UCI Heart disease dataset

Neural Networks of Machine Learning Applications

Spring 2023

Sakari Lukkarinen

Metropolia University of Applied Sciences

Heart disease (coronary artery disease)

Coronary artery disease (**CAD**), also known as **coronary heart disease** (**CHD**), **ischemic heart disease** (**IHD**), or simply **heart disease**, involves the reduction of blood flow to the heart muscle due to build-up of plaque (atherosclerosis) in the arteries of the heart.

It is the most common of the cardiovascular diseases. Types include stable angina, unstable angina, myocardial infarction, and sudden cardiac death.

A common symptom is chest pain or discomfort which may travel into the shoulder, arm, back, neck, or jaw. Occasionally it may feel like heartburn.

Coronary artery disease (Wikipedia)

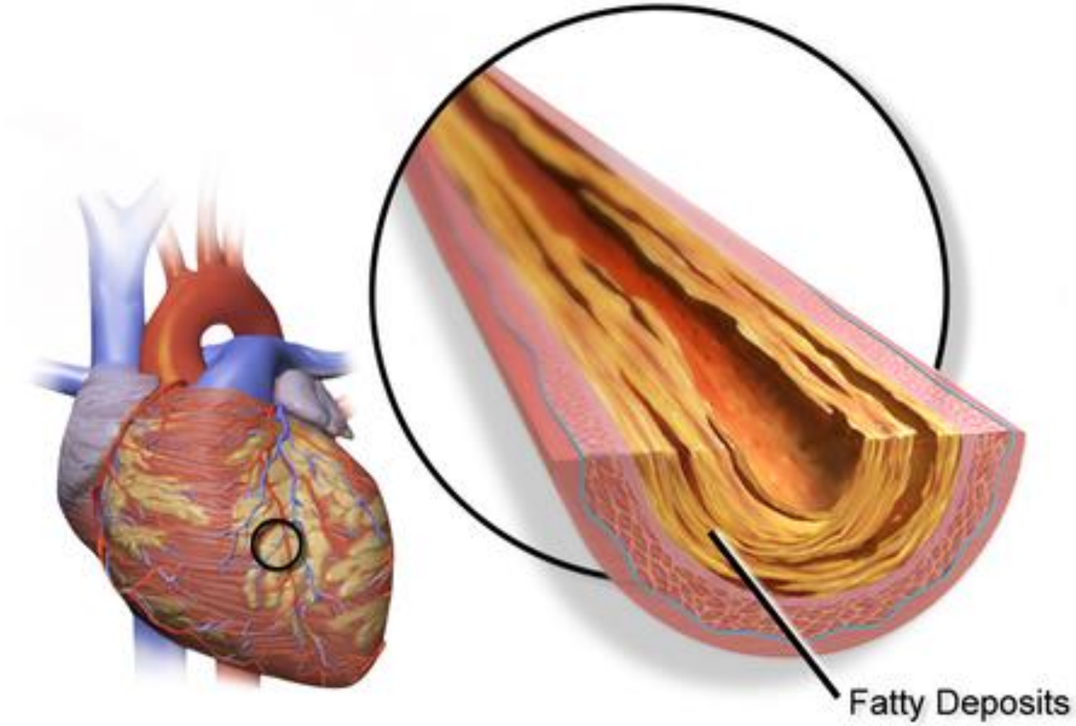


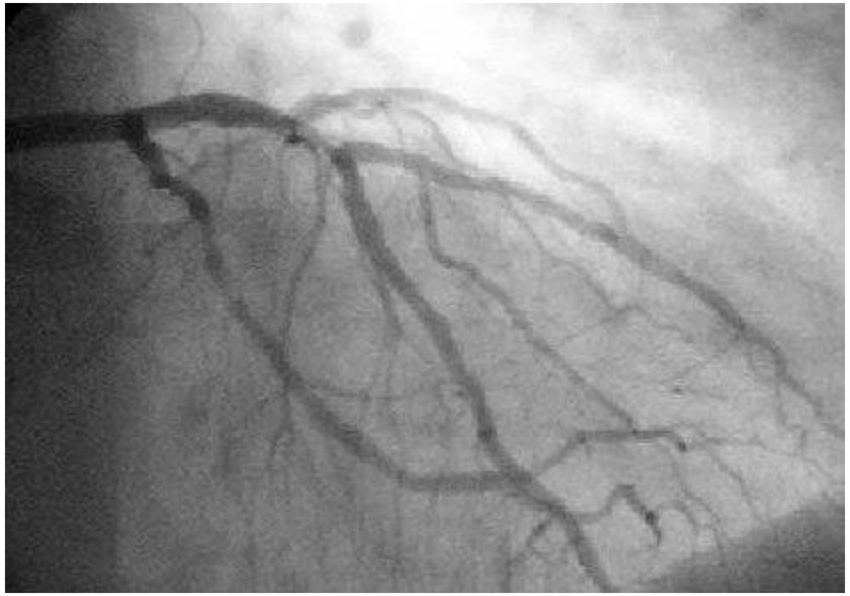
Illustration depicting atherosclerosis in a coronary artery.

Risk factors for heart disease

Risk factors include high blood pressure, smoking, diabetes, lack of exercise, obesity, high blood cholesterol, poor diet, depression, and excessive alcohol.

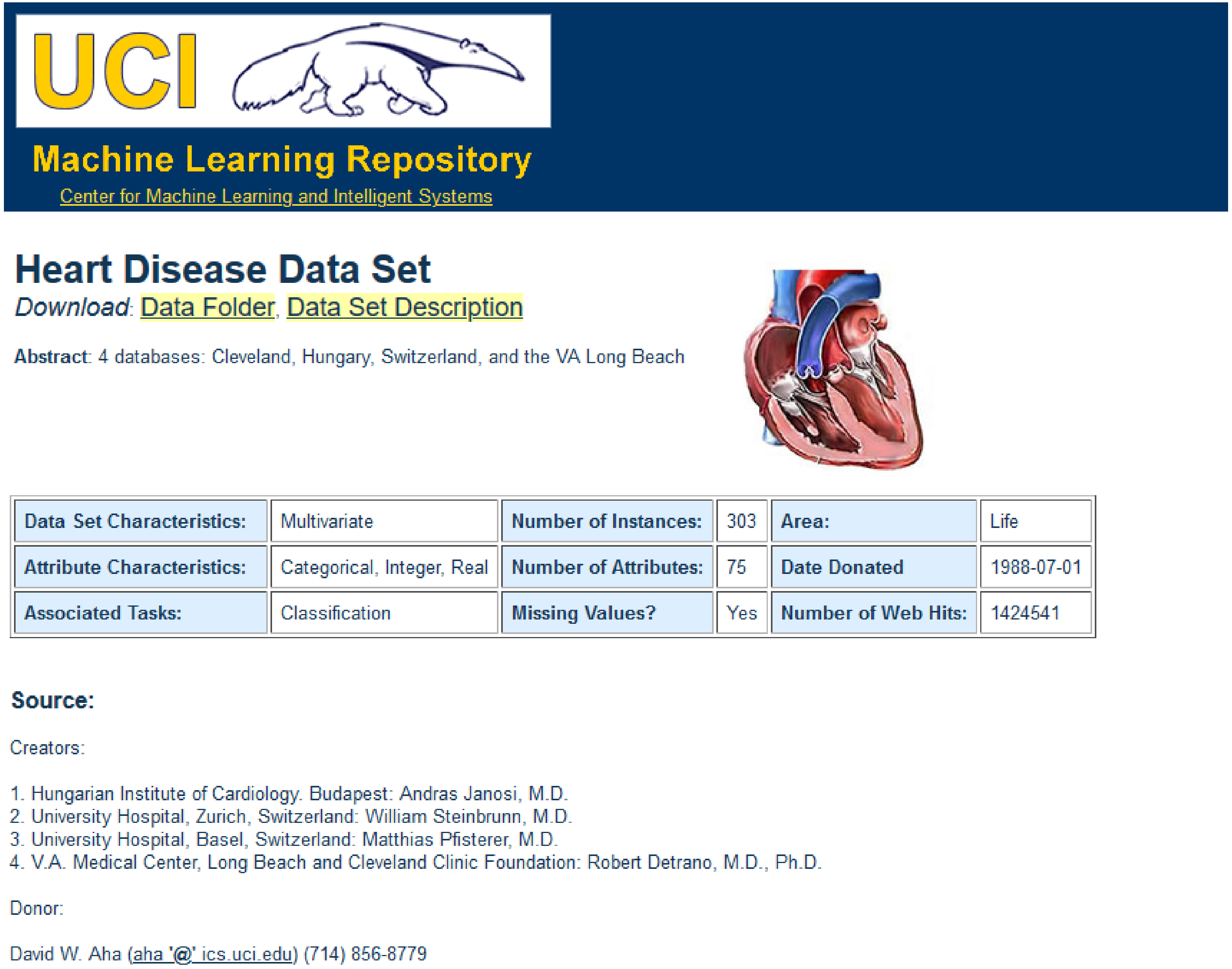
A number of tests may help with diagnoses including: electrocardiogram, cardiac stress testing, coronary computed tomographic angiography, and coronary angiogram, among others.

Coronary artery disease (Wikipedia)



A coronary angiogram (an X-ray with radiocontrast agent in the coronary arteries) that shows the left coronary circulation.

Coronary catheterization (Wikipedia)



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Heart Disease Data Set

Download: Data Folder, Data Set Description

Abstract: 4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach

Data Set Characteristics:	Multivariate	**Number of Instances:**	303	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	75	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	1424541

Source:

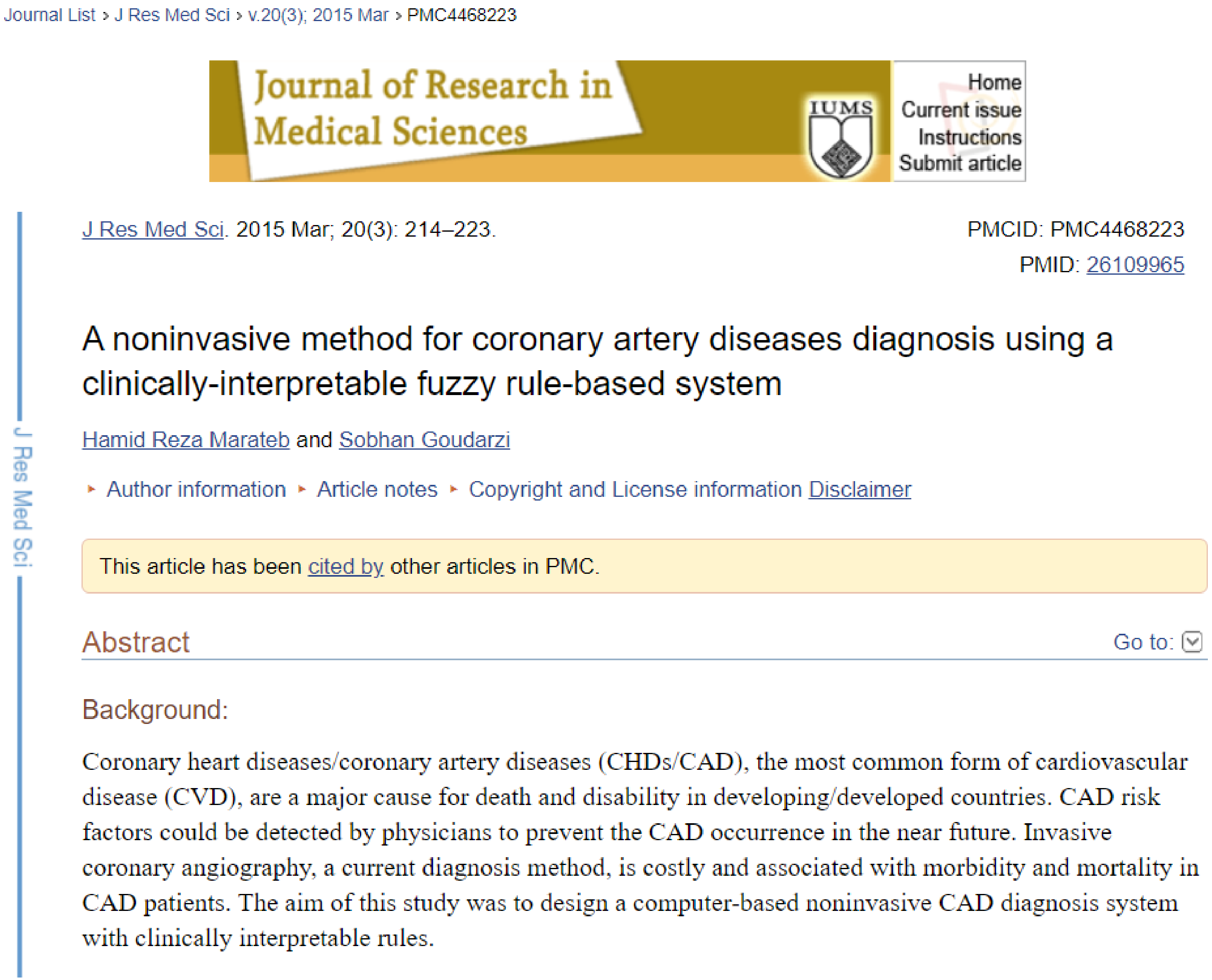
Creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor:

David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

UCI Machine Learning Repository: Heart Disease Data Set



Journal List › J Res Med Sci › v.20(3); 2015 Mar › PMC4468223

Journal of Research in Medical Sciences

J Res Med Sci. 2015 Mar; 20(3): 214–223.

PMCID: PMC4468223
PMID: 26109965

A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system

Hamid Reza Marateb and Sobhan Goudarzi

‣ Author information ‣ Article notes ‣ Copyright and License information Disclaimer

This article has been cited by other articles in PMC.

Abstract

Go to:

Background:

Coronary heart diseases/coronary artery diseases (CHDs/CAD), the most common form of cardiovascular disease (CVD), are a major cause for death and disability in developing/developed countries. CAD risk factors could be detected by physicians to prevent the CAD occurrence in the near future. Invasive coronary angiography, a current diagnosis method, is costly and associated with morbidity and mortality in CAD patients. The aim of this study was to design a computer-based noninvasive CAD diagnosis system with clinically interpretable rules.

A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system

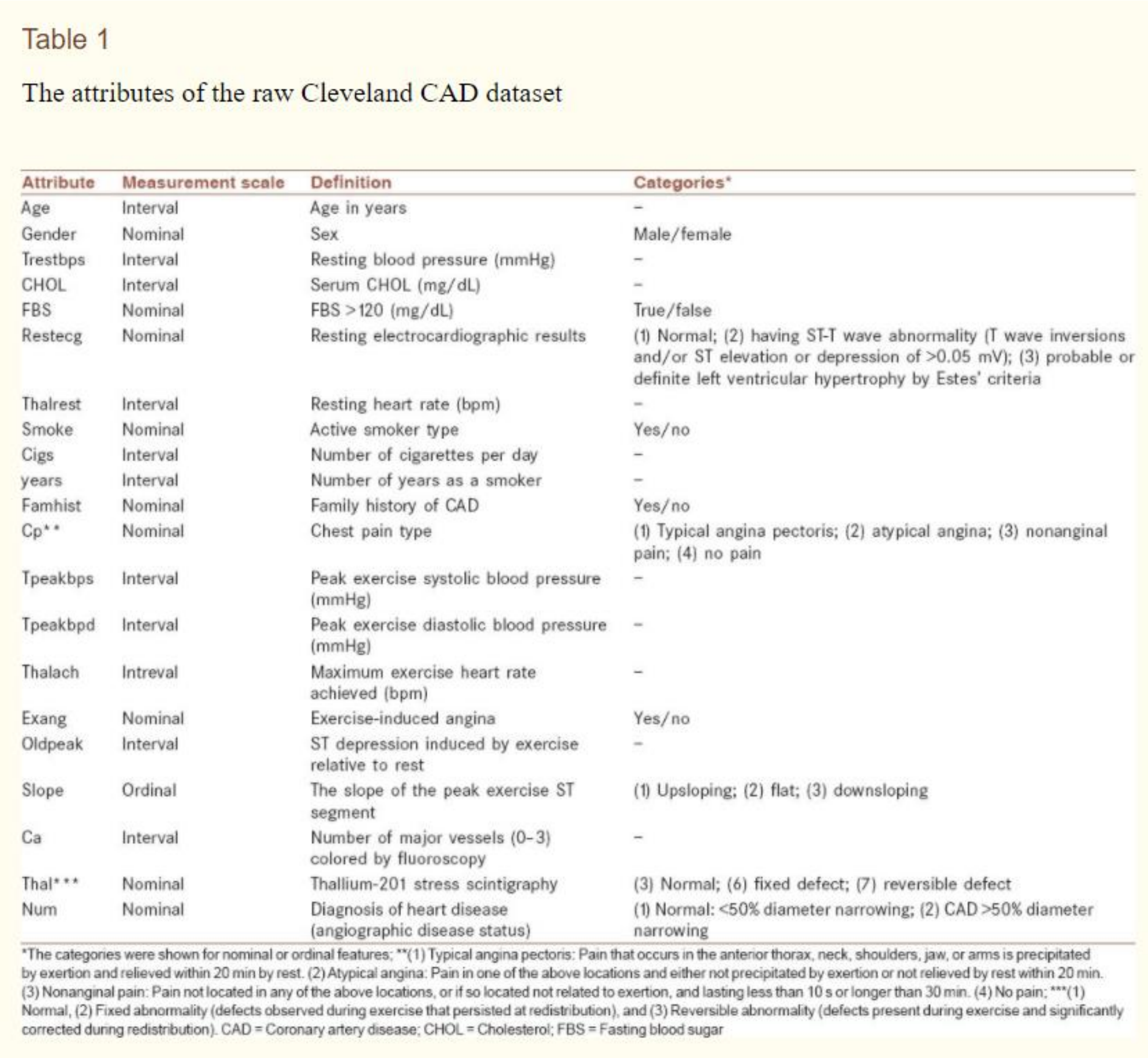


Table 1

The attributes of the raw Cleveland CAD dataset

Attribute	Measurement scale	Definition	Categories*
Age	Interval	Age in years	–
Gender	Nominal	Sex	Male/female
Trestbps	Interval	Resting blood pressure (mmHg)	–
CHOL	Interval	Serum CHOL (mg/dL)	–
FBS	Nominal	FBS >120 (mg/dL)	True/false
Restecg	Nominal	Resting electrocardiographic results	(1) Normal; (2) having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV); (3) probable or definite left ventricular hypertrophy by Estes' criteria
Thalrest	Interval	Resting heart rate (bpm)	–
Smoke	Nominal	Active smoker type	Yes/no
Cigs	Interval	Number of cigarettes per day	–
years	Interval	Number of years as a smoker	–
Famhist	Nominal	Family history of CAD	Yes/no
Cp**	Nominal	Chest pain type	(1) Typical angina pectoris; (2) atypical angina; (3) nonanginal pain; (4) no pain
Tpeakbps	Interval	Peak exercise systolic blood pressure (mmHg)	–
Tpeakbpd	Interval	Peak exercise diastolic blood pressure (mmHg)	–
Thalach	Intreval	Maximum exercise heart rate achieved (bpm)	–
Exang	Nominal	Exercise-induced angina	Yes/no
Oldpeak	Interval	ST depression induced by exercise relative to rest	–
Slope	Ordinal	The slope of the peak exercise ST segment	(1) Upsloping; (2) flat; (3) downsloping
Ca	Interval	Number of major vessels (0–3) colored by fluoroscopy	–
Thal***	Nominal	Thallium-201 stress scintigraphy	(3) Normal; (6) fixed defect; (7) reversible defect
Num	Nominal	Diagnosis of heart disease (angiographic disease status)	(1) Normal: <50% diameter narrowing; (2) CAD >50% diameter narrowing

*The categories were shown for nominal or ordinal features; **(1) Typical angina pectoris: Pain that occurs in the anterior thorax, neck, shoulders, jaw, or arms is precipitated by exertion and relieved within 20 min by rest. (2) Atypical angina: Pain in one of the above locations and either not precipitated by exertion or not relieved by rest within 20 min. (3) Nonanginal pain: Pain not located in any of the above locations, or if so located not related to exertion, and lasting less than 10 s or longer than 30 min. (4) No pain; ***(1) Normal, (2) Fixed abnormality (defects observed during exercise that persisted at redistribution), and (3) Reversible abnormality (defects present during exercise and significantly corrected during redistribution). CAD = Coronary artery disease; CHOL = Cholesterol; FBS = Fasting blood sugar

A noninvasive method for coronary artery diseases diagnosis ...

Example of confusion matrix

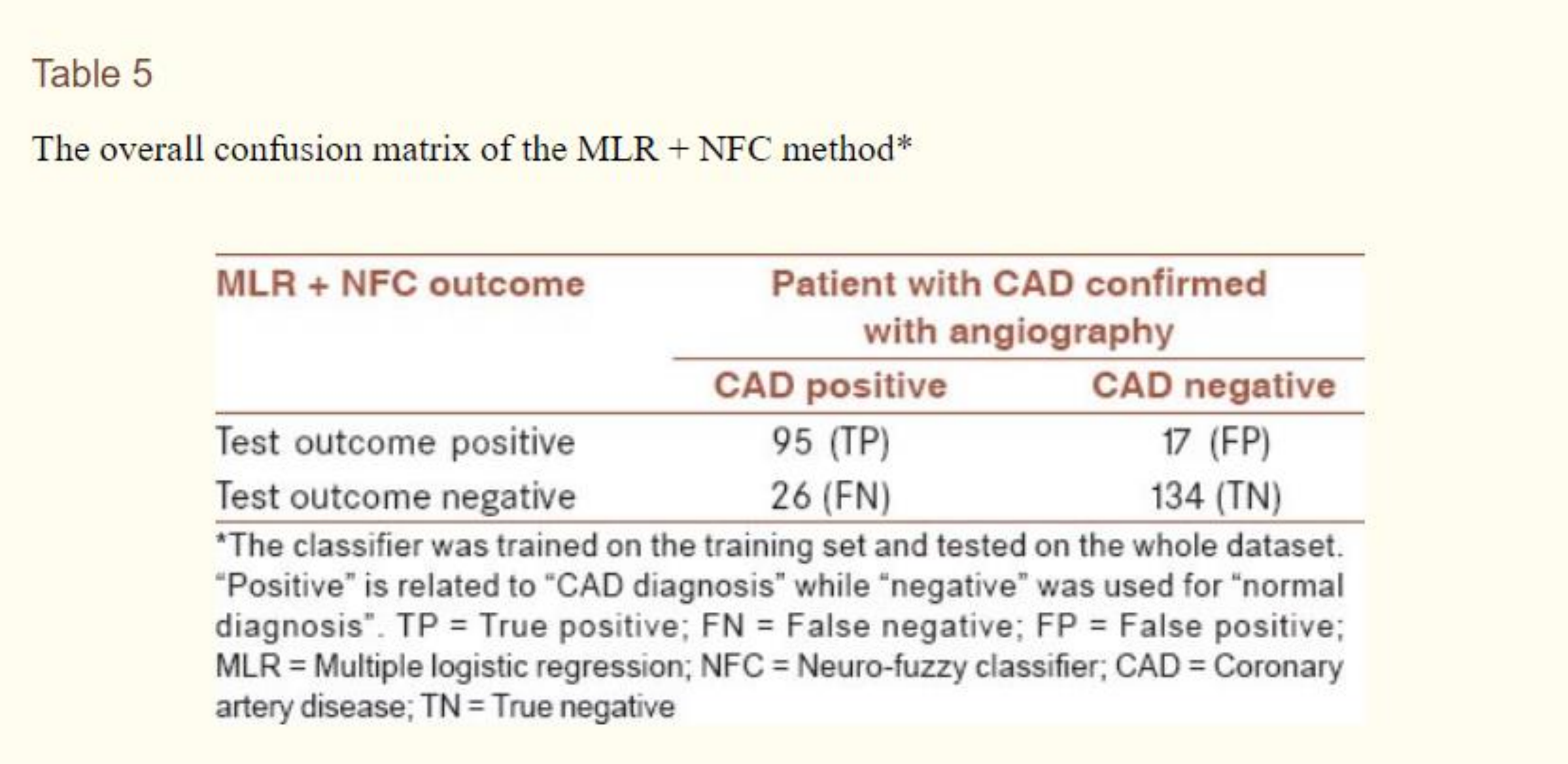


Table 5

The overall confusion matrix of the MLR + NFC method*

MLR + NFC outcome	Patient with CAD confirmed with angiography	
	CAD positive	CAD negative
Test outcome positive	95 (TP)	17 (FP)
Test outcome negative	26 (FN)	134 (TN)

*The classifier was trained on the training set and tested on the whole dataset. “Positive” is related to “CAD diagnosis” while “negative” was used for “normal diagnosis”. TP = True positive; FN = False negative; FP = False positive; MLR = Multiple logistic regression; NFC = Neuro-fuzzy classifier; CAD = Coronary artery disease; TN = True negative

A noninvasive method for coronary artery diseases diagnosis ...

Dataset information

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date.

The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory.

Attribute Information

Only 14 attributes used:

Variables (=Input)

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)

Label (=Output)

14. #58 (num) (the predicted attribute)

Angina (Chest Pain)

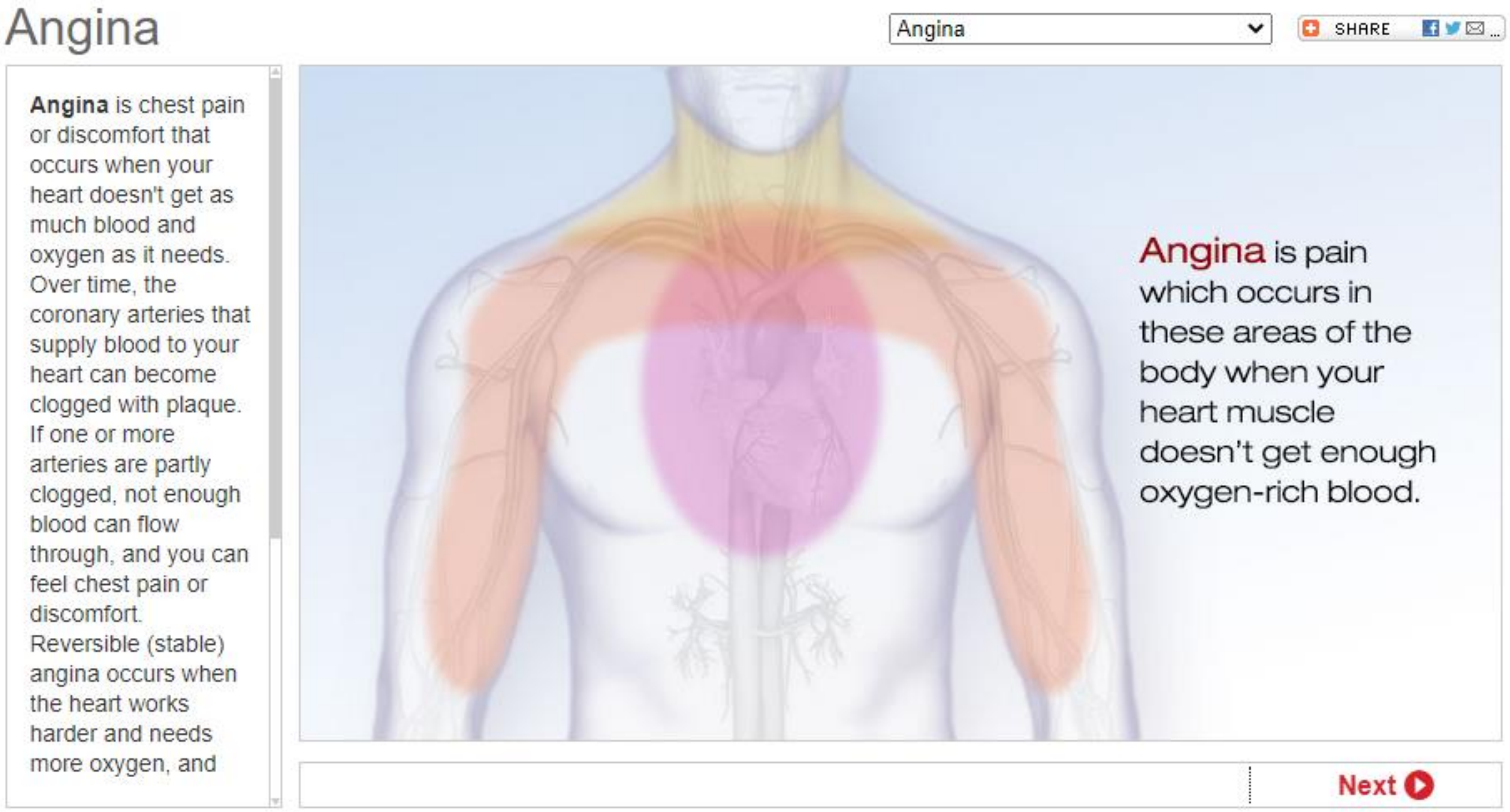
3. #9 (cp) chest pain type

-- Value 1: typical angina

-- Value 2: atypical angina

-- Value 3: non-anginal pain

-- Value 4: asymptomatic



Angina

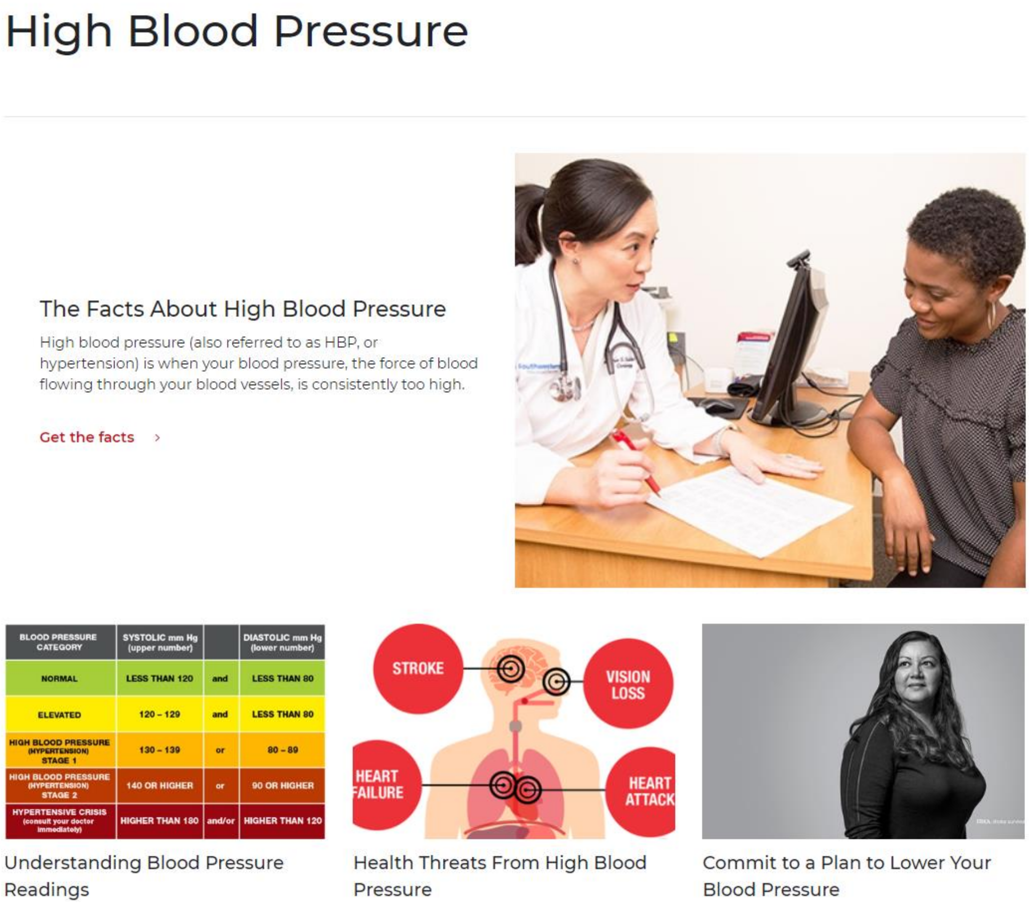
Angina is chest pain or discomfort that occurs when your heart doesn't get as much blood and oxygen as it needs. Over time, the coronary arteries that supply blood to your heart can become clogged with plaque. If one or more arteries are partly clogged, not enough blood can flow through, and you can feel chest pain or discomfort. Reversible (stable) angina occurs when the heart works harder and needs more oxygen, and

Angina is pain which occurs in these areas of the body when your heart muscle doesn't get enough oxygen-rich blood.

[Angina (American Heart Association)]

High Blood Pressure

4. #10 (trestbps) resting blood pressure (mmHg, systolic)



High Blood Pressure

The Facts About High Blood Pressure

High blood pressure (also referred to as HBP, or hypertension) is when your blood pressure, the force of blood flowing through your blood vessels, is consistently too high.

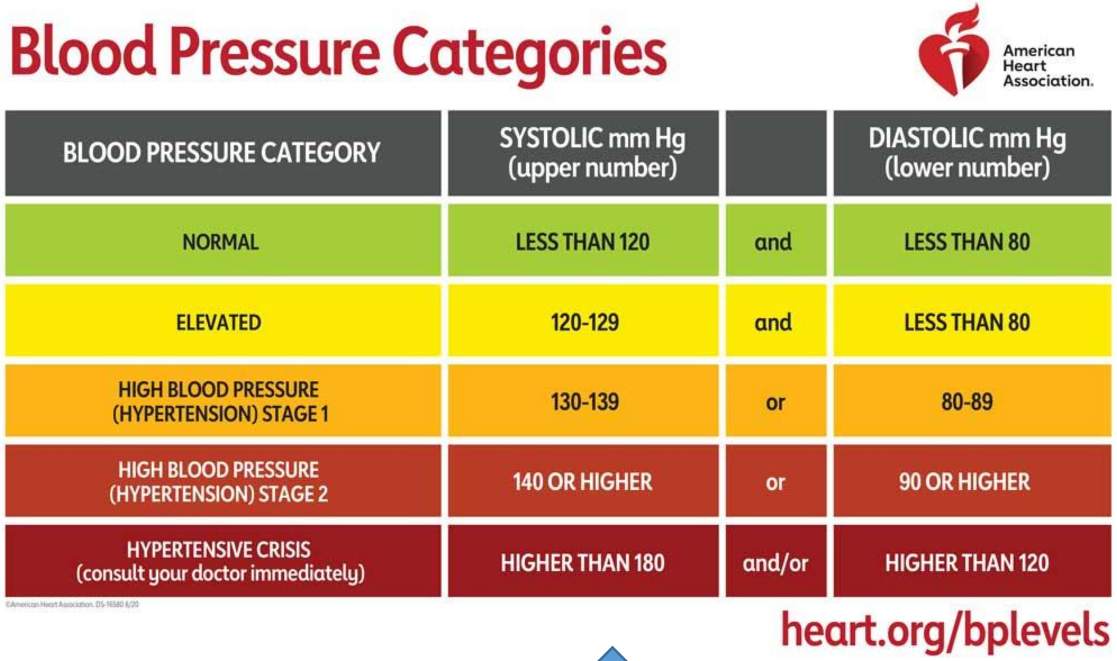
Get the facts ›

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

Understanding Blood Pressure Readings

Health Threats From High Blood Pressure

Commit to a Plan to Lower Your Blood Pressure



Blood Pressure Categories

American Heart Association.

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120-129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130-139	or	80-89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

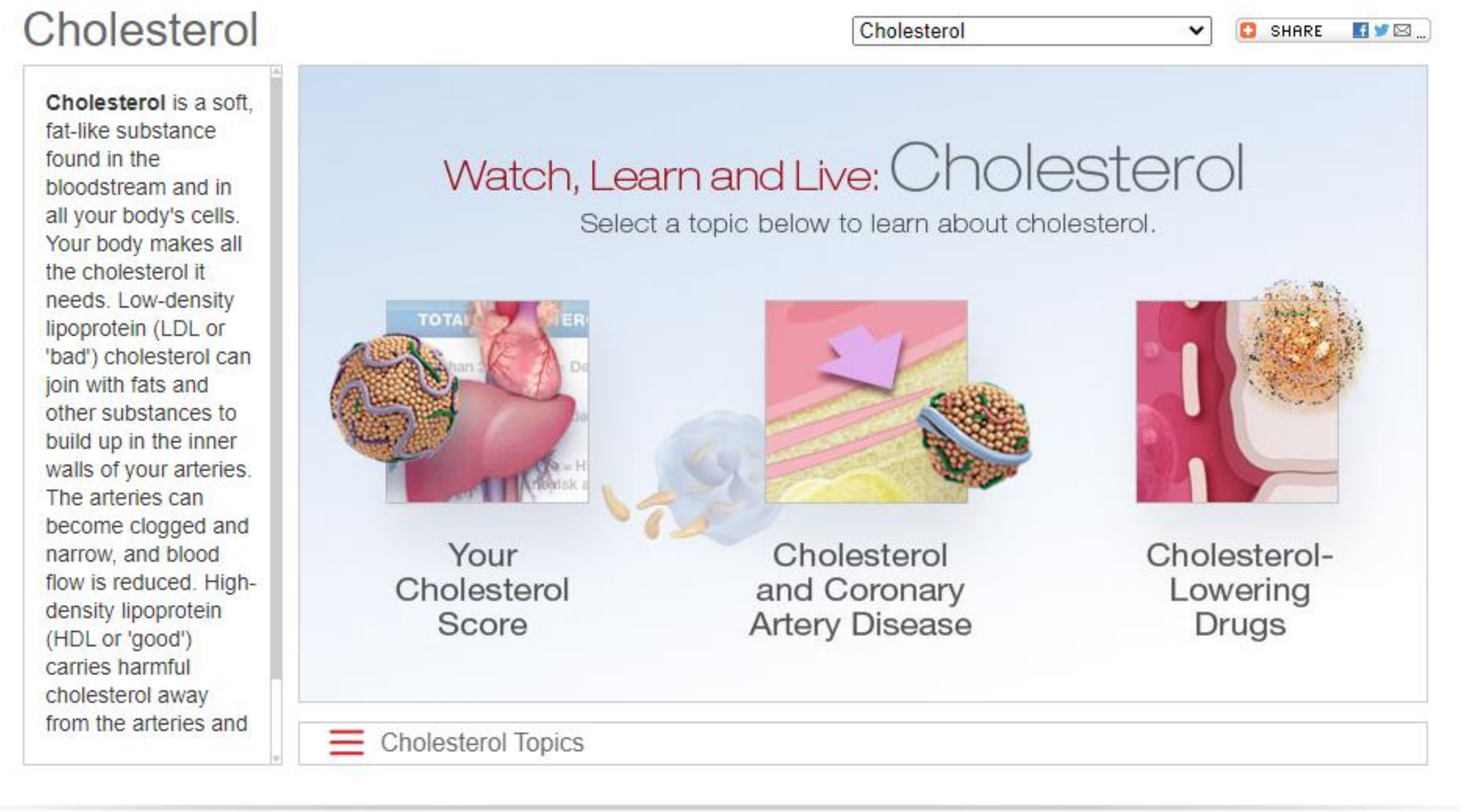
©American Heart Association. DS-16580 6/20

heart.org/bplevels

[High Blood Pressure (American Heart Association)](#)

Total blood (or serum) cholesterol

5. #12 (chol) serum cholesterol (mg/dl)



Cholesterol

Cholesterol is a soft, fat-like substance found in the bloodstream and in all your body's cells. Your body makes all the cholesterol it needs. Low-density lipoprotein (LDL or 'bad') cholesterol can join with fats and other substances to build up in the inner walls of your arteries. The arteries can become clogged and narrow, and blood flow is reduced. High-density lipoprotein (HDL or 'good') carries harmful cholesterol away from the arteries and

Watch, Learn and Live: Cholesterol

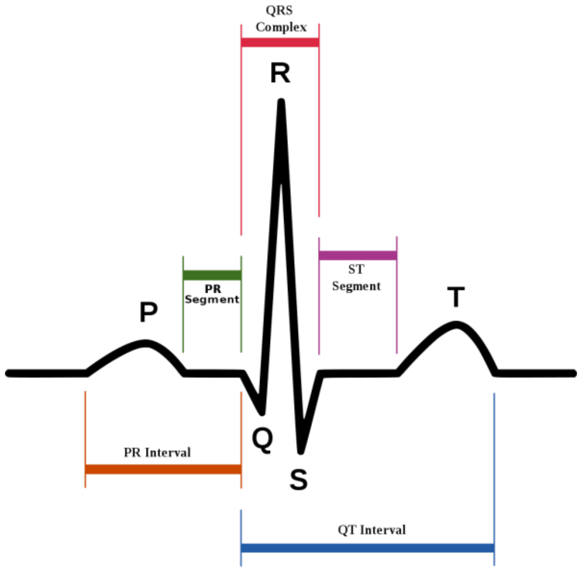
Select a topic below to learn about cholesterol.

Cholesterol
SHARE
Your Cholesterol Score
Cholesterol and Coronary Artery Disease
Cholesterol-Lowering Drugs
Cholesterol Topics

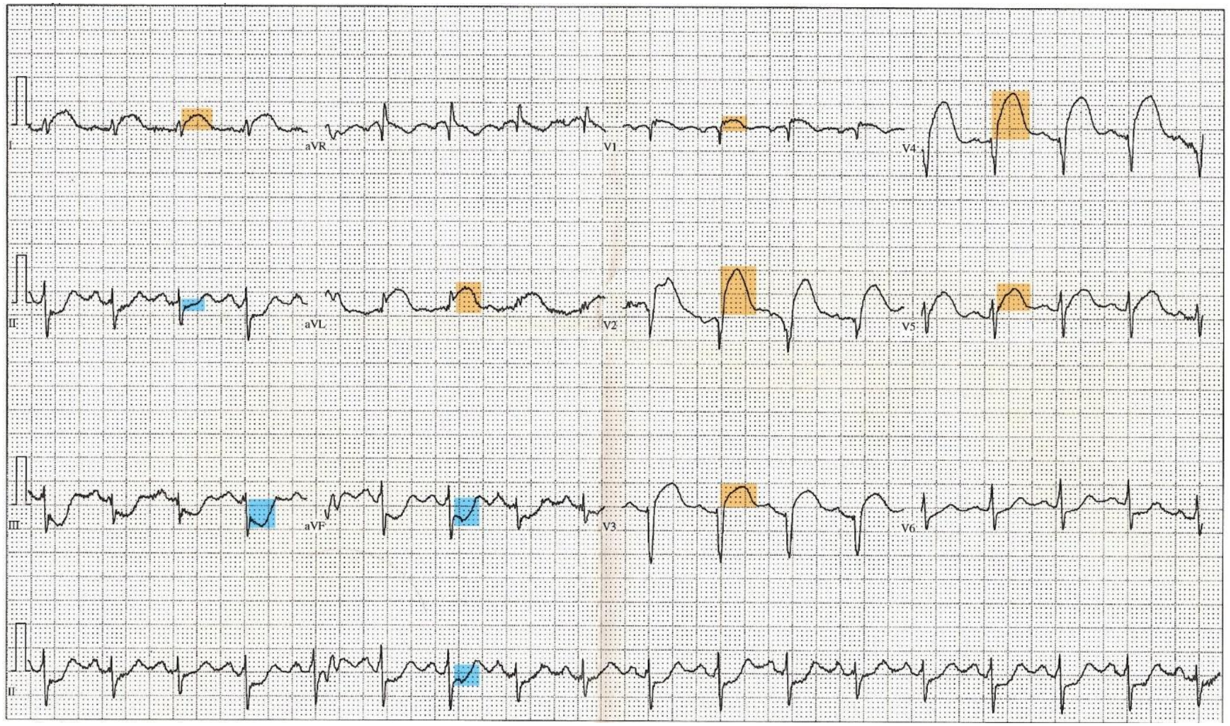
Watch, Learn, and Live: Cholesterol

Resting electrocardiogram results

7. #19 (restecg) : resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria



ST Segment (Wikipedia)



ST elevation (Wikipedia)

Maximum heart rate

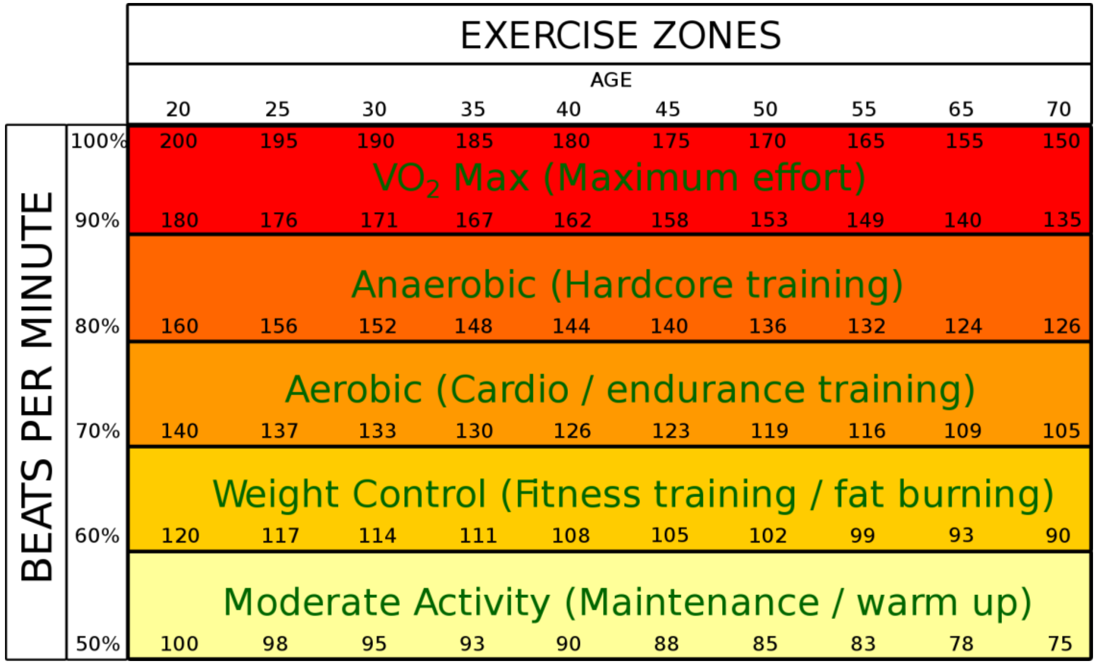
8. #32 (thalach) : maximum heart rate achieved (BPM = Beats Per Minute)

The *maximum heart rate* (HR_{max}) is the highest heart rate an individual can achieve without severe problems through exercise stress, and generally decreases with age.

Since HR_{max} varies by individual, the most accurate way of measuring any single person's HR_{max} is via a cardiac stress test.

The most widely cited formula for HR_{max} is: $HR_{max} = 220 - age$

Heart rate (Wikipedia)



EXERCISE ZONES

BEATS PER MINUTE	Zone	AGE 20	25	30	35	40	45	50	55	65	70
100%	VO_2 Max (Maximum effort)	200	195	190	185	180	175	170	165	155	150
90%	Anaerobic (Hardcore training)	180	176	171	167	162	158	153	149	140	135
80%	Aerobic (Cardio / endurance training)	160	156	152	148	144	140	136	132	124	126
70%	Weight Control (Fitness training / fat burning)	140	137	133	130	126	123	119	116	109	105
60%	Moderate Activity (Maintenance / warm up)	120	117	114	111	108	105	102	99	93	90
50%		100	98	95	93	90	88	85	83	78	75

Fox and Haskell formula; widely used.

Cardiac stress test

9. #38 (exang) : exercise induced angina (1 = yes; 0 = no)

10. #40 (oldpeak) ST depression induced by exercise relative to rest

11. #41 (slope) the slope of the peak exercise ST segment

-- Value 1: upsloping

-- Value 2: flat

-- Value 3: downsloping

A **cardiac stress test** (also referred to as a **cardiac diagnostic test**, **cardiopulmonary exercise test**, or abbreviated **CPX test**) is a cardiological test that measures the heart's ability to respond to external stress in a controlled clinical environment.

The stress response is induced by exercise or by intravenous pharmacological stimulation.

Cardiac stress test (Wikipedia)

Number of major vessels

12. #44 (ca) number of major vessels (0-3) colored by flouroscopy

Fluoroscopy is a type of medical imaging that shows a continuous X-ray image on a monitor, much like an X-ray movie.

During a fluoroscopy procedure, an X-ray beam is passed through the body.

The image is transmitted to a monitor so the movement of a body part or of an instrument or contrast agent ("X-ray dye") through the body can be seen in detail.

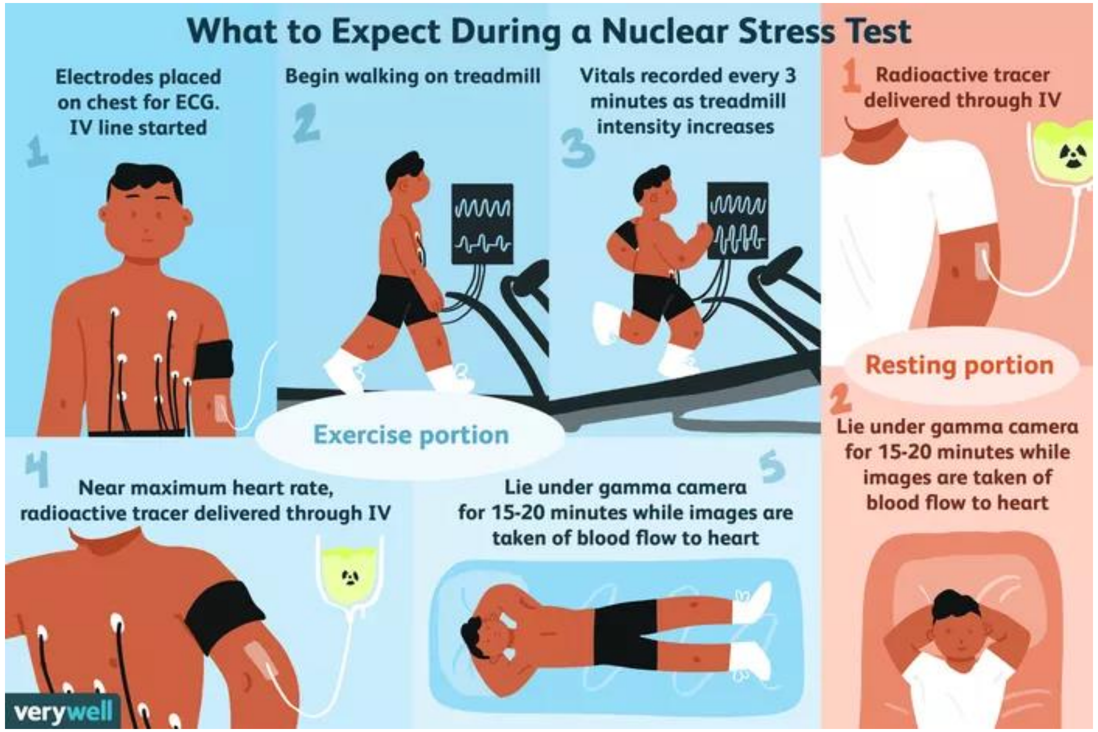
FDA: Fluoroscopy

Thallium stress test

13. #51 (thal) 3 = normal; 6 = fixed defect; 7 = reversable defect

A thallium stress test is a nuclear imaging test that shows how well blood flows into your heart while you're exercising or at rest. This test is also called a cardiac or nuclear stress test.

During the procedure, a liquid with a small amount of radioactivity called a radioisotope is administered into one of your veins. The radioisotope will flow through your bloodstream and end up in your heart. Once the radiation is in your heart, a special camera called a gamma camera can detect the radiation and reveal any issues your heart muscle is having.


What to Expect During a Nuclear Stress Test
Exercise portion
1 Electrodes placed on chest for ECG. IV line started
2 Begin walking on treadmill
3 Vitals recorded every 3 minutes as treadmill intensity increases
4 Near maximum heart rate, radioactive tracer delivered through IV
5 Lie under gamma camera for 15-20 minutes while images are taken of blood flow to heart
Resting portion
1 Radioactive tracer delivered through IV
2 Lie under gamma camera for 15-20 minutes while images are taken of blood flow to heart
verywell

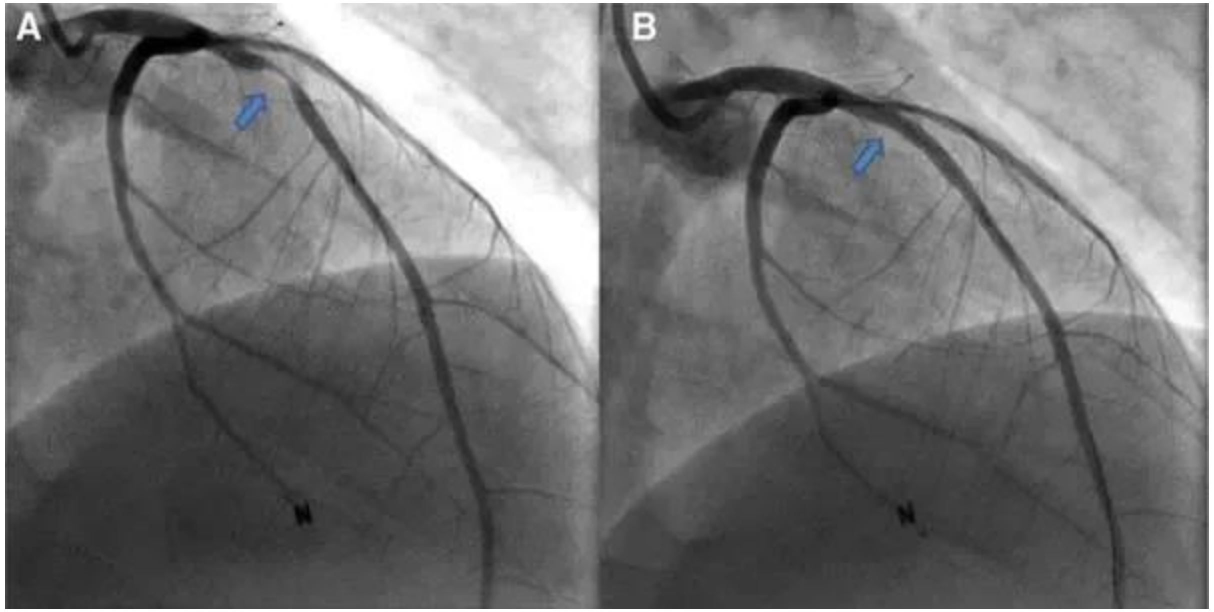
[Illustration by Emily Roberts, Verywell](#)

[Thallium stress test (Healthonline)](#)

Angiographic status (= Output)

14. #58 (num) (the predicted attribute)
diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing

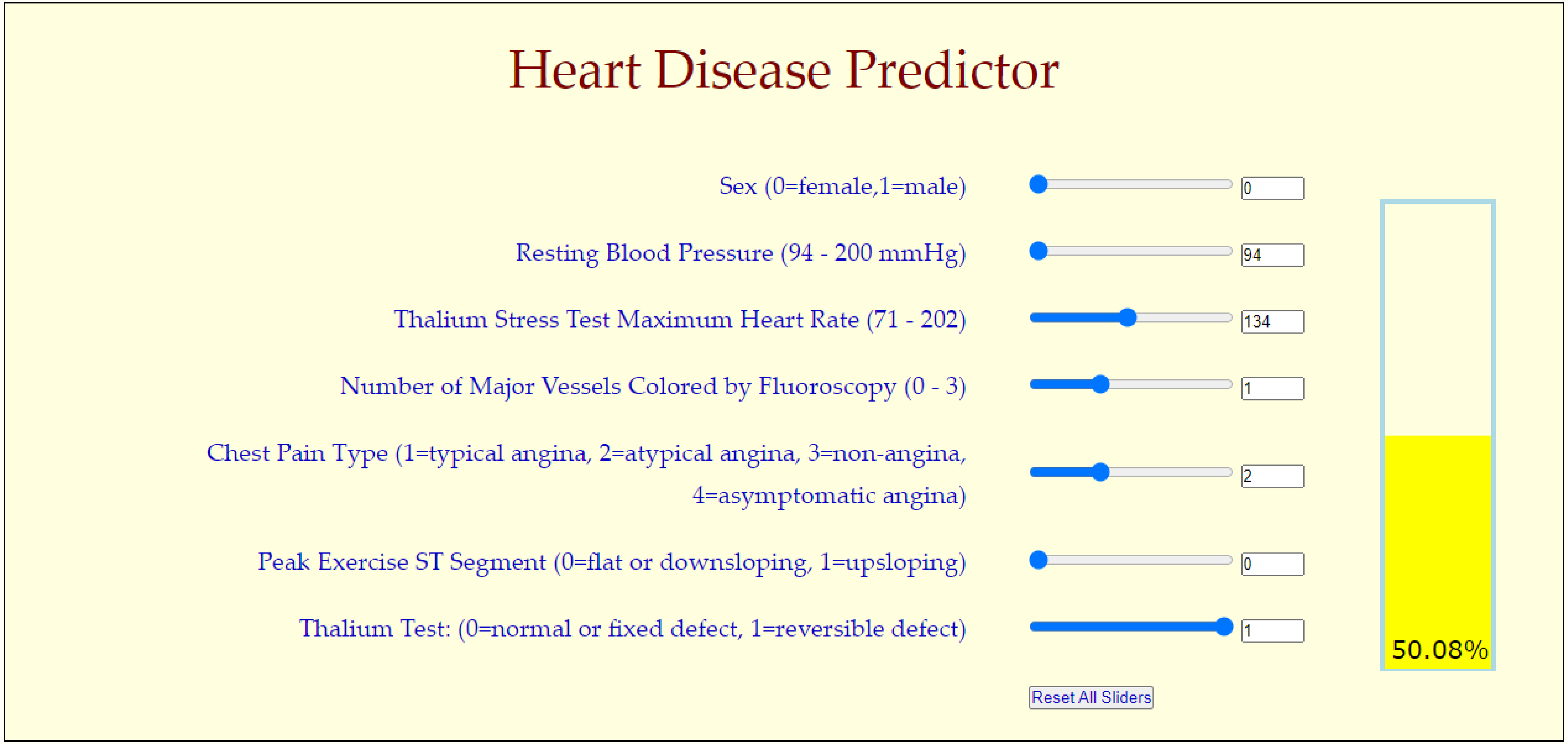
Angiography or **arteriography** is a medical imaging technique used to visualize the inside, or lumen, of blood vessels and organs of the body, with particular interest in the arteries, veins, and the heart chambers. This is traditionally done by injecting a radio-opaque contrast agent into the blood vessel and imaging using X-ray based techniques such as fluoroscopy.



How do cardiologists during an Angiogram determine what percentage of the coronary artery is blocked?

Angiography (Wikipedia)

Heart Disease Predictor (Online)



Heart Disease Predictor

Sex (0=female,1=male) 0

Resting Blood Pressure (94 - 200 mmHg) 94

Thalium Stress Test Maximum Heart Rate (71 - 202) 134

Number of Major Vessels Colored by Fluoroscopy (0 - 3) 1

Chest Pain Type (1=typical angina, 2=atypical angina, 3=non-angina, 4=asymptomatic angina) 2

Peak Exercise ST Segment (0=flat or downsloping, 1=upsloping) 0

Thalium Test: (0=normal or fixed defect, 1=reversible defect) 1

https://lucdemortier.github.io/projects/3_mcnulty

Practice with UCI heart disease dataset

- Basic skills
  1. Start with processed Cleveland data
  2. Make a straightforward preprocessing step and standard 3-layer classifier
     - **Normalize the dataset**, split into training, validation and test sets
     - Play with layers and number of neurons
  3. Learn to use the performance metrics
     - During training: **Accuracy**
     - After training and during testing: **Classification report, confusion matrix,** (ROC curve)
- Advanced skills
  1. Make a preprocessing plan
     - Study the variables
     - **Convert between categorical and numerical values**
     - **Use one-hot-coding**
     - Modify the model accordingly
  2. Try cross-evaluation techniques
  3. Use all processed data