# Summary/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 major cause of death worldwide. It is characterized by the gradual narrowing of the coronary arteries due to plaque buildup, which can lead to heart attacks, often without prior symptoms. This "silent killer" claims millions of lives each year, underscoring the importance of early detection and prevention.

This study uses a dataset containing 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, acquired from one of the multispecialty hospitals in India and first published in April 2021, is publicly available at [Mendeley Data](https://data.mendeley.com/datasets/dzz48mvjht/1).

The central research question addressed here is: Do individuals younger than 50 with normal cholesterol levels (within a "healthy" range) still face CVD risk if they have elevated blood pressure or ST depression? This question aims to challenge traditional risk-assessment frameworks by exploring whether non-traditional risk factors may still contribute to heart disease among younger patients.

# Methods

In this study, we will use “target” as the primary outcome variable indicating the absence or presence of heart disease and focus on key predictors including age, serum cholesterol, resting blood pressure, and ST depression (oldpeak). We will begin by restricting our analysis to patients under 50 years of age with serum cholesterol levels below 200 mg/dL, which defines a clinically normal range. Within this subgroup, individuals will be categorized based on secondary risk factors: elevated blood pressure (resting BP between 120 and 129 mmHg) and elevated ST depression (oldpeak > 1).

Descriptive analyses will be conducted to compare CVD prevalence across these risk groups, followed by chi-square tests to assess statistical significance. Logistic regression models, both unadjusted and adjusted for confounders such as gender and fasting blood sugar, will be used to quantify the independent and combined effects of the secondary risk factors on CVD risk. Model performance will be evaluated using odds ratios, relative risks, and receiver operating characteristic (ROC) curves with area under the curve (AUC) estimates.

Additionally, we will implement predictive models including linear regression, LASSO regression, and random forest, that will be tuned via grid search and cross-validation. Their performance will be measured by the root mean squared error (RMSE), thereby providing a comprehensive evaluation of model accuracy. Together, these methods aim to offer a nuanced assessment of cardiovascular risk among younger individuals.

# Results

## Exploratory and Descriptive Analysis

The data were first explored through summary statistics and visualizations, including histograms, density plots, boxplots, and correlation matrices. These initial analyses provided insights into the distributions of age, serum cholesterol, resting blood pressure, and ST depression, as well as the overall prevalence of heart disease.

Figure1.A graph showing the amount of age and percentage of the company

AI-generated content may be incorrect.

Figure 2.A graph showing a number of cholesterol distribution

AI-generated content may be incorrect.

Figure 3.A graph showing blood pressure distribution

AI-generated content may be incorrect.

Figure 4.A diagram of heart disease status

AI-generated content may be incorrect.

A focused subset was then created by filtering for patients under 50 with normal cholesterol levels (<200 mg/dL). Within this subset, individuals were further categorized into risk groups based on their blood pressure and ST depression status. Grouped summary tables and visualizations, such as heatmaps and faceted bar charts, were produced to illustrate the interaction between these secondary risk factors and heart disease prevalence.

## Inferential Analysis

Chi-square tests comparing the risk groups revealed no statistically significant association between elevated blood pressure or ST depression and the presence of heart disease. Logistic regression models, both with and without interaction terms, yielded odds ratios that were not statistically significant. For example, in the model without interaction, the odds ratio for elevated blood pressure was approximately 2.29 (95% CI: 0.55–12.00) and for ST depression about 1.83 (95% CI: 0.07–27.22). Similarly, relative risk estimates did not indicate significant differences across groups. The ROC curve for the logistic regression model produced an AUC of roughly 0.60, suggesting only modest discrimination between patients with and without heart disease.

Table 1. Chi-squared test:

No Heart Disease Heart Disease

Elevated 13 4

Otherwise 27 19

Pearson's Chi-squared test with Yates' continuity correction

data: table\_bp\_target

X-squared = 1.012, df = 1, p-value = 0.3144

No Heart Disease Heart Disease

Elevated 27 13

Normal 13 10

Pearson's Chi-squared test with Yates' continuity correction

data: table\_oldpeak\_target

X-squared = 0.35953, df = 1, p-value = 0.5488

Table 2. Odds Ratio:

|  |  |  |  |
| --- | --- | --- | --- |
| **Term** | **OR** | **CI\_lower** | **CI\_upper** |
| <chr> | <dbl> | <dbl> | <dbl> |
| (Intercept) | 0.2727273 | 0.06167846 | 0.873364 |
| elevated\_BPOtherwise | 2.2916667 | 0.5488216 | 11.996079 |
| elevated\_oldpeakNormal | 1.8333333 | 0.07089196 | 27.215198 |
| elevated\_BPOtherwise:elevated\_oldpeakNormal | 0.7140496 | 0.03790473 | 21.8166 |

Table 3. Logistic regression without interaction terms:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.2394 0.5834 -2.124 0.0336 \*

elevated\_BPOtherwise 0.7459 0.6611 1.128 0.2592

elevated\_oldpeakNormal 0.3217 0.5564 0.578 0.5631

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 82.692 on 62 degrees of freedom

Residual deviance: 80.588 on 60 degrees of freedom

AIC: 86.588

Number of Fisher Scoring iterations: 4

Table 4. Logistic regression with interaction terms:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.2993 0.6513 -1.995 0.0461 \*

elevated\_BPOtherwise 0.8293 0.7660 1.083 0.2790

elevated\_oldpeakNormal 0.6061 1.3872 0.437 0.6621

elevated\_BPOtherwise:elevated\_oldpeakNormal -0.3368 1.5129 -0.223 0.8238

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 82.692 on 62 degrees of freedom

Residual deviance: 80.539 on 59 degrees of freedom

AIC: 88.539

Number of Fisher Scoring iterations: 4

## Predictive Modeling

Predictive models using linear regression, LASSO, and random forest were developed to further assess the predictive power of these risk factors. The linear model yielded an RMSE of about 0.47, while the LASSO and random forest models produced RMSE values of approximately 0.48 and 0.47, respectively. Tuning procedures for LASSO and random forest indicated that moderate levels of regularization and a larger minimum node size, respectively, provided optimal performance. However, overall model performance remained modest.

# Discussion

## Summary and Interpretation

The analyses conducted in this study do not provide strong evidence that younger individuals with normal cholesterol levels are at an increased risk of cardiovascular disease due to the presence of elevated blood pressure or ST depression. Both the chi-square tests and the logistic regression models, including those adjusted for confounders, failed to reveal statistically significant differences in CVD prevalence between the risk groups. Predictive models further corroborated these findings, demonstrating only modest accuracy with RMSE values around 0.47–0.48 and an AUC of approximately 0.60.

## Strengths and Limitations

A key strength of this study is the comprehensive approach taken in integrating descriptive, inferential, and predictive analyses to evaluate the research question. The use of multiple modeling techniques and tuning strategies provides a robust assessment of model performance. However, limitations include potential issues with sparse data in some subgroups, which may lead to unstable parameter estimates and wide confidence intervals. These limitations suggest that further research with larger sample sizes may be necessary to confirm these findings.

## Conclusions

Based on the current analysis, there is little evidence to support the hypothesis that secondary risk factors such as elevated blood pressure and ST depression significantly increase the risk of heart disease among younger individuals with normal cholesterol levels. These results challenge conventional risk-assessment frameworks and highlight the need for further investigation to better understand the interplay of these risk factors in this population.