Let us discuss data sources in the health care industry.

So the industry is data-rich, but data may be hard to access.

Sometimes it involves unstructured data like doctor's notes.

Often the data is hard to get due to differences in technology.

Hospitals in southern Massachusetts versus California might use different technologies and different platforms.

Finally there are strong privacy laws, HIPAA, around health care data sharing.

So what is available?

Claims data is a major source.

Claims data are requests for reimbursement submitted to insurance companies or state-provided insurance from doctors, hospitals and pharmacies.

Another source of data is the eligibility information for employees.

And finally demographic information: gender and age.

Let me give you some examples on claims data.

So this shows six different claims.

Let's consider this one.

So this is the provider's name.

The corresponding diagnostic code.

This is about upper respiratory disorders.

This is another code associated with the diagnosis.

This is the scientific term for the diagnosis.

The specific code again.

This was an office visit, and it's an established patient.

The date.

And the amount of money that was claimed by the physician.

Others claims are similar.

As we see, the claims data is a rich, structured data source.

It is very high dimensional.

For example, claims involving diagnosis involve thousands of different codes.

Similarly with drugs, where there are tens of thousands, and procedures.

However, this collection of data does not capture all aspects of a person's treatment or health.

Many things must be inferred.

Unlike electronic medical records, we do not know the results of a test, only that the test was administered.

For example, we do not know the results of a blood test, but we do know that the blood test was administered.

The specific exercise we are going to see in this lecture is an analytics approach to building models starting with 2.4 million people over a three year span.

The observation period was 2001 to 2003.

This is where this data is coming from.

And then out of sample, we make predictions for the period of 2003 and 2004.

This was in the early years of D2Hawkeye.

Out of the 2.4 million people, we included only people with data for at least 10 months in both periods, both in the observation period and the results period.

This decreased the data to 400,000 people.