Emergency Medical Text Classifier: New system improves processing and classification of triage notes

Stephanie W. Haas[1], Debbie Travers[2,3], Anna Waller[3], Deepika Mahalingam[2], John Crouch[3], Todd A. Schwartz[2,4], Javed Mostafa[1,3]

1. School of Information and Library Science, University of North Carolina at Chapel Hill, NC
2. School of Nursing, University of North Carolina at Chapel Hill, NC
3. School of Medicine, University of North Carolina at Chapel Hill, NC
4. School of Public Health, University of North Carolina at Chapel Hill, NC

Abstract

Objective: Automated syndrome classification aims to aid near real-time syndromic surveillance to serve as an early warning system for disease outbreaks, using Emergency Department (ED) data. We present a system that improves the automatic classification of an ED record with triage note into one or more syndrome categories using the vector space model coupled with a 'learning' module that employs a pseudo-relevance feedback mechanism.

Materials and Methods: Terms from standard syndrome definitions are used to construct an initial reference dictionary for generating the syndrome and triage note vectors. Based on cosine similarity between the vectors, each record is classified into a syndrome category. We then take terms from the top-ranked records that belong to the syndrome of interest as feedback. These terms are added to the reference dictionary and the process is repeated to determine the final classification. The system was tested on two different datasets for each of three syndromes: Gastro-Intestinal (GI), Respiratory (Resp) and Fever-Rash (FR). Performance was measured in terms of sensitivity (Se) and specificity (Sp).

Results: The use of relevance feedback produced high values of sensitivity and specificity for all three syndromes in both test sets: GI: 90% and 71%, Resp: 97% and 73%, FR: 100% and 87%, respectively, in test set 1, and GI: 88% and 69%, Resp: 87% and 61%, FR: 97% and 71%, respectively, in test set 2.

Conclusions: The new system for pre-processing and syndromic classification of ED records with triage notes achieved improvements in Se and Sp. Our results also demonstrate that the system can be tuned to achieve different levels of performance based on user requirements.

Keywords: Disease outbreaks, electronic health records/classification, machine learning, natural language processing, public health informatics, public health surveillance/methods

Abbreviations: Centers for Disease Control and Prevention (CDC), Chief Complaint (CC), Emergency Department (ED), Emergency Medical Text Classifier (EMT-C), Emergency Medical Text Processor (EMT-P), Fever-Rash (FR), Gastro-Intestinal (GI), Master Term List (MTL), North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT), Respiratory (Resp), Sensitivity (Se), Specificity (Sp), Support Vector Machines (SVM), Structured Query Language (SQL), Triage Note (TN), term frequency-inverse document frequency (tf-idf), Unified Medical Language System (UMLS)

Correspondence: shaas@email.unc.edu

DOI: 10.5210/ojphi.v6i2.5469

Introduction and Background

The primary purpose of public health syndromic surveillance is to serve as an early warning system for disease outbreaks. In addition, syndromic surveillance systems provide situational awareness, allowing public health officials to monitor ongoing disease events, and identify related factors [1]. Syndromic surveillance can be framed as a classification problem: given syndrome definitions and information from the visit record, determine whether the record is likely to represent an instance of one, more than one, or none of the syndromes of interest. The costs of incorrect classification include unnecessary work for the public health official if a record is incorrectly classified as a syndrome instance requiring investigation or other action (a false positive), and not recognizing a potential outbreak in a timely manner if it is not classified (a false negative). Syndrome definitions include a text description (e.g., “acute infectious gastrointestinal illness of 7 days or less”) and a list of symptoms that are associated with the syndrome (e.g., diarrhea, fever). Therefore, a successful syndromic surveillance system must map symptoms as they are expressed in the visit record against those included in the syndrome definition, and determine if classification as an instance of the syndrome is warranted.

Early detection of outbreaks by syndromic surveillance systems depends on community-wide health-related data which are: 1) available in a timely manner and 2) quickly and accurately classified. Syndromes are typically created by local and state public health jurisdictions who may make use of bioterrorism syndromes defined by Centers for Disease Control and Prevention (CDC) (e.g., botulism-like, hemorrhagic illness) as well as other syndromes of interest (e.g., influenza-like illness, gastrointestinal) defined by the syndromic surveillance community through a consensus process [2,3]. Traditional surveillance systems use health-related data such as lab reports and final diagnoses to confirm the presence of diseases that meet case definitions such as *influenza* or *MRSA* [1,4]. In contrast, syndromic surveillance systems rely on symptom data recorded during Emergency Department (ED) patient visits. ED records contain timely clinical data which have been shown to act as an early warning system [5]. Although the diagnosis data and lab results from the ED record would provide the most accurate information for surveillance, these data elements are not available in a timely manner. In a study of diagnosis data available in a statewide public health surveillance system, the majority of ED diagnoses were not available for days to weeks after the ED visit [6]. For syndromic surveillance, "timeliness" is defined in terms of hours, and action by public health officials may be warranted before there is a definitive test result or diagnosis. Pre-diagnostic data used for surveillance include initial vital signs (e.g., measured temperature, heart rate) and chief complaint (CC) [7-9]. Researchers have also explored the addition of different portions of the electronic medical record to capture data in addition to chief complaint, such as discharge prescriptions, diagnostic test orders, structured clinical notes, and triage nurses' notes in narrative form [10-13].

Non-coded clinical data in unstructured (free) text form provide rich information for syndromic surveillance, but working with these data can be challenging. The triage note (TN) and chief complaint (CC) fields of an ED record capture the very initial interactions of clinicians with a patient. The CC describes the primary reason for the patient's ED visit in a few words or medical terms, while the TN, when present, contains a narrative with more detail about the history of the present illness and sometimes includes the nurse’s observations. Both fields may contain terms that represent syndrome-related symptoms, and can thus be used as evidence for the classification decision [14]. Because the CC is usually briefer and more focused, it is generally easier to automatically extract symptoms from the CC than from the TN. The TN is more wide ranging in the information it includes, as shown in Table 1. All three ED records have a CC of *fever*. When the triage note is added, each record meets a different syndrome definition, based on the highlighted keywords. Many syndromic surveillance systems use only the CC for syndromic classification, but the addition of data from clinical notes has been shown to increase the accuracy of syndromic surveillance [10]. In a pilot study, the addition of information from TNs led to improved sensitivity for three syndromes, Respiratory, Fever-Rash and Gastrointestinal, from 17% - 40% to over 81%, while specificity was maintained above 80% [11]. However, use of TNs can also increase the risk of false classification, since they can include syndrome-relevant information in non-standard terms, such as abbreviations, misspellings, negated forms, and other expressions characteristic of free-text notes [15].

Table 1: Examples of chief complaints and triage notes

Chief Complaint	Triage Note	Syndrome
Fever	**diarhea** then **febrile** x2d now **vom**	Gastrointestinal
Fever	Pt c/o wheezy **cough**, chest/throat sore, felt **hot**, w/**shaking chills**. Worse in last 48 hrs.	Respiratory
Fever	Mom reports child awoke w/fine **red rash** on trunk.. Lethargic, temp up to 103.2.Pt listless.	Fever-Rash

The development of syndromic surveillance systems using textual data is based on two major research areas: text mining of medical data to identify relevant concepts, and automated classification based on textual features. Meystre et al. [16] discuss some established techniques for information extraction from clinical unstructured text, such as incorporating supplemental information sources (e.g., the UMLS), mapping terms to a standard representation (e.g., ICD-9-CM codes), and cleaning and normalizing text using pre-processing software (e.g., Chief Complaint Processor (CCP), Emergency Medical Text Processor (EMT-P), NegEx, ConText) [3,7,17-27]. Although structured data would be easier to work with, fields such as the triage note are usually recorded in unstructured text. A system that is unable to extract concepts from these fields risks missing valuable information.

The North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT) is a state public health surveillance system which receives twice daily feeds of ED visit data. During the period covering the data used in this study, the number of hospitals providing data grew from 94 of 113 (83%) in 2006 to 110 of 112 (98%) by the beginning of 2009 [6,28]. All records contain the patient's chief complaint and, if available, also include initial vital signs and triage notes. Once received, CC text is cleaned and normalized using Emergency Medical Text processor (EMT-P) [23] to correct misspellings, expand abbreviations, map synonyms to preferred terms, etc. Negated concepts are identified using NegEx [17]. After this pre-processing, NC DETECT classifies the records as positive or negative for each of several syndromes. The syndrome definitions are represented as Structured Query Language (SQL) queries; classification is thus operationalized as a query to the NC DETECT database of pre-processed ED visit records. Approximately 13,000 new ED records are processed each day.

This translational research project created a system that can be deployed for near-real time syndromic surveillance. The system employs techniques drawn from natural language processing, text mining, and automatic classification, to extract symptoms from triage notes. The goal was to further increase sensitivity above that found in the pilot study [11], while not sacrificing specificity, in order to reduce manual intervention to identify false positives and increase the overall reliability of the system. NegEx [24] has been shown to help reduce false positives by identifying negated terms and concepts in unstructured text including TNs [29], but we found that it is not possible to compile a comprehensive list of all such terms. With our new system we sought to identify new syndrome-relevant terms automatically from a training set to produce a more robust classification that could be used by surveillance systems.

Previous Research

At the time the study data were collected, only 27% of all visits in the NC DETECT database contain a TN; however that percentage has been growing and the increase is expected to continue as more EDs capture the TN electronically. Previously, researchers found that the sensitivity for acute respiratory surveillance improved from 13% without a TN to 35% with a TN [14] and that a disproportionate percentage (46%) of visits that are flagged as positive for one or more syndromic surveillance reports include a TN [11]. Although the additional terms found in the TN may improve the system's recognition of syndrome-positive records, they also increase the possibility of false positives.

NC DETECT currently uses a rule-based classification approach which is triggered by the presence of specific terms or concepts. An advantage of this approach is that the rule is a relatively straight-forward translation of the expert knowledge in the syndrome definition into the query. For example, the GI symptom "vomiting" is implemented as the inclusion of "vomiting" as a term in the SQL query. However, such term-by-term transfer is insufficient: synonyms and other ways of expressing a symptom must also be included; the query must also include "emesis", "threw up" and so on. This type of rule maintenance must be performed by hand [14,30,31].

Table 2 shows results of testing NC DETECT queries against a manually classified sample [14,32].These results illustrate the challenge of syndromic surveillance: previous methods tend to generate low sensitivity and high specificity. Increasing sensitivity improves case detection, but is usually accompanied by a decrease in specificity, resulting in an increased burden for public health staff who must review positive signals.

Table 2: Baseline NC DETECT Performance

	Gastrointestinal [32]	**Respiratory** [14]	**Fever-Rash** [32]
# records manually classified (weighted to total # records)	3353 (2,418,168)	3699 (956,015)	3640 (2,418,168)
Sensitivity:	0.28	0.23	0.45
Specificity:	0.97	0.99	0.99

Classifiers can be developed by training machine learning algorithms, such as decision trees, support vector machines (SVM), and Bayesian classifiers (usually coupled with a weighting scheme), to identify patterns of relevant features that predict membership of a text file or patient case to a class [8,31,33]. One advantage of machine learning is that the system may learn patterns that are not obvious to a person. Features used by the classifier may include terms or sets of related terms (e.g., synonyms), and the features can be weighted since the presence of some features may provide stronger evidence than the presence of others. In machine learning, the classification model built during training is used to predict the correct classification for each incoming record. For syndromic surveillance, an accurate model will correctly identify records that are instances of a syndrome, based on the information available in the record, while minimizing false positives and false negatives. Selection of useful features to form the model is critical; preprocessing CC and TN improves the quality of features available for the model. Previous efforts on term expansion in medical information retrieval systems have used the Unified Medical Language System (UMLS) as a knowledge source to add only terms relevant to the context [34,35].

Relevance feedback is a strategy used to improve information retrieval performance by adding to and/or re-weighting the initial query [36-38]. The user identifies relevant documents returned by the initial query, which the system then uses as a source for additional query terms (e.g., synonyms) or as a basis for re-weighting existing query terms [39]. In machine learning, this technique may help identify new features or re-weight existing features by incorporating new information from the input stream or the incoming document set [39]. Fully automated systems use blind feedback, also known as pseudo-relevance feedback [34,40], which eliminates the need for human intervention. The system described here uses pseudo-relevance feedback to identify new features from classified ED records.

Study Objective

The objective of this study was to develop and test an automatic system for syndrome classification. Goals for the new system were to draw information from TNs as well as the CC, be easily adaptable to include new syndromes, and respond to changes in existing syndrome definitions and new ED data. Also, the new system should improve upon the performance of the baseline SQL queries currently in use by NC DETECT.

Methods

Our new system for processing and classifying ED records with triage notes was developed using syndrome definitions established by the CDC and expert consensus [2,3] as the basis for the initial query, making it comparable to the existing query or rule-based approach. The pseudo-relevance technique was based on the assumption that the top-ranked documents (i.e., ED records) returned by the initial query are relevant, and thus can serve as a source for additional terms. All terms were used as features in our vector space model for syndrome classification [41]. Our previous work [41] describes the basic vector space model, along with initial results obtained for GI syndrome. The version of the system described in this paper incorporates additional sources of features, including syndrome exclusion criteria and the UMLS.

We evaluated the new system in our informatics laboratory using data from the NC DETECT warehouse for ED visits from 2006-2008. This version was tested on 3 syndromes, Gastrointestinal Severe (GI), Respiratory (RESP), and Fever-Rash (FR), as defined by the NC DETECT Syndrome Definitions Workgroup.

System Components

Figure 1 illustrates the architecture of our new system, called Emergency Medical Text Classifier (EMT-C).

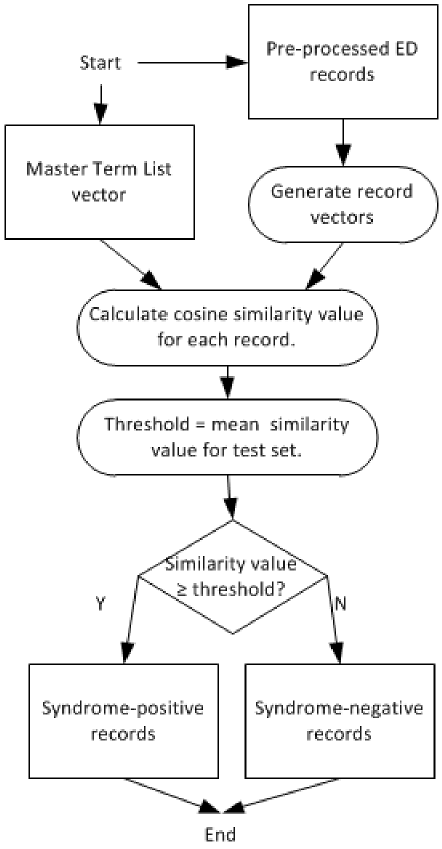
Start
Pre-processed ED records
Master Term List vector
Generate record vectors
Calculate cosine similarity value for each record.
Threshold = mean similarity value for test set.
Similarity value ≥ threshold?
Y
N
Syndrome-positive records
Syndrome-negative records
End

Figure 1. EMT-C Architecture

There are two inputs to the system. The first is a vector representing the Master Term List (MTL) for a syndrome; we describe its construction in the next section. The second is a set of pre-processed ED records which have been normalized using EMT-P modules [22,23]. These modules normalize terms commonly found in CC and TN (e.g., acronyms, abbreviations, truncations, misspellings and coordinate structures) to standardized terms from the UMLS. Note that although NC DETECT receives structured data from other fields of the ED record, EMT-C uses data only from the CC and TN.) A record's normalized terms are then translated to a binary vector, where 1 represents a term in the MTL also present in the record, and 0 represents a MTL term that is absent from the record. The record's vector is compared to the MTL vector using cosine similarity.

A mean similarity value is calculated from similarity values for a set of records, such as all records submitted to NC DETECT during a 24-hour period, and is used as the threshold for classifying the records in that set. Cosine similarity between a TN vector (v1) and syndrome vector (v2) is defined as:

$$\cos(\Theta) = \frac{\mathrm{dot}(v1, v2)}{\mathrm{norm}(v1) * \mathrm{norm}(v2)}$$

Records with a similarity value greater or equal to the mean are classified as syndrome positive, and those whose similarity value is below the threshold are classified as syndrome negative. This design decision is based on the assumption that syndromic surveillance is based on identifying surges in the number of records fitting a syndrome definition. The threshold is set to a lower similarity value if most records in the batch are dissimilar to the MTL; conversely if most records are highly similar, a higher threshold is needed to identify a surge.

Creating the Master Term List

The Master Term List (MTL) represents the syndrome definition to which each ED record is compared. The initial MTL consisted only of terms in the existing NC DETECT baseline SQL query. This list is pre-processed using the same EMT-P modules used for ED records, and represented as a binary vector. This vector contains only 1s, as all MTL terms are present in the MTL vector.

Training Set

A training set of ED records was pre-processed with EMT-P modules, and then translated into vectors for comparison with the MTL vector as shown in Fig. 2. The training set consisted of 259,365 ED records that had been classified as positive by the baseline NC DETECT SQL queries for one or more of three syndromes (Gastrointestinal, Respiratory, Fever-Rash) over a period of three years (2006-08).

Based on the assumption that the *n* records with the highest similarity values are true positives for the syndrome, terms from these pre-processed records not already in the MTL are added to it. This forms the pseudo-relevance feedback loop. The augmented MTL is translated into a new vector, as are the records, and their vectors are compared again. The feedback loop could iterate multiple times; as new records from the training set rise into the top *n* they can be harvested for new terms, (although over-fitting would be a risk if more than a couple of iterations were run). For this experiment we iterated once and performance improved. Upon a second iteration, performance degraded, so in the production version of EMT-C, the system iterated only once. Any future revision of the MTL would be triggered by degradation in performance.

The pseudo-feedback approach allows EMT-C to automatically incorporate terms from actual records that are not in the syndrome definition, thus expanding the evidence it uses for classification. For example, after the initial classification for the Respiratory syndrome, high-ranking records contained the terms "crackle" and "albuterol", which were not in the initial MTL list. "Crackle" describes abnormal lung sounds, and "albuterol" is a medication used to treat respiratory conditions. These terms were then added to the MTL list.

Evaluation

We tested EMT-C on two sets of ED records from NC DETECT that were previously manually annotated for syndromic surveillance research [14,22]. Each record contained the TN, CC and vital signs from the ED visit. A set of 485 was used for initial pilot testing of EMT-C [22]. The system was modified based on these results before final testing on a set of 3,053 [14]. The TN terms from this set were weighted by their term frequency-inverse document frequency (tf-idf) values, a standard information retrieval method used to identify terms that provide good discrimination between relevant and non-relevant documents [42]. A subset of the highly weighted terms was used to construct the master term list for classifying the records in the final testing dataset.

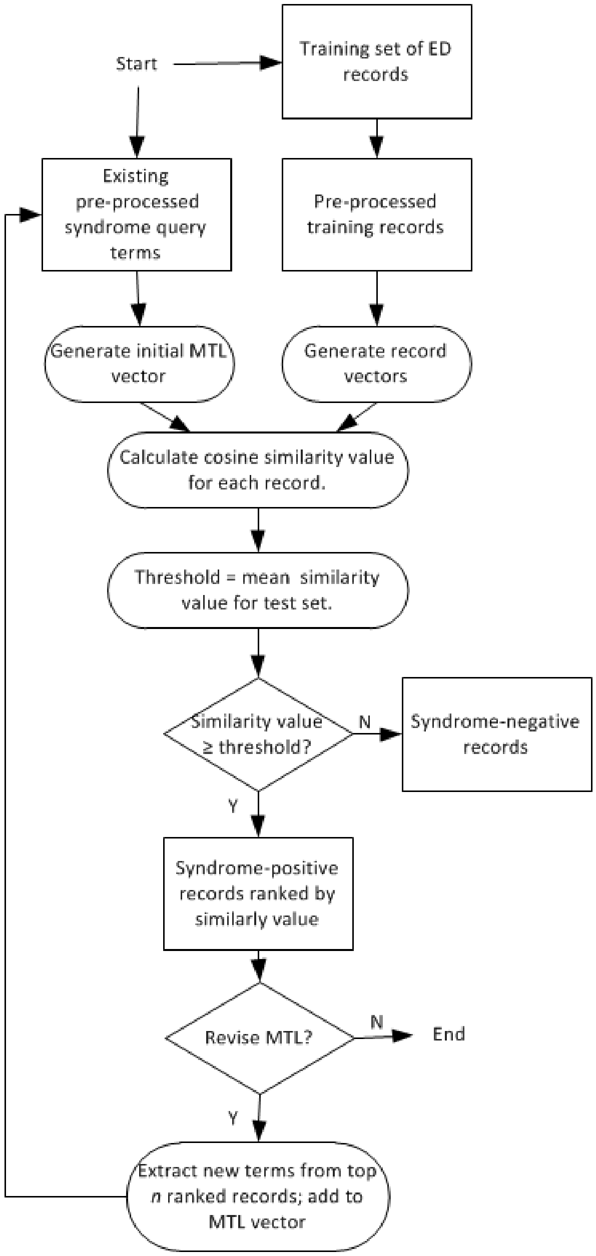
Start
Training set of ED records
Existing pre-processed syndrome query terms
Pre-processed training records
Generate initial MTL vector
Generate record vectors
Calculate cosine similarity value for each record.
Threshold = mean similarity value for test set.
Similarity value ≥ threshold?
N
Syndrome-negative records
Y
Syndrome-positive records ranked by similarly value
Revise MTL?
N
End
Y
Extract new terms from top n ranked records; add to MTL vector

Figure 2. Creation of Master Term List (MTL) Vector

*Gold Standard Set:*_Records in both test sets [14,22] had been manually classified by three clinicians as part of the previous studies. The clinical experts used the entire ED record (CC, TN, final diagnosis, measured temperature, admitted v. discharged) to retrospectively determine whether the visit conformed to one of the NC DETECT syndrome definitions (Gastrointestinal Severe, Respiratory, Fever-Rash). Though the 3 classes are not mutually exclusive in the production environment, the samples we used for this study contained records that were positive for no more than one syndrome. The initial agreement between two of the clinicians was measured with the kappa statistic and for the two studies was 0.76 and 0.82 respectively [14,22]. A third clinical expert adjudicated the cases with disagreement, and these final judgments were treated as the gold standard classifications for this study. Table 3 lists the number of positive and negative records for each syndrome in both test sets. TNs in the EMT-C pilot dataset contain an average of 22 words (128 characters) per TN. TNs in the final testing dataset contain an average of 25 words (151 characters) per TN.

Table 3: Distribution of records in pilot and final test sets

Syndrome	Pilot study dataset N=485		Final test dataset N=3053	
	Positive (%)	**Negative (%)**	**Positive (%)**	**Negative (%)**
Gastrointestinal	80 (16.5%)	405 (83.5%)	23 (0.8%)	3030 (99.2%)
Respiratory	87 (17.9%)	398 (82.1%)	171 (5.6%)	2882 (94.4%)
Fever-Rash	5 (1.0%)	480 (99.0%)	249 (8.2%)	2804 (91.8%)

Test Plan

EMT-C performance was measured in terms of sensitivity and specificity values calculated by comparing the system output with gold standard classification. Sensitivity is defined as the ratio of correctly classified syndrome positive records to gold standard syndrome positive records, while specificity is the ratio of correctly classified syndrome negative records to gold standard syndrome negative records.

Weighted versions of sensitivity and specificity were used for the final testing dataset to reflect the unbalanced stratified sampling used to select these records and make the results generalizable to the entire set of records. Several configurations of EMT-C were tested to assess the trade-off between sensitivity and specificity values. Configurations were based on 4 variations.

1. *Augmentation of the MTL with terms from the training set*. In one condition, the MTL contained only terms drawn from the syndrome SQL definition; in the other, it was augmented using terms extracted from the training set. These added terms were those with the highest tf-idf rankings.
2. *Source of terms for record vector*. In one condition, the vector was built using terms from both the TN and CC. In the other, terms were drawn only from the TN. In some records, the two fields contain the same terms: use of both fields does not add any additional information, nor does it add any "deceptive" terms. In other records, the two fields contain different terms. As mentioned earlier, this is a double-edged sword, providing stronger evidence of a true positive, or deceptive information leading to a false positive.
3. *Use of exclusion terms*. The NC DETECT syndromes include standalone exclusion criteria for each of the syndromes [7]. For example, the gastrointestinal syndrome excludes records with “Crohn” and “irritable bowel” since these chronic conditions are associated with symptoms that are identical to many of those in the GI syndrome, which could lead to false positive classification. The experiments were run with and without the use of exclusion terms.
4. *Use of pseudo-relevance feedback*. In one condition, the MTL was augmented through the pseudo-relevance feedback method described earlier; in the other, the feedback loop was omitted.

Results

Table 4 shows the maximum results of EMT-C processing on the final test set. Included in the table are the highest levels of Sensitivity (Se) and Specificity (Sp) produced for each syndrome, along with the settings for the variations (described in the previous section) for each value. Tables 5 and 6 provide the complete results for the pilot and final test sets, respectively.

Table 4: Maximum weighted Sensitivity and Specificity values, and the variation settings for the configurations that produced them, on Final Test Set (n = 3053 weighted to 1.34 million).

	Gastrointestinal	Respiratory	Fever-Rash
Maximum Sensitivity	0.97 1. MTL terms: syndrome 2. Record terms: TN + CC 3. Exclusion terms: either 4. Relevance feedback: with	0.91 1. MTL terms: syndrome 2. Record terms: TN + CC 3. Exclusion terms: either 4. Relevance feedback: with	>0.99 1. MTL terms: either 2. Record terms: TN + CC 3. Exclusion terms: either 4. Relevance feedback: either
Maximum Specificity	0.94 1. MTL terms: syndrome 2. Record terms: TN 3. Exclusion terms: with 4. Relevance feedback: without	0.91 1. MTL terms: either 2. Record terms: TN 3. Exclusion terms: either 4. Relevance feedback: without	0.93 1. MTL terms: syndrome 2. Record terms: TN 3. Exclusion terms: either 4. Relevance feedback: without

In all configurations of EMT-C in the final test set (Table 6), sensitivity was improved over the baseline NC DETECT queries (Table 2), while specificity decreased.

Augmenting the MTL with highly ranked (tf-idf) terms from the training set (variation 1) improved sensitivity for the GI syndrome in the absence of pseudo-relevance feedback, but decreased it for Respiratory syndrome in the same configurations. The effect of augmentation on specificity was mixed. For the GI syndrome, maximum values of both Se and Sp were produced without terms from the training set (see Table 4), but it decreased specificity in the absence of pseudo-relevance feedback. For the Respiratory syndrome, it had the opposite effect.

Sensitivity was higher in almost all configurations using pseudo-relevance feedback (variation 4) than in those without feedback for the GI and Respiratory syndromes. In contrast, specificity was higher in configurations without pseudo-relevance feedback. Maximum values of Se were produced using pseudo-relevance feedback for the GI and Respiratory syndromes; but maximum values of Sp were produced without feedback for all three syndromes (see Table 4).

Although drawing MTL terms from both TN and CC (variation 2) improved sensitivity in many configurations, in others, the opposite was true. This variation decreased specificity slightly in some configurations, and seemed to have no effect in others. Maximum Se values were produced using both TN and CC. In contrast, maximum Sp values were produced using TN alone (Table 4).

Table 5: Sensitivity and specificity values for all system configurations of EMT-C on Pilot Test Set (N = 485). Maximum values for sensitivity (Se) and specificity (Sp) are bolded. Configuration settings based on 4 system variations 1) Augmentation of Master Term List, 2) Source of terms for record vector, 3) Use of exclusion terms, 4) Use of pseudo-relevance feedback.

	Gastrointestinal				**Respiratory**				**Fever-Rash**			
	3. Without Exclusion Terms											
	4. Without relevance feedback		**4. With relevance feedback**		**4. Without relevance feedback**		**4. With relevance feedback**		**4. Without relevance feedback**		**4. With relevance feedback**	
Master Term List Configuration	**Se**	**Sp**	**Se**	**Sp**	**Se**	**Sp**	**Se**	**Sp**	**Se**	**Sp**	**Se**	**Sp**
1. MTL terms: syndrome 2. Record terms: TN+CC	0.89	0.86	0.90	0.71	0.87	0.86	**0.97**	0.73	**1**	**0.87**	**1**	.086
1. MTL terms: syndrome 2. Record terms: TN	0.76	0.88	**0.96**	0.56	0.83	0.86	**0.97**	0.73	**1**	**0.87**	**1**	**0.87**
1. MTL terms: syndrome + training 2. Record terms: TN+CC	0.88	0.86	0.94	0.57	0.86	0.86	**0.97**	0.73	**1**	0.58	**1**	0.57
1. MTL terms syndrome + training 2. Record terms: TN	0.75	0.88	0.90	0.59	0.82	**0.87**	0.95	0.73	**1**	0.59	**1**	0.58
	3. With Exclusion Terms											
1. MTL terms: syndrome 2. Record terms: TN+CC	0.83	0.88	0.84	0.78	0.87	0.86	**0.97**	0.73	**1**	**0.87**	**1**	**0.87**
1. MTL terms: syndrome 2. Record terms: TN	0.71	**0.90**	0.90	.063	0.83	0.86	**0.97**	0.73	**1**	**0.87**	**1**	**0.87**
1. MTL terms: syndrome + training 2. Record terms: TN+CC	0.83	0.88	0.88	0.67	0.86	0.86	**0.97**	0.73	**1**	0.59	**1**	0.58
1. MTL terms syndrome + training 2. Record terms: TN	0.71	**0.90**	0.85	0.66	0.82	**0.87**	0.95	0.73	**1**	0.60	**1**	0.58

Table 6: Weighted sensitivity and specificity for all system configurations of EMT-C on Final Test Set (N=3053 weighted to 1.34 million). Maximum values for sensitivity (Se) and specificity (Sp) are bolded. Configuration settings based on 4 system variations 1) Augmentation of Master Term List, 2) Source of terms for record vector, 3) Use of exclusion terms, 4) Use of pseudo-relevance feedback.

	Gastrointestinal				Respiratory				Fever-Rash			
	3. Without Exclusion Terms											
	4. Without relevance feedback		4. With relevance feedback		4. Without relevance feedback		4. With relevance feedback		4. Without relevance feedback		4. With relevance feedback	
Master Term List Configuration	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp
1. MTL terms: syndrome 2. Record terms: TN+CC	0.83	0.92	**0.97**	0.76	0.65	0.89	**0.91**	0.62	**>0.99**	0.91	**>0.99**	0.69
1. MTL terms: syndrome 2. Record terms: TN	0.77	0.93	0.86	0.79	0.54	**0.91**	0.87	0.64	0.96	**0.93**	0.98	0.79
1. MTL terms: syndrome + training 2. Record terms: TN+CC	0.94	0.76	0.94	0.78	0.64	0.90	0.81	0.80	**>0.99**	0.91	**>0.99**	0.73
1. MTL terms syndrome + training 2. Record terms: TN	0.92	0.80	0.86	0.81	0.54	**0.91**	0.76	0.80	0.96	**0.93**	0.98	0.76
	3. With Exclusion Terms											
1. MTL terms: syndrome 2. Record terms: TN+CC	0.77	0.93	0.91	0.79	0.65	0.89	**0.91**	0.62	**>0.99**	0.91	**>0.99**	0.69
1. MTL terms: syndrome 2. Record terms: TN	0.72	**.094**	0.80	0.81	0.54	**0.91**	0.87	0.64	0.96	**0.93**	0.98	0.79
1. MTL terms: syndrome + training 2. Record terms: TN+CC	0.89	0.80	0.89	0.81	0.64	0.90	0.81	0.80	**>0.99**	0.91	**>0.99**	0.73
1. MTL terms syndrome + training 2. Record terms: TN	0.86	0.83	0.83	0.80	0.54	**0.91**	0.76	0.80	0.96	**0.93**	0.98	0.76

Using exclusion terms (variation 3) decreased sensitivity for the GI syndrome, and had little or no effect for the other two syndromes. There was slight improvement in specificity using exclusion terms for the GI syndrome, and again, little or no effect in the other two.

Maximum sensitivity values for GI and Respiratory were produced using pseudo-relevance feedback without augmentation from the training set. Sensitivity scores for the Fever-Rash syndrome were excellent for all configurations in the final test. This is likely due to the nature of the syndrome itself: its symptoms, and thus terms representing them, are not associated with other syndromes, and thus their presence or absence provides relatively unambiguous evidence for classification. In contrast, other symptoms such as headache and dizziness are associated with more than one syndrome, as well as with other non-syndromic conditions.

Maximum specificity values, on the other hand, occurred in configurations without pseudo-relevance feedback except for the GI syndrome when the MTL was augmented with additional terms. However, the increase was slight.

Discussion

The results indicate that EMT-C, a syndromic classification system based on the vector space model with pseudo-relevance feedback, is a promising approach to syndromic surveillance. In a laboratory setting, we achieved improvements in sensitivity with moderate decreases in specificity. Terms drawn from a training set of ED records improve performance over terms drawn only from the syndrome definitions. The unstructured text expressions found in ED CC and TN contain synonymous terms and phrases (often more colloquial in nature) for symptoms in the syndrome definitions. In addition, they may contain information related to syndromic symptoms that are not necessarily symptoms, but describe the patient's situation. The risk is that the additional terms may also be associated with false positives, thus, adding them increases the noise in the system. These findings reflect the difficulty of the surveillance task. Early warning of an outbreak is crucial to successful management, but the data that are available in near-real-time are often of mixed quality. The ED CC and TN are available in a timely fashion, and are a rich source of information, but are difficult to mine [43,44].

Our results also indicate that there may not be a single model or configuration that is ideal for all syndromes. This is apparent in comparing the Fever-Rash results with those of the other two syndromes, but there were also differences between Gastrointestinal and Respiratory. One contributing factor is the precision of symptom terms in identifying one and only one syndrome. Another factor to consider is the frequency of occurrence of a syndrome in the population (prevalence rate) [44], as well as the relative risk of false positives and false negatives in managing an outbreak. Public health officials weigh the relative risks involved in missing a case; their preferences should be reflected in system design. For a syndrome such as Fever-Rash, the risk of false negatives may be deemed greater than the burden of dealing with false positives, thus an EMT-C configuration that maximizes sensitivity may be preferred. The converse may be true for Respiratory; thus, maximizing specificity may be the goal [31,32]. These requirements highlight the advantage to a system architecture such as that used in EMT-C that supports multiple configurations.

Limitations

This study was carried out in an informatics laboratory setting using data from previous years. In the actual production environment, batched data are uploaded into NC DETECT twice daily. ED visits are only accepted for upload if they contain several key data elements, including CC, so some visits are added in subsequent batches. Also, additional data elements can be added after the initial upload. These production constraints may influence system performance.

The training set for this research consisted of ED records from 2006-2008. We do not know if the vocabulary used by triage nurses in CC and TN has changed over time. If it has, retraining would reveal any changes that were significant enough to affect syndromic classification, as well as improving system performance.

Future Work

Next steps include EMT-C testing in the production environment on current ED visit records. In addition to examining the effects of the constraints described above, we also plan to gather user input on how EMT-C performs in this environment. For example, while our approach has been to improve the ability to detect signals with the syndromic surveillance system, users may have differing needs for the sensitivity/specificity tradeoff.

Future work will also focus on leveraging the system architecture to improve system performance in several ways. One advantage of the EMT-C model is the relative ease with which it can be trained. The work presented here developed models for 3 syndromes: Gastro-intestinal, Respiratory, and Fever-Rash. We plan to expand coverage of EMT-C to include more syndromes for acute infectious diseases of interest (influenza, neurological). In principle, the EMT-C model should be applicable to other types of classification problems and to other types of unstructured text, although we have not done so at this time.

This version of EMT-C uses single terms as features, which is common in the literature (e.g., [39]). It may be that using unordered bigrams (pairs of words) as features would improve performance. This would allow the model to represent terms that commonly co-occur (e.g., “vomiting and diarrhea”) or are used as phrases in the CC and TN, such as “rash all over”.

EMT-C currently uses binary vectors to represent the MTL and the records. That is, a vector represents only the presence or absence of terms. The vector space model also supports weighted vectors, in which some terms are represented by values between 0 and 1 to indicate the strength of evidence they provide. The presence of a stronger term is then represented with a higher weight than that of a weaker term. Thus, a small number of strong terms or a larger number of weak terms are equivalent in terms of similarity value to the MTL vector. Weights can also be negative, indicating that the presence of a term is evidence *against* a syndrome classification.

We can also experiment with different ways of calculating the similarity threshold, such as averaging similarity values from a larger number of records, or setting a lower bound for the threshold. Training could also incorporate ED data produced later in the ED visit, such as diagnostic ICD-9-CM codes. Although not available for the real-time classification desired for syndromic surveillance, they may be helpful in building the MTL [4].

One purpose of this study was to investigate ways of utilizing the information provided in TNs to improve syndromic surveillance. Therefore, the records included in the test sets all included TNs. In reality, over 50% of ED records sent to NC DETECT do not include a TN at this time. We expect more EDs to include TNs in the future, following the trend we have observed over the past several years. However, as EMT-C is put into production, we will determine whether it should classify all ED records regardless of the presence or absence of the TN, or whether two processing streams should be established, i.e., EMT-C for records with the TN, and the existing NC DETECT SQL query for records without the TN.

Conclusion

We developed a new system for pre-processing and syndromic classification of ED records with triage notes that achieved improvements in sensitivity and specificity over previous results in the laboratory setting. Our work also shows that EMT-C can be tuned to achieve different levels of performance based on user requirements. After field testing, EMT-C could be incorporated into NC DETECT to augment or substitute for the SQL queries currently in use (described in section 1.1). EMT-C could also be incorporated into other surveillance systems that receive as inputs text fields containing useful information that is unstructured, uncoded, and/or not easily quantified.

Human Subjects Protections

This study was approved by the Public Health-Nursing Institutional Review Board in the Office of Human Research Ethics at the University of North Carolina at Chapel Hill.

Acknowledgements

Data for this study were provided by The NC DETECT Data Oversight Committee and NC DETECT. The NC DETECT Data Oversight Committee and NC DETECT do not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented.

Funding

Funding for this project was provided by National Library of Medicine #1G08LM009787-01A1. Additional support for this research was provided by the NC TraCS Institute #UL1TR000083.

Conflict of Interest

The authors declare that they have no conflicts of interest in the research.

References

1. Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, et al. 2001. The emerging science of very early detection of disease outbreaks. *J Public Health Manag Pract*. (6), 51-59. PubMed http://dx.doi.org/10.1097/00124784-200107060-00006

2. Centers for Disease Control and Prevention [Internet]. Syndrome definitions for diseases associated with critical bioterrorism-associated agents October 23, 2003. [updated 2003 Oct 23; cited 2013 August 21]. Available from http://bt.cdc.gov/surveillance/syndromedef/index.asp.

3. Chapman WW, Dowling JN, Baer A, Buckeridge DL, Cochrane D, et al. 2010. Developing syndrome definitions based on consensus and current use. *J Am Med Inform Assoc*. 17, 595-601. PubMed http://dx.doi.org/10.1136/jamia.2010.003210

4. Betancourt JA, Hakre S, Polyak CS, Pavlin JA. 2007. Evaluation of ICD-9 codes for syndromic surveillance in the electronic surveillance system for the early notification of community based epidemics. *Mil Med*. 172, 346-52. PubMed

5. Zheng W, Aitken R, Muscatello DJ, Churches T. 2007. Potential for early warning of viral influenza activity in the community by monitoring clinical diagnoses of influenza in hospital emergency departments. *BMC Public Health*. 7, 250-59. PubMed http://dx.doi.org/10.1186/1471-2458-7-250

6. Travers DA, Barnett C, Ising A, Waller A. Timeliness of emergency department diagnoses for syndromic surveillance. Proceedings of the American Medical Informatics Association; 2006 Nov 11-15; Washington DC; 2006. p. 769-773. PubMed Central PMCID: PMC1839358.

7. Brown P, Halasz S, Goodall C, Cochrane DG, Milano P, et al. 2010. The ngram chief complaint classifier: A novel method of automatically creating chief complaint classifiers based on international classification of disease groupings. *J Biomed Inform*. 43, 268-72. PubMed http://dx.doi.org/10.1016/j.jbi.2009.08.015

8. Olszewski RT. Bayesian classification of triage diagnoses for the early detection of epidemics. Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference; 2003 May 12-14; St. Augustine, Florida. p. 412-416. http://www.aaai.org/Papers/FLAIRS/2003/Flairs03-080.pdf

9. Conway M, Dowling JN, Chapman WW. 2013. Using chief complaints for syndromic surveillance: a review of chief complaint based classifiers in North America. *J Biomed Inform*. 46, 734-43. PubMed http://dx.doi.org/10.1016/j.jbi.2013.04.003

10. Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. 2012. Comparison of natural language processing methods for identifying influenza from encounter notes. *Ann Intern Med*. 156, 11-18. PubMed http://dx.doi.org/10.7326/0003-4819-156-1-201201030-00003

11. Ising A, Travers DA, MacFarquhar J, Kipp A, Waller A. Triage note in emergency department-based syndromic surveillance. Adv Dis Surv. 2006;1:34. http://www.isdsjournal.org/articles/235.pdf

12. DeLisle S, South B, Anthony JA, Kalp E, Gundlapallli A, et al. 2010. Combining free text and structured electronic medical record entries to detect acute respiratory Infections. *PLoS ONE*. 6, e0013377. http://www.plosone.org/article/info:doi/10.1371/journal.pone.0013377. PubMed

13. Hripcsak G, Soulakis ND, Li L, Morrison FP, Lai AM, et al. 2009. Syndromic surveillance using ambulatory electronic health records. *J Am Med Inform Assoc*. 16, 354-61. PubMed http://dx.doi.org/10.1197/jamia.M2922

14. Scholer MJ, Ghneim GS, Wu S, Westlake M, Travers DA, et al. Defining and applying a method for improving sensitivity and specificity of an emergency department early event detection system. Proceedings of the American Medical Informatics Association; 2007 Nov 10-14; Chicago, Illinois; 2007. p. 651-5. PubMed PMID: 18693917; PubMed Central PMCID: PMC2655810.

15. Travers DA, Haas SW. 2003. Using nurses natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. *J Biomed Inform*. 36, 260-70. PubMed http://dx.doi.org/10.1016/j.jbi.2003.09.007

16. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 47, 128-44. PubMed

17. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. 2001. A simple algorithm for identifying negated findings and disease in discharge summaries. *J Biomed Inform*. 34(5), 301-10. PubMed http://dx.doi.org/10.1006/jbin.2001.1029

18. Chapman WW, Dowling JN, Wagner MM. 2005. Classification of emergency department chief complaints into seven syndromes: a retrospective analysis of 527,228 patients. *Ann Emerg Med*. 46, 445-55. PubMed http://dx.doi.org/10.1016/j.annemergmed.2005.04.012

19. Chapman WW, Dowling JN. Can chief complaints identify patients with febrile syndromes? Adv Dis Surv. 2007;3(6):1-9 http://www.researchgate.net/publication/228629434_Can_chief_complaints_identify_patients_with_febrile_syndromes

20. Dara J, Dowling JN, Travers DA, Cooper GF, Chapman WW. 2008. Evaluation of preprocessing techniques for chief complaint classification. *J Biomed Inform*. 41, 613-23. PubMed http://dx.doi.org/10.1016/j.jbi.2007.11.004

21. Dara J, Dowling JN, Travers D, Cooper GF, Chapman WW. Chief complaint preprocessing evaluated on statistical and non-statistical classifiers. Adv Dis Surv. 2007;2:4. http://www.researchgate.net/publication/5682247_Evaluation_of_preprocessing_techniques_for_chief_complaint_classification

22. Travers D, Kipp A, MacFarquhar J, Waller A. 2006. Evaluation of Emergency medical Text Processor for pre-processing chief complaint data for syndromic surveillance. *Adv Dis Surv*. 1, 71.

23. Travers DA, Haas SW. 2004. Evaluation of Emergency medical text processor: a system for cleaning chief complaint data. *Acad Emerg Med*. 11, 1170-76. PubMed http://dx.doi.org/10.1111/j.1553-2712.2004.tb00701.x

24. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan B. Evaluation of negation phrases in narrative clinical reports. Proceedings of the American Medical Informatics Association; 2001 Nov 3-7; Washington, DC; 2001. p. 105-9. PubMed Central PMCID: PMC2243578.

25. Harkema H, Dowling JN, Thornblade T, Chapman WW. 2009. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform*. 42, 839-51. PubMed http://dx.doi.org/10.1016/j.jbi.2009.05.002

26. Lu HM, Zeng D, Trujillo L, Komatsu K, Chen H. 2008. Ontology-enhanced automatic chief complaint classification for syndromic surveillance. *J Biomed Inform*. 41, 340-56. PubMed http://dx.doi.org/10.1016/j.jbi.2007.08.009

27. Yan P, Chen H, Zeng D. 2008. Syndromic surveillance systems: public health and biodefense. *Annu Rev Inform Sci*. 42, 425-95. http://dx.doi.org/10.1002/aris.2008.1440420117

28. NC DETECT [Internet]. North Carolina disease event tracking and epidemiologic collection tool. [cited 2013 Aug 21]. Available from www.ncdetect.org

29. Ising A, Travers D, Crouch J, Waller A. 2007. Improving negation processing in triage notes. *Adv Dis Surv*. 4, 50.

30. Cadieux G, Buckeridge DL, Jacques A, Libman M, Dendukuru N, et al. 2011. Accuracy of syndrome definitions based on diagnoses in physician claims. *BMC Public Health*. 11, 17. PubMed http://dx.doi.org/10.1186/1471-2458-11-17

31. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. 2011. Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection. *J Am Med Inform Assoc*. 18, 631-38. PubMed http://dx.doi.org/10.1136/amiajnl-2010-000022

32. Scholer M, Travers D, Waller A, McCalla A, Wetterhall S. Optimizing syndromic classification in biosurveillance systems. Proceedings of the International Society for Disease Surveillance; 2009 Dec 2-4; Miami FL; 2009.

33. Muscatello DJ, Churches T, Kaldor J, Zheng W, Chiu C, et al. 2005. An automated, broad-based, near real-time public health surveillance system using presentations to hospital Emergency Departments in New South Wales, Australia. *BMC Public Health*. 5, 141. PubMed http://dx.doi.org/10.1186/1471-2458-5-141

34. Liu Z, Chu W. 2005. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Inform Retrieval. 2007;10:173-202. http://www.cobase.cs.ucla.edu/tech-docs/vicliu/knowledge.base.query.expansion.pdf

35. Hersh W, Price S, Donohoe L. Assessing thesaurus-based query expansion using the UMLS metathesaurus. Proceedings of the American Medical Informatics Association; 2000. p. 344-8. PubMed PMID: 11079902; PubMed Central PMCID: PMC2244120.

36. Salton G, Buckley C. 1990. Improving retrieval performance by relevance feedback. *J Am Soc Inf Sci*. 41, 288-97. http://www.umiacs.umd.edu/~jimmylin/LBSC796-INFM718R-2006-Spring/papers/Salton90.pdf. http://dx.doi.org/10.1002/(SICI)1097-4571(199006)41:4<288::AID-ASI8>3.0.CO;2-H

37. Rocchio JJ. Relevance feedback in information retrieval. In: Salton G, editor. The Smart retrieval system: experiments in automatic document processing. Englewood Cliffs, NJ: Prentice Hall; 1971. p. 313-323.

38. Hiemstra D, Robertson S. Relevance feedback for best match term weighting algorithms in information retrieval. Proceedings of the Second DELOS Network of Excellence Workshop: Personalization and Recommender Systems in Digital Libraries; 2001 Jun 18-20; Dublin City University, Ireland; 2001. http://www.ercim.eu/publication/ws-proceedings/DelNoe02/hiemstra.pdf

39. Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. J Am Med Inform Assoc. 1999;6:393-411. PubMed PMID: 10495099; PubMed Central PMCID: PMC61382.

40. Mitra M, Singhal A, Buckley C. Improving automatic query expansion. Proceedings of the 21st Annual International Association of Computing Machinery Special Interest Group of Information Retrieval; 1998 Aug 24-28; Melbourne, Australia; 1998. http://dl.acm.org/citation.cfm?doid=290941.290995

41. Mahalingam D, Mostafa J, Travers D, Haas SW, Waller A. 2012. Automated syndrome classification using early phase emergency department data. Proceedings of the 2nd Association of Computing Machinery Special Interest Group of Health Information Technology International Health Informatics Symposium; 2012 Jan 28-30; Miami, Florida, USA; 2012. p. 373-378. http://dl.acm.org/citation.cfm?doid=2110363.2110406

42. Salton G, Buckley C. 1988. Term-weighting approaches in automatic retrieval. *Inf Process Manage*. 24, 513-23. http://www.sciencedirect.com/science/article/pii/0306457388900210. http://dx.doi.org/10.1016/0306-4573(88)90021-0

43. Buckeridge D, Burkom H, Moore A, Pavlin J, Cutchis P, et al. Evaluation of syndromic surveillance systems: development of an epidemic simulation model. MMWR CDC. 2004;53:137-143. http://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a27.htm

44. Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from

the CDC Working Group. MMWR CDC. 2004;53:125-129. http://www.cdc.gov/mmwr/PDF/rr/rr5305.pdf