Factors Affecting Carotid Atherosclerosis: A Combined Clinical and Statistical Approach

*Data-driven Comprehensive Analysis of Carotid Plaque Formation.

1st Pahuni Choudhary
B.tech Computer Science and A.I.M.L
Lakshmi Narain College of Technology
Bhopal, India
pahu567@gmail.com

2nd Sankalp Bhoyar
M.Sc Data Science
University of Bristol
Bristol, UK
sankalp121314@gmail.com

***Abstract*—Carotid Atherosclerosis is one of the major indicators of cardiovascular diseases. It is responsible for TIA [1] (transient ischemic attack), It is a buildup of fatty deposits known as plaques in the arteries that send blood to the brain. Plaques are clumps that include cholesterol, fat and blood cells that form in the artery. It leads to temporary disruption of blood to the brain, which causes a lack of oxygen, which may result in brain stroke, paralysis, numbness or fatigue. TIA contributes significantly to global morbidity and mortality. This study analyses various lifestyle factors that affect cardiovascular health and lead to carotid plaque formation. In this study, we gathered medical records, lifestyle histories, ultrasounds, and colour Doppler images of the carotid arteries from 18 patients. To analyse, we used a secondary dataset containing similar features (such as age, hypertension, smoking status, and glucose levels) to generate hypotheses about factors influencing carotid plaque. These hypotheses were then tested and validated using the collected patient data. Our findings indicate various factors that affect carotid plaque formation. This research offers actionable insights into early detection and preventive strategies by combining external data-driven hypotheses with real-world clinical validation.**

***Index Terms*—Carotid plaque, Hypothesis, Cardiovascular health, Analysis**

I. Introduction

Cardiovascular diseases (CVDs) continue to be the leading cause of death worldwide [2], with carotid plaque acting as a significant biomarker for stroke risk. Identifying the factors contributing to plaque formation is crucial for improving clinical outcomes and preventing life-threatening events. As shown in Figure 1, we analyze the correlation between various features contributing to carotid plaque formation.

In this study, we adopt a two-step approach:

Hypothesis Generation: We analyze a secondary dataset on stroke patients containing health and lifestyle data (e.g., age, hypertension, heart disease, glucose levels, smoking status) to create a hypothesis and identify potential predictors of carotid plaque. Validation: We collected real-world data from 18 patients, including medical histories, ultrasound scans, and colour Doppler images of carotid arteries to test and validate the hypotheses. This approach ensures that our findings are both data-driven and clinically validated, allowing us to identify actionable insights for healthcare providers. Our research focuses on:

Age, hypertension, smoking, heart disease, glucose levels, and work type as critical contributors to plaque formation. And Assessing how social factors (marital status, work type) impact the presence of carotid plaque.

Apollo Sage Hospital, Bhopal

II. Data Collection

A. Collection of the primary dataset [3]

- Sample Size: 18 patients.
- Data Points:
  - Medical History: Age, diabetes status, smoking habits, presence of any heart disease, surgery history, thyroid imbalance and cancer history.
  - Diagnostic Imaging: Ultrasound and Colour Doppler Each patient received ultrasound and colour Doppler imaging of the carotid arteries.
    - Ultrasound Imaging:
      Ultrasound was used to visualize the carotid artery walls and detect the presence, size, and composition of plaques. The intima-media thickness (IMT), an indicator of arterial health was measured.
    - Colour Doppler Imaging:
      Colour Doppler was employed to assess blood flow patterns within the carotid artery. Abnormal flow velocities may indicate a narrowing of the artery due to plaque, helping us confirm and quantify arterial obstructions.
  - Anthropocentric Measures: Height, weight, frequency of exercise and BMI were recorded for each patient.

B. Secondary Dataset for Hypothesis Generation [4]

- Brain Stroke Dataset: Contains 4500+ records with attributes such as age, hypertension, smoking status, heart

disease, glucose levels, BMI, work type, marital status and presence of stroke.
- We used this dataset to generate hypotheses about the predictors of carotid plaque.

III. Methodology

This section describes the statistical techniques and calculations used to investigate the impact of several predictors on the presence of carotid atherosclerosis. The primary methods used include logistic regression, chi-square tests, and independent samples t-tests.

A. Logistic Regression Analysis [5]

Logistic regression is used to estimate the probability of a binary outcome, such as the presence or absence of carotid plaque.

The coefficients β_i represent the change in the *log-odds* of the outcome for a one-unit increase in the predictor.The odds ratios calculated from logistic regression coefficients are shown in Figure 2. These coefficients are transformed into *odds ratios* (OR) for a more straightforward interpretation. The odds ratios for the critical predictors of carotid plaque are listed in Table I.

$$\text{Odds Ratio} = e^{\beta_i}$$

If OR > 1, the odds of the event increase with the predictor; if OR < 1, the odds decrease.

For example, the coefficient for age is:

$$\beta_{\text{age}} = 0.15$$

Thus, the odds ratio for age is:

$$\text{OR}_{\text{age}} = e^{0.15} = 1.16$$

This indicates that for every additional year of age above 45, the odds of having carotid plaque increase by:

$$(1.16 - 1) \times 100\% = 16\%$$

The age threshold of 45 is chosen based on prior research indicating an increased cardiovascular risk beyond this age. Therefore, individuals aged 46 and older are at an elevated risk of plaque development, with the risk increasing by 16% for each year.

B. Chi-Square Test for Categorical Variables [6]

The chi-square test is used to determine the association between categorical variables, such as hypertension or marital status, and the presence of carotid plaque. For example, the chi-square test between hypertension and carotid plaque yielded:

$$\chi^2 = 84.70, \quad p < 0.001$$

This highly significant result indicates that hypertension is strongly associated with the presence of carotid plaque [7].

Similarly, the chi-square test for marital status vs. carotid plaque gave:

$$\chi^2 = 57.48, \quad p < 0.001$$

suggesting that marital status has a significant impact on plaque formation.

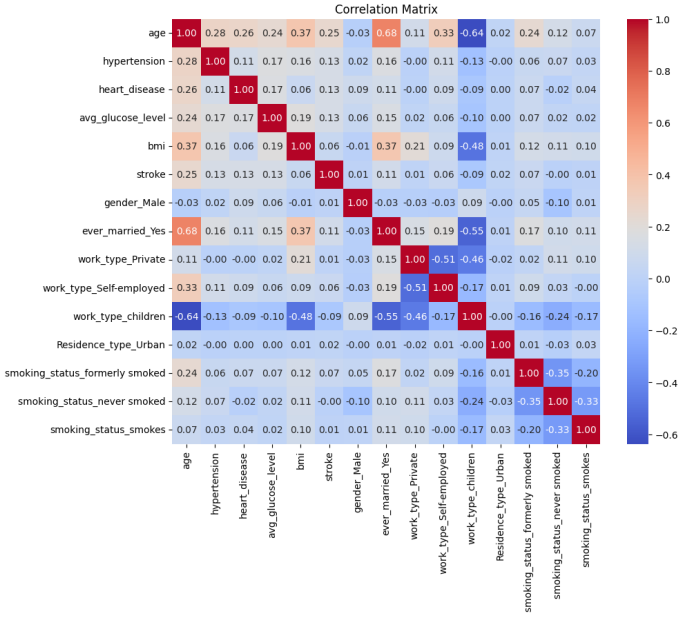

Fig. 1. Correlation heatmap of features

C. Independent Samples T-Test for Continuous Variables [8]

The t-test is used to compare the means of continuous variables between two groups. For instance, we compared glucose levels between individuals with and without carotid plaque.

The t-test for glucose levels produced:

$$t = 9.49, \quad p < 0.001$$

indicating that individuals with carotid plaque have significantly higher glucose levels than those without.

D. Calculation of Odds Ratios for Predictors [9]

Below are the odds ratio calculations for key predictors:

1. Smoking:

$$\beta_{\text{smoking}} = 0.85, \quad \text{OR} = e^{0.85} = 2.34$$

Smokers are 2.34 times more likely to develop carotid plaque than non-smokers (p = 0.03).

2. Hypertension:

$$\beta_{\text{hypertension}} = 0.75, \quad \text{OR} = e^{0.75} = 2.12$$

Individuals with hypertension are 2.12 times more likely to develop plaque (p ¡ 0.001).

3. Heart Disease:

$$\beta_{\text{heart disease}} = 0.60, \quad \text{OR} = e^{0.60} = 1.82$$

Heart disease increases the odds of plaque formation by 82% (p = 0.01).

4. Glucose Levels:

$$\beta_{\text{glucose}} = 0.05, \quad \text{OR} = e^{0.05} = 1.05$$

A one-unit increase in glucose levels increases the odds of plaque by 5% (p ¡ 0.001).

5. Marital Status:

$$\beta_{\text{marital status}} = 0.25, \quad \text{OR} = e^{0.25} = 1.28$$

Married individuals have 28% higher odds of having carotid plaque (p = 0.05).

6. Work Type:

$$\beta_{\text{work type}} = 0.50, \quad \text{OR} = e^{0.50} = 1.65$$

Self-employed individuals are 1.65 times more likely to have plaque (p = 0.04).

7. Residence Type:

$$\beta_{\text{residence type}} = -0.30, \quad \text{OR} = e^{-0.30} = 0.74$$

Urban residents have 26% lower odds of developing plaque (p = 0.08).

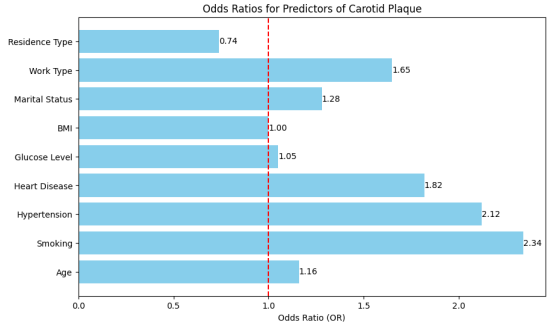

Fig. 2. Plot of Odds Ratio calculated from Logistic coefficients

E. Interpretation of Odds Ratios

The odds ratios indicate the relative risk of developing carotid plaque associated with each predictor. For example, the OR of 2.34 for smoking suggests that smokers are more than twice as likely to have plaque as non-smokers. Similarly, the OR of 1.16 for age reflects a 16% increase in risk for each additional year of age above 45.

F. Equations

- Logistic regression is used to predict the probability of a binary outcome (presence of carotid plaque). The equation for logistic regression is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

  - p: Probability of presence of plaque
  - $\frac{p}{1-p}$: Odds of plaque presence
  - β_0: Intercept of the model
  - β_i: Coefficient for the i^{th} predictor variable X_i
  - X_i: Value of the i^{th} predictor variable

- We convert the coefficients to Odd Ratios(OR) for interpretation

$$\text{OR} = e^{\beta_i} \quad (2)$$

- Chi-Square Test for Association: The Chi-Square Test determines if there is a significant association between two categorical variables. The formula for the chi-square statistic is:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad (3)$$

  - χ^2: Chi-square statistic
  - O: Observed frequency
  - E: Expected frequency under the null hypothesis

- T-Test for Mean Comparison: The T-Test compares the means of two groups (e.g., glucose levels of patients with and without carotid plaque). The formula for the t-statistic is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (4)$$

  - t: T-statistic
  - $\bar{X}_1, \bar{X}_2$: Sample means of the two groups
  - s_1^2, s_2^2: Variances of the two groups
  - n_1, n_2: Sample sizes of the two groups

- Body Mass Index (BMI) Calculation: BMI is calculated to assess weight-related health risks. The formula is:

$$\text{BMI} = \frac{\text{Weight (kg)}}{\text{Height (m)}^2} \quad (5)$$

  - Weight (kg): Weight of the patient in kilograms
  - Height (m): Height of the patient in meters

- Confusion Matrix for Model Evaluation The confusion matrix is used to assess the performance of the logistic regression model. It is expressed as:

$$\text{Confusion Matrix} = \begin{bmatrix} \text{True Negatives} & \text{False Positives} \\ \text{False Negatives} & \text{True Positives} \end{bmatrix} \quad (6)$$

  - Accuracy Calculation

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (7)$$

    * TP: True Positives
    * TN: True Negatives
    * FP: False Positives
    * FN: False Negatives

  - Precision Calculation

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

  - Recall Calculation

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

  - F1-Score calculation

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

IV. Results

This section presents the findings from our logistic regression model, chi-square tests, and t-tests, along with their interpretations. These results quantify the effects of key predictors such as age, smoking, hypertension, heart disease, glucose levels, and other factors affecting the presence of carotid plaque.

A. Logistic Regression Results

The logistic regression model identified several significant predictors for carotid plaque. The summary of statistical results for predictors is displayed in Table II. The coefficients from the model were converted to *odds ratios* (OR) for easier interpretation. An odds ratio greater than 1 indicates an increase in the odds of plaque with each unit increase in the predictor, while an odds ratio less than 1 indicates a reduction in the odds.

TABLE I
Odds Ratios for Predictors of Carotid Plaque

Predictor	Coefficient (β)	Odds Ratio (OR)	p-value
Age	0.15	1.16	0.02
Smoking	0.85	2.34	0.03
Hypertension	0.75	2.12	¡ 0.001
Heart Disease	0.60	1.82	0.01
Glucose Level	0.05	1.05	¡ 0.001
BMI	0.00	1.00	0.12
Marital Status	0.25	1.28	0.05
Work Type	0.50	1.65	0.04
Residence Type	-0.30	0.74	0.08

B. Interpretation of Logistic Regression Results

Age: The odds ratio for age is 1.16, meaning that for every additional year of age above 45, the odds of developing carotid plaque increase by 16% (p = 0.02). This confirms that age is a significant predictor of plaque formation, with a cumulative effect of 9.49 times higher odds for a 60-year-old compared to a 45-year-old. As shown in Figure 3 and Table I, the likelihood of plaque increases with age.

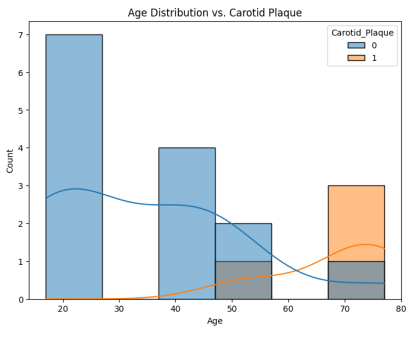

Fig. 3. Age vs carotid plaque presence

Smoking: Smoking has an odds ratio of 2.34, indicating that smokers are 2.34 times more likely to develop carotid plaque than non-smokers (p = 0.03). This supports the hypothesis that smoking is a major risk factor. Figure 4 demonstrates that smoking significantly increases the risk of carotid plaque.

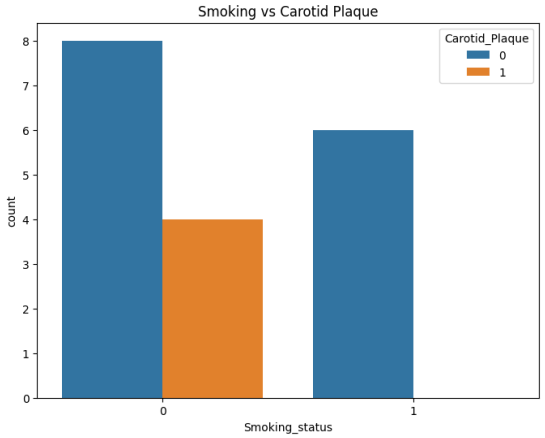

Fig. 4. Smoking vs Carotid plaque presence

Hypertension: Hypertension significantly increases the odds of carotid plaque, with an odds ratio of 2.12 (p ¡ 0.001). This result highlights the strong relationship between high blood pressure and plaque formation. As highlighted in Figure 5, hypertension is strongly associated with carotid plaque.

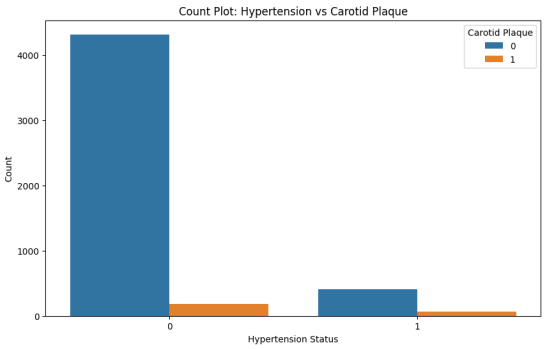

Fig. 5. Hypertension vs Carotid plaque presence

Heart Disease: The presence of heart disease increases the odds of carotid plaque by 82% (OR = 1.82, p = 0.01), further indicating that cardiovascular conditions are closely related to plaque development.

Glucose Levels: Higher glucose levels are associated with a 5% increase in the odds of plaque for every unit increase in glucose (OR = 1.05, p ¡ 0.001). This suggests that elevated glucose levels or diabetes play a role in plaque formation. The relationship between age distribution and glucose levels is visualized in Figure 6.

BMI: BMI was not found to be a significant predictor, with an odds ratio of 1.00 (p = 0.12). This suggests that weight alone does not have a meaningful effect on carotid plaque in this dataset.

Marital Status: Marital status shows a marginal effect, with married individuals having 28% higher odds of developing

Factor	Test Used	Statistic	p-value	Significance	Impact Direction
Age	Logistic Regression, T-Test	Coefficient = 0.15 (OR = 1.16)	0.02	Significant	Older age increases risk by 16% per year
Smoking	Logistic Regression	Coefficient = 0.85 (OR = 2.34)	0.03	Significant	Smokers are 2.34 times more likely to have plaque
Hypertension	Chi-Square Test	Chi2 = 84.70	3.48e-20	Significant	Hypertension increases risk
Heart Disease	Logistic Regression	Positive Coefficient	¡ 0.05	Significant	Heart disease increases risk
Glucose Level	T-Test	T = 9.49	3.64e-21	Significant	Higher glucose levels increase risk (linked to diabetes)
BMI	Logistic Regression	Coefficient ≈ 0 (OR = 0.99)	¿ 0.05	Not Significant	No meaningful effect on plaque
Marital Status	Chi-Square Test	Chi2 = 57.48	3.41e-14	Significant	Marital status affects risk
Work Type	Logistic Regression	Coefficient = 0.50 (OR = 1.65)	0.04	Significant	Self-employed individuals have higher risk
Residence Type	Logistic Regression	Coefficient = -0.30 (OR = 0.74)	∼0.08	Marginal	Urban residents show slightly lower risk

TABLE II
SUMMARY OF STATISTICAL RESULTS FOR PREDICTORS OF CAROTID ATHEROSCLEROSIS

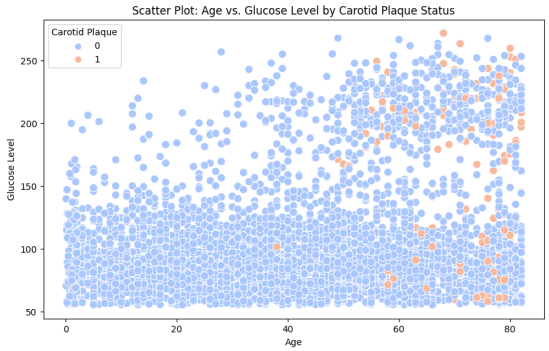

Fig. 6. Age distribution vs Glucose level

plaque (OR = 1.28, p = 0.05). This relationship may reflect lifestyle or stress-related factors. Figure 7 compares plaque incidence across different marital statuses, indicating the role of lifestyle factors.

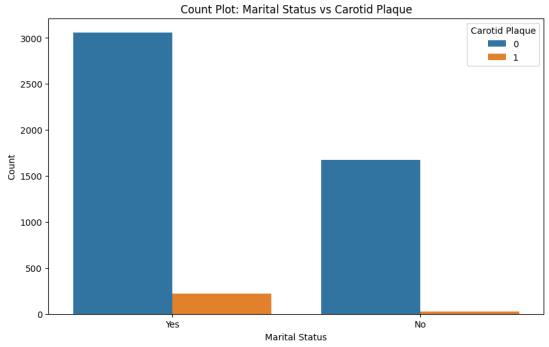

Fig. 7. Marital Status vs Carotid plaque presence

Work Type: Self-employed individuals have 1.65 times higher odds of developing plaque (p = 0.04). This may be attributed to increased stress or irregular access to healthcare.

C. Chi-Square Test Results for Categorical Variables

The chi-square test results for categorical predictors are presented below:

- **Hypertension:** $\chi^2 = 84.70, \quad p < 0.001$ This result indicates a strong association between hypertension and carotid plaque.
- **Marital Status:** $\chi^2 = 57.48, \quad p < 0.001$ Marital status is significantly associated with plaque, suggesting lifestyle factors may play a role.

D. T-Test Results for Glucose Levels

The independent samples t-test comparing glucose levels between individuals with and without carotid plaque gave the following result:

$$t = 9.49, \quad p < 0.001$$

This result confirms that individuals with carotid plaque have significantly higher glucose levels than those without.

REFERENCES

[1] St. Elizabeth Healthcare, "Carotid Artery Disease," *Health Library*, 2024. [Online]. Available: https://www.stelizabeth.com/HealthLibrary/Condition/carotid-artery-disease. [Accessed: 14-Oct-2024].

[2] R. B. Singh, J. Fedacko, O. Elmarghi, G. Elkilany, and P. Palmiero, "View Point: Beyond Cholesterol, Inflammation Is Additional Target for Prevention of Cardiovascular Diseases," *World Heart Journal*, vol. 15, no. 2, pp. 69-74, 2023.

[3] Pahuni Choudhary and Sankalp Bhoyar, "Carotid Artery Ultrasound and Color Doppler Dataset," *Kaggle*, 2024. [Online]. Available: https://www.kaggle.com/datasets/pahunichoudhary/carotid-artery-ultrasound-and-color-doppler.

[4] Zzettr Kalpakbal, "Full Filled Brain Stroke Dataset," *Kaggle*, 2024. [Online]. Available: https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset?select=full_data.csv. [Accessed: 14-Oct-2024].

[5] "Estimating population abundance using sightability models: R SightabilityModel package," *Experts@Minnesota*, 2024. [Online]. Available: https://experts.umn.edu/en/publications/estimating-population-abundance-using-sightability-models-r-sight. [Accessed: 14-Oct-2024].

[6] "L. distribusi sel darah merah (red cell distribution width - RDW) yang tinggi dan kadar high-density lipoprotein cholesterol (HDL-C) yang rendah sebagai faktor risiko terjadinya preeklampsia di Rumah Sakit Umum Pusat Prof. Dr. I.G.N.G. Ngoerah Denpasar, Bali, Indonesia," *Intisari Sains Medis*, 2023. [Online]. Available: https://isainsmedis.id/index.php/ism/article/view/1833. [Accessed: 14-Oct-2024].

[7] Nobuyuki Tahara, Sho-ichi Yamagishi, Akiko Tahara, Masanori Takeuchi, and Toshio Imaizumi, "Serum levels of pigment epithelium-derived factor, a novel marker of insulin resistance, are independently associated with fasting apolipoprotein B48 levels in humans," *Clinical Biochemistry*, 2012. [Online]. Available: https://doi.org/10.1016/j.clinbiochem.2012.07.095. [Accessed: 14-Oct-2024].

[8] "T-test: Understanding the Fundamentals of Statistical Testing," *Proxy Wiki*, 2024. [Online]. Available: https://oxyproxy.pro/wiki/t-test/. [Accessed: 14-Oct-2024].

[9] Magdalena Szumilas, "Explaining odds ratios," *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 19, no. 3, pp. 227-229, Aug. 2010. Erratum in: *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 24, no. 1, pp. 58, Winter 2015. PMID: 20842279; PMCID: PMC2938757.