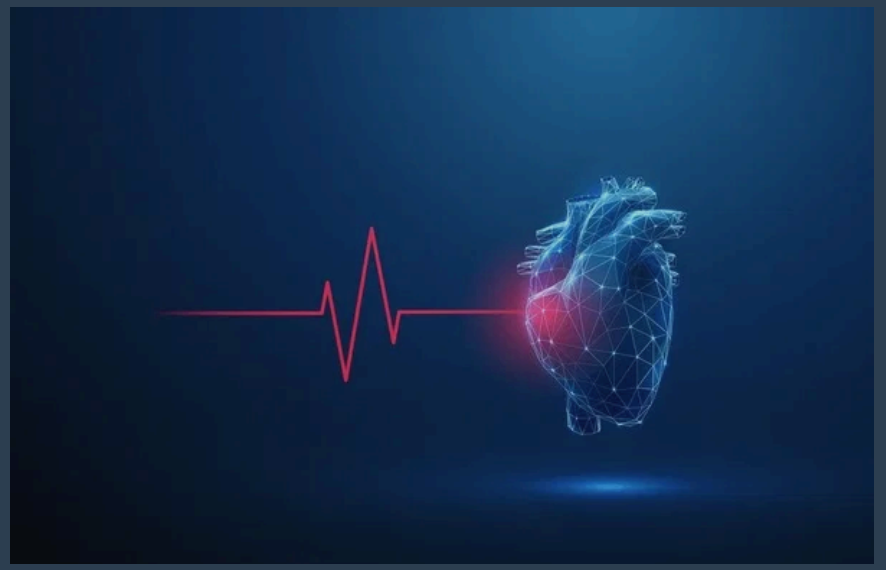
Heart Disease Prediction: 10-Year Risk Analysis

Predicting Cardiovascular Risk with Machine Learning



Objective: Identify high-risk individuals using clinical and behavioral factors.

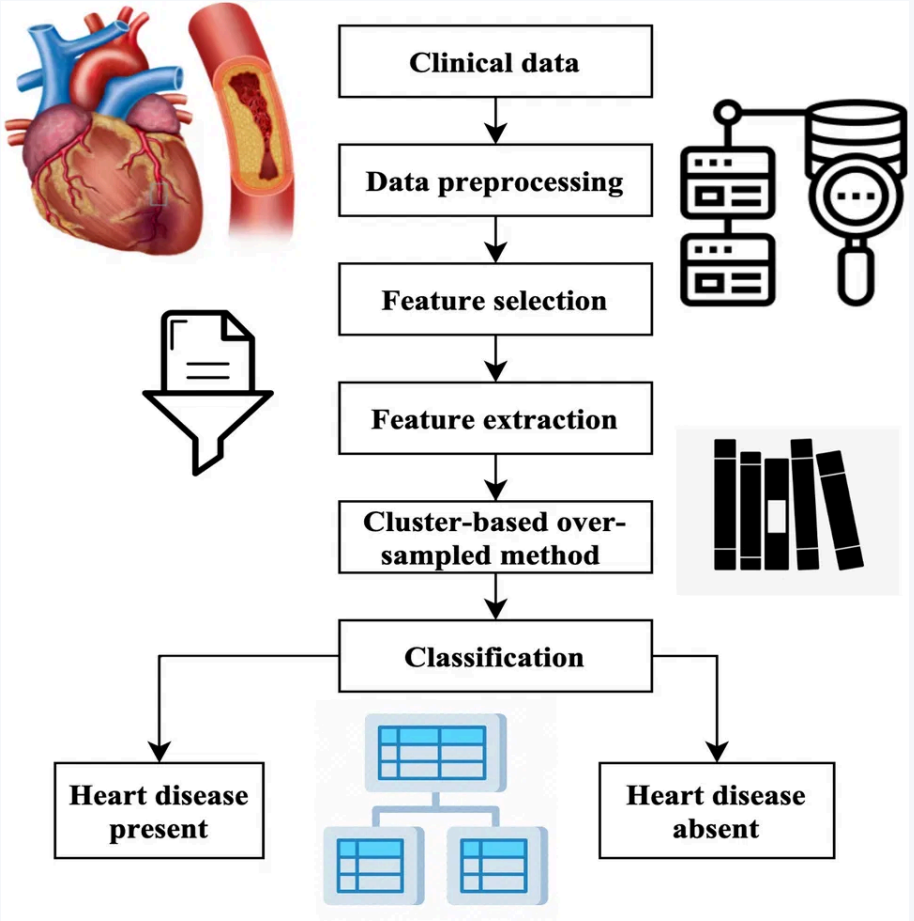
Significance: Early detection for life-saving medical interventions.

Methodology: Data preprocessing, EDA, and Logistic Regression modeling.

The Framingham Dataset: A Clinical Foundation

The dataset originates from the Framingham Heart Study, a renowned long-term cardiovascular cohort study. It initially contains 4,240 records, each characterized by 15 clinical and behavioral features.

Feature Category	Key Variables
Demographics	Age, Gender, Education
Behavioral	Current Smoker, Cigs Per Day
Medical History	BP Meds, Stroke, Hypertension, Diabetes
Clinical Metrics	Cholesterol, BP, BMI, Heart Rate, Glucose



Data Preprocessing and Quality Assurance

Feature Standardization

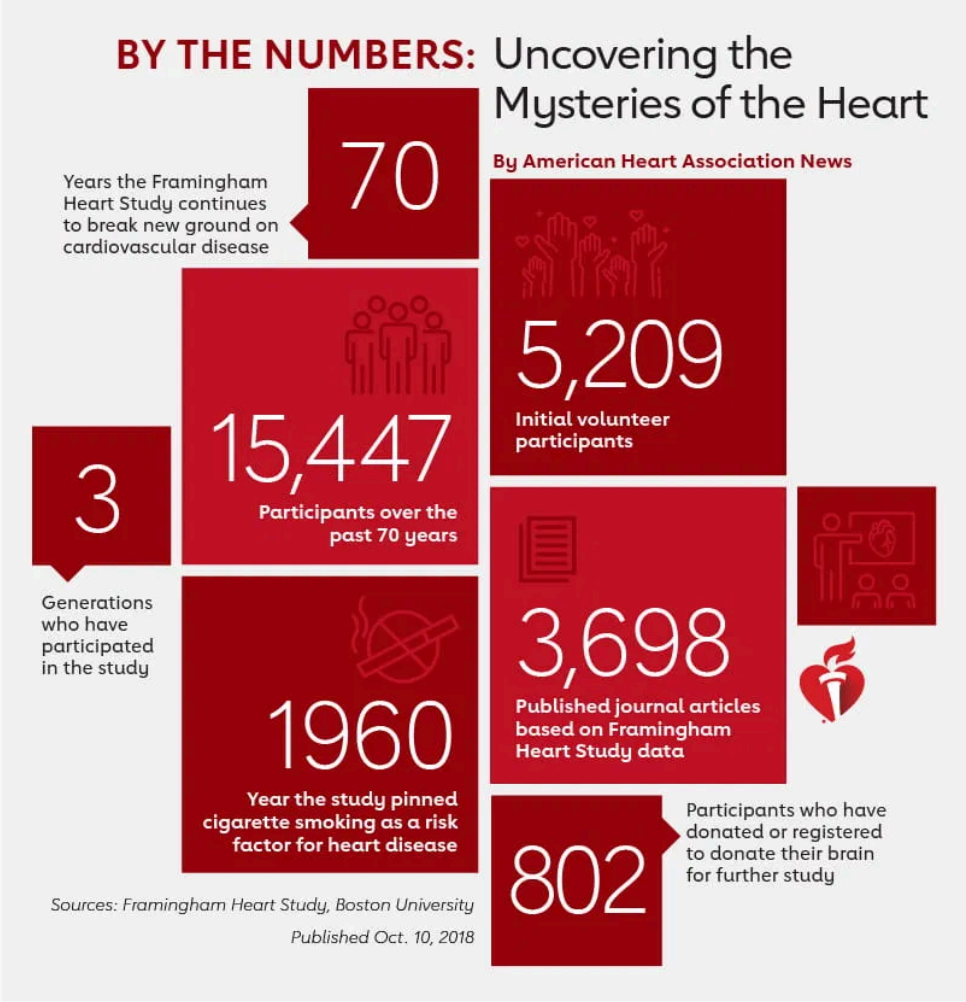
Standardized column names for clarity, such as renaming 'male' to 'gender' and 'totChol' to 'total_cholesterol'.

Missing Value Management

Identified 489 records with missing values, primarily in 'glucose' and 'BPMeds' columns.

Data Cleaning Strategy

Opted for row deletion to maintain data integrity, resulting in a refined dataset of 3,751 records for final analysis.



Analyzing Heart Disease Distribution

Significant Class Imbalance

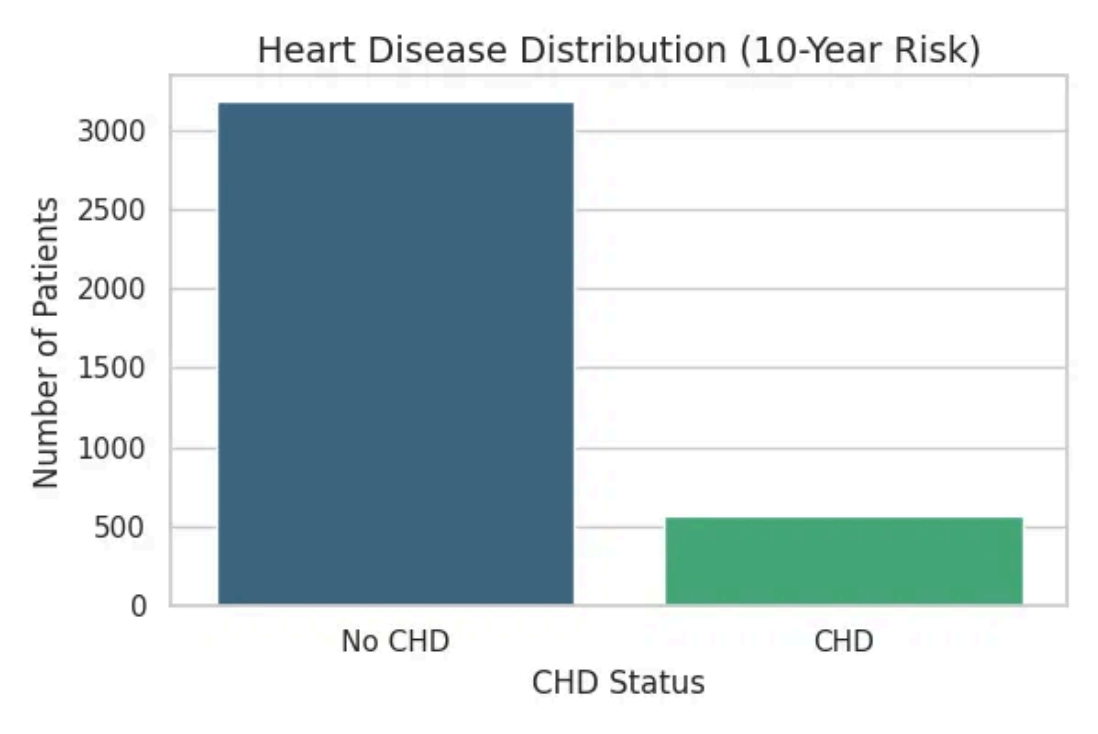
Approximately 15% of participants developed CHD within the 10-year window, while 85% remained healthy.

Impact on Modeling

This disparity necessitates careful evaluation to avoid biased accuracy, as models may favor the majority class.

Evaluation Strategy

Emphasis is placed on robust metrics like precision, recall, and F1-score rather than simple accuracy alone.



Demographic and Behavioral Insights

Age Factor

Heart disease risk shows a strong positive correlation with increasing age. The highest risk concentration is observed among individuals in the **50-70 age bracket**, highlighting age as a primary driver of cardiovascular risk.

Smoking Habits

Analysis of smoking intensity reveals significant variation. The **cigs_per_day** variable shows a wide range of habits, providing a critical behavioral metric for assessing long-term heart health outcomes.

Health Indicators

The study maintains a balanced gender representation. Baseline health indicators, including **BMI and heart rate**, provide a comprehensive overview of the cohort's general physical status prior to modeling.

Identifying Clinical Outliers

Detection Method

Outlier detection was performed using **boxplots** to identify extreme values in clinical metrics. This process is essential for ensuring the robustness of the predictive model against extreme data points.

Systolic Blood Pressure

Significant outliers were observed in **systolic blood pressure**, indicating cases of severe hypertension. These extreme values could disproportionately influence the model if not properly addressed.

Cholesterol Levels

Some participants exhibited exceptionally high **total cholesterol levels**, exceeding 400 mg/dL. Understanding these outliers is crucial for both clinical risk assessment and model training.

Logistic Regression: The Predictive Engine

Model Selection

Logistic Regression was selected for its high interpretability and proven effectiveness in binary classification tasks within clinical research.

Feature Scaling

Applied **StandardScaler** to normalize numerical variables, ensuring that all clinical metrics contribute equally to the model's decision-making process.

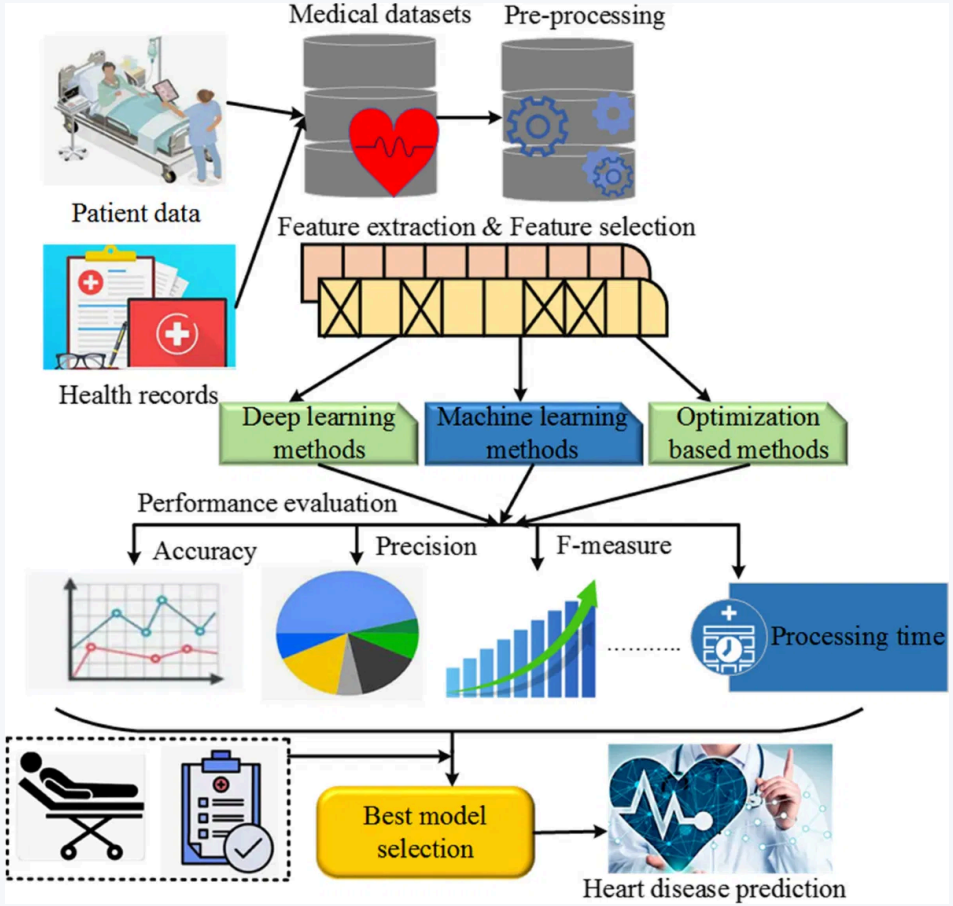
Validation Strategy

The dataset was partitioned into a **training set (80%)** and a **testing set (20%)** to facilitate rigorous validation and assess generalization.

Evaluating Model Performance

The Logistic Regression model achieved an overall accuracy of **86.41%** on the test dataset. While accuracy is high, a detailed metric analysis reveals class-specific performance variations.

Metric	Score
Overall Accuracy	86.41%
Precision (No CHD)	0.87
Recall (CHD)	0.02
Weighted F1-Score	0.81



Confusion Matrix: Prediction Breakdown

TRUE NEGATIVES

841

The model correctly identified individuals who did not develop CHD, showing high reliability for negative cases.

FALSE NEGATIVES

127

Individuals with CHD were missed by the model, representing a critical area for future clinical optimization.

TRUE POSITIVES

3

Only a small fraction of actual CHD cases were correctly flagged in the current model configuration.

FALSE POSITIVES

8

The model is highly conservative, resulting in very few false alarms for healthy individuals.

> *Insight: The model prioritizes the avoidance of false alarms over the detection of all risk cases, making it a conservative screening tool.*

ROC Curve and Diagnostic Ability

Diagnostic Power

The **Area Under the Curve (AUC)** serves as a measure of the model's diagnostic power. An AUC significantly above 0.5 indicates that the model performs better than random guessing in distinguishing between risk classes.

Visual Analysis

The ROC curve illustrates the trade-off between **sensitivity** (True Positive Rate) and **specificity** (False Positive Rate) across various decision thresholds for clinical assessment.

Model Utility

While the overall accuracy is high, the ROC curve suggests that there is significant room for **optimization** to improve the model's sensitivity to CHD cases in a screening context.

Conclusions and Future Directions

Key Findings

Clinical factors such as age and blood pressure are potent predictors of long-term heart disease risk. The study confirms the high diagnostic value of these baseline metrics.

Model Assessment

Logistic Regression provides a solid baseline with 86.41% accuracy, but its performance is limited by the inherent class imbalance of the dataset.

Future Recommendations

Future research should explore advanced techniques such as SMOTE for oversampling or ensemble methods like Random Forest and XGBoost to improve sensitivity.

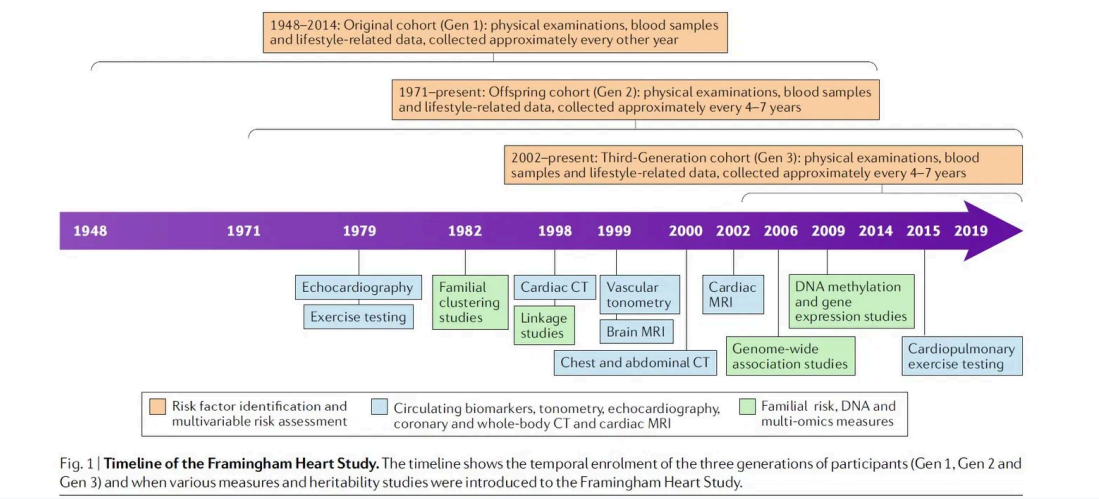


Fig. 1 | **Timeline of the Framingham Heart Study.** The timeline shows the temporal enrolment of the three generations of participants (Gen 1, Gen 2 and Gen 3) and when various measures and heritability studies were introduced to the Framingham Heart Study.

The Framingham Heart Study: Decades of clinical significance and data evolution.