To build an analytics model, let us discuss the variables we used.

First, we used the 13,000 diagnoses.

It's for the codes for diagnosis that claims data utilize.

There were also 22,000 different codes for procedures and 45,000 codes for prescription drugs.

To work with this massive amount of variables, we aggregated the variables as follows.

Out of the 13,000 diagnoses, we defined 217 diagnosis groups.

Out of the 20,000 procedures, we aggregated the data to develop 213 procedure groups.

And, finally, from 45,000 prescription drugs, we developed 189 therapeutic groups.

To illustrate an example of how we infer further information from the data, the graph here shows on the horizontal axis, time, and on the vertical axis, costs in thousands of dollars.

So patient one is a patient who, on a monthly basis, has costs on the order of $10,000 to $15,000, a fairly significant cost but fairly constant in time.

Patient two has also an annual cost of a similar size to patient one.

But in all but the third month, the costs are almost $0.

Whereas in the third month, it cost about $70,000.

In fact, this is additional data we defined indicating whether the patient has a chronic or an acute condition.

In addition to the initial variables, the 217 procedure groups, and 189 drugs, and so forth, we also defined in collaboration with medical doctors, 269 medically-defined rules.

For example, the first type of rule indicates the interaction between various indices.

For example, obesity and depression.

Then new variables regarding interaction between diagnosis and age.

For example, more than 65 years old and coronary artery disease.

Noncompliance with treatment.

For example, non-fulfillment of a particular drug order.

And, finally, illness severity.

For example, severe depression as opposed to regular depression.

And the last set of variables involve demographic information like gender and age.

An important aspect of the variables are the variables related to cost.

So rather than using costs directly, we bucketed costs and considered everyone in the group equally.

So we defined five buckets.

So the buckets were partitioned in such a way so that 20% of all costs is in bucket five, 20% is in bucket four, and so forth.

So the partitions were from 0 to 3,000, from 3,000 to 8,000, from 8,000 to 19,000, from 19,000 to 55,000, and above 55,000.

The number of patients that were below 3,000 was-- 78% of the patients had costs below 3,000.

Just to remind you, we created a bucket so that the total cost in each bucket was 20% of the total.

But the number of patients in bucket one, for example, is very high (78%).

Let us interpret the buckets medically.

So this shows the various levels of risk.

Bucket one consists of patients that have rather low risk.

Bucket two has what is called emerging risk.

In bucket three, moderate level of risk.

Bucket four, high risk.

And bucket five, very high risk.

So from a medical perspective, buckets two and three, the medical and the moderate risk patients, are candidates for wellness programs.

Whereas bucket four, the high risk patients, are candidates for disease management programs.

And finally bucket five, the very high risk patients, are candidates for case management.