Sensitivity and Specificity of an *Ngram* Method for Classifying Emergency Department Visits into the Respiratory Syndrome in the Turkish Language

Philip Brown[1], Cem Oktay MD[2], Arif Alper Cevik MD[3],
Isa Kilicaslan MD[2], Colin Goodall PhD[1], Sylvia Halasz PhD[1], Dennis G Cochrane MD[4,5],
John R Allegra MD, PhD[4,5], Cemal Bilgin[1], Guy Jacobson PhD[1], Simon Tse PhD[1]
AT&T Labs – Research[1], Akdeniz Üniversitesi, Turkey[2],
Eskisehir Osmangazi Üniversitesi, Turkey[3],
Emergency Medical Associates of NJ Research Foundation[4],
Morristown Memorial Hospital Residency in Emergency Medicine[5]

Introduction: Previously we developed an “Ngram” classifier for syndromic surveillance of emergency department (ED) chief complaints (CC) in Turkish for bioterrorism. The classifier is developed from a set of ED visits for which both the ICD diagnosis code and CC are available. A computer program calculates the associations of text fragments within the CC (e.g. 3 characters for a “3-gram”) with a syndromic group of ICD codes. The program then generates an algorithm which can be deployed to evaluate chief complaint data in real-time. However, the N-gram method differs from most other classifiers in that it assigns a probability that each visit falls within the syndrome rather than ruling the visit “in” or “out” of the syndrome. It is possible to dichotomize visits “in” or “out” using N-grams by choosing a cut-off sensitivity for the n-grams used, but this affects the specificity of the method. The effect of this trade-off is best measured by a receiver-operator curve.

Objective: Our objective was to determine the sensitivity and specificity of the Ngram CC classifier for individual ED visits. We also wish to compare these results to those obtained when we substituted anglicized characters for 6 problematic Turkish characters.

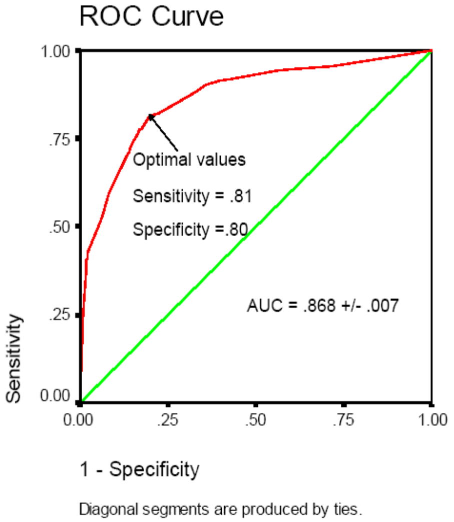
Methods: Design: Retrospective cohort. Setting: University hospital ED in Turkey. Participants: All ED visits for 2002. Protocol: We developed a respiratory (RESP) grouping of ICD10 codes chosen to be similar to the ESSENCE-CDC RESP ICD9 codes. We then used an *Ngram* method adapted from AT&T Labs applied to the first 10 months of data as a training set to create a Turkish CC RESP classifier. We next applied the classifier to a test set of visits for the last two months and generated a probability that each *Ngram* was associated with the RESP syndrome. We then generated an ROC curve using different values of the cut-off probability to dichotomize the visits in or out of the RESP syndrome. We did the ROC analysis both with the original Turkish character set and with the anglicized approximation.

Results: Table 1 summarizes the results of the ROC analysis which were identical when we substituted anglicized typed characters for the Turkish typed characters.

Table 1

AUC	Sensitivity*	Specificity*	Sensitivity**
0.87+/-0.01	0.81	0.80	0.55

* optimized from ROC analysis ** specificity set at 93%
AUC = area under the curve



Conclusion: There was no difference between the ROC results when we used Turkish characters or the anglicized alternatives. The sensitivity and specificity of an Ngram CC classifier for the RESP syndrome compares favorably to manually created CC classifiers. This approach has promise in that it may offer a complementary method to using manual and natural-language techniques to create CC classifiers. It has the advantages of rapid automated creation of CC classifiers based on ICD9 groupings and is independent of the spoken language or dialect.