hemoglobin levels and Gamma-GTP concentrations, where smoking status appears to correlate with higher values. Current smokers demonstrate notably elevated hemoglobin levels compared to both former smokers and those who never smoked. Similarly, Gamma-GTP levels show a gradient effect with highest values in current smokers, intermediate values in former smokers, and lowest values in never-smokers. For the remaining physiological parameters (blood pressure, total cholesterol, and fasting blood glucose), differences across smoking status categories appear less pronounced in the visual examination.

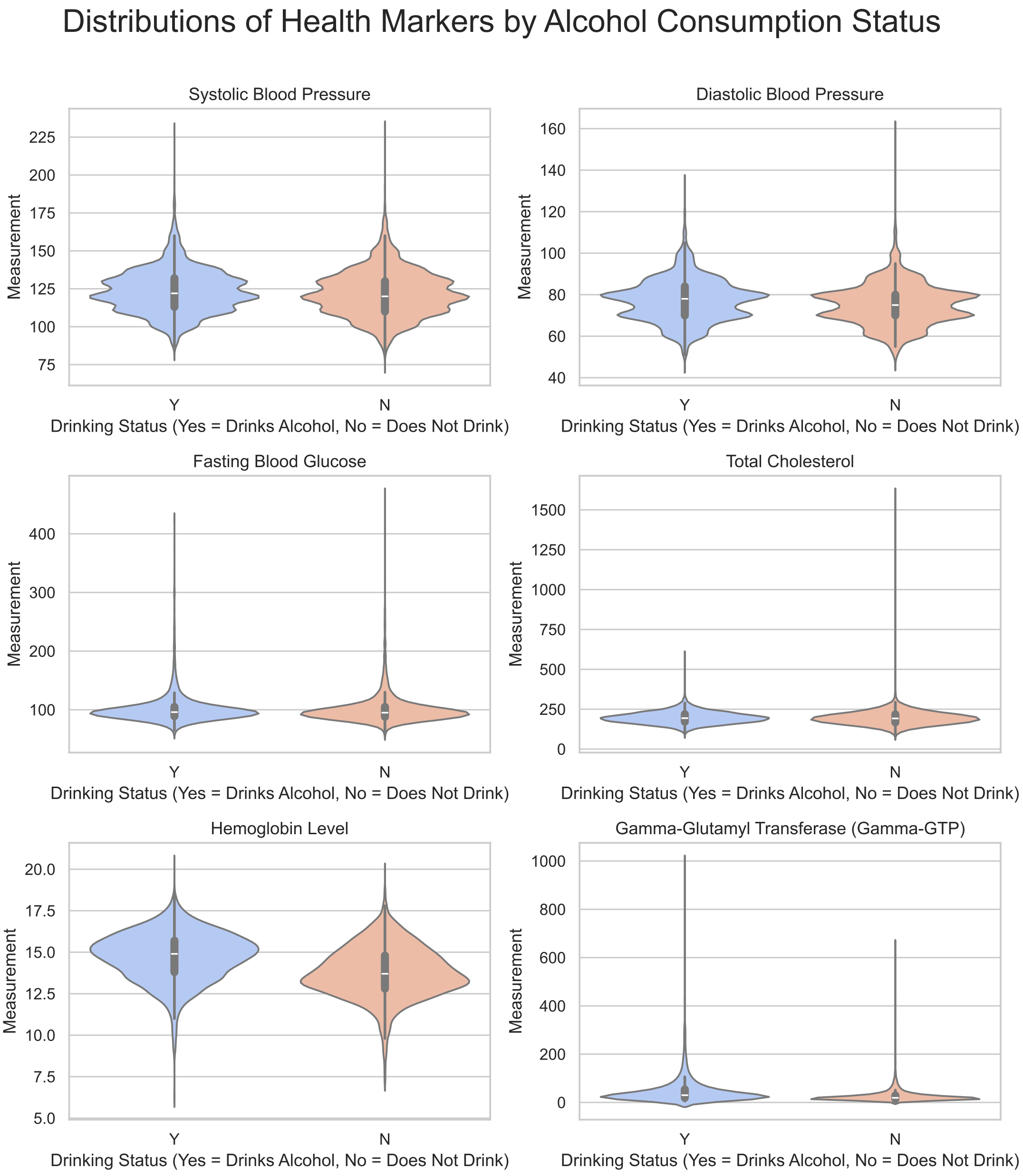
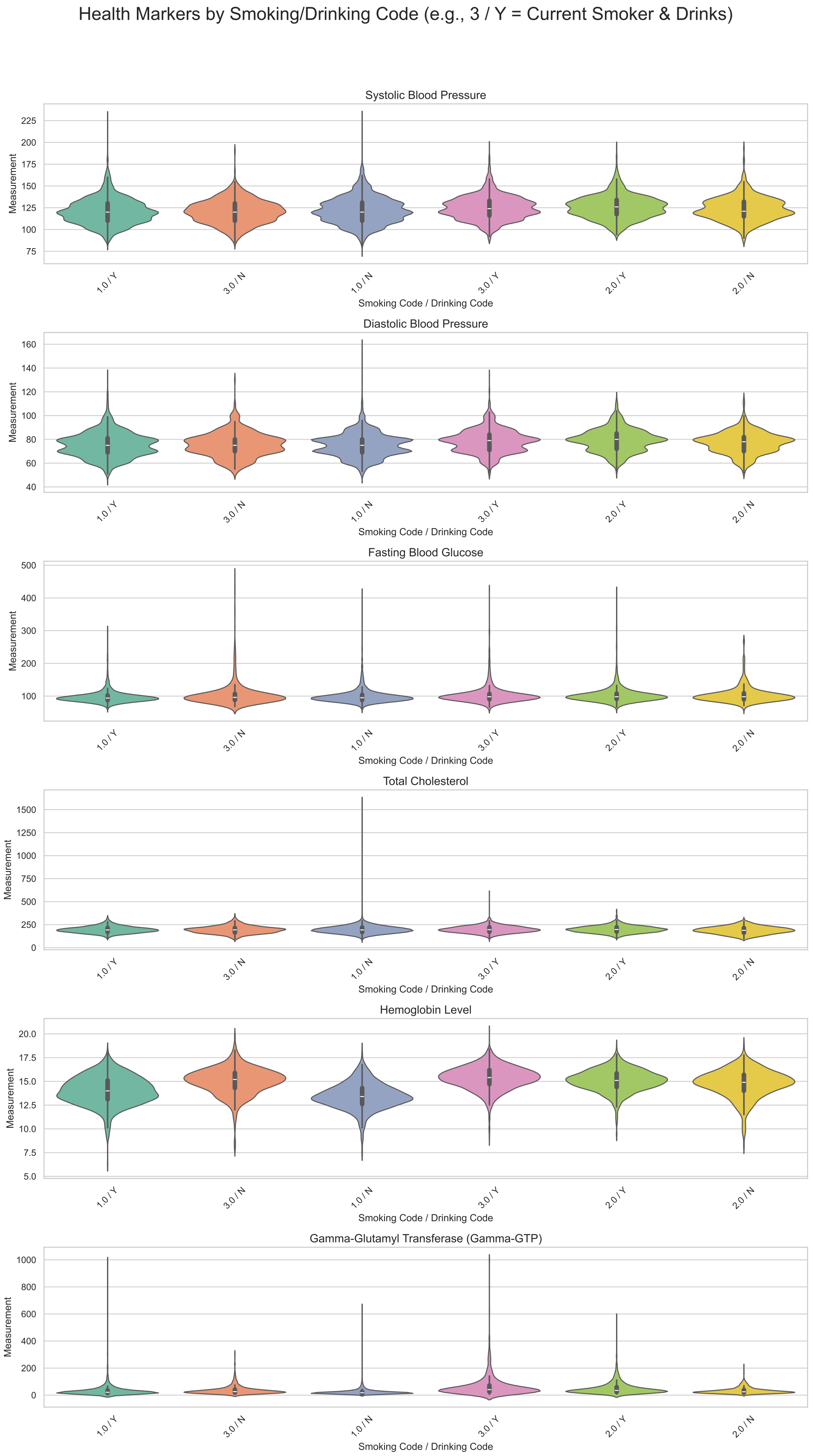


Figure 2

Figure 2 presents violin plots illustrating the distributions of health markers stratified by alcohol consumption status. A notable difference is observed where alcohol drinkers appear to have higher hemoglobin levels compared to non-drinkers. Even more pronounced is the difference in Gamma-GTP levels, with drinkers showing substantially higher values and greater variance compared to non-drinkers. This is consistent with the established role of Gamma-GTP as a sensitive marker of alcohol consumption and potential liver damage. For blood pressure measurements, both systolic and diastolic values show slight elevations in the drinking group, though the differences appear less dramatic than those observed for hemoglobin and Gamma-GTP.


Health Markers by Smoking/Drinking Code (e.g., 3 / Y = Current Smoker & Drinks)

Systolic Blood Pressure

Diastolic Blood Pressure

Fasting Blood Glucose

Total Cholesterol

Hemoglobin Level

Gamma-Glutamyl Transferase (Gamma-GTP)

Figure 3

Figure 3 displays a series of violin plots showing the distributions of health markers categorized by combined smoking and drinking status. Six distinct groups are presented, representing all possible combinations of smoking (never, former, current) and drinking (yes/no) statuses. The most pronounced differences appear in hemoglobin levels and Gamma-GTP values, suggesting that the combination of smoking and drinking may have more significant effects on these particular health markers than either behavior alone. Notably, individuals who both smoke and drink exhibit the highest Gamma-GTP values, indicating potential synergistic effects of these behaviors on liver function.

3 Statistical Analysis and Results

3.1 Significance of Variable Differences by Drinking Status

To assess the significance of differences in physiological parameters across lifestyle groups, we employed appropriate statistical tests. For comparisons between two groups (drinkers vs. non-drinkers), Welch's t-test with unequal variances was used because it provides robust results even when the assumption of equal variances may be violated, which is common in health data where different lifestyle groups often exhibit different levels of variability in physiological measures.

Our analysis revealed that most physiological variables showed statistically significant differences between drinkers and non-drinkers. The parameters with the highest statistical significance (lowest p-values) were:

- Total Cholesterol ($p \approx 1.36 \times 10^{-3}$)
- Systolic Blood Pressure ($p \approx 1.24 \times 10^{-6}$)
- Diastolic Blood Pressure ($p \approx 1.89 \times 10^{-25}$)
- Hemoglobin ($p \approx 1.31 \times 10^{-219}$)
- Gamma-GTP ($p \approx 4.36 \times 10^{-103}$)

Only one variable showed no significant difference between drinking groups: fasting blood glucose (BLDS) (p = 0.447). This suggests that while alcohol consumption is associated with alterations in most physiological parameters measured, glucose metabolism as measured by fasting blood glucose appears to be less directly affected by drinking status in this population.

3.2 Significance of Variable Differences by Smoking Status

For comparisons across smoking status categories (never, former, current), Kruskal-Wallis tests were employed. This non-parametric test was selected because it doesn't require the assumption of normal distribution and is appropriate for comparing three or more independent groups, making it well-suited for analyzing the three smoking categories in our dataset.

All measured physiological parameters showed statistically significant differences across smoking status categories:

- Total Cholesterol ($p \approx 0.045$)

- Systolic Blood Pressure ($p \approx 2.2 \times 10^{-16}$)
- Diastolic Blood Pressure ($p \approx 2.2 \times 10^{-16}$)
- Hemoglobin ($p \approx 2.2 \times 10^{-16}$)
- Gamma-GTP ($p \approx 2.2 \times 10^{-16}$)
- Fasting Blood Glucose ($p \approx 2.2 \times 10^{-16}$)

The extremely small p-values (represented as 2.2×10^{-16}, which is the default minimum value reported by many statistical software packages) indicate strong evidence against the null hypothesis of no difference between smoking groups for most variables. Total cholesterol showed a less dramatic but still statistically significant difference ($p \approx 0.045$).

4 Linear Regression

To further examine the effects of smoking and drinking on the six physiological indicators, we constructed linear regression models with these lifestyle factors as primary independent variables. Linear regression was selected for its interpretability and ability to quantify both direction and magnitude of associations. The adjusted R-squared values reported provide an assessment of each model's explanatory power, identifying which physiological parameters are most strongly associated with lifestyle behaviors. The analysis included six dependent variables: hemoglobin, diastolic blood pressure (DBP), systolic blood pressure (SBP), fasting blood glucose, total cholesterol, and gamma-GTP.

4.1 Regression Results

4.1.1 Linear Model for DBP

$$\begin{aligned}\text{DBP} &= 4.09 \;+\; 0.59\,(\text{DRK\_YN}) \;+\; 0.67\,(\text{hemoglobin}) \\ &\quad + 0.49\,(\text{SBP}) \;-\; 0.01\,(\text{tot\_chole}) \;+\; 0.005\,(\text{gamma\_GTP}).\end{aligned}$$

Summary: Based on the F-test and R-squared indicator (Adjusted $R^2 \approx 0.57$), the model explains approximately 57% of the variation in diastolic blood pressure, indicating moderate predictive power with high statistical significance ($p < 2.2 \times 10^{-16}$). Notably, the drinking group exhibits significantly higher mean DBP compared to non-drinkers, and hemoglobin shows a pronounced effect on DBP. In contrast, smoking shows no direct impact on DBP.

4.1.2 Linear Model for SBP

$$\begin{aligned}\text{SBP} &= 38.5391 \;-\; 0.40\,(\text{SMK\_stat\_type\_cd}) \;-\; 0.844\,(\text{DRK\_YN}) \\ &\quad + 1.07\,(\text{DBP}) \;+\; 0.04\,(\text{BLDS}) \;-\; 0.0006\,(\text{tot\_chole}) \;+\; 0.0115\,(\text{gamma\_GTP}).\end{aligned}$$

Summary: The model explains approximately 56% of SBP variation (Adjusted $R^2 \approx 0.56$) with high statistical significance ($p < 2.0 \times 10^{-16}$), indicating strong collective association between these predictors and systolic blood pressure. Both smoking status and drinking status show statistically significant effects on SBP, with results suggesting both smoking and alcohol consumption are associated with lower SBP levels.

4.1.3 Linear Model for Hemoglobin

$$\begin{aligned}\text{hemoglobin} = 10.07 \;+\; 0.688\,(\text{SMK\_stat\_type\_cd}) \;+\; 0.468\,(\text{DRK\_Y})\\ +\,0.251\,(\text{DBP}) \;+\; 0.157\,(\text{BLDS}) \;-\; 0.351\,(\text{tot\_chole}) \;+\; 0.184\,(\text{gamma\_GTP}).\end{aligned}$$

Summary: The model explains approximately 28% of hemoglobin variation (Adjusted $R^2 \approx 0.28$), with 72% of variance unexplained and potentially attributable to other factors or random error, indicating relatively weak explanatory power. Both smoking and drinking significantly impact hemoglobin levels, with smoking showing a stronger influence than drinking.

4.1.4 Linear Model for BLDS

$$\begin{aligned}\text{BLDS} = 63.48 \;+\; 1.45\,(\text{SMK\_stat\_type\_cd}) \;-\; 2.77\,(\text{DRK\_Y})\\ +\,0.27\,(\text{SBP}) \;+\; 0.58\,(\text{hemoglobin}) \;-\; 0.07\,(\text{DBP})\\ -\,0.01\,(\text{tot\_chole}) \;+\; 0.07\,(\text{gamma\_GTP}).\end{aligned}$$

Summary: With an adjusted $R^2 \approx 0.05$, the model explains only 5% of BLDS variation. The remaining 95% of variance is likely attributable to unaccounted factors or random error, indicating very weak explanatory power. While smoking and drinking show some impact on fasting blood glucose, their influence appears minimal compared to other potential factors not included in this model.

4.1.5 Linear Model for Total Cholesterol

$$\begin{aligned}\text{tot-chole} = 134.77 \;-\; 2.84\,(\text{SMK\_stat\_type\_cd}) \;-\; 0.11\,(\text{SBP})\\ +\,0.47\,(\text{DBP}) \;+\; 3.13\,(\text{hemoglobin}) \;-\; 0.03\,(\text{BLDS}) \;+\; 0.04\,(\text{gamma\_GTP}).\end{aligned}$$

Summary: With an adjusted $R^2 \approx 0.02$, the model explains merely 2% of total cholesterol variation. The remaining 98% of variance likely stems from unaccounted factors or random error, indicating extremely weak explanatory power. Smoking shows a modest impact on total cholesterol, while drinking demonstrates no significant effect. The model suggests that total cholesterol is primarily influenced by factors outside these variables.

4.1.6 Linear Model for Gamma-GTP

$$\begin{aligned}\text{gamma-GTP} = 46.55 \;+\; 9.41\,(\text{SMK\_stat\_type\_cd}) \;+\; 13.37\,(\text{DRK\_Y})\\ +\,0.29\,(\text{SBP}) \;+\; 0.27\,(\text{DBP}) \;+\; 2.48\,(\text{hemoglobin})\\ -\,0.26\,(\text{BLDS}) \;+\; 0.05\,(\text{tot-chole})\end{aligned}$$

Summary: With an adjusted $R^2 \approx 0.13$, the model explains 13% of gamma-GTP variation, with 87% of variance unexplained. While this indicates limited explanatory power, both smoking and drinking show significant positive associations with gamma-GTP levels. The coefficients suggest alcohol consumption (13.37) has a stronger impact than smoking (9.41), consistent with gamma-GTP's established role as a biomarker for liver function affected by these behaviors.

5 Logistic Regression

To further explore relationships between physiological parameters and lifestyle behaviors, we conducted logistic regression analyses to predict drinking and smoking status based on measured health indicators.

We selected binary logistic regression for drinking status due to its appropriateness for dichotomous outcomes, and multinomial logistic regression for smoking status given its three categories. Standardization in the multinomial model facilitated direct comparison of effect sizes across different measurement scales.

5.1 Predicting Alcohol Consumption Status

We developed a binary logistic regression model to predict drinking status (DRK_YN) based on all measured physiological parameters:

$$P(\text{DRK}_{YN} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{SBP} + \beta_2 \cdot \text{DBP} + \beta_3 \cdot \text{BLDS} + \beta_4 \cdot \text{tot\_chole} + \beta_5 \cdot \text{hemoglobin} + \beta_6 \cdot \text{gamma\_GTP})}}$$

The estimated coefficients were:

- Intercept (β_0): -4.9099
- Systolic Blood Pressure (β_1): -0.0105
- Diastolic Blood Pressure (β_2): 0.0127
- Fasting Blood Glucose (β_3): -0.0048
- Total Cholesterol (β_4): -0.0008
- Hemoglobin (β_5): 0.3775
- Gamma-GTP (β_6): 0.0131

The model achieved 0.666 accuracy with balanced precision (0.69) and recall (0.64) for drinkers, yielding an F1-score of 0.66.

Hemoglobin emerged as the strongest predictor of drinking status, with each unit increase associated with 1.46 times higher odds of being a drinker. Gamma-GTP also showed a positive association with drinking, consistent with its role as an alcohol consumption biomarker. Interestingly, systolic blood pressure showed a slight negative association when controlling for other parameters, contradicting the positive association in univariate analyses.

5.2 Predicting Smoking Status

We employed multinomial logistic regression to predict smoking status (1 = never smoked, 2 = former smoker, 3 = current smoker) using standardized physiological parameters:

$$P(Y = k|X) = \frac{e^{z_k}}{\sum_j e^{z_j}}$$

Where:

$$\begin{aligned} z_k = \beta_{k0} &+ \beta_{k1}\frac{\text{SBP} - \mu_{\text{SBP}}}{\sigma_{\text{SBP}}} + \beta_{k2}\frac{\text{DBP} - \mu_{\text{DBP}}}{\sigma_{\text{DBP}}} \\ &+ \beta_{k3}\frac{\text{BLDS} - \mu_{\text{BLDS}}}{\sigma_{\text{BLDS}}} + \beta_{k4}\frac{\text{tot\_chole} - \mu_{\text{tot\_chole}}}{\sigma_{\text{tot\_chole}}} \\ &+ \beta_{k5}\frac{\text{hemoglobin} - \mu_{\text{hemoglobin}}}{\sigma_{\text{hemoglobin}}} + \beta_{k6}\frac{\text{gamma\_GTP} - \mu_{\text{gamma\_GTP}}}{\sigma_{\text{gamma\_GTP}}} \end{aligned}$$

The coefficients revealed distinct patterns: for never-smokers, hemoglobin had the strongest negative association ($\beta = -0.8118$); for former smokers, diastolic blood pressure showed the strongest positive association ($\beta = 0.0772$); for current smokers, hemoglobin demonstrated the strongest positive association ($\beta = 0.5471$).

The model achieved 0.6425 overall accuracy but performed unevenly across classes: strong for never-smokers (precision = 0.69, recall = 0.90), moderate for current smokers (precision = 0.49, recall = 0.47), and poor for former smokers (precision = 0.17, recall $\approx$ 0.00).

5.3 Implications of Logistic Regression Results

These analyses statistically confirm relationships observed in our exploratory data analysis. Hemoglobin emerges as a critical predictor for both drinking and smoking status, supporting research suggesting these behaviors affect erythropoiesis and oxygen-carrying capacity.

The positive association between Gamma-GTP and both drinking and current smoking (stronger for drinking) aligns with its role as a liver stress marker, induced by both alcohol consumption and cigarette toxin metabolism.

The difficulty in classifying former smokers suggests some smoking-induced changes may be reversible after cessation, resulting in physiological profiles distinct from both never-smokers and current smokers—implying potential benefits of smoking cessation even for long-term smokers.

6 Conclusion

This comprehensive analysis of the Korean NHIS dataset has revealed significant associations between lifestyle factors (smoking and alcohol consumption) and various physiological parameters. The most pronounced effects were observed for hemoglobin levels and Gamma-GTP, both of which showed strong positive associations with smoking and drinking behaviors, with evidence of synergistic effects when these behaviors co-occur. Blood pressure measurements (both systolic and diastolic) also showed significant associations with lifestyle factors, though the magnitude of these effects was smaller compared to hemoglobin and Gamma-GTP. Total cholesterol showed modest but statistically significant associations with both smoking and drinking status, while fasting blood glucose was significantly associated with smoking status but not with drinking status.

7 Acknowledgment

Jierui Li was responsible for writing the draft. Hongan Zhu was responsible for data visualization and paper writing. Zhen Tang was responsible for statistical analysis by

drinking status. Haoyang Wang was responsible for statistical analysis by smoking status and linear regression models. Shuyang Chen was responsible for the logistic regression analysis.