Article from

Predictive Analytics and Futurism

June 2017
Issue 15

Using Predictive Modeling to Risk-Adjust Primary Care Panel Sizes

By Anders Larson

Most health actuaries are familiar with the concept of risk adjustment. Some of the most well-known uses in the health insurance industry include using risk scores to help determine payment rates for Medicare Advantage plans, transferring funds between commercial plans on the ACA exchanges, and adjusting capitation rates for managed Medicaid plans. It is also common for insurers, self-funded employers and providers to use risk scores to account for differences in morbidity between different populations for a variety of other purposes.

In many cases, risk adjustment models use diagnosis codes and other information from claim and enrollment data to produce risk scores that predict total costs, or at least predict a significant portion of total costs (for instance, medical or pharmacy costs only). However, risk adjustment does not necessarily need to be defined so narrowly. Depending on the intended purpose, "risk scores" are not required to be based strictly on diagnosis code information, and they are not required to predict total costs. For purposes of this article, we will define a risk score as a quantitative model that makes a prediction about health care utilization or expenditures. For some applications, it may be important for the model to make the predictions based on patient characteristics that are not controlled by the parties at financial risk (often a payer). One example of a risk score predicting something other than claims costs is the Framingham Risk Score, which predicts the 10-year cardiovascular risk of an individual, based on age, gender, cholesterol levels, smoking status and blood pressure.

This article discusses another nontraditional use of risk adjustment that incorporates modern predictive modeling techniques: risk-adjusting primary care panel sizes. We will describe the business problem, available data sources and challenges specific to this assignment, as well as the statistical techniques used to develop the risk scores.

THE BUSINESS PROBLEM

Provider reimbursement has shifted from a largely fee-for-service model in recent years to include value-based contracts between payers and providers. This paradigm shift has also extended to compensation models within provider organizations. For instance, primary care physicians are often compensated based on the number and intensity of services they provide, regardless of the number of unique patients they serve. In that case, seeing a single patient 10 times generates roughly the same income as seeing 10 patients once each. This system can create an incentive for physicians to bring patients in for more services than are necessary. In turn, this also limits the physician's ability to open the practice to additional new patients.

If the goal of the primary care organization is to provide appropriate care to the maximum number of patients, the organization needs a way to determine the appropriate number of patients for each physician (panel size). Of course, all physicians do not serve the same type of patients, and it would be unrealistic to expect all physicians to have the same panel size, even if they work the same number of hours. So what **is** the appropriate panel size for each physician?

The way we approached this problem was to develop a customized model to predict the number of primary care visits each patient should require over the next six months. The prediction was based on a wide variety of patient characteristics, including demographic information, clinical conditions and historical utilization of certain health care services, such as emergency room visits and inpatient admissions. The model did **not** base the predictions on each patient's historical office visit utilization or which physician they were assigned to. If these features were included, physicians who had been seeing their patients too frequently in the past would have their patients receive predictions that were higher than similar patients who saw other physicians. It is true that excluding these features reduced the predictive power of our model, but this was necessary to achieve the specific business objectives.

Ultimately, the predicted office visits were converted to office visit time for the physician's current patient panel, and the

predicted office visit time was compared to the physician's scheduled working hours over the next six months to determine if the physician had capacity to add new patients. The predicted office visit time for each patient could also be used to help facilitate more useful comparisons of "risk-adjusted panel sizes" between physicians. For instance, if an average patient required 30 minutes of office visit time per six months and a physician's current panel of patients was estimated to require 30,000 minutes of office visit time over the next six months, we would say this physician had (30,000 / 30) = 1,000 risk-adjusted patients. The number of risk-adjusted patients divided by the number of actual patients represented the panel's average risk score.

CHALLENGES WITH AVAILABLE DATA SOURCES

Providers, including primary care physicians, typically see patients who are insured by a variety of payers (and some patients who are uninsured). Therefore, using paid claims data from insurers, which actuaries most commonly rely on for analysis, was not a viable data source in this case. Instead, we used billing data from the provider organization, which included some of the same fields as paid claims data: service dates, provider ID, ICD diagnosis codes, CPT codes, place of service and billed charges (but not plan paid or allowed amounts). Our analysis incorporated billing data from three years for more than 200,000 patients, which allowed us to develop a very robust model.

One challenge with this data source was that there was no concept of "enrollment," which would typically exist with paid claims data. This presented two problems:

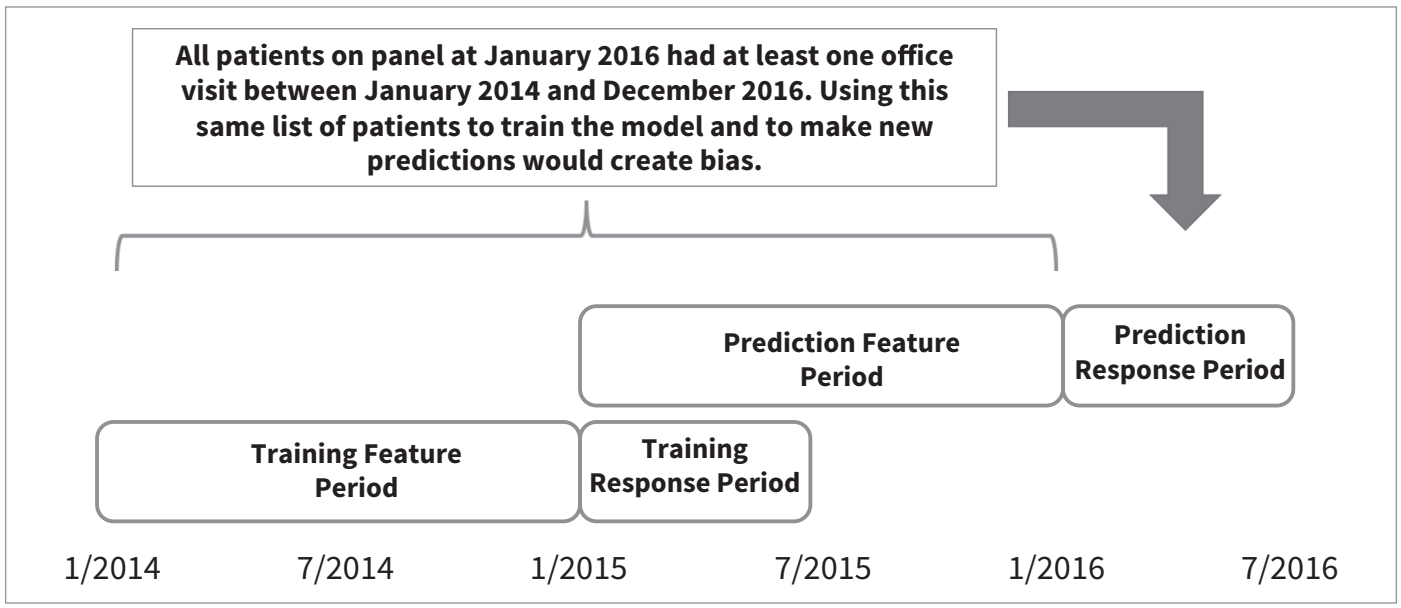
1. The data included all services that had occurred with the provider organization over a specific period, regardless of whether the patient was seeing a primary care physician within the provider organization. Selecting which patients and services should be included in our analysis was critical.

2. If a patient did not have any services over a period, there was no clear way to determine whether a patient was "eligible" for services and simply did not have any, or whether the patient was not really "eligible" to receive services. For instance, a patient who moved to the area in January 2016 would not have received any services in our data in 2015, but it would not be accurate to say this patient was not receiving any services at all in 2015.

To address the first problem, we limited our analysis to data for two sets of patients: all patients on the current primary care panel[1] and all patients on the primary care panel as of a specific date in the past. The data for the first set of patients was needed to determine the characteristics of the current panel of patients, for whom we would be making predictions. The data for the second set of patients, however, was equally critical: this would be the data used to train and calibrate our predictive model.

In predictive modeling, the data used to train the model should be a reasonable representation of the data used to make predictions. Figure 1 shows the time periods used in our analysis. In our case, we trained the model by looking at the relationship between patient characteristics in 2014 (training feature period) and utilization in the first half of 2015 (training response period).

Figure 1
Training and Prediction Periods



For this provider organization, the patients were included in the primary care panel only if they had seen this group of primary care physicians in the past two years. If we used data for the January 2016 panel of patients to train the model, we would necessarily exclude anyone who dropped off the panel in the past year. Certainly some patients on the current panel would drop off in the future, and these types of patients needed to be represented in the training data set.

It was more difficult to address the second problem (interpreting periods of inactivity). One option was to consider a person "eligible" for all months after their first observed service. Although this approach was reasonable, we were concerned that utilization rates would be distorted for patients whose first visit occurred relatively recently due to the low amount of "eligibility." In the end, we opted not to estimate periods of eligibility at all. Patients on the primary care panel were not differentiated based on the date of their first service, although we did include a binary variable indicating whether the patient was appearing on the primary care panel for the first time (i.e., their first service had been in the most recent month, since the primary care panel was updated monthly). These patients were clearly very new and might require extra office visit time in the next few months.

SELECTING THE PREDICTIVE ALGORITHM

Many popular risk-scoring algorithms are based on some type of linear model. For instance, the CMS-HCC model used in the Medicare program assigns a coefficient to each of approximately 80 conditions, and each patient's risk score can be calculated by summing the coefficients for the conditions observed for that patient, plus an additional value related to the person's age, gender and enrollment category. Although there are some exceptions, the model generally does not account for interactions between conditions or differences in how a condition might impact patients differently at different ages. For example, the value of congestive heart failure is the same for a 90-year-old male and a 65-year-old female.

While linear models have the advantage of being relatively easy to understand and interpret, they are often outperformed by other modern machine learning algorithms. In many cases, industry standards and generally accepted practices also limit the ability for many risk-scoring algorithms to use more complex models. Since this was not the case for this assignment, we were open to different approaches. We found early in our work that decision-tree-based models produced more accurate results than a generalized linear model (GLM), even when the two models used the same features. Among the reasons for this were multicollinearity between features, the large number of available features, and clear nonlinear relationships between certain variables, such as age and office visits.

Given the computing power available today, it is rare to use a single decision tree algorithm in modern predictive modeling. Instead, predictions are often derived from large numbers of decision trees, referred to as ensembles. The two most common ensemble techniques are boosting and bagging. In our case, we opted for a boosted decision tree model known as a gradient boosting machine (GBM). Although using a bagging algorithm such as a random forest would have likely produced satisfactory results, the GBM had the advantage of being able to properly model a conditionally Poisson response variable. In our case, we were interested in predicting a count of office visits for each patient, which was commonly zero, one or two.

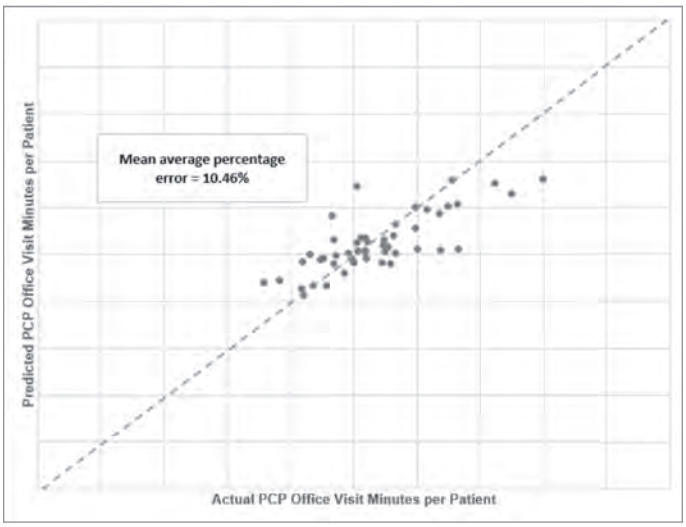
To avoid overfitting, we used a technique known as cross-validation. Cross validation involves training the model on a portion of the training data and testing the fit of the model on the remaining training data. This is repeated for other slices of the training data to get a realistic estimate of the model fit with different hyperparameters. In our case, we used 10-fold cross validation, meaning we split the training data into 10 cohorts to perform the cross validation.

Figures 2 and 3 show the model fit for the physicians with a credible number of assigned patients, both with the GLM and GBM models. The green dotted line indicates where "perfect" predictions should fall. Although the predictions are similar, the GLM model has more "big misses" where the predicted results were far from actuals, several of which can be seen on the far right of Figure 3.

The value of our model was not derived solely from its predictive accuracy. A "black box" model would be unlikely to get buy-in from physicians, regardless of how impressive the error metrics might be. We needed to provide some indication of what features were driving the results. Since decision-tree–based models do not have coefficients in the same way that linear models do, other techniques are needed for determining feature importance. In our case, we utilized a relative influence method that is based on how much each feature reduced the Poisson loss function. One way of interpreting this metric is that it indicates how much predictive power would be lost by removing each feature.

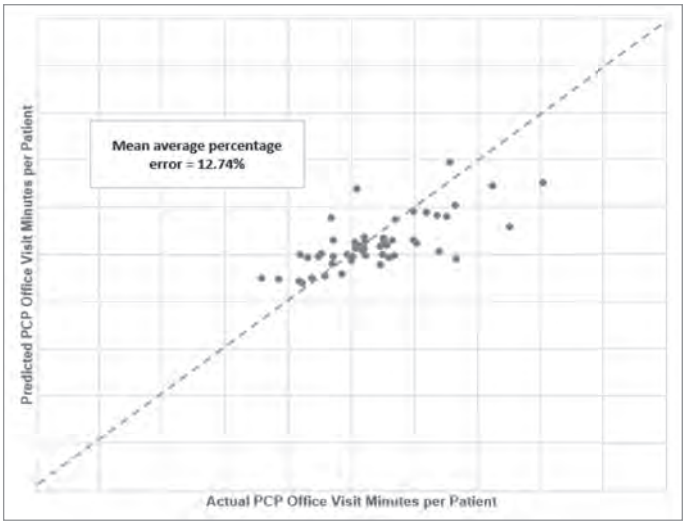
We also removed many features that appeared to have low relative influence. We found that instead of using a list of approximately 120 clinical conditions as features, we could achieve almost identical predictive accuracy by using only eight specific conditions, plus a simple count of the number of other conditions. Limiting the number of features allowed us to communicate our results more easily to physicians, who could verify whether the relationships identified by the model were intuitive.

Figure 2
GBM Model Fit



Note: Mean average percentage error was calculated including all PCPs, including those not shown in Figure 2.

Figure 3
GLM Model Fit



Note: Mean average percentage error was calculated including all PCPs, including those not shown in Figure 3.

CONCLUSION

There is no one-size-fits-all solution to risk adjustment. As the health care delivery system continues to evolve, the applications of risk adjustment are likely to evolve as well. The concept of risk adjustment can be applied to specific types of services and can be used to achieve a variety of business objectives. However, more innovative or nontraditional uses of risk adjustment sometimes require models that are customized for the particular situation. That may mean applying modern machine learning algorithms, as we did in this case, but that is not always required. If a simpler model is able to achieve a similar level of predictive accuracy, there may not be a need to use a more complex model. Even with a simpler model, however, care must be taken to calibrate the model on a data set that appropriately reflects the data that will be used to develop predictions in the future and to take steps to ensure the model is not overfitting the calibration data. In many cases, this is the most challenging and crucial part of the process.

Despite the challenges (or perhaps because of the challenges), actuaries with a combination of health care subject matter expertise and strong predictive modeling abilities are well positioned to be leaders with risk adjustment. ■

Anders Larson, FSA, MAAA, is an actuary at Milliman in Indianapolis. He can be reached at *Anders.larson@milliman.com.*

ENDNOTES

1 The primary care panel is a list of all current patients assigned to any primary care physician in the organization. This list is updated on a regular basis to add new patients and remove patients who are no longer considered current. At the time of our analysis, the "current" panel was from January 2016.