So let us explain what claims data is.

So medical claims are generated when a patient visits a doctor.

Medical claims include diagnosis code, procedures codes, as well as costs.

Pharmacy claims involve drugs, the quantity of these drugs, the prescribing doctor, as well as the medication costs.

Claims data are electronically available, they are standardized, they use well-established codes.

However, since humans generate them, they are not 100% accurate.

And often, under-reporting is common in the sense that it's a tedious job to record these claims, and as a result, often people under-report them.

Also, claims for hospital visits can be vague.

In creating a data set, our objective was to assess quality, health care quality.

So we used a large health insurance claims database, and we randomly selected 131 diabetes patients.

The ages ranged between 35 to 55 and the costs were in the neighborhood of $10,000 to $20,000.

The period in which these claims were recorded were September 1, 2003 to August 31, 2005.

An expert physician reviewed the claims and wrote descriptive notes, like "ongoing use of narcotics"; "only on Avandia, not a good first choice drug"; "had regular visits, mammogram, and immunizations"; "was given home testing supplies".

After this review, this expert physician rated the quality of care on a two-point scale, poor or good.

Examples included, I'd say care was poor.

Poorly treated diabetes.

Not an eye exam, but overall I'd say high quality.

So based on these comments, we extracted variables.

The dependent variable was the quality of care.

The independent variables involve the ongoing use of narcotics; only on Avandia, not a good first choice drug; had regular visits, mammogram, and immunizations; was given home testing supplies.

Overall, the independent variables involved diabetes treatment variables, patient demographics, health care utilization, providers, claims, and prescriptions.

The dependent variable was modeled as a binary variable -- 1 for low-quality care and 0 for high-quality care.

This is by its nature a categorical variable.

It only takes two possible values.

We have seen linear regression as a way of predicting continuous outcomes.

Of course, we can utilize linear regression to predict quality of care here, but then we have to round the outcome to 0 or 1.

Instead, we will explain in this lecture how we can use logistic regression, which is an extension of linear regression, to environments where the dependent variable is categorical.

In our case, 0 or 1.