Predicting Heart Attack Risk

— A Data Science Approach to Forecasting Cardiovascular Risk

BANA 288–PREDICTIVE ANALYTICS
Professor– Ken Murphy
Group 5: Becky Wang, Vy Nguyen,
Xinying Wu, Yanan Sun

INTRODUCTION

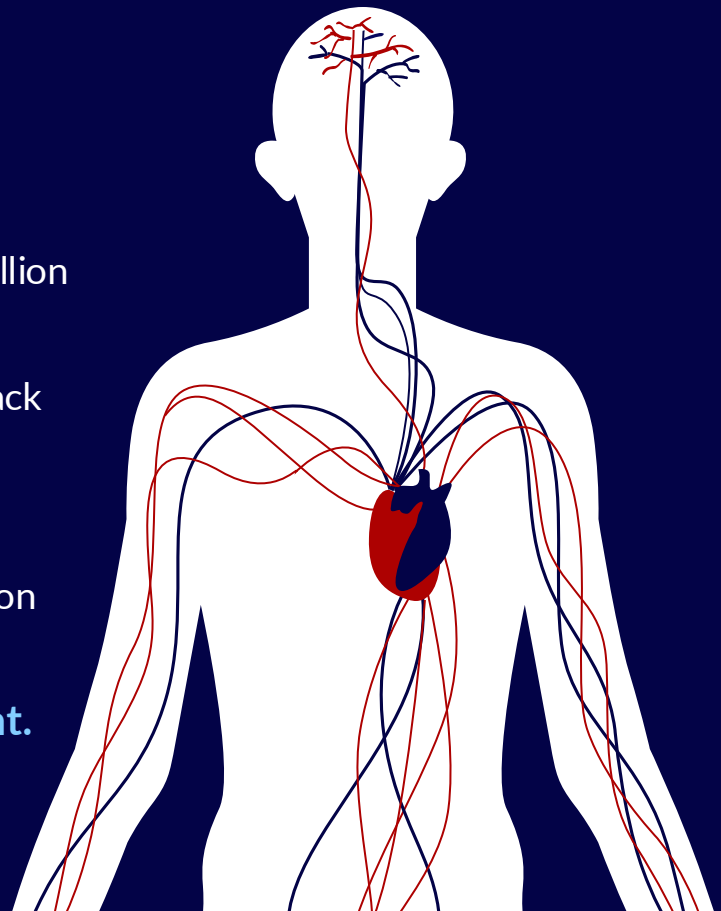
Why This Matters?

Surprising Factors

- leading cause of death worldwide, claiming over **18** million lives annually
- Every **40** seconds, someone in the U.S. has a heart attack
- More than **50%** of heart attack victims had no prior symptoms.
- Heart disease costs the U.S. economy over **$200** billion annually.

A heart attack doesn’t wait for a doctor’s appointment.

But what if data could warn you before it strikes?



Research Question & Hypotheses

Research Question

- What are the key factors that significantly influence heart attacks?
- Can heart attacks be predicted based on these factors?
- Is it possible to prevent heart attacks by addressing them?

Hypothese

By analyzing physiological, lifestyle, and historical data, we can accurately predict heart attack risk.

Alternative

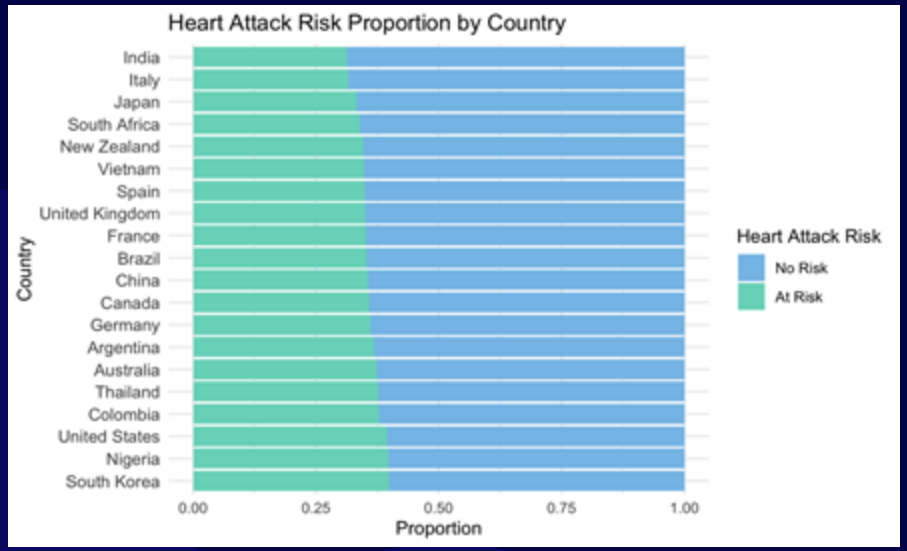
Heart attacks are unpredictable and depend mainly on genetics.

Overview of Data

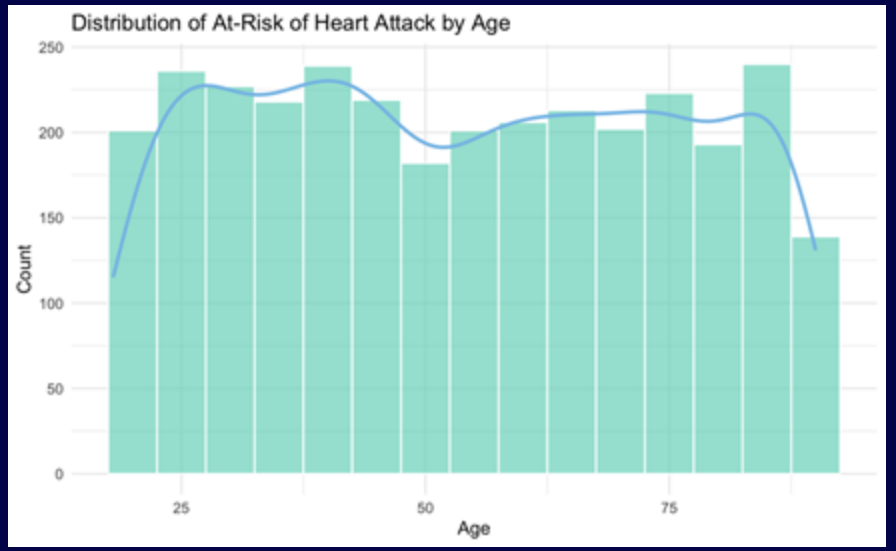
- **Dataset Source:** Kaggle's Heart Attack Risk Prediction Dataset
- **Sample Size:** 8,763 observations, 26 columns
- **Response Variable:** Heart Attack Risk (Binary: 1 = A t Risk, 0 = No Risk)
- **25 Predictor Variables:**
    1. **Demographic Factors:** Sex, Age, Country, Continent, Hemisphere, Income
    2. **Physiological & Clinical Variables:** Cholesterol, Blood.Pressure (Systolic, Diastolic), Triglycerides, Heart.Rate, BMI
    3. **Medical History:** Diabetes, Family.History, Previous.Heart.Problems, Medication.Use, Stress.Level
    4. **Lifestyle Factors:** Smoking, Alcohol.Consumption, Diet, Exercise.Hours.Per.Week, Physical.Activity.Days.Per.Week, Sedentary.Hours.Per.Day, Sleep.Hours.Per.Day

Descriptive Analysis

Heart Attack Risk by Country

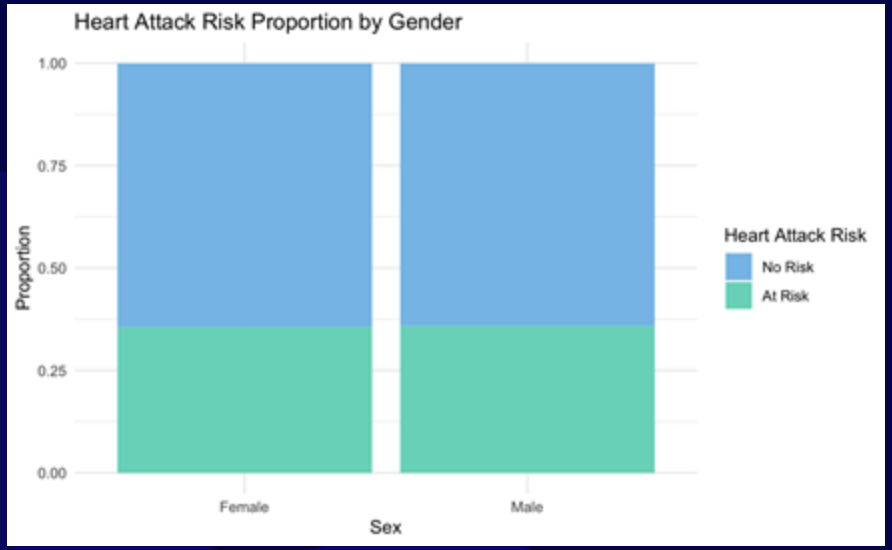


Heart Attack Risk by Age

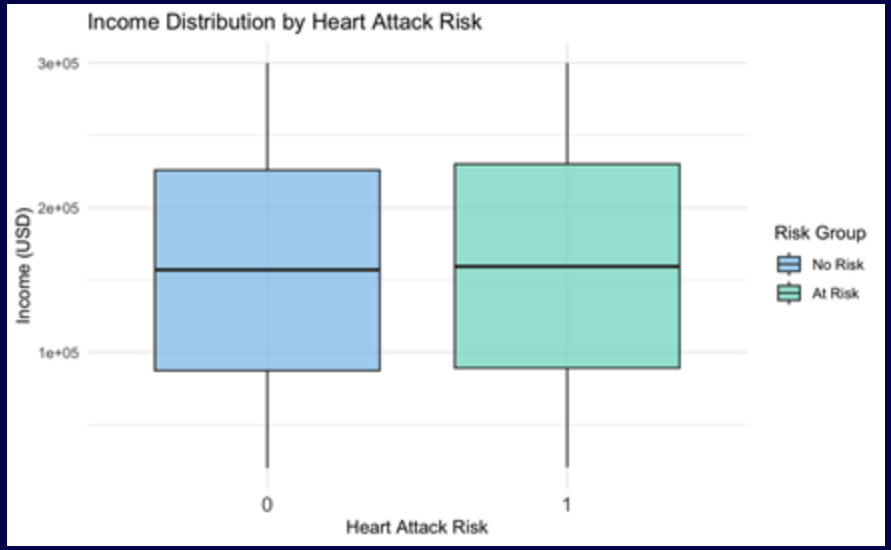


Descriptive Analysis

Heart Attack Risk by Gender

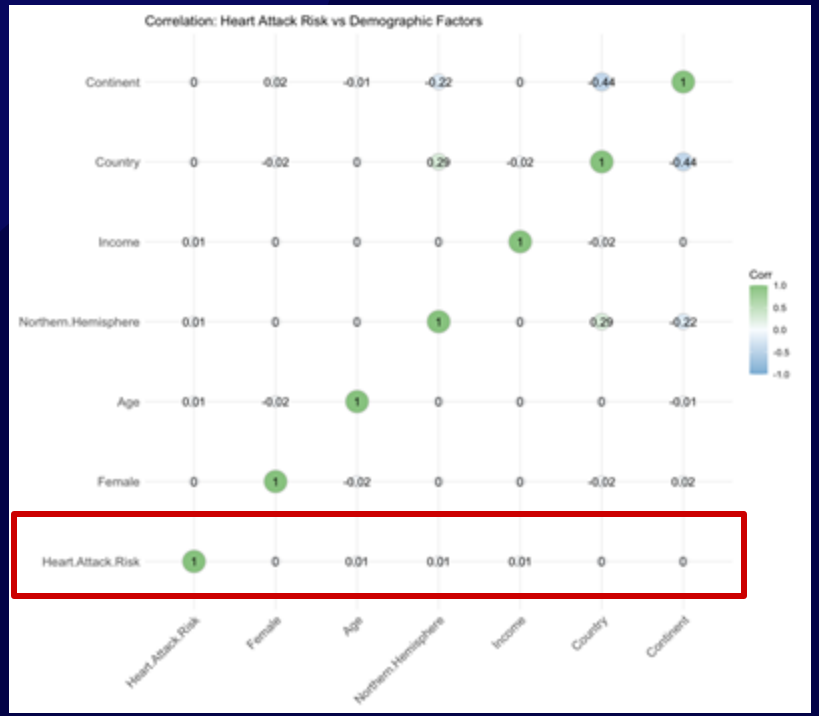


Heart Attack Risk by Income



Correlation Analysis

Demographic Factors



No correlation!

Physiological & Clinical Factors

Highest positive correlation: Cholesterol & Systolic

Correlation Analysis

Medical History Factors

Highest positive correlation: Diabetes

Lifestyle Factors

Highest negative correlation: Sleep.Hours.Per.Day

Correlation Analysis - Key Insights

- The strongest correlation is only around 0.02, suggesting no strong linear relationship.

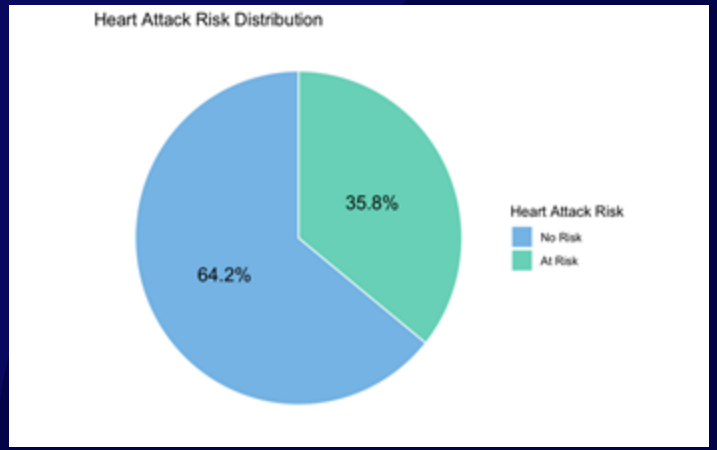
- In reality, the relationship between these factors and heart attack risk may be more complex, involving non-linear effects or interactions between multiple variables.

- This could also be due to incomplete data from Kaggle.

Data Preprocessing

1. Drop Patient.ID and move Heart.Attack.Risk to the first column
2. Handle missing values and outliers
    - No missing values and outliers detected based on z-score method
1. Split Blood.Pressure variable into Systolic and Diastolic
    - i.e. 158/88 to 158 Systolic and 88 Diastolic
1. Encode categorical variables into 0/1
    - 9 binary variable: only convert Sex into Female and Hemisphere into Northern.Hemisphere
    - 3 multi-category variables: convert to factors and create dummies for Diet, Country, Continent
1. Scale numeric features
    - Mean of 0 and Standard Deviation of 1.

Data Preprocessing



Original Dataset Distribution

Train-Test split (handling class imbalance)

- Original class distribution: ~36% "At risk", ~64% "No risk"
- Training set class distribution: 50% "At risk", 50% "No risk"
- Test set class distribution: ~36% "At risk", ~64% "No risk"

Final datasets (keep all 8,763 observations)

- data (26 columns) has multi-category variables as factors only
- data1 (52 columns) has multi-category variables converted to dummies

Predictive Models Selected

1

Logistic Regression

Linear, interpretable, probability-based classification

2

Discriminant Analysis

Statistical, assumes normality, linear/quadratic separation

3

k-NN

Distance-based, non-parametric, memory-intensive

4

Decision Tree

Rule-based, interpretable, prone to overfitting

5

Random Forest

Ensemble, reduces overfitting, high accuracy

6

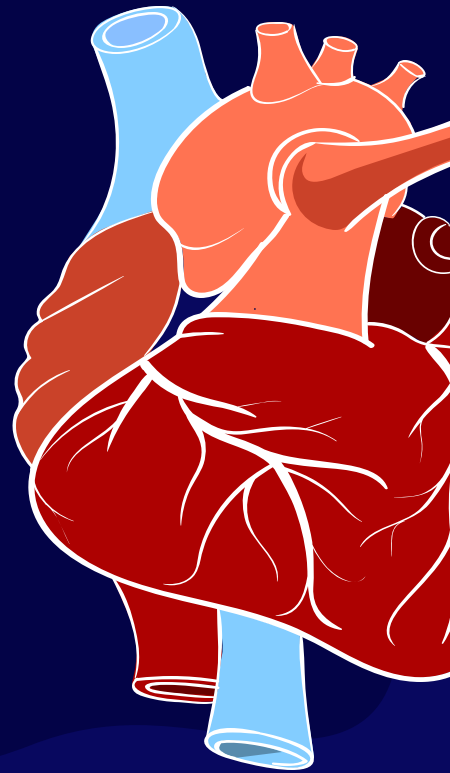
SVM

Hyperplane-based, good for high-dimensional data

Model Performance - Worldwide Data

Model performance on all dataset:

	Test Data Accuracy
Logistic Regression	55%
Discriminant Analysis	50%
k-NN	51.67%
Decision Tree	51.32%
Random Forest	51.37%
SVM	64.23%



Support Vector Machine

Data Processing

- Transfer all binary variables to factors
- Standardization using scale()

Tune Model

- gamma = c(0.5, 1, 2, 5)
- cost = c(0.01, 0.1, 1)

Best Model Selection

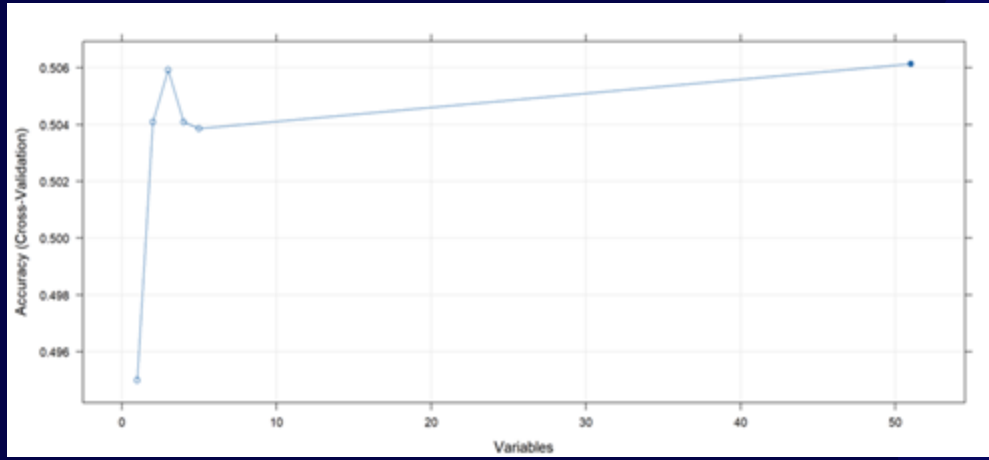
- Kernel = “radial”, gamma = 5, cost = 0.01

Prediction Result

- Accuracy: 64.23%

Summary

- Influential variables suggest that geographical and lifestyle factors might play a significant role in predicting the risk of heart attacks.



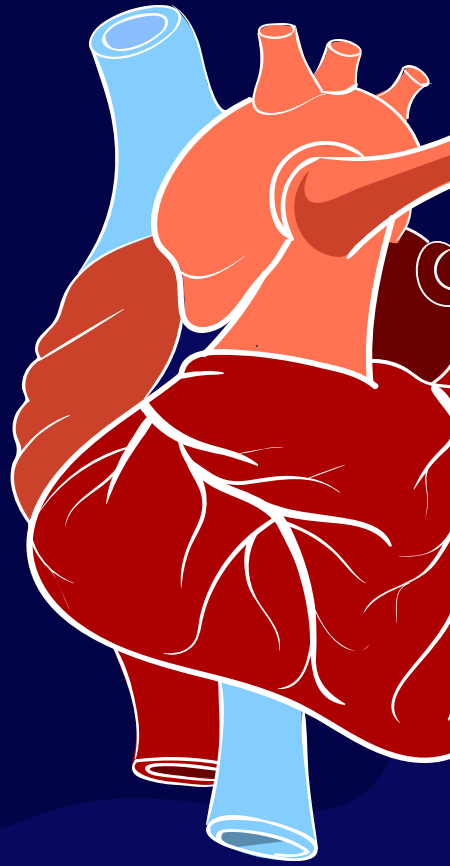
Top 5 Most Influential Variables:

- Country_CountryNigeria
- Country_CountryAustralia
- Continent_ContinentSouth.America
- Diet_DietHealthy
- Northern.Hemisphere

Model Performance - Subset Data

k-NN Model performance on different countries:

	Test Data Accuracy
Italy	68%
United States	49%
Japan	66%
China	48%



KNN

Data Processing

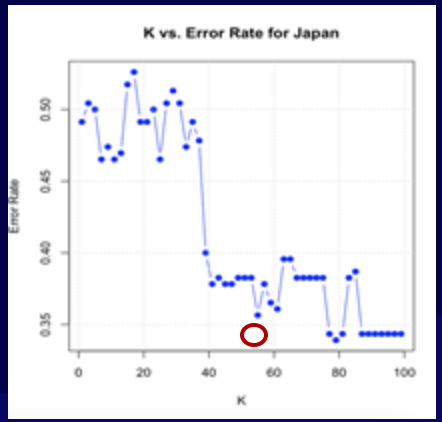
- Normalization method:
  Min-Max Normalization (first attempt:50.4%)
  Z-Score Standardization (final approach:51.7%)

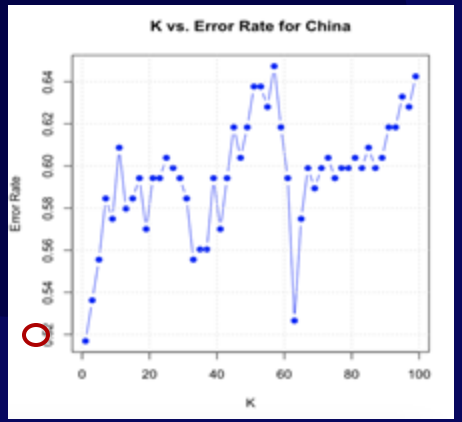
Tune Model

- Tested K values: k = seq(1, 99, by = 2)
- Selected the best K based on lowest test error

Best Model Selection

- **Italy:** K = 89, **Accuracy: 67.97%**
- **United States:** K = 7, **Accuracy: 48.63%**
- **Japan:** K = 79, **Accuracy: 66.09%**
- **China:** K = 1, **Accuracy: 48.31%**





CONCLUSIONS-Key Findings

- Our **SVM** is the best model for predicting all dataset. However we are not quite confident about using this model to predict heart attack risk worldwide.

- We are confident that using **k-NN** to predict heart attack risk in countries like **Italy and Japan**. We observed a moderate linear correlation between certain factors and heart attack risk, and the accuracy of the models is moderately high.

Top 3 factors in 4 Different Countries

1.**Physical.Activity.Days.Per.Week**(-0.1806)
2.**Systolic blood pressure**(+0.1449)
3.**Incomes**(-0.0952)

1.**BMI** (0.1694)
2.**Previous Heart Problems** (-0.1424)
3.**Income** (0.0942)

1.**Stress.Level** (-0.1589)
2.**Triglycerides** (+0.1219)
3.**Cholesterol** (+0.1023)

1.**Previous.Heart.Problems** (-0.3429)
2.**Heart.Rate** (+0.1970)
3.**Unhealthy Diet**(-0.1574)

CONCLUSIONS-Practical Implications

- Italy's heart attack prevention strategies should focus on tackling obesity, improving healthcare access for low-income groups, and prioritizing high-risk individuals with past heart problems.
- Japan should focus on preventive healthcare by encouraging heart rate monitoring, improving diet habits, and providing specialized care for those with previous heart issues.

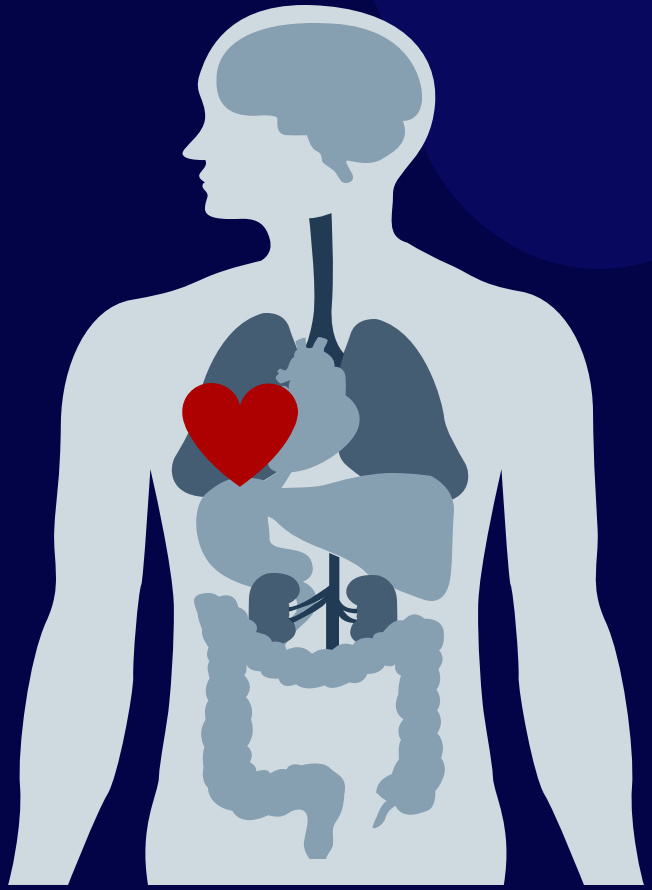
CONCLUSIONS-Limitations

- Current dataset may not match the actual situation
- Subsets of different country data are small

Next Steps:

- Real-time health data could improve the model's predictive power
- Larger, more diverse dataset
- Predict heart attack risks in other countries

Q & A



BANA 288–PREDICTIVE ANALYTICS
Professor– Ken Murphy
Group 5: Becky Wang, Vy Nguyen, Xinying Wu, Yanan Sun