The *N*Gram CC classifier: A Novel Method of Automatically Creating CC Classifiers Based on ICD9 Groupings

Sylvia Halasz[1] PhD, Philip Brown[1], Colin Goodall[1] PhD, Dennis G. Cochrane[2] MD, John R. Allegra[2] MD, PhD,

[1]AT&T Labs Research; [2]Emergency Medical Associates of New Jersey Research Foundation

Introduction

Syndromic surveillance of emergency department (ED) visit data is often based on computer algorithms which assign patient chief complaints (CC) to syndromes. ICD9 code data may also be used to develop visit classifiers for syndromic surveillance but the ICD9 code is generally not available immediately, thus limiting its utility. However, ICD9 has the advantages that ICD9 classifiers may be created rapidly and precisely as a subset of existing ICD9 codes and that the ICD9 codes are independent of the spoken language. If a classifier based on ICD9 codes could be used to automatically create the code for a chief-complaint assignment algorithm then CC algorithms could be created and updated more rapidly and with less labor. They could also be created in multiple spoken languages. We had developed a method for doing this based on an "*n*gram" text processing program adapted from business research technology (AT&T Labs). The method applies the ICD9 classifier to a training set of ED visits for which both the CC and ICD9 code are known. A computerized method is used to automatically generate a collection of CC substrings with associated probabilities, and then generate a CC classifier program. The method includes specialized selection techniques and model pruning to automatically create a compact and efficient classifier.

Objectives

Our objective was to determine how closely the performance of an *n*gram CC classifier for the gastrointestinal (GI) syndrome matched the performance of the ICD9 classifier.

Methods

We used a computerized database of consecutive visits seen by ED physicians from 1-1-2000 to 12-31--2004 (2.7 million visits). We used as our ICD9 classifier an existing ESSENCE filter for "lower GI" modified by removing undifferentiated abdominal pain. The ICD9 classifier was applied to a training set of visits for the year 2000 to create the *n*gram based CC algorithm. We then used the *n*gram CC, and ICD9 classifiers to categorize the test set of visits (2001-2005). We generated a time series graph of the daily GI visit estimate by each of the two methods. We then analyzed the agreement between the *n*gram CC classifier and the ICD9 classifier using a correlation coefficient.

Results

Visual inspection of the time series graph (Figure 1) demonstrates that the *n*gram CC algorithm identified the same seasonal gastroenteritis peaks found by the ICD classifier. The scatter plot of *n*gram vs. ICD is shown in Fig 2 with a correlation coefficient, R = 0.9.

Conclusion

The *n*gram CC algorithm performed similarly to the ICD classifier. This approach has promise in that it may offer a complementary method to using manual and natural-language processing techniques to create CC classifiers. It has the advantages that it allows the rapid automated creation and updating of CC classifiers based on ICD9 groupings and may be independent of the spoken language or dialect.

Fig. 1 Time Series of Daily Counts for *N* gram and ICD

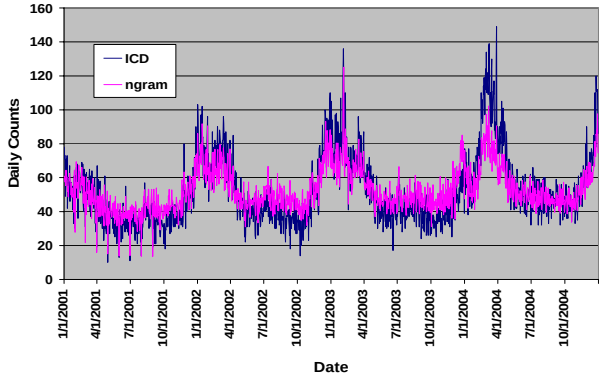

Fig. 2 Daily Counts *N* gram vs ICD