INTRODUCTION

The problem with using the diagnosis codes is that there are just too many to be able to use them all in a predictive model or regression. The requirements of a predictive model are that categorical data have just a small number of levels; this requirement will lead to the need to compress the number of levels in the variable. Therefore, thus far, there is a predetermined list of codes that count in risk adjustment, leaving many codes not included (as in the case of the Charlson Index). Otherwise, consensus panels are used to determine categories of severity, as in the case of the APRDRG Index. We have shown that in many cases, some of the omitted codes include as much, if not more, risk compared to those codes that are included; patients with the omitted conditions will be identified as less severe compared to patients with included conditions. In this chapter, we will introduce a method that can compress the diagnoses into clusters while still using all of the codes, without relying upon consensus panels. Moreover, outcomes are not used to define the severity index, so they can be used to validate the model; outcomes can then be used to consider the quality of providers.

Perhaps the major reason to use the modeling here is that the methodology described does not require that the diagnosis codes used are independent as is required for regression models; in fact, the modeling
Text Mining and Patient Severity Clusters

makes use of the dependency in the codes as defined by co-morbidities. Moreover, the methodology does not require the assumption of the uniformity of data entry. Again, the method described in this chapter has the capacity to ignore the upcoding of some healthcare providers; moreover, it can be used to identify those providers who are upcoding. Since uniformity is a false assumption that is required to use regression models, it is possible to generate a patient severity index using text mining that is superior to those in current use defined by using logistic regression models.

Text mining diagnosis codes takes advantage of the linkage across patient conditions instead of trying to force the assumption of independence. Combinations of diagnoses are used to define groups of patients. For example, patients with diabetes have a high probability of heart disease and kidney failure compared to the general population. Instead of relying on these three conditions and assuming that the general population is just as likely to acquire them in combination, text mining examines the combinations of diabetes, diabetes with kidney failure, diabetes with heart failure, and diabetes with both conditions.

BACKGROUND

Nominal data have always been difficult to use in quantitative analyses. There always has to be some way to compress text into a small number of categories. Under recent development, methods have been devised that can use grammar, syntax, and natural language to quantify information that is locked in text format. The methodologies developed are labeled under the general topic of text mining.

The process of text analysis generally involves the following steps:

1. Transpose the data so that the observational unit is the identifier and all nominal values are defined in the observational unit.
2. Tokenize the nominal data so that each nominal value is defined as one token.
3. Concatenate the nominal tokens into a text string such that there is one text string per identifier. Each text string is a collection of tokens; each token represents a noun.
4. Use text mining to cluster the text strings so that each identifier belongs to one cluster.
5. Use other statistical methods to define a natural ranking in the clusters.
6. Use the clusters defined by text mining in other statistical analyses.

The first step in analyzing text data is to define a term by document matrix. Each document forms a row of the matrix; each term forms a column. The resulting matrix will be extremely large, but very sparse, with most of the cells containing zeros. The matrix can be compressed using the technique of singular value decomposition with the matrix restricted to a maximum of N dimensions. In our example, the document consists of a text string of ICD9 codes, and each of the codes forms a column. Each cell contains a count of the number of times a code appears in a document. In this case, the document is defined as the patient’s condition, and each code defines one diagnosis of the condition. Since there are only a handful of diagnoses connected to any one patient, most of the cells will contain the value, ‘0’. Therefore, the matrix will be very large, but most of the cells will be empty.

Singular value decomposition is based upon an assignment of weights to each term in the dataset. Terms that are common and appear frequently, such as ‘of’, ‘and’, ‘the’ are given low or zero weight while terms that appear in only a handful of documents are given a high weight (entropy). Other weighting schemes take into consideration target or outcome variables (information gain, chi-square). In our