Introduction:

The data set we were dealing with was known by the acronym CAIR CVD 2025. The sample size included 1529 patient. Collection of data took place in Bangladesh. The location was Jamalpur Medical College Hospital. The variable we tested in our data were systolic BP, total cholesterol, smoking status, diabetes status, physical activity, BMI, age, sex, and the risk score of CVD. I just love the data in this one. There's just enough data in here to be scientifically valid, and just enough data to actually be meaningful when we talk about prevention.

Thus, in our situation, the manner by which we were able to formulate our plan was first by examining the distribution of systolic blood pressure, then the distribution of their cholesterol, and then, by means of steps three and four, we had to examine if smokers and people with diabetes had indeed higher heart disease risk scores. Lastly, we had to examine if, divided by their levels of physical activity, their levels of blood pressure and overall levels of cholesterol were important in deciding their risks.

One of the principal causes of deaths due to CVD, and the principal risks, relate to modifiable factors such as blood pressure, lipids, smoking, diabetes, and physical inactivity. We intend to explain the influence of each of the factors on the overall risk factor score in CAIR-CVD-2025. These results need to be treated as associational rather than causal.

Methods:

Rstudio was used in doing all the tasks. Data was converted to clinical categories of BP and levels of cholesterol, then plots were made based on each question, pie charts to show distributions, histogram and density plots were needed in comparisons of distributions, then simple scatter plots with simple regression lines were needed in cases where we had to display correlation between variables. Pearson r and p-value were added in the scatter plots.

A pie chart was used to show how normal, elevated, stage 1 HTN, and stage 2 HTN systolic blood pressure categories distribute within the study population. Another pie chart was used to show how Heart Healthy, At-Risk, and Dangerous total cholesterol categories distribute within the study population.

Two Facet Wrapped Histograms, one histogram showing frequency of CVD Score of Smokers and the other histogram showing frequency of CVD score of non-smokers, will be used to answer the question of weather or not there is a difference in CVD risk scores between smokers and non-smokers. Two Facet Wrapped Density Plots, one density plot showing density of of CVD Risk Score of Diabetics and the other density plot showing density of CVD Risk Score of Non-Diabetics, will be used to answer the question of weather or not there is a difference in CVD risk scores between diabetics and non-diabetics.

Three facet wrapped scatterplots with trend lines and pearson correlation tests was plotted. One scatterplot was made for low physical activity, one scatterplot for Moderate Physical Activity, and one scatterplot for High Physical Activity. These three scatterplots helped answer the question of how systolic blood presssure relates to CVD risk score accrosss different levels of physical activity. Another set of three facet wrapped scatterplots with trend lines and pearson correlation tests was created. One scatterplot for low physical activity, one for moderate physical activity, and one for high physical activity. These three scatterplots helped answer the question of how total cholesterol relates to CVD risk score across different levels of physical activity.

We used clinical thresholds to represent BP and cholesterol values, pie charts to represent proportions, histogram/density plots to compare group distributions, and stratified scatter plots with Pearson's r to evaluate strengths and directions of associations. Our approach offers the best compromise between understandability and scientific rigor in terms of audience. Furthermore, all plots in our manuscript have been generated from one data set, and the complete R code ensures our work's replicability.

Table 1: Summary Report of the Dataset

	Overall (N=1529)
Age	
N-Miss	78
Median	46.0
Q	
1,Q3	37.0, 55.0
Range	25.0 - 79.0
Sex	

F	773 (50.6%)
M	756 (49.4%)
BMI	
N-Miss	64
Median	28.2
Q	
1,Q3	22.6, 34.0
Range	15.0 - 46.2
Smoking.Status	
Y	789 (51.6%)
N	740 (48.4%)
Diabetes.Status	
N	752 (49.2%)

Y	777 (50.8%)
Systolic.BP	
N-Miss	71
Median	125.0
Q	
1,Q3	107.0, 141.0
Range	90.0 - 179.0
Blood.Pressure.Category	
Elevated	100 (6.5%)
Hypertension Stage 1	497 (32.5%)
Hypertension Stage 2	632 (41.3%)
Normal	300 (19.6%)
CVD.Risk.Score	

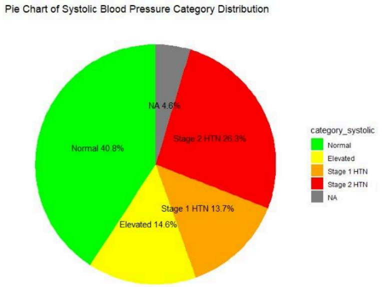
N-Miss	70
Median	16.9
Q	
1,Q3	15.2, 18.6
Range	10.5 - 24.2
Total.Cholesterol..mg.dL.	
N-Miss	73
Median	197.0
Q	
1,Q3	150.0, 249.0
Range	100.0 - 300.0

Description of Summary Table:

Table 1 shows a summary of the baseline characteristics of the CAIR CVD 2025 cohort (N = 1,529), which is roughly balanced for both sex (50.6% female and 49.4% male) and a median age of 46 (IQR: 37–55 years). On the whole, the population has a relatively high profile of cardiometabolic risk, with a median BMI of 28.2, a prevalence of smoking and/or diabetes of over half the population, and a degree of hypertension in over 70% of the population at Stages 1 and 2. The median total cholesterol level is 197 mg/dL, and the median CVD risk score is 16.9.

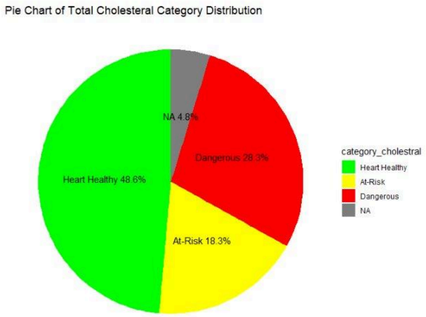
Analysis and Visuals:

Figure 1: Pie Chart of Systolic Blood Pressure Category Distribution



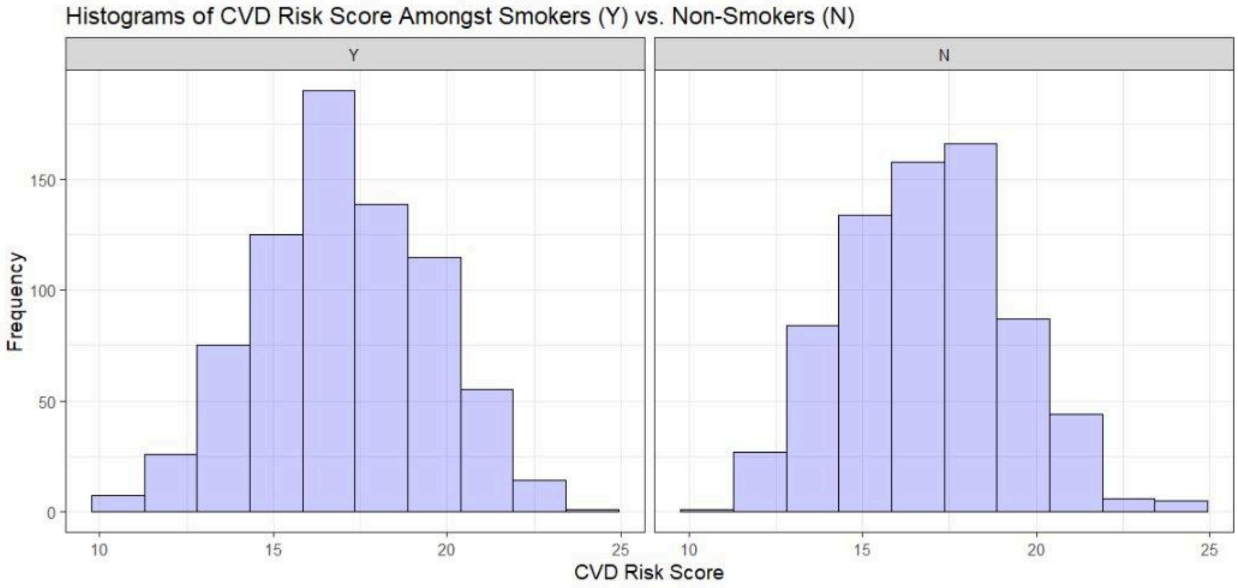
When BP is disaggregated, out of every five people, two fall in the normal category, whereas almost one-fourth of the people actually fall in Stage 2 hypertensive level. The remaining people fall in the higher and crisis levels. All in all, this pie chart shows that high BP is a big problem.

Figure 2: Pie Chart of Total Cholesterol Category Distribution



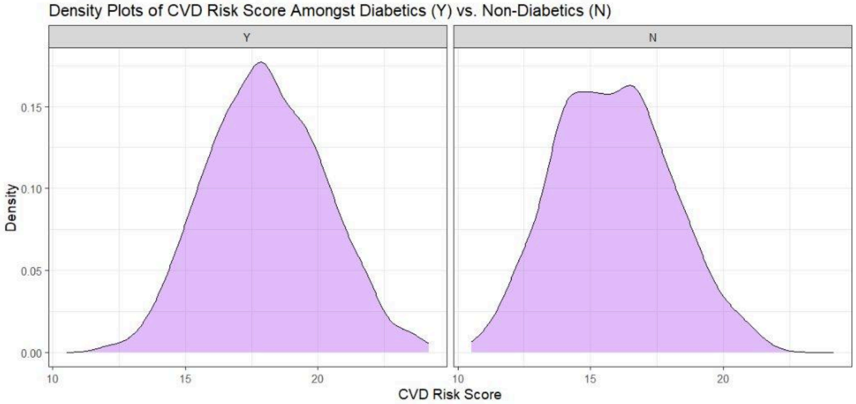
As far as levels of cholesterol, just short of half of patients where “heart healthy,” one fourth where “at risk,” and almost one third where in the “dangerous” category. There is enough percent of patients in the ‘dangerous’ category for there to be a significant health issue.

Figure 3: Histograms of CVD Risk Score Among Smokers vs. Non-Smokers



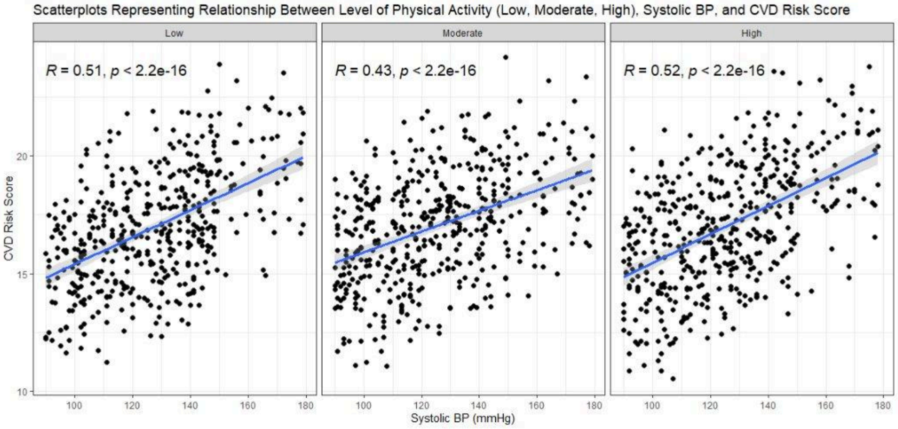
Looking at both histograms, one can easily discern what is happening. The smokers just tend to bunch up at higher levels of cardiovascular risk. One does not need any complicated analysis in order to tell one what's obvious in the data. Smoking cessation should be an intervention, rather than tip, in terms of promoting overall well-being.

Figure 4: Density Plots of CVD Risk Score Among Diabetics vs. Non-Diabetics



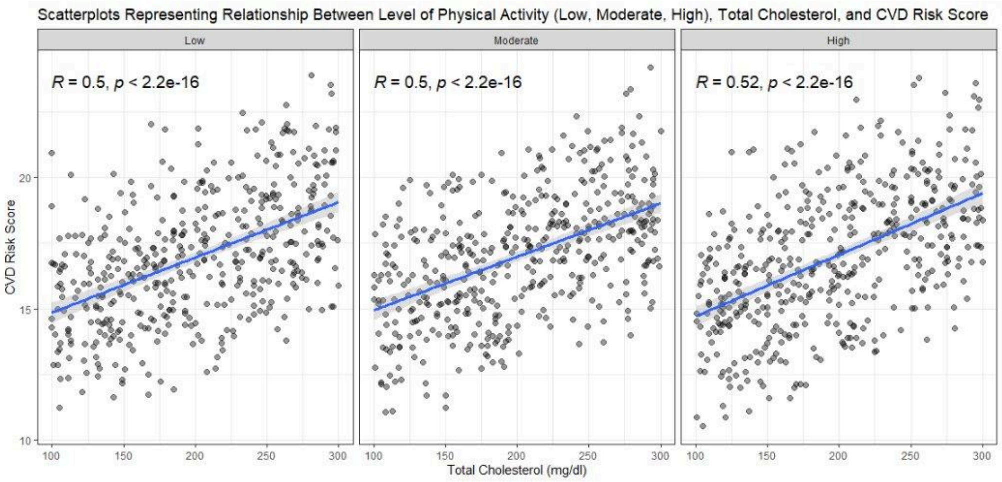
One density plot shows the density of CVD Risk Score of Diabetics and the other density plot shows the density of CVD Risk Score of Non-Diabetics. The Diabetes density curve reaches a greater max height compared to the non-diabetes density curve. The fact that diabetes amplifies cardiovascular risk score is proven when comparing these two density plots.

Figure 5: Scatterplots Representing Relationship Between Level of Physical Activity, Systolic BP, and CVD Risk Score



Across the three scatterplots, the straight upward lines signified that higher systolic BP resulted in higher risk score. R = 0.43-0.52 in all three plots signaling a significant enough relationship. Exercising does help in the reduction of risk score but BP control is an absolute necessity.

Figure 6: Scatterplots Representing Relationship between Level of Physical Activity, Total Cholesterol, and CVD Risk Score



We also observe the same pattern if we examine the overall cholesterol. Regardless of the levels of exercise, the correlation between one's own cholesterol and one's own risk is high, with correlation coefficients of 0.5 in all plots. This is the moment to understand the small benefit of exercise, but we must not forget the overarching importance of the management of cholestrol.

Figures 1-2 answer the distribution questions (groups BP and cholesterol, respectively). Figures 3-4 demonstrate the change in distribution of the risk of CVD in smokers versus non-smokers, and in DMs versus non-DMs, respectively. Figures 5-6 show the positive correlation with the risk of CVD (≈ 0.43-0.52) regardless of physical activity.

Discussion and Conclusion:

There are two themes. One, BP and cholesterol are doing the heavy lifting when it comes to predictive factors in this group of patients. Two, risk factors of smoking and Diabetes increase risk notabely, while exercise lowers risk to a very small degree. The overall message is to first control BP and cholestrol, then prevent smoking and diabetes, then focus on getting sufficient exercise after all that.

Several notable limitations of the dataset included the fact that the data was collected only from individuals in only one setting. Also, there were not any variables in the data related to medication and diet. A final limitation could have been potential self report bias. We were dealing with cross-sectional data, and therefore we cannot conclude anything significant with respect to causation. Next time, what we would do, would be to gather more data with respect to their diet and drugs, and then longitudinal data with respect to the effect of treatment and lifestyle change on risk.

Thus, in conclusion, the higher the person's blood pressure, or the higher his or her levels of cholesterol, the higher his or her chances of cardiovascular diseases even if he or she is active. But lifestyle also plays an important part, and if one actually does desire to work on what can be altered, then one must work on lifestyle change and then screening.

Taken together, the results strongly emphasize the need for BP and lipid management, with added benefit from stopping smoking, controlling diabetes, and getting physically active. The primary limitations involve the fact that data comes from one center, data does not include specifics on medications and diets, and data was obtained in a cross-sectional manner. Future trials should involve data on treatment and diets, especially prospective studies tracking patients longitudinally.

References

Centers for Disease Control and Prevention (2023). Heart Disease Facts. U.S. Department of Health and Human Services. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

Sharker Nirob, et al, CAIR-CVD-2025: An Extensive Cardiovascular Disease Risk Assessment Dataset from Bangladesh. Kaggle. https://www.kaggle.com/datasets/jocelyndumlao/cair-cvd-2025-cardiovascular-risk-from-bangladesh

World Health Organization. (2024). WHO. Prevention of cardiovascular disease : guidelines for assessment and management of total cardiovascular risk. World Health Organization.
https://www.who.int/publications/i/item/9789241547178

R Packages Used: ggplot2, tidyverse, ggpubr, ggsignif, arsenal

Rstudio version 4.5.1 was used to create plots and summary table