The Performance of a *N*Gram Classifier for Patients' Chief Complaint Based on a Computerized Pick List Entry and Free Text in an Italian Emergency Department

Philip Brown[1], Gemma Morabito MD[2], Sylvia Halasz PhD[1], Colin Goodall PhD[1], Dennis G Cochrane MD[3,4], Bruno Tartaglino MD[5], Mauro Giraudo[5], Olivia Cerrina[5] , John R Allegra MD, PhD[3,4]

[1]AT&T Labs - Research, [2]UOC Medicina d'Urgenza e Pronto Soccorso, Azienda Ospedaliera "Sant'Andrea" Roma, Italy, [3]Emergency Medical Associates of NJ Research Foundation, [4]Morristown Memorial Hospital Residency in Emergency Medicine, [5] Dipartimento di Emergenza e Accettazione e Medicina d'Urgenza. Azienda Ospedaliera "Santa Croce e Carle" Cuneo, Italy.

Introduction: Syndromic surveillance of emergency department (ED) visit data is often based on computer algorithms which assign patient chief complaints (CC) and ICD code data to syndromes. The triage nurse note (NN) has also been used for surveillance. Previously we developed an "*N*Gram" classifier for syndromic surveillance of emergency department (ED) (CC) in Italian for detection of natural outbreaks and bioterrorism. The classifier is developed from a set of ED visits for which both the ICD diagnosis code and CC are available by measuring the associations of text fragments within the CC (e.g. 3 characters for a "3-gram") with a syndromic group of ICD codes. We found good correlation between daily volumes by the ICD10 classifier and estimated by *N*Grams. However, because the CC was limited to 23 options based on the pick list, it might be possible to obtain results as good as the *N*Gram method or better using a simpler probabilistic approach. Also, in addition to the CC, the Italian data included a free-text NN note. We might be able achieve improved performance by applying the n-gram method to the NN or the CC supplemented by the NN.

Objectives: Our objective was to compare the performance of the *N*Gram CC classifier to two discrete classifiers based on probabilistic associations with the CC pick list items. Also, we wished to determine the performance of the *N*Gram method applied to CC alone, NN alone, and CC plus NN.

Methods: We used a computerized database of consecutive visits in an Italian ED in 2005 and 2006. We limited our study to visits which had both an entry from the 23 item CC pick list and the free text NN. These data included 21 of the 23 CC pick list categories. We used the 2005 data for training the classifiers and the 2006 data for testing the classifiers. We used an existing ESSENCE classifier for RESP as the criterion standard for training and testing. To test the probabilistic methods, we created two discrete classifiers (1) comprising 21 distinct probability estimates (Prob) (2) using recursive partitioning in R, a classification tree (RPart) that identified 3 subgroups of the 21 CC's. To compare the NGram method alone to the probabilistic methods, and to test the effect of adding NN data to the NGram method, we created three NGram based CC classifiers based on the following: CC alone, NN alone, and CC plus NN. We created 5 sets of receiver operator curves and determined the areas under the curves (AUC). We also examined the sensitivities of the NGram method with specificity set to 93%.

Results: Of the 102,215 visits in the database, 69,021 (68%) had both a CC and NN. Table 1 gives the results of the ROC analysis: area under the curve [AUC], optimal sensitivity based on least-squares distance to (1,1), and sensitivity at the (higher) specificity of 93%. Table 1 includes the three NGram and two probabilistic classifiers, Prob and RPart.

Data	**method**	**AUC**	**Sens***	**Spec***	**Sens****
CC	***N*Gram**	0.91	0.87	0.90	0.76
CC	Prob	0.92	0.88	0.88	0.77
CC	RPart	0.89	0.85	0.91	0.77
NN	***N*Gram**	0.84	0.78	0.75	0.57
CC+NN	***N*Gram**	0.92	0.89	0.89	0.77

* optimized from ROC Curve ** specificity set at 93%

Conclusion: The *N*Gram CC classifier performed successfully when applied to Italian language data, and performed similarly to simple probabilistic classifiers when applied to CC data that was based on a 23 item pick list. Adding free text nursing note data to the CC data improved the sensitivity and specificity slightly. The *N*Gram methodology is not affected by very large numbers of distinct text strings, and may have the most benefit when applied to data where both CC and NN are free text.