

# Inside Text Mining

Text mining provides a powerful diagnosis of hospital quality rankings.

By Patricia Cerrito, Ph.D.

Physicians don't always have an opportunity to fully document every patient risk factor when updating medical charts. Even when they do, errors can be made when these risk factors are later entered into hospital databases by administrative staff. While the underreporting of patient risk factors may not always directly affect a patient's medical care, it can have a serious, negative effect on hospital rankings created by industry groups such as the Joint Commission on Accreditation of Healthcare Organizations and the National Committee for Quality Assurance.

Hospitals that underreport patient risk factors will have lower predictions for patient mortality. Even if their success rates are equal to other hospitals, their rankings will be lower because their actual patient outcomes were worse than what should have been expected from the reported risk factors.

How can hospitals and other healthcare providers improve the accuracy of their reported patient risk factors? Text mining software can play a key role in helping analysts automatically deduce predicted patient risk factors by examining ICD-9 codes in patient billing data.

## Patient Risk Factors

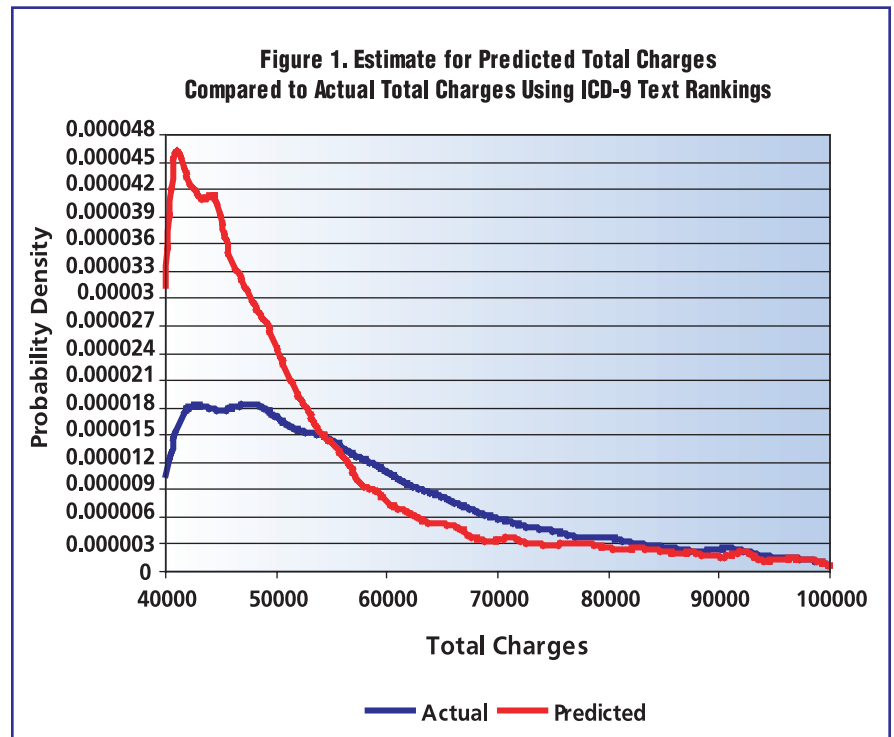
The standard procedure used by insurance providers, industry watchdog groups and professional societies to examine hospital quality and cost-effectiveness has been to compare risk-adjusted estimated length of stay, healthcare costs, mortality rates or complication rates against the actual values of these rates.

If the actual patient outcomes were better than expected, then the healthcare facility will be ranked higher. The opposite also is true: If the actual patient outcomes were worse than expected, then the healthcare facility will be ranked lower. Once a final ranking is computed, hospital administrators can gauge how their facilities are performing compared to other facilities by consulting publicly available quality rankings. Some healthcare quality organizations base their rankings on publicly available Medicare billing data, while others, including insurance companies and national societies, establish their rankings using clinical information.

To create models that predict the quality of care at any given hospital, analysts have tradition-

ally assumed that patient risk factors are uniformly entered by all clinical providers, but we know this can't possibly be true. If one hospital regularly underreports the risk factors facing its patients, then the expected patient outcomes will be based on fewer risks compared to other hospitals that overreport on risk factors. The difference between expected and actual patient outcomes will be greater compared to other hospitals, resulting in a low ranking. Therefore, the process rewards hospitals with clinicians who tend to overreport on risk factors.

One reason why it is invalid to assume the uniform entry of data is that there are differing levels of detail in the coding, but these differences are not well-defined. For example, there are 51 ICD-9 codes



related to diabetes. Consider just the following five:

- 250 diabetes mellitus without mention of complications;
- 25000 type II diabetes mellitus without mention of complications;
- 25001 type I diabetes mellitus without mention of complications;
- 25002 type II diabetes mellitus without mention of complications uncontrolled;
- 25003 type I diabetes mellitus without mention of complications uncontrolled.

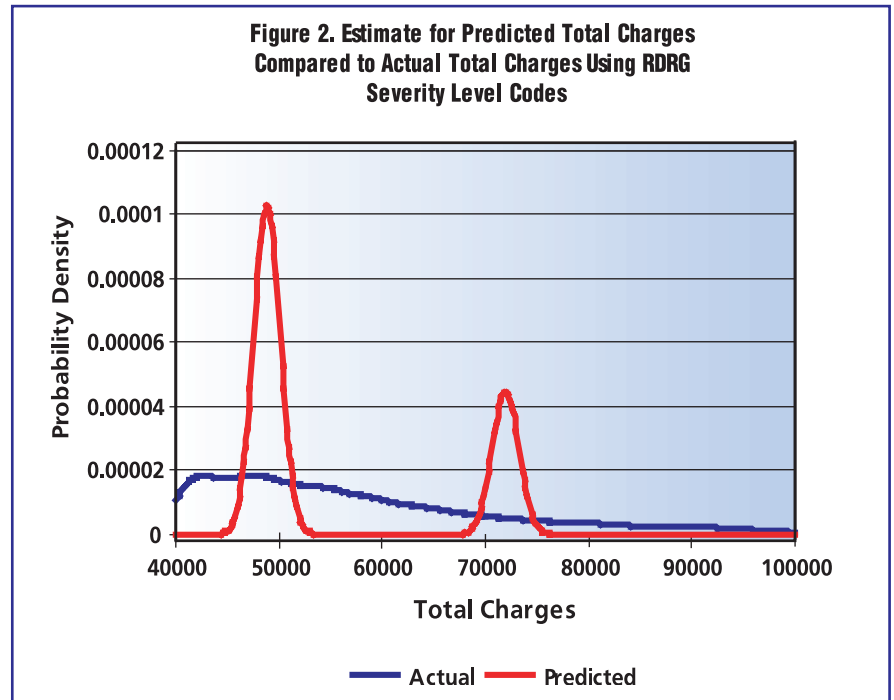
A physician has sole discretion in documenting “uncontrolled diabetes,” because there are very few guidelines that physicians may consult when making this diagnosis. Thus, a hospital where physicians liberally define “uncontrolled diabetes” will ultimately score better in those quality rankings where that value is used in their predictor models.

Once hospital analysts create their own quality predictor models, these models should be validated. Validation requires an examination of the model assumptions to determine their reliability. It also requires that consistent results are achieved when new data are entered. However, consistency can only occur if hospitals do not change reporting practices from one year to the next—and with inevitable changes in personnel, changes in reporting practices are virtually guaranteed.

We have found that a better method of validating quality predictor models is to compare the results defined using one method to results using a different method. By taking this approach, a model that utilizes text mining can be a validated means of examining healthcare quality.

## Using Text Mining

Text mining transforms unstructured data into a format that allows for many traditional data mining analytical techniques. With text



mining, analysts can examine the text found in medical reports, Web pages, research articles and billing data in many of the same ways as they can delve into structured data like age, gender, weight, blood pressure and cholesterol level.

Through text mining, every word in a text document is pre-processed and categorized through a variety of techniques: stemming, creating multiword tokens and feature extraction, to name a few. Meaningless words are then tossed out, and, finally, traditional data mining techniques such as clustering, neural networks, decision trees and regression can be applied to analyze large volumes of text.

Text mining goes far beyond the capabilities of a simple text search engine, though. For instance, text mining places synonymous words like “teach,” “instruct,” “educate” and “train” in the same category for later analysis. That’s why text mining is a novel approach for analyzing ICD-9 codes in medical records: It treats these codes as representations of text rather than as numerical categories. By doing so, similarities between codes can

be related to similarities in patient conditions.

A recent analysis of almost 15,000 patients from 13 different hospitals illustrates this point. One cardiac patient was coded as follows: “4271 42731 42781 4019 41401 412 2724.” These codes indicate that this patient suffered from unspecified paroxysmal tachycardia, atrial fibrillation, cardiac dysrhythmia, unspecified essential hypertension, coronary atherosclerosis, old myocardial infarction and lipid metabolism disorder. A second patient was coded “4271 412 4280 2724 4019 27800,” which represents unspecified paroxysmal tachycardia, old myocardial infarction, congestive heart failure, lipid metabolism disorder, unspecified essential hypertension and obesity.

In looking at these two examples, it is apparent that certain ICD-9 codes tend to be linked to each other from patient to patient. Diagnoses for “unspecified paroxysmal tachycardia,” “unspecified essential hypertension” and “old myocardial infarction” appear in the billing data for both these patients. It is reasonable to as-

sume that these patients shared other medical conditions as well, but it is possible that these conditions weren't recorded in their medical charts and subsequent billing data because the doctors treating these two patients employ different practices for recording diagnoses.

Analysts can determine which diagnoses are linked by looking for patterns in large volumes of text strings of ICD-9 codes. Using the SAS Text Miner software package, nine patient clusters were defined using the ICD-9 codes from the patient sample of 15,000 patients. To make an accurate comparison to standard ranking models, these nine clusters were then merged into four severity rankings, and then a second patient group of 8,000 from nine different hospitals was analyzed to examine and validate the relationship between these clusters and total charges.

Figure 1 shows that this severity ranking is very effective in estimating charges of the most costly patients. This is contrasted with Figure 2, which shows predicted costs using the more standard logistic regression model. Note that logistic regression cannot predict total costs for patients much beyond an average level.

Figure 3 shows how we classified the predicted outcomes for each of the 13 hospitals we evaluated. Hospital No. 12 reported only 26 percent of its patients in the lowest severity category, while hospital No. 11 reported 43 percent there. Similarly, hospital Nos. 1 and 3 have 24 percent of their patients in the most severe risk category.

Because hospital Nos. 1, 3, 6 and 12 reported higher percentages of their patients in the most severe risk category, they will have higher predicted values for length of stay and total charges. Con-

versely, because hospital Nos. 7 and 11 reported higher percentages of their patients in the least severe risk category, they will have low predicted values for length of stay and total charges. For hospital Nos. 7 and 11 to achieve high quality rankings, they

looked and underappreciated. Industry groups generate quality rankings to judge the success of hospitals in treating their patients, but some hospitals are ranked low only because their coding practices are not accurate and complete. Hospitals with

Figure 3. Comparison of Hospitals by Rank Category for Dataset Containing Billing Data From 13 Hospitals.

Hospital Number	Low Severity	Low/Medium Severity	Medium/High Severity	High Severity
1	31.2	17.7	27.1	24.1
2	36.6	17.5	28.6	17.3
3	28.0	22.9	25.1	24.0
4	32.9	21.8	34.3	11.0
5	33.0	20.0	33.1	13.9
6	28.1	25.4	26.1	20.4
7	34.2	23.0	33.5	9.3
8	31.1	17.4	40.9	10.6
9	32.0	18.2	34.2	10.7
10	37.0	18.2	34.2	10.7
11	43.3	24.7	22.6	9.4
12	26.2	14.5	40.0	19.3
13	35.5	20.1	32.3	12.1

will have to have experienced extremely low actual values.

Because all these hospitals serve similar patient populations, it is not likely that hospital Nos. 7 and 11 are consistently treating patients with less severe conditions. More likely, the differences in predicted outcomes can be attributed to different coding practices at these healthcare facilities. To achieve a more accurate representation of the quality of healthcare they provide, hospital Nos. 7 and 11 should make an effort to encourage their practicing physicians and clinical staff to adopt improved, more accurate methods of coding.

Hospital administrators, doctors and clinicians work very hard to provide their patients with the best possible treatment, and their work is often over-

low quality rankings should analyze their coding processes by using text mining to determine where the coding can be improved.

Text mining provides a powerful means for hospitals to examine their own coding practices. The results can be used to demonstrate to clinicians how the improved coding of diagnoses and risk factors in patient medical charts can increase the hospital's ranking in quality models and improve overall patient care.



Patricia Cerrito, Ph.D., is a professor of mathematics at the University of Louisville, Louisville, Ky.

HMT

For more information about Text Miner from SAS, [www.sas.com](http://www.sas.com)

101588US.0304