

Developer: Ethan Nolet

Sources:

- Kaggle: <https://www.kaggle.com/search?q=heart+data>
- UCI repo: <https://archive.ics.uci.edu/datasets/?search=Heart+Disease>
- Github: <https://github.com/caravanuden/cardio?tab=readme-ov-file>
- Publication: <https://www.nature.com/articles/s41598-024-80919-9#Sec24>

Resources Explored

- [Cardiovascular Disease Dataset](#)
- [CDC Diabetes Health Indicators](#)
- [Heart Attack Risk & Prediction Dataset In India](#)
- [Heart Disease](#)
- [Risk Factors for Cardiovascular Heart Disease](#)
- <https://link.springer.com/article/10.1007/s10916-016-0536-z>
- <https://www.nature.com/articles/s41598-024-80919-9#Sec24>
- [Hospital Admissions Data: Two Year Hospital Admissions and Discharge Data from Hero DMC Heart Institute](#)

Data to use for the project

[Hospital Admissions Data: Two Year Hospital Admissions and Discharge Data from Hero DMC Heart Institute](#)

Data Exploration Notes

1. <https://link.springer.com/article/10.1007/s10916-016-0536-z> and <https://www.nature.com/articles/s41598-024-80919-9#Sec24> : From these, I established an initial list of risk factors for CVD, HTN, and CAD:

CVD: Radiation exposure, age, ambient noise, irregular work hours, body weight and BMI, elevated triglyceride, history of HTN and CVD, elevated FBS levels, poor sitting posture, dust exposure, and smoking.

HTN: Continuous work shifts, BMI, body weight, family history of HTN, age, improper posture, elevated cholesterol levels, low HDL levels, gas exposure, prolonged sitting, smoking, dust exposure, and a history of HTN.

CAD: Smoking, Diabetes Mellitus, High Density Lipoprotein, Duke Treadmill Score, and Duration Recovery (from Duke Treadmill test)

2. [Cardiovascular Disease Dataset](#): This focuses on clinical data obtained from patient exercise tests, which is the most useful for identifying CAD. I believe this would be best applied to the Plan A+ for identifying additional diseases.
3. [CDC Diabetes Health Indicators](#): This database does contain a field HeartDiseaseorAttack, which tracks the presence of a coronary heart disease or myocardial infarction. Additional factors include high blood pressure, high cholesterol, BMI, smoker status, history of stroke, and physical activity, with all but the BMI stored as a binary value. I believe the physical activity field would be useful based on the relationship between sedentary lifestyle and cardiovascular diseases.
4. (Potential Best) [Heart Attack Risk & Prediction Dataset In India](#): This database records the presence of HTN, but does not give heart disease status directly; it provides only an evaluation for 'Heart Attack Risk'. Additional factors include obesity status, smoking status, physical activity, cholesterol level, triglyceride level, LDL level, HDL level, blood pressure, exposure to air pollution, family history, and stress level. Across all of the datasets I found, this has the greatest overlap of factors with the publication, particularly due to including triglyceride levels, HDL levels, and exposure to air pollution.
5. (Potential Best if Verifiable) [Heart Disease](#): This database records heart disease status directly. Additional factors include blood pressure, cholesterol level, exercise habits, smoker status, family heart disease history, diabetes, BMI, high blood pressure, low HDL, high LDL, stress level, sleep hours, sugar consumption, triglyceride level, CRP level, and homocysteine level. As with the [Heart Attack Risk & Prediction Dataset in India], this database has a great overlap of factors with the publication. However, I cannot find any information about the origin of this data, and the use of an AI-generated cover image increases my doubts about its integrity.
6. [Risk Factors for Cardiovascular Heart Disease](#): This dataset records cardiovascular disease status directly. Additional factors include weight, blood pressure, cholesterol level, glucose level, smoker status, alcohol consumption, and active status. This data is notable for tracking 70,000 individuals, which I expect could be valuable.

7. [Hospital Admissions Data: Two Year Hospital Admissions and Discharge Data from Hero DMC Heart Institute](#): This database contains information collected from patients admitted to the Hero DMC Heart Institute, such as their age, sex, patient history, smoking status, alcohol use, hemoglobin, total lymphocyte count, platelets, glucose, urea, creatinine, brain natriuretic peptide, raised cardiac enzymes, and ejection fraction.