Journal of Biomedical Informatics 43 (2010) 268–272

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The *n*gram chief complaint classifier: A novel method of automatically creating chief complaint classifiers based on international classification of diseases groupings

Philip Brown [a], Sylvia Halász [a], Colin Goodall [a], Dennis G. Cochrane [b,c], Peter Milano [b], John R. Allegra [b,c,*]

[a] *AT&T Labs – Research, Florham Park, NJ, USA*
[b] *Emergency Medical Associates of New Jersey, Livingston, NJ, USA*
[c] *Morristown Memorial Hospital Residency in Emergency Medicine, Morristown, NJ, USA*

ARTICLE INFO

Article history:
Received 13 May 2009
Available online 27 August 2009

Keywords:
Disease outbreaks
Epidemiology
Public health
Surveillance

ABSTRACT

Introduction: The *n*gram classifier is created by using text fragments to measure associations between chief complaints (CC) and a syndromic grouping of ICD-9-CM codes. *Objectives:* For gastrointestinal (GI) syndrome to determine: (1) *n*gram CC classifier sensitivity/specificity. (2) Daily volumes for *n*gram CC and ICD-9-CM classifiers. *Methods:* Design: Retrospective cohort. Setting: 19 Emergency Departments. Participants: Consecutive visits (1/1/2000–12/31/2005). Protocol: (1) Used an existing ICD-9-CM filter for "lower GI" to create the *n*gram CC classifier from a training set and then measured sensitivity/specificity in a test set using an ICD-9-CM classifier as criterion. (2) Compare daily volumes based on ICD-9-CM with that predicted by the *n*gram classifier. *Results:* For a specificity of 0.96, sensitivity was 0.70. The daily volume correlation for *n*gram vs. ICD-9-CM was $R = 0.92$. *Conclusion:* The *n*gram CC classifier performed similarly to manually developed CC classifiers and has advantages of rapid automated creation and updating, and may be used independent of language or dialect.

1. Introduction

Early detection of disease outbreaks, whether from bioterrorism or from natural or accidental causes, is important for effective treatment, containment, and minimizing morbidity and mortality [1]. Once detected, the size, spread and tempo of outbreaks can also be monitored [2].

Although many types of routinely collected clinical, administrative, pharmacy and laboratory data have been explored as possible data sets for this monitoring and detection [3], emergency department (ED) databases containing patient chief complaint (CC) and/or International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) [4] codes have proved especially promising [5–9]. These data sets are valuable resources for epidemiologists and public health officials as they are widely available and can be monitored to alert public health officials if the current incidence of disease exceeds a threshold established from historical data. A CC is usually a one-line statement about why the patient came to the emergency department. It is frequently written in the patient's own words, and is generally entered by a triage nurse or registration clerk.

The ICD-9-CM codes are based on the ED physicians' diagnoses. Frequently, these CCs and ICD-9-CM codes are grouped into syndromes for the purposes of surveillance. "Syndromes" are chosen to identify disease processes. Typical categories of syndromes include: respiratory, gastrointestinal, rash, fever, and neurological. This is done because the definitive diagnosis may rely on delayed confirmatory laboratory studies and may not be available at the time of the disease outbreak [10]. An upward trend in the volume of visits within a syndrome may give an early warning of an outbreak, and may also be used to identify an outbreak when specific diagnostic testing is not available. The assignment of an ED visit to a syndrome is based on CC and ICD-9-CM code, both of which are available from the ED record. These are not affected by laboratory culture results, possible inpatient evaluation, discharge information or follow-up data.

This syndrome classification is currently done through several methods. In some, such as the Tally Sheet System used by the Santa Clara County Public Health Department [11], data are manually collected. In others the data collection is automated. These include the ICD-9-CM based Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE) [12] which downloads ICD-9-CM code data from U.S. Department of Defense health care facilities around the world and performs daily analyses, and the natural-language CC based New York City Department of Health and Mental Hygiene system which codes complaints into syndromes on the basis of matching keywords [13]. Although automated systems have the advantage of being rapidly deployed to new data sets, natural-language systems [3] are appropriate only for the language and dialect in which they were developed. Also,

* Corresponding author. Address: 7 Valley View Dr., Montville, NJ 07045, USA. Fax: +1 973 290 7209.
E-mail address: JOHNALLE@verizon.net (J.R. Allegra).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.08.015

their development is often a long and labor-intensive process [3,14,15].

It would be useful if rapidly available CC data sets could also be used across the world, in "event-based" or "drop-in" surveillance [11]. Current systems of this nature are of limited utility as existing manual and natural-language based syndrome classifiers cannot be used across languages or dialects.

It has been shown that whenever possible, syndromic assignments should be made based on a combination of CC and ICD-9-CM codes [1]. However, while CCs are often available in "real-time", ICD-9-CM codes may not be available in real-time. In some systems such as ESSENCE, the ICD-9-CM code is assigned at the time of the patient visit, but in many systems the ICD-9-CM codes are used for billing, and are not assigned until days after the visit. Thus, there is motivation to develop syndromic assignment methods that use CCs exclusively as an input, for the times that CCs are the only data available in a timely fashion.

It has been suggested that integrated surveillance approaches be developed which incorporate different data sources as they become available [1]. In surveillance systems where both the CC and the ICD-9-CM code are collected for each visit but the ICD-9-CM code is collected later, the ICD-9-CM code may nevertheless be useful. For example, the ICD-9-CM code groupings could be used by the computer algorithm generating the CC classifier. Using that method, the CC classifier might be created and updated more rapidly and with less labor than a manual method based on the CC alone [3].

ICD-9-CM also has an advantage over CC of being independent of the spoken language, dialect or local idiosyncrasies of CC usage. Existing natural-language [3,15–17] or manual [18–20] CC classifiers are language dependent and require considerable labor to develop and maintain. CC classifiers developed automatically from the ICD-9-CM code could be more easily created in multiple spoken languages and dialects. This would be particularly useful for rapid development and deployment of a CC classifier in a "drop-in" surveillance situation.

Previously we have described in an abstract the use of an automated method for producing a CC classifier based on ICD-9-CM code groupings, called the "*n*Gram Classifier" [34]. This classifier is trained on a set of ED visits for which both the ICD-9-CM diagnosis code and CC are available by measuring the associations of text fragments within the CC with a syndromic group of ICD-9-CM codes (e.g. 4 characters for a "4-gram": "iarh", "diah" or "rrea", parts of common misspellings for diarrhea). The choice of the automated method was based on its speed and effectiveness. The *n*gram classifier creation is a heuristic approach most closely related to stepwise regression variable selection. The method requires three passes through the training set (one to generate absolute predictive values and two "pruning" passes to reduce the overall *n*gram count and to create a decision tree of relative predictive values). For the training sets used in this study (one year of ED visits—roughly half a million visits), an *n*gram classifier can usually be built in less than half an hour. Applying the classifier to a test set of equivalent size typically takes less than 10 min.

Our objectives in this study were twofold: (1) to characterize the sensitivity and specificity of an *n*gram CC classifier for a gastrointestinal (GI) syndrome through designating each CC as "in" or "out" of the syndrome by setting various thresholds to the probability that a given *n*gram is associated with the ICD-9-CM classifier for the GI syndrome, and (2) to determine how closely the daily volume estimates of an *n*gram CC classifier matched the ICD-9-CM classifier for a GI syndrome. We make the daily volume comparison omitting the ICD-9-CM code for undifferentiated abdominal pain to demonstrate the seasonal peaks of gastroenteritis which would otherwise be obscured by the large number of visits for undifferentiated abdominal pain. This code is assigned when the source or cause of the abdominal pain is unknown at the time of ED disposition.

2. Methods

The method developed for the assignment of patient CC to syndromes is based on an "*n*gram" text processing program adapted from business research technology (AT&T Labs). The method applies the ICD-9-CM classifier to a training set of ED visits for which both the CC and ICD-9-CM 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.

For each objective, we used a computerized database of consecutive visits seen by ED physicians in 19 emergency departments in New Jersey and New York from 1/1/2000 to 12/31/2005 (approximately 3.5 million visits).

For the first objective, to characterize the sensitivity and specificity of an *n*gram CC classifier for a GI syndrome, we used an existing ESSENCE syndromic grouping of ICD-9-CM codes as our ICD-9-CM classifier [12]. Prevalence in the test population was 13.7%.

The ICD-9-CM classifier was applied to a training set (approximately half a million visits for the year 2004) to create the *n*gram based CC classifier. This generated classifier, limited to 4-grams that appeared in at least 100 CCs, contained 83 *n*grams (or CC substrings) and their associated probabilities. We found that classifiers based on 4-grams performed the best. We then used the *n*gram CC and ICD-9-CM classifiers to categorize individual visits from the test set (approximately half a million visits for the year 2005). We were able to characterize each visit as "in" or "out" of the syndrome by setting different cutoffs, or thresholds, to the probability of the *n*gram being associated with the ICD-9-CM classifier. We then determined the sensitivity and specificity for each threshold using the ICD-9-CM classifier as the criterion standard, creating a receiver-operating characteristics curve for the method based on these different thresholds.

For the second objective, we used a modified version of the ESSENCE ICD-9-CM classifier for "lower GI". To highlight seasonal variations, we removed the ICD-9-CM code corresponding to "undifferentiated abdominal pain" from the "lower GI" syndrome and extended the test set over a five-year period. Prevalence in the test population was 3.4%.

The ICD-9-CM classifier was applied to a training set (approximately half a million visits for the year 2000) to create a new *n*gram based CC classifier. We chose the year 2000 as the training set for the second objective, because it was the first year of the dataset, and we wished to examine seasonal variations in subsequent years. This generated classifier contained 17 CC substrings and their associated probabilities. We then used this *n*gram CC and ICD-9-CM classifier to categorize visits from the test set (approximately three million visits for the years 2001 through 2005). We obtained the daily predicted volumes for *n*gram CC classifier by calculating the probability that each visit belonged to the syndromic group. We then added up all the probabilities for all visits for the day to determine the predicted volume. The daily volumes for the ICD-9-CM classifier were obtained simply by categorizing each visit as "in" or "out" of the syndrome. We generated a time series graph of the daily visit volume estimates for each of the two methods. We then analyzed the agreement between the *n*gram CC classifier and the ICD-9-CM classifier using a correlation coefficient. We compared daily visit volumes for these classifiers rather than visit-by-visit agreement, as this is what would be more useful to epidemiologists looking for disease outbreaks.

We received IRB approval for this study.