CardioDetect Data Dictionary

Complete Feature Reference

---

1. Dataset Overview

1.1 Summary Statistics

Attribute	Value
Total Patients	16,123
Total Features	34
Target Variable	Binary (CHD 0/1)
Missing Values	0 (complete dataset)
Time Period	1948-2023

1.2 Data Sources

Source	Patients	Percentage	Description
Framingham Heart Study	~4,000	25%	Longitudinal cardiovascular cohort from Framingham, MA
NHANES	~10,000	62%	National Health and Nutrition Examination Survey
Custom Collection	~2,000	13%	Supplementary clinical records
Total	**16,123**	**100%**	Combined dataset

1.3 Geographic Coverage

- **Primary**: United States
- **Cohort Demographics**: Representative of US adult population
- **Age Range**: 18-100 years

1.4 Data Collection Period

Source	Collection Years
Framingham	1948-2023 (ongoing)
NHANES	2015-2020
Custom	2020-2023

---

2. Feature Descriptions

2.1 Demographic Features (2)

Feature	Type	Range	Unit	Description	Source
age	Integer	18-100	years	Patient age at time of examination	All sources
sex	Integer	0-1	binary	Biological sex: 0=Female, 1=Male	All sources

2.2 Clinical Measurements (6)

Feature	Type	Range	Unit	Description	Source
systolic_bp	Float	80-250	mmHg	Systolic blood pressure (upper reading)	All sources
diastolic_bp	Float	40-150	mmHg	Diastolic blood pressure (lower reading)	All sources
bmi	Float	15-60	kg/m²	Body Mass Index = weight(kg) / height(m)²	All sources
heart_rate	Integer	40-150	bpm	Resting heart rate (beats per minute)	All sources
total_cholesterol	Float	100-400	mg/dL	Total serum cholesterol	All sources
fasting_glucose	Float	50-300	mg/dL	Fasting blood glucose level	All sources

2.3 Risk Factor Indicators (5)

Feature	Type	Range	Unit	Description	Source
smoking	Integer	0-60	cigs/day	Average cigarettes smoked per day	All sources
bp_meds	Integer	0-1	binary	Currently on blood pressure medication	All sources
hypertension	Integer	0-1	binary	Diagnosed with hypertension	All sources
diabetes	Integer	0-1	binary	Diagnosed with diabetes	All sources

2.4 Derived Measurements (4)

Feature	Type	Range	Unit	Description	Formula
pulse_pressure	Float	20-120	mmHg	Difference between systolic and diastolic BP	systolic_bp - diastolic_bp
mean_arterial_pressure	Float	50-150	mmHg	Average arterial pressure during cardiac cycle	(systolic_bp + 2×diastolic_bp) / 3
metabolic_syndrome_score	Integer	0-5	count	Count of metabolic syndrome components	Sum of flags

2.5 Clinical Flags (4)

Feature	Type	Range	Description	Threshold
hypertension_flag	Integer	0-1	Elevated blood pressure indicator	systolic_bp 140 OR diastolic_bp 90
high_cholesterol_flag	Integer	0-1	Elevated cholesterol indicator	total_cholesterol 240
high_glucose_flag	Integer	0-1	Elevated glucose indicator	fasting_glucose 126
obesity_flag	Integer	0-1	Obesity indicator	bmi 30

2.6 Log Transformations (3)

Feature	Type	Range	Description	Formula
log_total_cholesterol	Float	4.6-6.0	Natural log of total cholesterol	ln(total_cholesterol)
log_fasting_glucose	Float	3.9-5.7	Natural log of fasting glucose	ln(fasting_glucose)
log_bmi	Float	2.7-4.1	Natural log of BMI	ln(bmi)

Purpose: Log transformations normalize right-skewed distributions and reduce outlier influence.

2.7 Interaction Terms (3)

Feature	Type	Description	Formula
age_sbp_interaction	Float	Age × systolic BP interaction	age × systolic_bp
bmi_glucose_interaction	Float	BMI × glucose interaction	bmi × fasting_glucose
age_smoking_interaction	Float	Age × smoking interaction	age × smoking

Purpose: Capture non-linear relationships between risk factors.

2.8 Age Group Categories (5)

Feature	Type	Description	Age Range
age_group_<40	Boolean	Young adult indicator	age $<$ 40
age_group_40-49	Boolean	Middle-aged (early)	40 age $<$ 50
age_group_50-59	Boolean	Middle-aged (late)	50 age $<$ 60
age_group_60-69	Boolean	Senior (early)	60 age $<$ 70
age_group_70+	Boolean	Senior (late)	age 70

Encoding: One-hot encoded, mutually exclusive categories.

2.9 BMI Categories (4)

Feature	Type	Description	BMI Range
bmi_cat_Underweight	Boolean	Underweight indicator	bmi $<$ 18.5
bmi_cat_Normal	Boolean	Normal weight indicator	18.5 bmi $<$ 25
bmi_cat_Overweight	Boolean	Overweight indicator	25 bmi $<$ 30
bmi_cat_Obese	Boolean	Obese indicator	bmi 30

Encoding: One-hot encoded based on WHO BMI classification.

3. Data Quality Metrics

3.1 Completeness

Metric	Value
Total Records	16,123
Complete Records	16,123
Missing Values	0
Completeness Rate	100%

3.2 Validity

All features are validated against clinical ranges:

Feature	Valid Range	Invalid Count	Validity Rate
age	18-100	0	100%
systolic_bp	80-250	0	100%
diastolic_bp	40-150	0	100%
bmi	15-60	0	100%
total_cholesterol	100-400	0	100%
fasting_glucose	50-300	0	100%

3.3 Outlier Handling

Approach	Description
Detection	IQR method (1.5× interquartile range)
Treatment	Winsorization to 1st/99th percentile
Verification	Clinical range validation

3.4 Data Encoding

Feature Type	Encoding Method
Binary (0/1)	Direct numeric
Categorical	One-hot encoding
Continuous	StandardScaler normalization
Boolean	True/False → 1/0

4. Target Variable

4.1 Definition

Attribute	Value
Name	`risk_target`
Type	Binary (Integer)

Attribute	Value
Values	0 = No CHD, 1 = CHD Present
Definition	10-year cardiovascular disease event

4.2 Event Definition

A positive target (risk_target = 1) indicates occurrence of any of the following within 10 years:

- Myocardial infarction (heart attack)
- Coronary heart disease death
- Angina pectoris
- Coronary insufficiency

4.3 Distribution

Class	Count	Percentage
Negative (0)	12,253	76.0%
Positive (1)	3,870	24.0%
Total	**16,123**	**100%**

4.4 Class Balance by Split

Split	Negative	Positive	Positive Rate
Training	8,573	2,713	24.0%
Validation	1,839	580	24.0%
Test	1,841	577	23.9%

Stratified splitting preserves class distribution across all splits.

4.5 Clinical Significance

Outcome	Clinical Meaning
risk_target = 0	Low 10-year CVD risk, routine follow-up
risk_target = 1	Elevated 10-year CVD risk, intervention recommended

---

5. Data Splits

5.1 Split Configuration

Split	Purpose	Patients	Percentage
Training	Model training	11,286	70%
Validation	Hyperparameter tuning	2,419	15%
Test	Final evaluation	2,418	15%

5.2 Split Strategy

```python
from sklearn.model_selection import train_test_split

# First split: 70% train, 30% temp
X_train, X_temp, y_train, y_temp = train_test_split(
    X, y,
    test_size=0.30,
    stratify=y,
    random_state=42
)

# Second split: 50/50 of temp → 15% val, 15% test
X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp,
    test_size=0.50,
    stratify=y_temp,
    random_state=42
)
```

5.3 File Locations

Split	File Path
Training	`data/split/train.csv`
Validation	`data/split/val.csv`
Test	`data/split/test.csv`

5.4 Split Verification

Check	Training	Validation	Test
Sample Count	11,286	2,419	2,418
Positive Rate	24.0%	24.0%	23.9%
No Overlap			
Reproducible	(seed=42)		

---

6. Feature Statistics

6.1 Continuous Features (Training Set)

Feature	Mean	Std	Min	25%	50%	75%	Max
age	49.5	13.2	18	40	49	59	100
systolic_bp	132.4	21.3	80	117	130	145	250
diastolic_bp	82.7	11.8	40	75	82	90	142
bmi	27.3	5.4	15.5	23.6	26.5	30.1	56.8
total_cholesterol	236.7	44.1	107	206	234	263	394
fasting_glucose	97.8	27.3	52	81	91	106	295
heart_rate	75.6	11.9	42	68	75	83	143

6.2 Binary Features (Training Set)

Feature	Count (1)	Rate
sex (Male)	5,082	45.0%
bp_meds	1,243	11.0%
hypertension	3,612	32.0%
diabetes	678	6.0%
hypertension_flag	4,515	40.0%
high_cholesterol_flag	2,483	22.0%
high_glucose_flag	565	5.0%
obesity_flag	3,160	28.0%

6.3 Categorical Features (Training Set)

Age Groups:

Category	Count	Percentage
<40	2,257	20.0%
40-49	2,821	25.0%
50-59	2,821	25.0%
60-69	2,257	20.0%
70+	1,130	10.0%

BMI Categories:

Category	Count	Percentage
Underweight	226	2.0%
Normal	3,612	32.0%
Overweight	4,176	37.0%
Obese	3,272	29.0%

---

7. Preprocessing Pipeline

7.1 Feature Engineering Steps

```
Raw Data


  1. FEATURE DERIVATION
  - Calculate derived metrics
  - Create interaction terms
  - Apply log transformations



  2. CATEGORICAL ENCODING
  - One-hot encode age groups
```

```
- One-hot encode BMI cats
- Binary encode flags



3. TRAIN/VAL/TEST SPLIT
- Stratified 70/15/15
- Preserve class balance



4. FEATURE SCALING
- Fit StandardScaler on train
- Transform all splits
```

7.2 Scaling Configuration

```python
from sklearn.preprocessing import StandardScaler

# Fit on training data only
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)

# Transform validation and test with same parameters
X_val_scaled = scaler.transform(X_val)
X_test_scaled = scaler.transform(X_test)
```

Scaling Formula:

$$z = \frac{x - \mu}{\sigma}$$

Where: - x = original value - μ = training set mean - σ = training set standard deviation - z = scaled value

---

Summary

The CardioDetect dataset contains **16,123 patients** with **34 clinical features** designed for cardiovascular risk prediction. All features are validated, complete, and properly encoded for machine learning.

Quick Reference

Category	Count
Demographics	2
Clinical Measurements	6
Risk Factors	5
Derived Features	4
Clinical Flags	4
Log Transforms	3
Interactions	3
Age Categories	5

Category	Count
BMI Categories	4
Total	**34**

File Structure

```
data/
  final/
     final_risk_dataset.csv    # Complete dataset
  split/
      train.csv                  # Training set (70%)
      val.csv                    # Validation set (15%)
      test.csv                   # Test set (15%)
```

Data Dictionary - CardioDetect v2.0

Page count: 5 pages