Claims data offers an expansive view of the patients health history.

Specifically, claims data include information on demographics, medical history, and medications.

They offer insights regarding a patient's risk.

And as I will demonstrate, may reveal indicative signals and patterns.

We'll use health insurance claims filed for about 7,000 members from January, 2000 until November, 2007.

We concentrated on members with the four main attributes.

At least five claims with coronary artery disease diagnosis, at least five claims with hypertension diagnostic codes, at least 100 total medical claims, at least five pharmacy claims, and data from at least five years.

These selections yield patients with a high risk of heart attack.

And a reasonably rich medical history with continuous coverage.

Let us discuss how we've aggregated this data.

The resulting data sets includes about 20 million health insurance entries, including individual, medical, and pharmaceutical records.

Diagnosis, procedures, and drug codes in the data set comprised tens of thousands of attributes.

The codes were aggregated into groups.

218 diagnosis groups, 180 procedure groups, 538 drug groups.

46 diagnosis groups were considered by clinicians as possible risk factors for heart attacks.

Let us discuss how we view the data over time.

It is important in this study to view the medical records chronologically, and to represent a patient's diagnosis profile over time.

So we record the cost and number of medical claims and hospital visits by a diagnosis.

All the observations we have span over five years of data.

They were split into 21 periods, each 90 days in length.

We examine nine months of diagnostic history, leading up to heart attack or no heart attack event, and align the data to make observations date-independent, while preserving the order of events.

We recorded the diagnostic history in three periods.

Zero to three months before the event, three to six months before the event, and six to nine months before the event.

What was a target variable we're trying to predict?

The target prediction variable is the occurrence of a heart attack.

We define this from a combination of several claims.

Namely, diagnosis of a heart attack, alongside a trip to the emergency room, followed by subsequent hospitalization.

Only considering heart attack diagnosis that are associated with the visits to an emergency room, and following hospitalization helps ensure that the target outcome is in fact a heart attack event.

The target variable is binary.

It is denoted by plus 1 or minus 1 for the occurrence or non-occurrence of a heart attack in the targeted period of 90 days.

How's the data organized?

There were 147 variables.

Variable one is the patient's identification number, and variable two is the patient's gender.

There were variables related to the diagnoses group counts nine, six, and three months before the heart attack target period.

There were variables related to the total course nine, six, and three months before the heart attack target period, and the final variable for 147, includes the classification of whether the event was a heart attack or not.

Cost of medical care is a good summary of a person's health.

In our database, the total cost of medical care in the three 90 day periods preceding the heart attack target event

ranged from $0.00 to $636,000 and approximately 70% of the overall cost were generated by only 11% of the population.

This means that the highest patients with high medical expenses are a very small proportion of the data, and could skew our final results.

According to the American Medical Association, only 10% of individuals have projected medical expenses of approximately $10,000 or greater per year, which is more than four times greater than the average projected medical expenses of 2,400 per year.

To lessen the effects of these high-cost outliers, we divided the data into different cost buckets, based on the findings of the American Medical Association.

We did not want to have too many cost bins because the size of the data set.

The table in the slide gives a summary of the cost bucket partitions.

Patients with expenses over $10,000 in the nine month period were allocated to cost bucket 3.

Patients with less than 2,000 in expenses were allocated to cost bucket 1.

And the remaining patients with costs between 2,000 and 10,000 to cost bucket 2.

Please note that the majority of patients, 4,400 out of 6,500, or 67.5% of all patients fell into the first bucket of low expenses.