In this lecture, we'll be using analytical models to prevent heart disease.

The first step is to identify risk factors, or the independent variables, that we will use in our model.

Then, using data, we'll create a logistic regression model to predict heart disease.

Using more data, we'll validate our model to make sure it performs well out of sample and on different populations than the training set population.

Lastly, we'll discuss how medical interventions can be defined using the model.

We'll be predicting the 10-year risk of coronary heart disease or CHD.

This was the subject of an important 1998 paper introducing what is known as the Framingham Risk Score.

This is one of the most influential applications of the Framingham Heart Study data.

We'll use logistic regression to create a similar model.

CHD is a disease of the blood vessels supplying the heart.

This is one type of heart disease, which has been the leading cause of death worldwide since 1921.

In 2008, $7.3 million people died from CHD.

Even though the number of deaths due to CHD is still very high, age-adjusted death rates have actually declined 60% since 1950.

This is in part due to earlier detection and monitoring partly because of the Framingham Heart Study.

Before building a logistic regression model, we need to identify the independent variables we want to use.

When predicting the risk of a disease, we want to identify what are known as risk factors.

These are the variables that increase the chances of developing a disease.

The term risk factors was actually coined by William Kannell and Roy Dawber from the Framingham Heart Study.

Identifying these risk factors is the key to successful prediction of CHD.

In this lecture, we'll focus on the risk factors that they collected data for in the original data collection for the Framingham Heart Study.

We'll be using an anonymized version of the original data that was collected.

This data set includes several demographic risk factors-- the sex of the patient, male or female; the age of the patient in years; the education level coded as either 1 for some high school, 2 for a high school diploma or GED, 3 for some college or vocational school, and 4 for a college degree.

The data set also includes behavioral risk factors associated with smoking-- whether or not the patient is a current smoker and the number of cigarettes that the person smoked on average in one day.

While it is now widely known that smoking increases the risk of heart disease, the idea of smoking being bad for you was a novel idea in the 1940s.

Medical history risk factors were also included.

These were whether or not the patient was on blood pressure medication, whether or not the patient had previously had a stroke, whether or not the patient was hypertensive, and whether or not the patient had diabetes.

Lastly, the data set includes risk factors from the first physical examination of the patient.

The total cholesterol level, systolic blood pressure, diastolic blood pressure, Body Mass Index, or BMI, heart rate, and blood glucose level of the patient were measured.

In the next video, we'll use these risk factors to see if we can predict the 10-year risk CHD.