Biomedical Informatics Insights

ORIGINAL RESEARCH

Using n-Grams for Syndromic Surveillance in a Turkish Emergency Department Without English Translation: A Feasibility Study

Sylvia Halász[1], Philip Brown[2], Cem Oktay[3], Arif Alper Çevik[4], Isa Kılıçaslan[3], Colin Goodall[2], Dennis G Cochrane[5,6], Thomas R Fowler[6], Guy Jacobson[2], Simon Tse[2] and John R Allegra[5,6]

[1]AT&T Interactive, Glendale, CA, USA. [2]AT&T Labs, Research, Florham Park, NJ, USA. [3]Akdeniz Üniversitesi, Antalya, Turkey. [4]Eskisehir Osmangazi University, Eskisehir, Turkey. [5]Emergency Medical Associates of NJ Research Foundation, Livingston, NJ, USA. [6]Morristown Memorial Hospital Residency in Emergency Medicine, Morristown, NJ, USA.
Corresponding author email: tfowler2@health.usf.edu

Abstract

Introduction: Syndromic surveillance is designed for early detection of disease outbreaks. An important data source for syndromic surveillance is free-text chief complaints (CCs), which are generally recorded in the local language. For automated syndromic surveillance, CCs must be classified into predefined syndromic categories. The n-gram classifier is created by using text fragments to measure associations between chief complaints (CC) and a syndromic grouping of ICD codes.

Objectives: The objective was to create a Turkish n-gram CC classifier for the respiratory syndrome and then compare daily volumes between the n-gram CC classifier and a respiratory ICD-10 code grouping on a test set of data.

Methods: The design was a feasibility study based on retrospective cohort data. The setting was a university hospital emergency department (ED) in Turkey. Included were all ED visits in the 2002 database of this hospital. Two of the authors created a respiratory grouping of International Classification of Diseases, 10th Revision ICD-10-CM codes by consensus, chosen to be similar to a standard respiratory (RESP) grouping of ICD codes created by the Electronic Surveillance System for Early Notification of Community-based Epidemics (ESSENCE), a project of the Centers for Disease Control and Prevention. An n-gram method adapted from AT&T Labs' technologies was applied to the first 10 months of data as a training set to create a Turkish CC RESP classifier. The classifier was then tested on the subsequent 2 months of visits to generate a time series graph and determine the correlation with daily volumes measured by the CC classifier versus the RESP ICD-10 grouping.

Results: The Turkish ED database contained 30,157 visits. The correlation (R^2) of n-gram versus ICD-10 for the test set was 0.78.

Conclusion: The n-gram method automatically created a CC RESP classifier of the Turkish CCs that performed similarly to the ICD-10 RESP grouping. The n-gram technique has the advantage of systematic, consistent, and rapid deployment as well as language independence.

Keywords: disease outbreaks, epidemiology, public health, surveillance, n-gram

Biomedical Informatics Insights 2013:6 29–33

doi: 10.4137/BII.S11334

This article is available from http://www.la-press.com.

Introduction

While early detection is paramount in minimizing the effects and amount of mortality of an epidemic, such monitoring can be difficult in foreign settings.[1] For monitoring disease outbreaks in emergency departments (EDs), databases with patients' chief complaints (CC) or the International Classification of Diseases, Tenth Revision, Clinical Modification codes (ICD-10-CM) have been widely used for monitoring by public health officials.[2–7] Epidemiologists can analyze and use such data sets by recording high or low incidences of disease relative to historical data. A CC is a short description of the reason the patient is visiting the ED, frequently written in the patient's words, whereas the ICD-10-CM code is based on the physician's diagnosis. A standard procedure is to group CCs and ICD-10-CM codes into syndromes in order to simplify the process of surveillance. This way, an unusual spike in a syndrome can be detected without specific diagnoses or lab results, but merely through available CCs and ICD-10-CM codes from ED records.[8]

However, this classification process can be problematic. The traditional manual method of keeping tally sheets, though effective, can be very laborious, exemplified by the Tally Sheet system used by the Santa Clara County Public Health Department.[9] Some automated systems have been developed, such as the the natural-language CC-based New York City Department of Health and Mental Hygiene system, which matched keywords for complaints with syndromes, but these were difficult to develop and were limited in scope, as they could only be used in the target language for which they were designed.[10]

Using CCs as exclusive inputs simplifies syndromic assignment; CCs may be the only data pieces available because ICD-CM codes may not be assigned until well after the ED visit during billing processing. Ideally, integrated surveillance that utilized both CCs and ICD-CM codes, as they became available, could be utilized. We previously used a novel "n-gram classifier" method to greatly increase the efficiency and speed of ED syndromic surveillance using a computer algorithm that trained on a set of matched CCs and ICD-9-CM codes in order to generate a CC classifier.[11] Examples of n-grams for cough would include 3-gram "cou" or "ugh" and 4-gram "coug" or "ough".

Using ICD-CM is helpful in this process because it is language independent and more uniform than CCs, which may contain colloquial terms and errors. Thus CC classifiers developed from ICD codes would be much easier to create in other languages and deploy in foreign settings.

While manual data collection has been utilized in the past, an automated n-gram classifier has been described that uses data from a set of ED visits for which both ICD diagnosis code and CCs are available.[12] Text fragments (3 to 6 characters long) are found, which are then associated with syndromic ICD code groupings. This method was found to be efficient and feasible for epidemiological, large scale assessments in English[11] but theoretically would be language-independent and could be implemented anywhere, in any language, as long as its characters are able to be processed by the program.

The objective of this study was to create a Turkish n-gram CC classifier for the respiratory syndrome and then compare daily volumes between the n-gram CC classifier and RESP ICD on a test set of data.

Methods

Adapted from business research technology developed by AT&T Labs, we used an n-gram text processing program for the assignment of patient CCs to syndromes. The classifier is trained on a set of ED visits for which both the ICD diagnosis code and CC are available. A computerized method is used to determine the probability that a given text fragment within the CC is associated with a syndromic group of ICD codes. The n-Gram method has been tested with n-Grams varying from 3 to 6 characters. For example, the 3-gram classifiers for "cough" would be; "cou", "oug", "ugh". Thus we obtain a collection of CC substrings with associated probabilities, which produces our CC classifier program. The method includes selection techniques and model pruning to automatically create a compact and efficient classifier.

The study was conducted using data from all visits to a university hospital ED in Turkey during 2002. Two of the authors created a respiratory grouping of ICD-10-CM codes chosen to be similar to a standard respiratory (RESP) grouping of ICD-9 codes created by the Electronic Surveillance System for Early Notification of Community-based Epidemics (ESSENCE), a project of the United States Centers for Disease Control and Prevention.[13]

Once this ICD-10-CM classifier was defined, the new n-gram-based Turkish CC RESP classifier was created using the training set of the first 10 months of data. This was subsequently tested on the patient data from the last 2 months of ED visit data. The results were graphed as a time series to show the correlation between daily volumes of the n-gram CC classifier and RESP ICD-10 grouping. We then analyzed the agreement between the n-gram CC classifier and the ICD-10-CM classifier using a correlation coefficient.

Results

The 2002 ED database contained 30,157 visits. Figure 1 is the time series graphing of daily visit volumes for both the Turkish n-gram and the ICD-10-CM classifiers for respiratory syndrome over the last 2 months of data. Visual inspection indicates that the CC classifier peaks closely match those of the ICD-10-CM classifier. Figure 2 is a scatter plot of n-gram versus ICD-10-CM classifiers for the test set visits; the correlation coefficient is R^2 = 0.78 (y = 0.642x + 3.0498)

Discussion

As seen in Figure 1, the n-gram CC classifier identified the same peaks for the respiratory syndrome found by the ICD-10-CM classifier, with a good degree of correlation.

Natural-language CC classifying systems have been used successfully, but never outside of a familiar language. The Bayesian CC natural-language classifier was used for surveillance of free-text English CC from 10 EDs and 20 walk-in clinics in Salt Lake City during the 2002 Winter Olympics.[14] This technique was also used in Pennsylvania and Ohio to detect a group of carbon monoxide exposures within 4 hours of the presentation of the first case to an ED.[15]

The CC classifying system is difficult to use in surveillance because its free-text nature requires the ability to recognize a multitude of word fragment variations to accommodate misspelling, abbreviations, and acronyms. It is crucial that such variations are detected so that cases are not overlooked. Historically, detection of such errors has been shown to be extremely laborious and difficult.[16] Natural-language processing systems, such as the Emergency Medical Text Processor (EMT-P)[17] have been used to clean ED text for natural-language system CC classifiers. The n-gram system has a unique advantage of being independent of language or regional idiosyncrasies. Searching MEDLINE, we found only one other publication employing a similar method using Chinese data for syndromic surveillance. A novel Chinese CC classification system was used although the study used translation services as part of their methodology, and their methods were not truly language independent.[18]

An important characteristic of Center for Disease Control (CDC) guidelines for public health surveillance systems is acceptability, which they define as being “reflected by the willingness of participants and stakeholders to contribute to the data collection, analysis, and use.”[19] Most syndromic surveillance systems today require much additional effort from

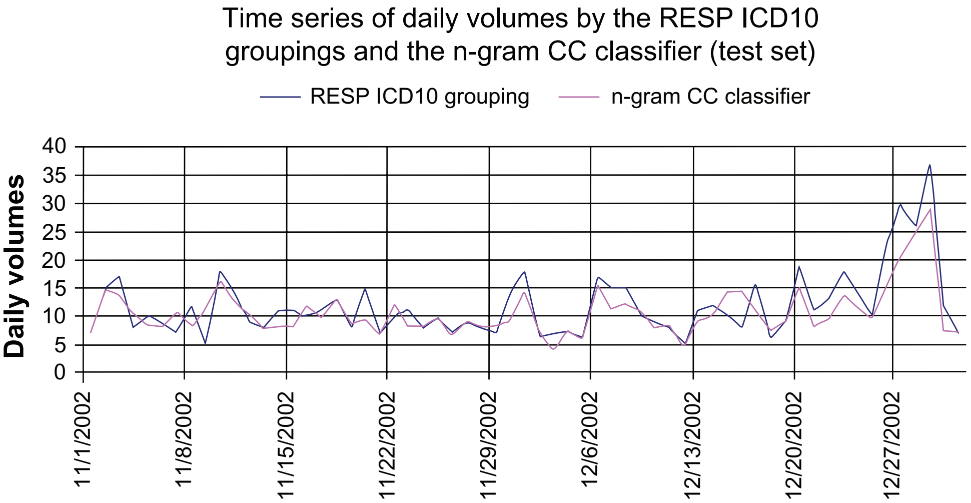

Figure 1. A comparison over time of the n-gram CC classifiers detected and the associated ICD code (test set, November 1, 2002 through December 31, 2002).

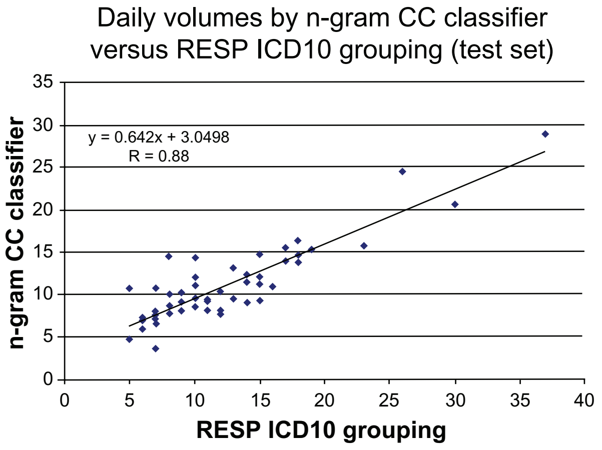

Figure 2. Daily volumes by n-gram CC and ICD10 RESP classifiers.

physicians and nurses. Common tasks needed to collect sufficient data include nurses filling out paper tally sheets, touch screen data entry by physicians, or staff completing web-based forms.[20,21] The n-gram, however, uses existing clinical data and does not necessitate additional effort by medical personnel, which makes this method significantly more feasible as far as amount of work hours required by medical personel.

The n-gram method of classification may also be expanded to classify triage notes to syndromes. Here we only captured the primary reasons of visits with CC data. Higher sensitivity may be obtained through use of triage notes in the syndrome queries, as shown in a study by Ising et al, where triage note data was used to enhance syndrome queries in the North Carolina Bioterrorism and Emerging Infection Prevention System.[22]

Limitations

One limitation of our study was that ESSENCE ICD-10-CM codes were used to create the n-grams CC classifier. Manual chart review would have been more accurate; however, review of the chart of every visit would have required too much labor for the scope of this study.

To apply this method, large training sets such as those used in this study must be available, which could prove to be problematic in emergency situations where such data may not be readily available.

This study was done in one emergency department in one region of Turkey. Using a set of n-grams across various emergency departments in different regions of Turkey would be helpful in understanding the scope of applicability for this method.

Only one syndrome was examined. Other studies examining multiple syndromes are needed.

This study was performed for one foreign language, Turkish. Further work is needed for other languages and alphabets.

We restricted our analysis to comparing daily volumes using the two classifiers as this would be most useful to epidemiologists looking for increases in volumes indicating disease outbreaks. One could also examine the agreement between the two classifiers for each visit. In this case, a receiver operator curve would be generated to determine the best cutoff to maximize sensitivity and specificity for including or excluding a visit in the syndrome.

We used visual inspection of a time series graph and the correlation coefficient of daily volumes to compare the agreement between the classifiers. There may be better ways of testing whether the n-gram CC method is useful such as determining whether "signals" for spikes in daily volumes occur on the same day.

Conclusions

The n-gram method automatically created a RESP classifier of the Turkish CCs that performed like the ICD-10-CM RESP grouping without knowledge of the Turkish language and without translating the Turkish CCs into English. This approach has great potential as it offers an additional technique beyond using manual and natural-language techniques. Its language-independent character is advantageous for rapid deployment and practical use in syndromic surveillance.

Author Contributions

Conceived and designed the experiments: SH, PB, CO, AAC, IK, CG, DGC, JRA. Analyzed the data: SH, PB, CO, AAC, IK, CG, DGC, TRF, GJ, ST, JRA. Wrote the first draft of the manuscript: SH, PB, CO, CG, DGC, TRF, GJ, ST, JRA. Contributed to the writing of the manuscript: SH, DGC, TRF, JRA. Agree with manuscript results and conclusions: SH, PB, CO, AAC, IK, CG, DGC, TRF, GJ, ST, JRA. Jointly developed the structure and arguments for the paper: SH, PB, CO, AAC, IK, CG, DGC, TRF, GJ, ST, JRA. Made critical revisions and approved final version: SH, PB, CO, AAC, IK, CG, DGC, TRF, GJ, ST, JRA. All authors reviewed and approved of the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References

1. Reis BY, Mandl KD. Syndromic surveillance: the effects of syndrome grouping on model accuracy and outbreak detection. *Ann Emerg Med*. 2004;44(3):235–41.
2. American Medical Association. *International Classification of Diseases*, *9th Revision, Clinical Modification*. Chicago, IL: American Medical Association; 2002.
3. Beitel AJ, Olson KL, Reis BY, Mandl KD. Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. *Pediatr Emerg Care*. 2004;20:355–60.
4. Espino JU, Wagner MM. Accuracy of ICD-9–coded chief complaints and diagnoses for the detection of acute respiratory illness. *Proc AMIA Symp*. 2001:164–8.
5. Ivanov O, Wagner MM, Chapman WW, Olszewski RT. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. *Proc AMIA Symp*. 2002:345–9.
6. Mocny M, Cochrane DG, Allegra JR, Nguyen T, Heffernan R. A comparison of two methods for biosurveillance of respiratory disease in the emergency department: chief complaint vs. ICD-9 diagnosis code. *Acad Emerg Med*. 2003;10:513.
7. Tsui FC, Wagner MM, Dato V, Chang CC. Value of ICD-9 coded chief complaints for detection of epidemics. *Proc AMIA Symp*. 2001:711–5.
8. Begier EM, Sockwell D, Branch LM, et al. The national capitol region's emergency department syndromic surveillance system: do chief complaint and discharge diagnosis yield difference results? *Emerg Infect Dis*. 2003;9:393–6.
9. Bravata DM, McDonald KM, Smith WM, et al. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. *Ann Intern Med*. 2004;140(11):910–22.
10. Mikosz CA, Silva J, Black S, Gibbs G, Cardenas I. Comparison of two major emergency department-based free-text chief-complaint coding systems. *MMWR Morb Mortal Wkly Rep*. 2004;53(Suppl):101–5.
11. Brown P, Halász S, Goodall C, Cochrane DG, Milano P, Allegra JR. The ngram chief complaint classifier: A novel method of automatically creating chief complaint classifiers based on international classification of diseases groupings. *J Biomed Inform.* 2010;43:268–72.
12. Brown P, Halasz S, Cochrane DG, Allegra JR, Goodall CR, Tse S. Optimizing Performance of an Ngram Method for Classifying Emergency Department Visits into the Respiratory Syndrome. *Adv Dis Surveill*. 2006; 1-30.
13. Pavlin JA, Kelley PW. *ESSENCE: Electronic Surveillance System for Early Notification of Community-based Epidemics*. Silver Spring, MD: US Department of Defense, Global Emerging Infections Surveillance and Response System; 2001.
14. Tsui FC, Espino JU, Wagner MM, et al. Data, network, and application: technical description of the Utah RODS Winter Olympic Biosurveillance System. *Proc AMIA Symp*. 2002:815–9.
15. Wagner MM, Espino J, Tsui FC, et al. Syndrome and outbreak detection using chief-complaint data—experience of the Real-Time Outbreak and Disease Surveillance project. *MMWR Morb Mortal Wkly Rep*. 2004;53(Suppl): 28–31.
16. Shapiro AR. Taming variability in free text: application to health surveillance. *MMWR Morb Mortal Wkly Rep*. 2004;53(Suppl):95–100.
17. Travers DA, Haas SW. Evaluation of emergency medical text processor, a system for cleaning chief complaint text data. *Acad Emerg Med*. 2004; 11(11):1170–6.
18. Lu H, Chen H, Zeng D, et al. Multilingual chief complaint classification for syndromic surveillance: An experiment with Chinese chief complaints. *Int J of Med Informatics*. 2009;78:308–20.
19. Guidelines for evaluating surveillance systems. *MMWR Morb Mortal Wkly Rep*. 1988;37(Suppl 5):1–18.
20. Bravata DM, Rahman MM, Luong N, Divan HA, Cody SH. A comparison of syndromic incidence data collected by triage nurses in Santa Clara County with regional infectious disease data. *J Urban Health*. 2003;80:122.
21. Zelicoff A, Brillman J, Forslund DW, et al. The Rapid Syndrome Validation Project (RSVP). Albuquerque, NM: Sandia National Laboratories; 2001.
22. Ising AI, Travers DA, MacFarquhar J, Kipp A, Waller AE. Triage note in emergency department-based syndromic surveillance. *Adv Dis Surv*. 2006;1:34.