Accuracy of Three Classifiers of Acute Gastrointestinal Syndrome for Syndromic Surveillance

Oleg Ivanov M.D., M.P.H., Michael M. Wagner M.D., Ph.D., Wendy W. Chapman Ph.D., Robert T. Olszewski Ph.D.
The RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh and Biomedical Security Institute, University of Pittsburgh and Carnegie Mellon University

ICD-9–coded emergency department (ED) diagnoses and free-text triage diagnoses are routinely collected data elements that have potential value for public health surveillance and early detection of epidemics.

We constructed and measured performance of three classifiers for the detection of cases of acute gastrointestinal syndrome of public health significance: one used ICD-9–coded ED diagnosis as input data; the other two used free-text triage diagnosis. We measured the performance of these classifiers against the expert classification of cases based on review of ED reports. The sensitivity of the ICD-9–code classifier was 0.32, and the specificity was 0.99. The sensitivity of a naïve Bayes classifier using triage diagnoses was 0.63, the specificity was 0.94, and the area under the ROC curve was 0.82. A bigram Bayes classifier had sensitivity 0.38, specificity 0.94, and area under the ROC of 0.69.

We conclude that a naive Bayes classifier of free-text triage diagnosis data provides more sensitive and earlier detection of cases of acute gastrointestinal syndrome than either a bigram Bayes classifier or an ICD-9 code classifier. The sensitivity achieved should be sufficient for syndromic surveillance system designed to detect moderate to large epidemics.

INTRODUCTION

Valid and reliable automatic disease classifiers are essential components of computerized early epidemic detection system. Such classifiers could be considered as objects with two primary attributes: the first attribute defined by what disease or condition a given classifier is designed to classify, the second by its input data.

Recently, to ensure timeliness of an epidemic detection and to provide extensive coverage of population, the concept of syndromic surveillance was developed. Syndrome is usually defined as a group of related symptoms or diseases. Syndromic surveillance is a practice of monitoring temporal and spatial trends of syndrome rates in population. Syndromic surveillance is considered as a complementary surveillance strategy to reportable disease surveillance.[1] In this study we measured performance of classifiers designed to detect cases of acute gastrointestinal syndrome of public health significance.

We defined *acute gastrointestinal syndrome of public health significance* to be a set of symptom complexes caused by a group of acute gastrointestinal disorders, which share common features of preventability, high morbidity and relatively low mortality. These disorders are caused by such agents as *pathogenic E. coli, Non-typhoidal Salmonellas, Vibrio cholerae, Campylobacter jejeuni, Cryptosporidium parvum, Giardia. lamblia, Shigellas, Yersinia enterocolitica, Entamoeba histolytica, Staphylococcus aureus, Bacillus cereus, Clostridium perfringens.*[2] With regard to the threat of bioterroristic attacks by agents causing acute gastrointestinal syndrome, experts from Centers for Disease Control and Prevention placed *ε-toxin of Clostridium perfringens, ricin toxin* and *Staphylococcus enterotoxin B* into Category B, which includes second highest priority potential bioterroristic agents. Category B agents are characterized by moderate ease of dissemination, moderate morbidity and low mortality and requirement for specific enhancements of diagnostic capacity and for enhanced disease surveillance.[3]

The second key attribute of classifiers is the data that they use for classification.[4] Some classifiers may not be able to handle free text, others may not be able to handle coded data. In this study we had available from the same source free text and coded data. The goal of automatic *early* epidemic detection encourages the use of data that represent the earliest electronically available medical summaries of person's health status.

As our data source we used data generated by the ED work process. Data collected were ICD-9–coded ED diagnoses and free-text triage nurse diagnoses. Currently, in the ED at the Presbyterian University Hospital (PUH ED), UPMC Health System, a triage nurse interviews each newly arriving patient and enters a triage diagnosis into the registration computer. The triage diagnosis describes the reason for the patient's visit to the emergency department. It may contain a patient's subjective description of the reason for admission or a set of medical terms used by triage nurse to describe condition of the patient on admission. Triage nurses usually limit the length of a triage diagnosis to no more than 50 characters. After the

clinical work-up, the ED physician assigns a clinical diagnosis. At the time the study was conducted, clinical ED diagnosis was manually encoded into corresponding ICD-9 code by the physician. Currently, ED diagnoses are chosen by the physician from the computer's drop-down menu. ICD-9 codes are automatically assigned to them. Triage diagnosis string and ICD-9—coded diagnosis are available electronically in real-time from the medical center's Admission Discharge Transfer system.

Because of the public health importance of timely automatic detection of acute gastrointestinal syndrome outbreaks we measured performance of three different classifiers designed to detect cases of acute gastrointestinal syndrome using the data just discussed.

METHODS

ICD-9 Classifier

A classifier for ICD-9—coded ED diagnoses was developed by two internists by reviewing all ICD-9 codes that had been used to encode diagnosis in the EDs of nine UPMC Health System hospitals during the previous three years.[5] The internists included a code in the code set of the gastrointestinal syndrome classifier if the ICD-9 code might be used for a patient presenting with gastrointestinal symptoms of interest to public health officials. The final ICD-9 code set included 16 unique ICD-9 codes (see Table 1).

Naïve Bayes Classifier

We measured performance of two Bayesian classifiers of free-text triage diagnosis data that had been built for the production Real-time Outbreak and Disease Surveillance (RODS) system.[6] Both were created by probabilistic machine-learning method.[7]

The naïve Bayes classifier assumes that words in a triage diagnosis are conditionally independent given the syndrome. It was trained from a dataset created by manual review of 16,880 unique free-text triage diagnoses, representing 46,723 triage diagnoses recorded for all visits to the ED of another hospital in Pittsburgh during January to December 2000. The length of the triage diagnosis string in the training dataset was limited to 20 characters due to truncation by the hospital information system. A given triage diagnosis was classified by one of the authors [OI], a licensed physician, as belonging to the acute gastrointestinal syndrome class if it was possible that the patient had an acute gastrointestinal syndrome of public health significance. The goal was to create sensitive classifiers; thus, there was a bias towards overinclusion. We processed the training dataset using a custom application to create a naive Bayes classifier of acute gastrointestinal syndrome.

Table 1. ICD-9 codes used in the ICD-9 classifier

ICD-9	Description
003.0	Salmonella gastroenteritis
005.9	Food poisoning, unspecified
008.5	Bacterial enteritis, unspecified
008.8	Other organism, not elsewhere classified
009.2	Infectious diarrhea
007.4	Other protozoal intestinal diseases, Cryptosporidiosis
276.0	Hyperosmolality and/or hypernatremia
276.9	Electrolyte and fluid disorders not elsewhere classified
558.9	Other and unspecified noninfectious gastroenteritis and colitis
564.89	Other functional disorders of intestine
569.69	Other colostomy/enterostomy
569.9	Unspecified disorder of intestine
579.9	Unspecified intestinal malabsorption
785.50	Shock, unspecified
787.01	Nausea with vomiting
787.91	Diarrhea

Bigram Bayes Classifier

Using the same training set, we created a bigram Bayes classifier. A bigram Bayes classifier is based on the assumption of probabilistic dependence between adjacent words. We hypothesized that its performance would be better than that of a naive Bayes classifier.

Test Set

We created a test set by expert classification of cases based on information in medical records. We drew a simple random sample of 1425 dictated reports from all ED visits in the year 2000 to the PUH ED. Each dictated report typically provided information about the reason for admission, chief complaint, clinical history, physical examination, laboratory results, progress notes, prescriptions, and diagnosis. Identifying information, such as names of patients, their physicians, relatives etc was removed from reports using De-Identifier, a tool that extracts such information from the free-text reports according to the criteria set by HIPAA (Health Insurance Portability and Accountability Act, 1996).[8]

The group of expert judges included seven physicians, among them four infectious disease fellows and three emergency medicine residents. Six experts reviewed the ED reports and the seventh expert—an infectious disease fellow—resolved cases receiving

Table 2. Contents of the questionnaire used by experts for classification of test set ED reports*

Questions	Possible answers
Diarrhea present?	Yes, No
Diarrhea duration?	Acute (<2 weeks), Chronic, Unknown
Diarrhea etiology?	Infectious, Non-infectious, Unknown
Stool culture taken?	Yes, No
Vomiting present?	Yes, No
Vomiting duration?	Acute, Chronic, Unknown
Vomiting etiology?	Infectious, Non-infectious, Unknown
Case of an acute infectious GI disorder?	Yes, No

*Some details are not included due to space limitations

conflicting judgments. We defined the following criteria for a case of acute gastrointestinal syndrome: a patient presenting with a set of clinical and laboratory findings, requiring from the emergency department physician diagnostic and therapeutic management of this patient as a case of acute infectious diarrhea *or* food poisoning *or* dysentery. This disjunction attempts to cover the diverse presentations of the agents listed above. Reports were presented to the experts in a standardized way using a custom-designed computer interface. The interface also provided a structured questionnaire (see Table 2) for the experts to record their responses. The experts received two training sessions with the computer system. To improve reliability of the experts' judgments, we used two sets of questions. The first set of questions asked whether specific gastrointestinal symptoms were present in the patient. The purpose of these questions was to help alert experts to the relevant reports. The second set confronted an expert with a decision to classify a patient as a case of acute infectious gastrointestinal disorder based on the disjunction of acute infectious diarrhea *or* food poisoning *or* dysentery. Each expert reviewed and classified 475 reports and the distribution of the reports among the experts was designed in such a way that every report was reviewed and classified two times. To minimize possible effects of expert pairs' bias, reports were assigned so that every physician classified 95 reports in common with every other physician. Interrater reliability coefficient (Cronbach's alpha) between experts was calculated using SAS® statistical package, procedure CORR .[9, 10]

Performance Measure for ICD-9 Classifier

We calculated sensitivity, specificity, positive predictive value, negative predictive value, efficiency (correct classification rate) and corresponding 95% confidence intervals for ICD-9 code classifier for all 1425 reports in the test set.[11]

One of the authors performed an error analysis of the misclassifications made by ICD-9 classifier by analyzing the reported ICD-9—coded diagnoses for the false negatives and false positives using approach described in [5]. If an error was a result of an ICD-9 code being omitted from or erroneously included into the set of ICD-9 codes used in the classifier, then we classified it as "correctable." Otherwise, we classified it as "uncorrectable."

Performance Measure for Bayesian Classifiers

Triage diagnosis strings were not available for PUH ED before May 2000. Thus, the test set for measuring performance of Bayesian classifiers included 886 records, which represented a simple random sample of all PUH ED admissions from May to December 2000. Using these 886 records, we calculated the sensitivity, specificity, positive predictive value, negative predictive value, accuracy (correct classification rate) and corresponding 95% confidence intervals of the naive and bigram Bayes classifiers. We report these metrics for one point on the ROC curve created by varying the probability threshold used for classification. The point on the ROC curve was selected so as to be representative of a reasonably sensitive and specific classifier. We also determined area under ROC curve and 95% confidence interval according to the method described by *Hanley*. [12]

Measurement of Timeliness of Data Availability

We measured minimum, maximum and average time delay in availability of free-text triage diagnosis and ICD-9—coded diagnosis for electronic processing.

RESULTS

Of 1425 patients, experts classified 22 as cases of acute gastrointestinal syndrome that satisfied the operational case definition, which corresponded to a prevalence of 0.015 (95% CI 0.01 – 0.02). Interrater reliability coefficient (Cronbach's alpha) was 0.86.

Classifier Performance

Table 3 shows the sensitivity, specificity, positive predictive value, negative predictive value, efficiency (correct classification rate), area under ROC, and 95% confidence intervals for ICD-9, naïve Bayes and bigram Bayes classifiers. Performance of ICD-9 classifier was similar to the results of a previous study

Table 3. Performance of classifiers

Metrics	Classifiers					
	ICD-9		Naïve Bayes		Bigram Bayes	
		95% CI		95% CI		95% CI
Sensitivity	0.32	0.14 – 0.54	0.63	0.35 – 0.85	0.38	0.15 – 0.65
Specificity	0.99	0.98 – 0.99	0.94	0.92 – 0.96	0.94	0.92 – 0.95
PPV	0.37	0.16 – 0.61	0.16	0.08 – 0.28	0.10	0.04 – 0.20
NPV	0.99	0.98 – 0.99	0.99	0.98 – 1.0	0.99	0.98 – 0.99
Accuracy	0.98	0.97 – 0.99	0.94	0.92 – 0.95	0.93	0.91 – 0.94
AUC	N/A	N/A	0.82	0.75 – 0.90	0.69	0.59 – 0.79

PPV-Positive predictive value, NPV-Negative predictive value, AUC-Area under the ROC curve

of ICD-9 classifier for acute respiratory syndrome—sensitivity (and positive predictive value) were lower than expected and specificity (and negative predictive value) were higher than expected given that the classifiers were designed by the expert physicians to be sensitive.[5] The Naïve Bayes classifier provides higher sensitivity, but at some loss of specificity, and a notable reduction in positive predictive value.

Error Analysis of the ICD-9 Classifier

The ICD-9 classifier produced 15 false negatives. The explanation for the low sensitivity in the previous study and this one is ICD-9 miscoding. [5] 11 false negatives were coded with ICD-9 codes not related to symptoms of gastrointestinal disorders and, consequently these codes could not be included into the classifier's code set. Another four false negatives were potentially "correctable" errors caused by the absence of codes in the ICD-9 classifier code set that potentially could have been included: 789.00 (abdominal pain-site nos), 276.5 (hypovolemia), 787.03 (vomiting alone), and 789.07 (generalized abdominal pain).

The ICD-9 classifier also produced 12 false positives, which did not satisfy operational case definition criteria. 11 false positives were assigned ICD-9 code 787.01 (nausea and vomiting) and these errors were considered "uncorrectable" - we would not want to remove this ICD-9 code from the classifier because it contributed two true positives. One report was assigned ICD-9 code 558.9 (noninfectious gastroenteritis). It is not clear what effect removal of this code would have on overall performance since it did not produce a true positive.

Timeliness of Data

Free-text triage diagnoses were available immediately on a patient's admission to the ED. The average time delay between availability of the free-text triage diagnosis and ICD-9—coded diagnosis was 6.3 hours. The maximum time delay was 14.5 hours.

DISCUSSION

The classifiers not only differed in their underlying algorithms (probabilistic versus ICD-9 sets) but also in their input data, so the experiments measured the performance of the combinations of classifier and data, not just the performance attributable to the data or classifier alone.

The naïve Bayes classifier was more sensitive than the ICD-9 classifier. Sensitivity in a case classifier is desirable for detection of epidemics. A sensitivity of 0.63 means that in the presence of an actual outbreak of acute gastrointestinal syndrome, 63% of the affected patients presenting to an ED will be identified automatically on the basis of triage diagnosis alone, which seems to be acceptable for the purpose of detection of moderate to large-scale epidemics caused by accidental or intentional contamination of food or/and water supplies.

It is important to distinguish sensitivity of a detector when measured on a single case and sensitivity of the same detector when it is used to detect an epidemic. If a detector can detect an individual case with a sensitivity of 0.63, then the probability that it will detect at least one case in an outbreak of size two is 1-(0.37*0.37)= 0.86 (one minus the probability that both cases will be missed) and the probability that it will detect at least one case in an outbreak of size four is 0.98. Of course this is an overly simplistic example because detecting a single case is usually not sufficient to say that there is an epidemic and the specificity of the detector also needs to be taken into account. Nevertheless, the point that moderate sensitivity for single case detection can still produce extremely good sensitivity for moderate to large outbreaks is still valid.

For these reasons, we think that sensitivity of 63% for individual case detection would be sufficient not only to detect large outbreaks, but also to provide reassurance that large outbreaks were not occurring. This type of reassurance in the face of hoaxes or false alarms is expected to have significant value.

The ICD-9 classifier was more specific than the naïve Bayes classifier even when the probabilistic threshold of the naïve Bayes classifier was adjusted for

optimal combination of sensitivity and specificity. In epidemic detection, higher specificity and correct classification rate allow public health officials to be more confident that when the surveillance system identifies an epidemic, this is actually an epidemic of acute gastrointestinal syndrome and not of some other condition. Such confidence would allow using rather narrow list of differential diagnoses during epidemic investigation when the primary task is to ascertain the cause of an epidemic.

An advantage of Bayesian classifiers over the ICD-9 classifiers is the ability to arbitrarily specify probability threshold for classification, thereby creating more sensitive or, conversely more specific classifiers. The fact that Bayesian classifiers use free-text input also provides significant room for improvement of their performance without trading off sensitivity against specificity. Such improvement may be achieved by addressing truncation in the training set (see description of the training data), and the use of unsupervised spell-checkers. Truncation and misspellings unnecessarily expand domain specific vocabulary thereby worsening performance of classifiers. Performance of free text classification may also be improved by exploring Bayesian classifiers that use more complex network structures that take into account semantic features of triage diagnosis strings. The performance of the bigram Bayes classifier was worse than that of the naive Bayes classifier, which we explain by the fact that bigram model is sparser than naïve model and requires larger amount of training data.

Error analysis of ICD-9 classifier performance allowed us to identify points where performance of the classifier might be improved without trying to improve the quality of the ICD-9 coding process. Such tuning is achievable by inclusion or removal of certain ICD-9 codes into or from the classifier's code set. Inclusion of ICD-9 codes causing "correctable" errors for false negatives would increase sensitivity of the ICD-9 classifier but might reduce its specificity because of unspecific nature of the symptoms that they refer to. Removal of ICD-9 codes that were responsible for the majority of false positives would increase specificity but also reduce sensitivity of the ICD-9 classifier. Therefore we concluded that the ICD-9 classifier is less amenable to performance improvement than the Bayesian classifiers.

It is interesting to consider whether ICD-9—coded diagnoses inherently contain more information than free-text triage diagnoses. Hypothetically, clinical ED diagnoses should be more accurate than free-text triage diagnoses because they are formulated by physicians. The results of this study, however, do not support this conjecture. We believe that the main reason for such discrepancy was miscoding of ED diagnosis. Recent introduction of automatic encoding of ED diagnoses may improve the coding quality in PUH ED.

The earlier availability of free-text triage diagnosis increases its value for an *early* epidemic detection system and, consequently, suggests the future research focus on improvement of free-text classification methods.

CONCLUSIONS

A naive Bayes classifier of free-text triage diagnosis data provides more sensitive and much earlier detection of cases of acute gastrointestinal syndrome than either a bigram Bayes classifier or an ICD-9 code classifier. The sensitivity achieved should be sufficient for syndromic surveillance system designed to detect moderate to large epidemics.

ACKNOWLEDGEMENTS

This work was supported by contract 290-00-0009 from the Agency for Healthcare Research and Quality and CDC grant UPO/CCU 318753-02.

REFERENCES

1. Pavlin, JA. Electronic surveillance system for the early notification of community-based epidemics (ESSENCE). Conference and workshop on syndromic and other surveillance methods for emerging infections including bioterrorism. Gaithersburg, MD, 2000.
2. Chin JE. Control of Communicable Diseases, Manual. 2000. Washington, DC: American Public Health Association.
3. http://www.bt.cdc.gov/Agent/Agentlist.asp
4. Wagner, MM. Availability and Comparative Value of Data Elements Required for an Effective Bioterrorism Detection System. AHRQ Second Interim Report, November 2001.
5. Espino, JU, Wagner, MM. Accuracy of ICD-9-coded Chief Complaints and Diagnoses for the Detection of Acute Respiratory Illness. AMIA 2001 Annual Symposium. 2001. Washington, DC.
6. Tsui, F et al. Data, Network, and Application: Technical Description of the Utah RODS Winter Olympic Biosurveillance System. Submitted to AMIA Symp. 2002
7. Mitchell, T. Machine learning. 1997. Burr Ridge, IL: McGraw Hill.
8. Cooper GF et al. An Evaluation of a Computer Program for De-Identifying Textual Patient Records. CBMI, University of Pittsburgh. September 19, 2001.
9. SAS/Stat User's Guide,Version 8, Raleigh, NC: SAS Institute, 2000
10. Friedman CP, Wyatt JC. Evaluation Methods for Medical Informatics. 1997. New York, NY: Springer-Verlag.
11. Kraemer HC. Evaluating Medical Tests. 1992. Newbury Park, CA: Sage.
12. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. 1982. Radiology 143, 29-36.