**Abstract**

This study examines whether younger individuals with clinically normal cholesterol levels are at an increased risk of cardiovascular disease (CVD) when they also present with elevated blood pressure or ST depression. Using a dataset of anonymized patient records from a multispecialty hospital in India, we assess associations between these secondary risk factors and heart disease through descriptive statistics, inferential tests, and predictive modeling. The findings, based on chi-square tests, logistic regression, and machine learning approaches, suggest that these additional risk factors do not significantly alter the risk of CVD in this population.

**Introduction**

Coronary artery disease (CAD), often referred to as heart disease, is a leading cause of death worldwide. It is characterized by the gradual narrowing of the coronary arteries due to plaque buildup, which can result in myocardial infarction, often without prior symptoms. This "silent killer" claims millions of lives each year, highlighting the importance of early detection and prevention.

This study utilizes a dataset comprising anonymized patient records related to cardiovascular risk factors and diagnoses. Each record includes variables such as age, gender, blood pressure, cholesterol levels, and various clinical indicators like ST depression and chest pain type. The data, collected from a multispecialty hospital in India and published on Mendeley Data in April 2021, offers an opportunity to examine risk beyond conventional markers.

The central research question is whether individuals under 50 with normal cholesterol levels (defined as <200 mg/dL) remain at risk for CVD if they also present with elevated blood pressure or ST depression. The goal is to challenge traditional frameworks and explore whether secondary risk factors contribute meaningfully to heart disease risk in this population.

**Methods**

**Study Population and Objective**

The study focused on individuals under 50 years of age with clinically normal cholesterol levels (<200 mg/dL). These thresholds were chosen based on clinical relevance. The dataset includes anonymized demographic, diagnostic, and physiological data on more than 1,000 patients.

The primary outcome, target, indicates presence (1) or absence (0) of heart disease. Key predictors included age, resting blood pressure, serum cholesterol, ST depression (oldpeak), gender, and fasting blood sugar.

**Risk Variable Derivation**

Two dichotomous predictors were engineered:

* Elevated Blood Pressure: Defined as resting systolic BP between 120–129 mmHg
* Elevated ST Depression: Defined as oldpeak > 1.0

All categorical variables were converted to factors for modeling.

**Statistical Procedures**

Descriptive statistics summarized clinical variables, while histograms, boxplots, and correlation matrices described distributions and associations. A filtered dataset containing only younger individuals with normal cholesterol was used in all analyses.

Chi-square tests evaluated whether elevated blood pressure or ST depression was associated with higher CVD prevalence. Logistic regression models with and without interaction terms were fit, along with a log-binomial model for estimating relative risk. Models were further adjusted for gender and fasting blood sugar.

**Predictive Modeling**

Three predictive models—linear regression, LASSO regression, and random forest—were trained using the filtered dataset. Model performance was assessed using RMSE. Tuning procedures for LASSO (penalty parameter) and random forest (mtry and min\_n) were performed via grid search and cross-validation. ROC analysis and AUC were used to evaluate classification performance.

Results

**Descriptive Statistics**

The dataset was explored using summary statistics and visualizations (see Table 1, table1\_summary\_statistics.rds). Age ranged widely, and resting blood pressure and cholesterol also showed broad variability. Correlation analysis revealed moderate relationships among several predictors.

**Inferential Findings**

Chi-square tests indicated no statistically significant association between elevated blood pressure or ST depression and heart disease (p-values = 0.31 and 0.55, respectively). Summary tables for group frequencies are provided in the appendix.

Logistic regression models showed no statistically significant odds ratios for either risk factor. In the interaction model, elevated blood pressure had an OR of 2.29 (95% CI: 0.55–12.00), and ST depression an OR of 1.83 (95% CI: 0.07–27.22) (Table 2, table2\_odds\_ratios.rds). These results were consistent with the non-interaction model (Table 3), and interaction terms were also nonsignificant (Table 4).

Relative risk estimates from the log-binomial model were similarly non-significant, with RR = 1.66 for elevated blood pressure and RR = 1.21 for elevated ST depression (Table 3, table3\_relative\_risks.rds).

**Adjusted Model Results**

After adjusting for gender and fasting blood sugar, model coefficients became unstable, likely due to sparse data and collinearity. Although gender showed elevated risk (adjusted OR = 4.69), other estimates were unreliable (Table 4, table4\_adjusted\_odds\_ratios.rds).

**Model Discrimination**

ROC curve analysis showed poor model performance (AUC ≈ 0.60), indicating limited ability to discriminate between individuals with and without CVD (Figure 1, figure1\_roc\_logistic\_model.rds).

**Predictive Modeling and Tuning**

The linear, LASSO, and random forest models produced RMSEs of approximately 0.47, 0.48, and 0.47, respectively (Table 5, table5\_model\_rmse\_comparison.rds). Tuning for LASSO showed little improvement across penalty values (Figure 2), and random forest tuning identified an optimal configuration of mtry = 2, min\_n = 14 with similar performance (Figure 3).

**Discussion**

**Summary and Interpretation**

The analyses presented here do not support the hypothesis that younger individuals with normal cholesterol face elevated CVD risk due to the presence of high-normal blood pressure or ST depression. Chi-square tests, logistic regression models, and machine learning approaches consistently failed to show statistically or clinically significant results. RMSEs for all predictive models remained similar, and AUC metrics hovered around 0.60, further suggesting limited discriminatory power.

**Strengths and Limitations**

A key strength of this study is its comprehensive approach, incorporating multiple analytical methods across inferential and predictive domains. Grid search and cross-validation further enhanced model assessment.

However, limitations include small subgroup sizes and potential instability of parameter estimates due to sparse data. Adjusted models in particular exhibited wide confidence intervals and signs of overfitting. These findings emphasize the need for larger, prospective studies to validate the observed trends.

**Conclusion**

Based on the current analysis, there is limited evidence to suggest that elevated blood pressure or ST depression meaningfully increases the risk of cardiovascular disease in younger individuals with normal cholesterol. While point estimates suggest possible risk elevation, statistical significance was not achieved. These findings raise important questions about the sufficiency of traditional risk thresholds and underscore the need for more personalized risk assessment in early prevention.

**References**

1. American Heart Association. Understanding Blood Pressure Readings.
2. Mayo Clinic. Prehypertension: Symptoms and Causes.
3. Cleveland Clinic. Cholesterol Numbers: What Do They Mean?
4. Johns Hopkins Medicine. Lipid Panel Overview.
5. Goff DC Jr., et al. Stroke Risk in Young Adults (PMC6376358).
6. American Heart Association Journal. ST-T Abnormalities and CVD Risk.