produced." He has equally praised the 2003 vintage. Without tasting a single drop of wine, Ashenfelter was able to come to the same conclusion about the quality of the vintages as the man considered to be the most influential wine critic of our time. Additionally, his method does not require any tasting of the wine; just the use of a simple equation.

1.2 Assessing Quality in Healthcare

Perhaps no other domestic policy topic in the United States has spawned more debate in recent years than healthcare. However, even though there is a significant amount of disagreement about healthcare policies, the ultimate goal of politicians, physicians, hospitals, and patients is the same: *quality* healthcare. But what exactly is quality in healthcare? How is it defined, measured, and improved? One possibility is for an expert physician with many years of experience to look at cases and assess the quality of healthcare that patients have received. Clearly, this is impractical if one wants to make this assessment available for every patient in the healthcare system, as it takes an expert physician approximately an hour to read and understand each case before making an appropriate assessment. Also, how does the physician get the information to make this assessment?

Dimitris Bertsimas and David Czerwinski (two analytics professors), together with Michael Kane (a physician), built a model that assesses the quality of healthcare received by a group of patients. In this work, they define good quality care as healthcare that improves outcomes, educates patients, coordinates care among all doctors that see a patient, and controls costs. The goal is to capture the concerns of the patient, physician, and hospital.

This model is built using the opinion of a physician, and it is intended to accurately predict good or bad quality of care on cases not seen by the physician. With this method, an expert's opinion is used to assess the quality of care, but a model is developed to extend this opinion to all patients, without requiring the expert to evaluate each individual case. For this model, the researchers set the goal of predicting the opinion of Dr. Michael Kane, an internal medicine physician with over 40 years of experience. Keep in mind that the model could easily be extended to predict the average opinion of a committee of physicians, or to predict the opinion of a set of guidelines (this is discussed more later in this section). The key observation here is that the model predicts the opinion of a *domain expert*; this concept could be extended to many other applications and problems.

Claims data, the data that healthcare providers submit to insurance companies to be paid for various services, provides the data for this model. This was the most easily accessible and sufficiently large set of data available electronically and up to date. This data is not 100% accurate, and underreporting is common, but other data sources are not very accessible. With the increasing use of electronic medical records, there is potential for more accurate and complete data in future calibrations of the model. We will also see claims data used for different applications later in this book.

The first step in building the model was asking a physician, in this case Dr. Kane, to rate the quality of care of a set of 101 patients. The patients selected were diabetic and between the ages of 35 and 55 with annual healthcare costs between $10,000 and $20,000. The creators of the model decided to test it with diabetic patients since there is a wide range of tests, medications, and complications associated with diabetes. The age range was to ensure that the patients should receive similar care (an 18 year old and an 81 year old will probably need very different care, partially due to their age difference), and the cost range was to insure that the patient had enough data to make an accurate assessment, but not so much data that it was impractical for Dr. Kane to review.

Dr. Kane rated the quality of care for each patient on a two point scale: poor care, or good care. He also gave his level of confidence that the patient was indeed receiving that quality of care. These ratings are summarized in Table 1.2. He also wrote a paragraph for each patient, explaining his reasoning. The following paragraph gives an example:

> Male on glucophage, had sporadic medical visits and labs. Did have eye exam 4/05. His primary problems were back pain and narcotic use. He had monthly percocet prescriptions in addition to an NSAID and a muscle relaxer. No diagnostic studies. Had a few physical therapy visits in October and November 2003. No other significant diagnostic or therapeutic initiatives. Poor care with high confidence.

From Dr. Kane's assessments, 80 different variables were defined that fell into six different categories: (1) diabetes treatment (e.g. number of glycated hemoglobin tests); (2) utilization (e.g. number of office visits); (3) markers of good care (e.g. mammogram); (4) markers of poor care (e.g. narcotics); (5) providers (e.g. number of doctors); and (6) prescriptions (e.g. number of different drugs). Since blindly classifying every patient as receiving good care gives an accuracy percentage of 78%, the goal was to be more accurate than this. In addition to accuracy, the types of errors that occur are also significant: the percentage of cases classified as poor care that are actually good care and the percentage of cases classified as good care that are actually poor care. Errors of the first type (good care mistakenly classified as poor care) may cause unnecessary expenses to investigate and treat perfectly fine patients. But errors of the second type (poor care mistakenly classified as good care) can be very dangerous. These errors mean that we are overlooking patients that need help, and might cause serious health issues in the future. Understanding the trade-off between the different types of errors for any model is critical in deciding how useful the model will be in practice.

The model uses logistic regression to predict the binary variable "Quality," which takes value 1 if the patient received good quality care according to the

Table 1.2: The quality of care ratings given by Dr. Kane, along with his level of confidence.

	Low Confidence	High Confidence
Poor Quality	5	17
Good Quality	19	60

assessment of Dr. Kane, and takes value 0 otherwise. Logistic regression is a predictive method used when the outcome variable is binary, and produces a vector of coefficients similarly to linear regression. For more about logistic regression, see Chapter 21.

In this case, the best logistic regression model uses only three different independent variables: a binary variable for whether or not the patient was started on a combination of drugs to treat their diabetes (Started on Combination), a variable for the number of glycated hemoglobin tests the patient was given (Hemoglobin Tests), and a variable for the fraction of time the patient refilled a prescription to treat an acute condition within 30 days of the prescription running out (Acute Refills). The predictive equation is given by:

$$\begin{aligned}\text{Logit(Quality)} =& 1.66 - 4.23 \times \text{(Started on Combination)}\\ &+ 0.34 \times \text{(Hemoglobin Tests)} - 0.26 \times \text{(Acute Refills)}.\end{aligned}$$

This is the *log-odds*, or *logit*, of the model, and is described in more detail in Chapter 21. If we look at this equation, we can see that Started on Combination has a negative coefficient. This means that if everything else is equal, a patient who is started on a combination of drugs (versus a single drug) to treat their diabetes is more likely to have poor quality care. Similarly, Acute Refills has a negative coefficient, so a patient who is repeatedly refilling a drug to treat an acute condition is more likely to have poor quality care. On the other hand, Hemoglobin Tests has a positive coefficient, meaning that more hemoglobin tests is predictive of good quality care. This makes sense, since it is recommended that even healthy diabetics receive this test at least twice a year.

Outcomes

The accuracy of the model for the quality of healthcare is given by the classification matrix in Table 1.3. The rows are labeled by the actual classifications and the columns are labeled by the predicted classifications. Of the cases that were actually classified as poor care, 12 of them were predicted to be poor care and 10 of them were predicted to be good care. Of the cases that were actually classified as good care, 4 of them were predicted to be poor care while 75 of them were predicted to be good care. Approximately half of the poor care patients were classified correctly, and almost all of the good care patients were classified correctly. This gives an 86% success rate, compared to the baseline of 78%.

To test the model out-of-sample, Dr. Kane rated the care of 30 additional patients that were not used to build the model. The baseline success rate (by blindly rating all patients as good care) in this case was 63%. The accuracy of the model was 80%. This shows that only a simple model is needed to capture half of the cases of poor care.

This model can be used on all patients in the system without having the expert rate all of the patients and classify them accurately. As mentioned earlier, this model could also be extended to predict the quality ratings of a committee of physicians, by having them each rate the quality of care of every patient. The ultimate ratings could then be the most common rating across all physicians. This work shows that a simple model has the potential to be used to replicate the opinion of experts when assessing quality of care.

Table 1.3: The classification matrix for the logistic regression model, comparing the actual outcomes to the predicted outcomes.

	Predicted Poor	Predicted Good
Actual Poor	12	10
Actual Good	4	75

1.3 Forecasting Supreme Court Decisions

In many political decisions, predictions are abundant. In the months leading up to a presidential election, predictions are made over and over about who will win. But when it comes to the Supreme Court, predictions are not typically made. However, in 2002, Theodore Ruger, Pauline Kim, Andrew Martin, and Kevin Quinn (two political science professors and two law professors), set out to predict Supreme Court rulings. Most people thought that models did not stand a chance against expert predictions, including the authors themselves. They thought that knowledge of the details of each case and the qualitative aspects of the Court would enable experts to predict much better than any statistical model.

These professors decided to use classification trees as their statistical model, due to the need for a flexible method for pattern detection in a situation with many variables that might not have a linear relationship. For more about classification trees, see Chapter 21.

The ultimate goal was to predict, for each case argued by the Supreme Court in the October 2002 term, whether the Supreme Court Justices would affirm the case (uphold the lower court's decision) or reverse the case (overturn the lower court's decision). The predictions made by the model were only