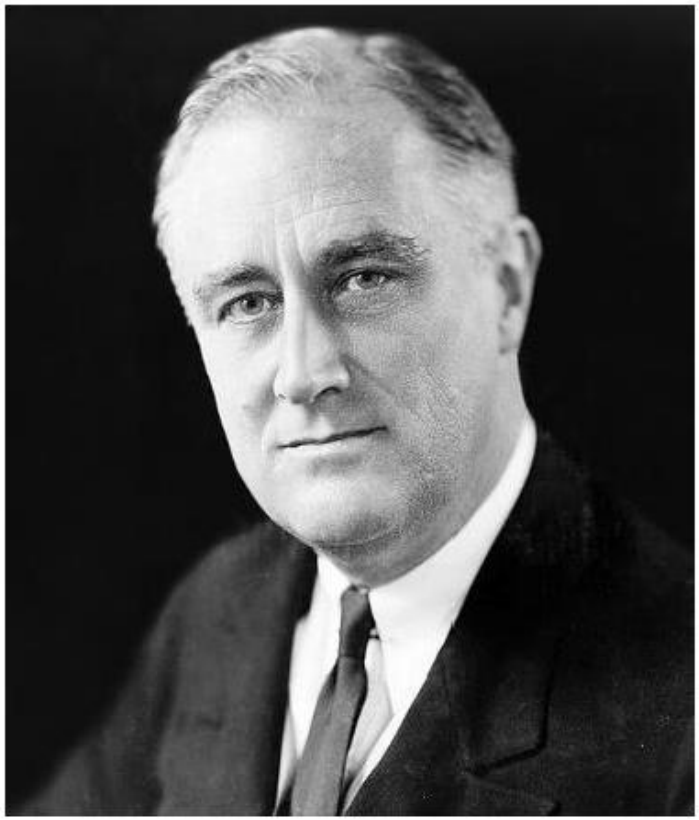
DIMENSIONLESS
TECHNOLOGY

The Framingham Heart Study

Franklin Delano Roosevelt

- President of the United States, 1933-1945
  - Longest-serving president
  - Led country through Great Depression
  - Commander in Chief of U.S. military in World War II
- Died while president, April 12, 1945



FDR's Blood Pressure

- Before presidency, blood pressure of 140/100
  - Healthy blood pressure is less than 120/80
  - Today, this is already considered high blood pressure
- One year before death, 210/120
  - Today, this is called Hypertensive Crisis, and emergency care is needed
  - FDR's personal physician:

    "A moderate degree of arteriosclerosis, although no more than normal for a man of his age"
- Two months before death: 260/150
- Day of death: 300/190

Early Misconceptions

- High blood pressure dubbed *essential hypertension*
  - Considered important to force blood through arteries
  - Considered harmful to lower blood pressure
- Today, we know better

"Today, presidential blood pressure numbers like FDR's would send the country's leading doctors racing down hallways … whisking the nation's leader into the cardiac care unit of Bethesda Naval Hospital."
-- Daniel Levy, Framingham Heart Study Director

How did we learn ?

- In late 1940s, U.S. Government set out to better understand cardiovascular disease (CVD)
- Plan: track large cohort of initially healthy patients over time
- City of Framingham, MA selected as site for study
  - Appropriate size
  - Stable population
  - Cooperative doctors and residents
- 1948: beginning of Framingham Heart Study

The Framingham Heart Study

- 5,209 patients aged 30-59 enrolled
- Patients given questionnaire and exam every 2 years
  - Physical characteristics
  - Behavioral characteristics
  - Test results
- Exams and questions expanded over time
- We will build models using the Framingham data to predict and prevent heart disease

Quick Question

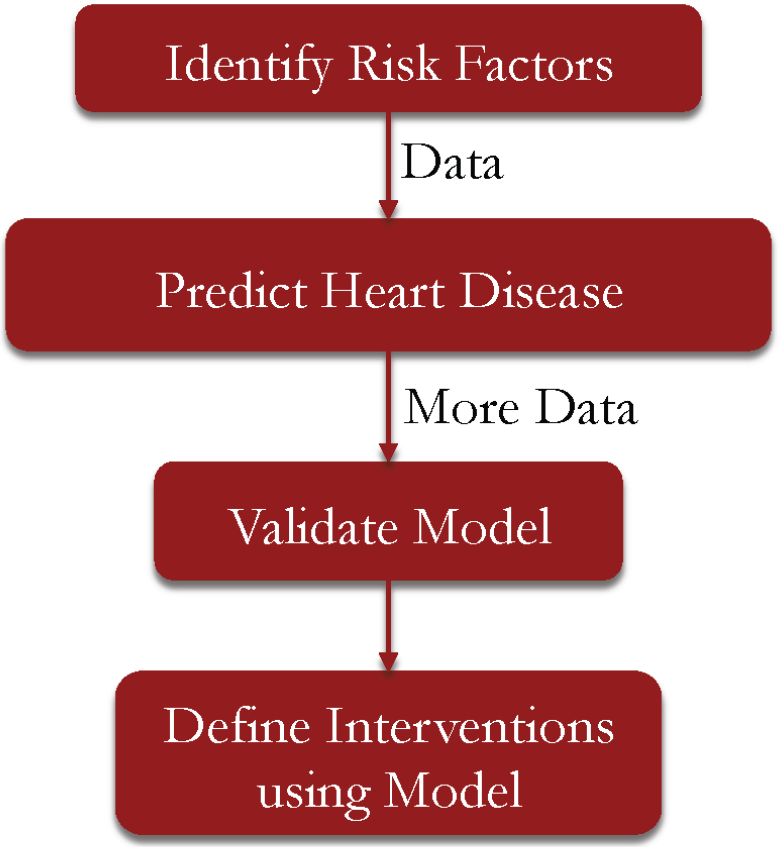
- Why was the city of Framingham, Massachusetts selected for this study? Select all that apply.

- [ ] It represented all types of people in the United States.
- [ ] It had an appropriate size.
- [ ] It had a stable population to observe over time.
- [ ] It contained an abnormally large number of people with heart disease.
- [ ] The doctors and residents in Framingham were willing to participate.

DIMENSIONLESS
TECHNOLOGY

Risk Factors

Analytics to prevent Heart Disease



Coronary Heart Disease

- We will predict 10-year risk of CHD
  - Subject of important 1998 paper, introducing the Framingham Risk Score
- CHD is a disease of the blood vessels supplying the heart
- Heart disease has been the leading cause of death worldwide since 1921
  - 7.3 million people died from CHD in 2008
  - Since 1950, age-adjusted death rates have declined 60%

Risk Factors

- *Risk factors* are variables that increase the chances of a disease
- Term coined by William Kannell and Roy Dawber from the Framingham Heart Study
- Key to successful prediction of CHD: identifying important risk factors

Hypothesized CHD Risk Factors

- We will investigate risk factors collected in the first data collection for the study
  - Anonymized version of original data
- Demographic risk factors
  - *male*: sex of patient
  - *age*: age in years at first examination
  - *education*: Some high school (1), high school/GED (2), some college/vocational school (3), college (4)

Hypothesized CHD Risk Factors

- Behavioral risk factors
  - *currentSmoker*, *cigsPerDay*: Smoking behavior
- Medical history risk factors
  - *BPmeds*: On blood pressure medication at time of first examination
  - *prevalentStroke*: Previously had a stroke
  - *prevalentHyp*: Currently hypertensive
  - *diabetes:* Currently has diabetes

Hypothesized CHD Risk Factors

- Risk factors from first examination
  - *totChol*: Total cholesterol (mg/dL)
  - *sysBP*: Systolic blood pressure
  - *diaBP*: Diastolic blood pressure
  - *BMI*: Body Mass Index, weight (kg)/height $(m)^2$
  - *heartRate*: Heart rate (beats/minute)
  - *glucose*: Blood glucose level (mg/dL)

Quick Question

- Are "risk factors" the independent variables or the dependent variables in our model?

- Independent Variables
- Dependent Variables
- Neither

Quick Question

In many situations, a dataset is handed to you and you are tasked with discovering which variables are important. But for the Framingham Heart Study, the researchers had to collect data from patients. In a situation like this one, where data needs to be collected by the researchers, should the potential risk factors be defined before or after the data is collected?

- Before
- After

A Logistic Regression Model

An Analytical Approach

- Randomly split patients into training and testing sets
- Use logistic regression on training set to predict whether or not a patient experienced CHD within 10 years of first examination
- Evaluate predictive power on test set

Building the Dataset

- Read the data file "framingham.csv" from https://storage.googleapis.com/dimensionless/Analytics/framingham.csv
- Save it in data frame "framingham"
- Look at the structure of the data frame.
- We have data for 4,240 patients and 16 variables.
- Explore each of the variables
- The last variable is the outcome or dependent variable, whether or not the patient developed CHD in the next 10 years.

Building the Dataset

- Split the data into a training set and a testing set.
- Set the seed to 1000.
- Use split ratio to be 0.65.
  - When you have more data like we do here, you can afford to put less data in the training set and more in the testing set.This will increase our confidence in the ability of the model to extend to new data since we have a larger test set, and still give us enough data in the training set to create our model.
- We'll call our training set "train" and testing set "test.

Building the model

- Build our logistic regression model using the training set. We'll call it framinghamLog
- We will be using all of the other variables in the data set as independent variables.
- We can do this by
    - Glm(TenYearCHD~.,data=train,family=binomial)
- Look at the summary of the model.
- Male, age, prevalent stroke, total cholesterol, systolic blood pressure, and glucose are all significant in our model.
- Cigarettes per day and prevalent hypertension are almost significant.
- All of the significant variables have positive coefficients, meaning that higher values in these variables contribute to a higher probability of 10-year coronary heart disease.

Making Predictions

- We'll call our predictions predictTest
- Run the predict() on test data.
- Use a threshold value of 0.5 to create a confusion matrix.
- With a threshold of 0.5, we predict an outcome of 1, the true column, very rarely. This means that our model rarely predicts a 10-year CHD risk above 50%.
- Calculate the accuracy of this model.
- Calculate the accuracy of the baseline model.
- Compute the out-of-sample AUC.

Model Strength

- Model rarely predicts 10-year CHD risk above 50%
  - Accuracy very near a baseline of always predicting no CHD
- Model can differentiate low-risk from high-risk patients (AUC = 0.74)
- Some significant variables suggest interventions
  - Smoking
  - Cholesterol
  - Systolic blood pressure
  - Glucose

Validating the Model

Risk Model Validation

- So far, we have used *internal validation*
  - Train with some patients, test with others
- Weakness: unclear if model generalizes to other populations
- Framingham cohort white, middle class
- Important to test on other populations

Framingham Risk Model Validation

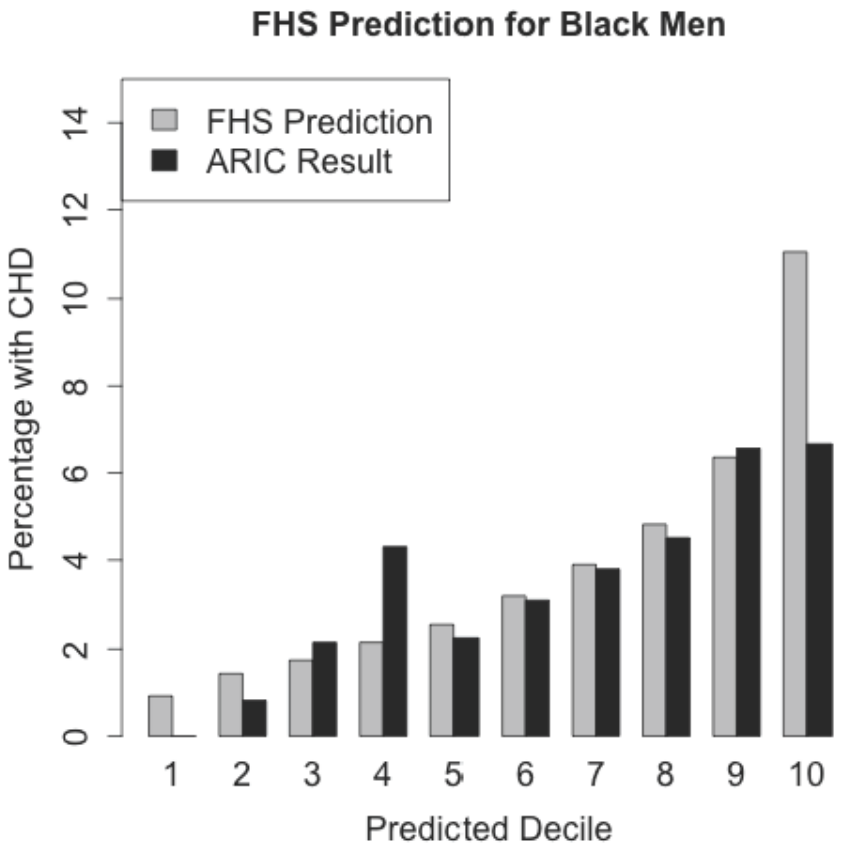
- Framingham Risk Model tested on diverse cohorts

Study	Population
Atherosclerosis Risk in Communities (ARIC) Study	White and Black
Honolulu Heart Program (HHP)	Japanese American
Puerto Rico Heart Health Program (PR)	Hispanic
Strong Heart Study (SHS)	Native American

- Cohort studies collecting same risk factors
- Validation Plan
  - Predict CHD risk for each patient using FHS model
  - Compare to actual outcomes for each risk decile

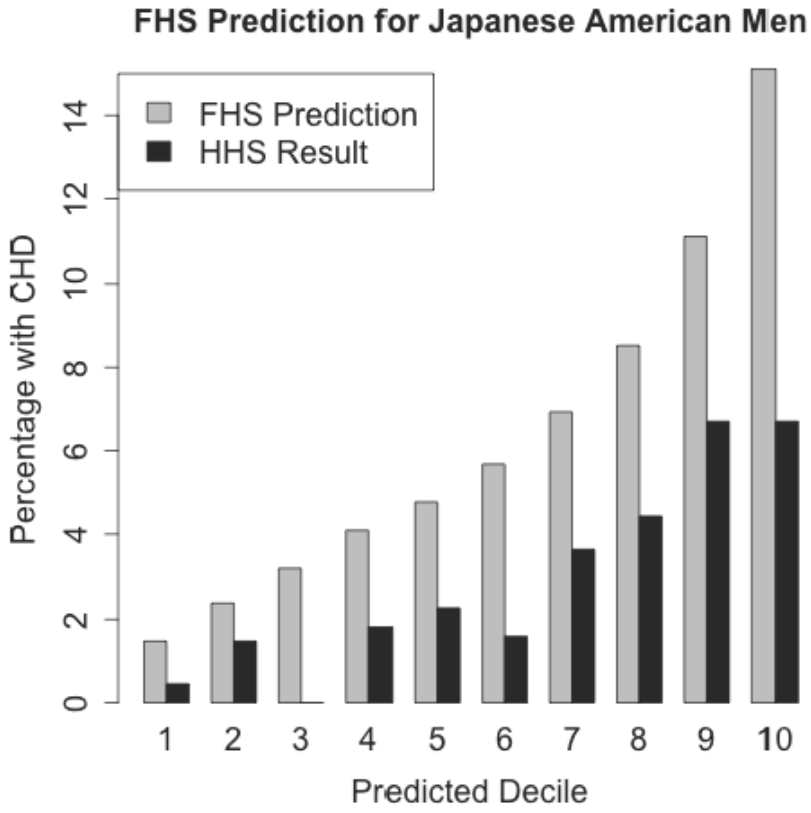
Validation for Black Men

- 1,428 black men in ARIC study
- Similar clinical characteristics, except higher diabetes rate
- Similar CHD rate
- Framingham risk model predictions accurate



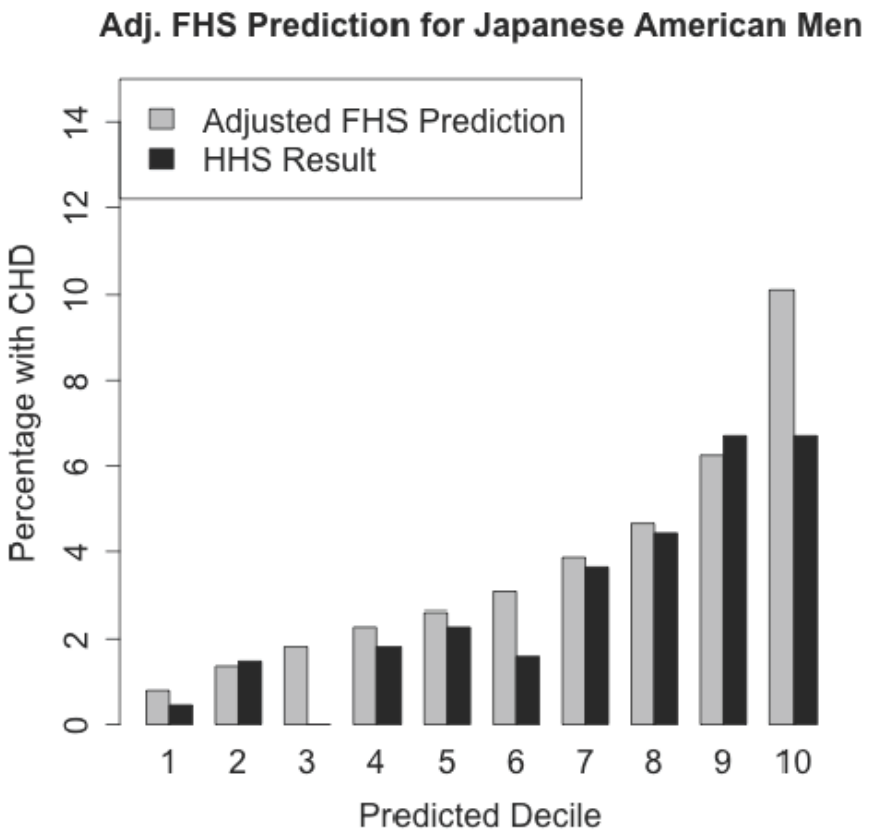
Validation for Japanese American Men

- 2,755 Japanese American men in HHS
- Lower CHD rate
- Framingham risk model systematically overpredicts CHD risk



Recalibrated Model

- Recalibration adjusts model to new population
- Changes predicted risk, but does not reorder predictions
- More accurate risk estimates



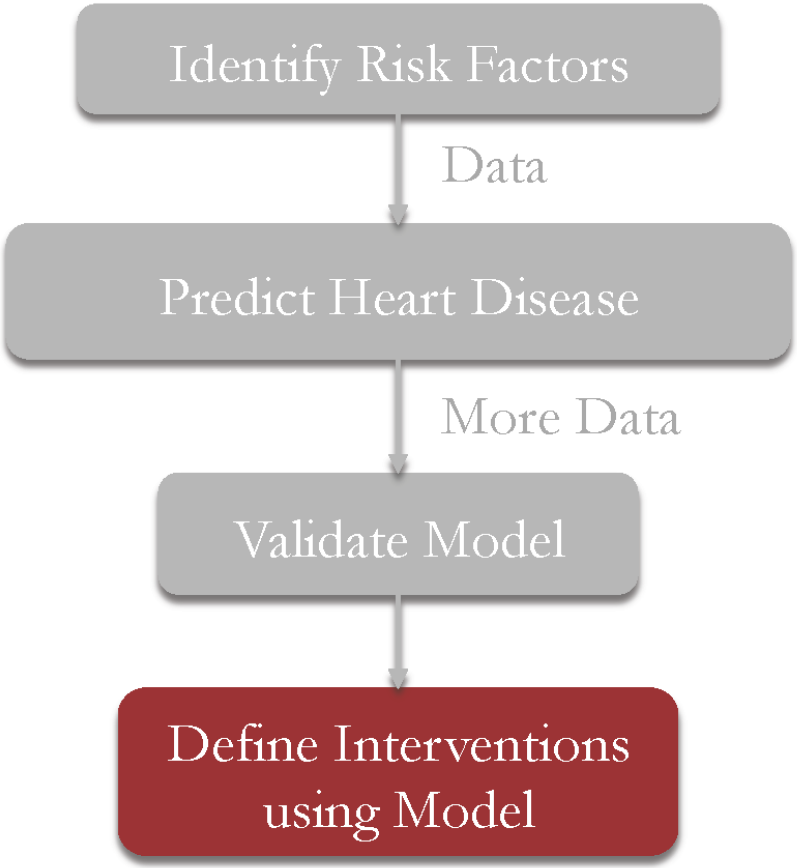
Quick Question

- For which of the following models should external validation be used? Consider both the population used to train the model, and the population that the model will be used on. (Select all that apply.)

- [ ] A model to predict obesity risk. Data from a random sample of California residents was used to build the model, and we want to use the model to predict the obesity risk of all United States residents.
- [ ] A model to predict the stress of MIT students. Data from a random sample of MIT students was used to build the model, and we want to use the model to predict the stress level of all MIT students.
- [ ] A model to predict the probability of a runner winning a marathon. Data from all runners in the Boston Marathon was used to build the model, and we want use the model to predict the probability of winning for all people who run marathons.

Interventions

Intervention



Drugs to Lower Blood Pressure

- In FDR's time, hypertension drugs too toxic for practical use
- In 1950s, the diuretic chlorothiazide was developed
- Framingham Heart Study gave Ed Freis the evidence needed to argue for testing effects of BP drugs
- Veterans Administration (VA) Trial: randomized, double blind clinical trial
- Found decreased risk of CHD
- Now, >$1B market for diuretics worldwide

Drugs to Lower Cholesterol

- Despite Framingham results, early cholesterol drugs too toxic for practical use
- In 1970s, first statins were developed
- Study of 4,444 patients with CHD: statins cause 37% risk reduction of second heart attack
- Study of 6,595 men with high cholesterol: statins cause 32% risk reduction of CVD deaths
- Now, > $20B market for statins worldwide

Quick Question

We had built a logistic regression model and found that the following variables were significant (or almost significant) for predicting ten year risk of CHD: male, age, number of cigarettes per day, whether or not the patient previously had a stroke, whether or not the patient is currently hypertensive, total cholesterol level, systolic blood pressure, and blood glucose level.

Quick Question

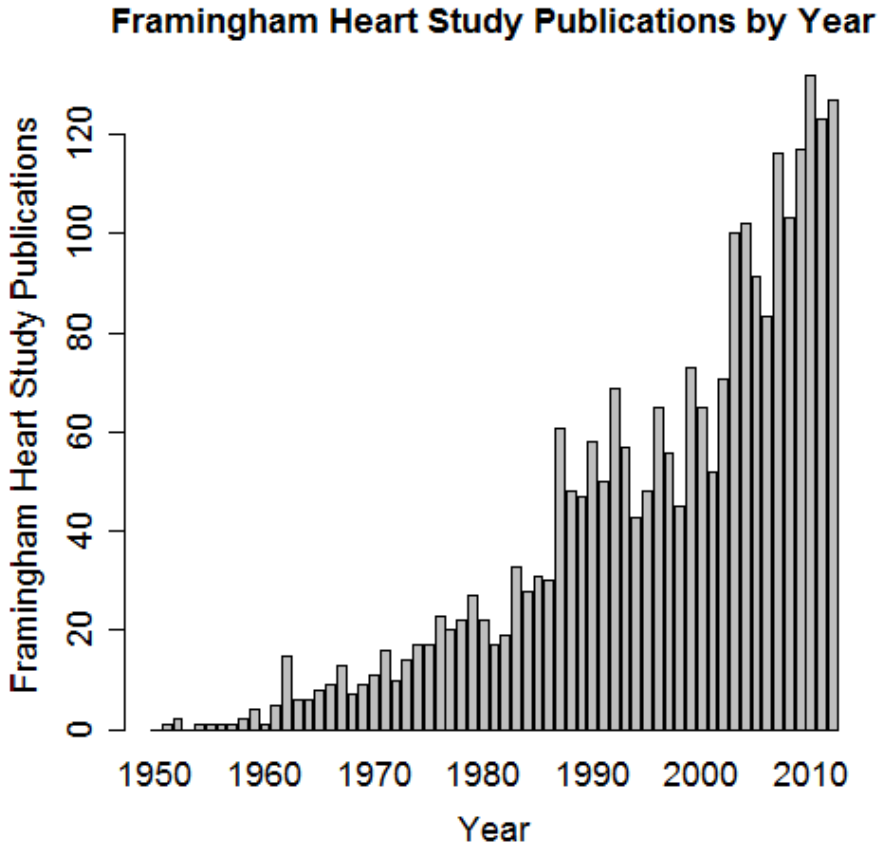
Which **one** of the following variables would be the most dramatically affected by a behavioural intervention? HINT: Think about how much control the patient has over each of the variables.

- Male
- Age
- Number of Cigarettes per day
- Previously had a Stroke
- Hypertensive
- Total Cholesterol Level
- Systolic Blood Pressure
- Blood Glucose Level

Overall Impact

The Heart Study Through the years

- More than 2,400 studies use Framingham data
- Many other risk factors evaluated
  - Obesity
  - Exercise
  - Psychosocial issues
  - …
- *Texas Heart Institute Journal*: top 10 cardiology advances of 1900s



Available Online

Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack

The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and older who do not have heart disease or diabetes. To find your risk score, enter your information in the calculator below.

Age:	years
Gender:	Female Male
Total Cholesterol:	mg/dL
HDL Cholesterol:	mg/dL
Smoker:	No Yes
Systolic Blood Pressure:	mm/Hg
Are you currently on any medication to treat high blood pressure.	No Yes

Calculate Your 10-Year Risk

TOP **Total cholesterol** - Total cholesterol is the sum of all the cholesterol in your blood. The higher your total cholesterol, the greater your risk for heart disease. Here are the total values that matter to you:

Less than 200 mg/dL 'Desirable' level that puts you at lower risk for heart disease. A cholesterol level of 200 mg/dL or greater increases your risk.

200 to 239 mg/dL 'Borderline-high.'

[Framingham Risk Score Calculator]

Research Direction and Challenges

- Second generation enrolled in 1971, third in 2002
  - Enables study of family history as a risk factor
- More diverse cohorts begun in 1994 and 2003
- Social network analysis of participants
- Genome-wide association study linking studying genetics as risk factors
- Many challenges related to funding
  - Funding cuts in 1969 nearly closed study
  - 2013 sequester threatening to close study

Clinical Decision Rules

- Paved the way for *clinical decision rules*
- Predict clinical outcomes with data
  - Patient and disease characteristics
  - Test results
- More than 75,000 published across medicine
- Rate increasing