Evaluation of a Chief Complaint Pre-Processor for Biosurveillance

**Debbie Travers[1], PhD, RN, Shiying Wu[2], PhD, MS, Matthew Scholer[1], MD, PhD,
Matt Westlake[2], MS, Anna Waller[1], ScD, Anne-Lyne McCalla[2], MS, RD
[1]University of North Carolina, Chapel Hill, NC
[2]Research Triangle Institute, Research Triangle Park, NC**

Abstract
Emergency Department (ED) chief complaint (CC) data are key components of syndromic surveillance systems. However, it is difficult to use CC data because they are not standardized and contain varying semantic and lexical forms for the same concept. The purpose of this project was to revise a previously-developed text processor for pre-processing CC data specifically for syndromic surveillance and then evaluate it for acute respiratory illness surveillance to support decisions by public health epidemiologists. We evaluated the text processor accuracy and used the results to customize it for respiratory surveillance. We sampled 3,699 ED records from a population-based public health surveillance system. We found equal sensitivity, specificity, and positive and negative predictive value of syndrome queries of data processed through the text processor compared to a standard keyword method on raw, unprocessed data.

Introduction

Biosurveillance systems are used for early detection of disease outbreaks of public health interest. The systems monitor electronic clinical data for clusters of symptoms from patient data that may be indicative of particular syndromes, such as influenza, acute gastrointestinal illness, or fever/rash outbreaks. Many biosurveillance systems utilize ED data because the data are timely and widely available in electronic form. One data element collected in EDs that has been widely used for biosurveillance is the patient's chief complaint (CC), which is described by the Centers for Disease Control and Prevention (CDC) as "(the) patient's reason for seeking care or attention (in the ED)". Unfortunately the quality of CC data from in the U.S. is highly variable, given the lack of a standardized terminology for ED CC.[1-2] Most ED CCs are entered electronically, either as free text entries or as terms from locally-developed or vendor-supplied controlled CC lists. Both free text and controlled CC list terms contain lexical and semantic variants that have proven challenging for surveillance. Some biosurveillance systems include text variants into keyword searches, while others pre-process the data using a variety of natural language processing techniques. [1, 3-5] There is no consensus on which approach is best.

We developed the Emergency Medical Text Processor (EMT-P) to address the variability in textual CC data from EDs.[6] The system includes modules that: 1) clean CC data by replacing acronyms, abbreviations, misspellings and truncations with standard terms; 2) expand coordinate constructions and other syntactic structures; and 3) map the cleaned CCs to standard concepts from the Unified Medical Language System (UMLS). EMT-P generates cleaned CC terms that correspond to standard UMLS concept unique identifiers (CUIs). Table 1 shows EMT-P inputs and outputs examples.

Table 1- Examples of EMT-P Processing

Raw CC	Cleaned CC(s)	CUI
cp	chest pain	C0008031
chert pain	chest pain	C0008031
chest/back pain	chest pain back pain	C0008031 C0004604
chst pn/sob	chest pain shortess of breath	C0008031 C0392680

The original version of EMT-P was validated for general clinical purposes and found to be 96% accurate.[7] However it was not designed specifically for use with biosurveillance systems.

In 2005, EMT-P Version 2.1 (v.2.1) was implemented for use in cleaning CC data for the North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT), a population-based biosurveillance system for North Carolina.[8] As of March 2007, it includes timely visit data from 88% of hospital based EDs in the state. NC DETECT monitors a variety of signs and symptoms, looking for patterns that may indicate an infectious disease outbreak or bioterrorism event of public health importance. Epidemiologists use the system daily and monitor 8 syndromes that were developed locally and based on the CDC infection-related syndrome definitions.[9] The syndrome queries search for keywords in CC and (when available) other clinical data such as temperature and triage notes. The queries are designed to take abbreviations and misspellings in text data into account, as well as deal with basic negation terms.

EMT-P output is used for aggregate reports of the most common CCs for North Carolina EDs, but it has yet to be utilized for NC DETECT syndrome reports. In a pilot study of 717 ED visits, we evaluated EMT-P for pre-processing of CC data for biosurveillance and found that EMT-P v.2.1 performed as well, but not better than, keyword queries.[10]

In response to the pilot study findings, we decided to further evaluate EMT-P on a larger, more representative sample of ED records. We initiated the current project to customize EMT-P for use with respiratory surveillance. A team of clinicians and technical staff systematically modified EMT-P by addressing CC data that are most relevant for early detection of acute respiratory illness (ARI). The ARI query syndrome is designed to identify acute (<14 days) illnesses of lower respiratory tract disease such as influenza, SARS, anthrax and plague. We used a training set of NC DETECT visit data to train the revised version, EMT-P v.2.2. We then tested EMT-P v.2.2 on a testing set of NC DETECT data.

The purpose of this study was to revise EMT-P and then to evaluate the revised version for pre-processing of CC data prior to syndromic classification for acute respiratory illness.

Methods

Sampling

We began the study by drawing a sample of data from the NC DETECT static dataset, which is comprised of visits from Oct 1, 2004 through Sep 30, 2005 (in order to include an entire influenza season,). We utilized a training set to develop revisions to EMT-P for respiratory surveillance, and a test set to evaluate EMT-P v.2.2 as a pre-processor for syndromic classification of ARI.

The 2005 NC DETECT static dataset contained 1,121,691 ED visit records. After removing 1,000 pilot study cases and excluding injury related visits (ICD-9 codes 800-959), there were 956,015 visit records from which to draw the two samples for this study. The testing set was drawn first. The static dataset was stratified into four strata according to their less stringent respiratory syndrome query result (i.e., respiratory symptoms only as opposed to the standard NC DETECT acute respiratory illness query which requires both a respiratory and a constitutional symptom) and the availability of triage notes. Given our constraints on available resources, we selected the sample size and sample allocation to obtain a standard deviation of 1% for the estimated sensitivity and a standard deviation of 0.5% for the estimated specificity and prevalence. Using the estimated prevalence of ARI for the 4 strata based on a pilot study of 1,000 cases sampled from the 2005 data, we applied the algorithm developed by Chromy[11] to obtain the optimal sample size and sample allocation. A stratified random sample was drawn accordingly from the 2005 dataset to create the testing set and included 3,699 ED records. The testing set was used for the current EMT-P evaluation as well as related studies of syndromic surveillance methods.

The sample was manually reviewed by clinical experts which became the gold standard ratings for these studies. Each of the 3,699 records was reviewed by two clinicians who independently judged whether the ED visits met the NC DETECT respiratory case definition. Any discrepancies were adjudicated by a third clinical expert, a physician epidemiologist with over 22 years experience with disease surveillance. Inter-rater agreement between the initial two clinicians was measured with the kappa statistic and found to be 0.76. [12] The final gold standard sample of 3,699 records included 505 which were deemed positive and 2611 deemed negative for acute respiratory illness.

The training set was then drawn from the remaining 952,026 records left in the dataset. The goal was to take a sample of approximately 1,000 records since we only had the resources to manually examine about 1,000 CCs. The following factors were considered in designing the sample:

1. EMT-P treats the same CC exactly the same, thus we only need a sample of unique CCs.
2. A correction of a more common error would improve the performance of EMT-P more.
3. To improve the performance of EMT-P for ARI detection, we should aim to maximize sampling of true positives by focusing on the positive cases identified by the SQL query for the less stringent respiratory syndrome.
4. We should not ignore false negative cases which are likely among the less stringent respiratory syndrome negative cases.

With the above considerations, we decided to take 2/3 of the most frequent unique CCs from the less stringent respiratory positive cases so as to focus our efforts on that group, and draw the remaining 1/3 from the most frequent unique CCs among the less stringent respiratory negative cases, in an effort to detect key terms among false negatives. Prior to the sample selection, we converted all CCs to lower case and removed all unnecessary spaces (leading, trailing, multiple). We selected all the unique CCs from the less stringent respiratory positive cases with a frequency larger than 8, obtaining 643 unique CCs. We then selected all the unique CCs from the less stringent respiratory negative cases that have not been selected with a frequency larger than 82, obtaining 360 unique CCs. Thus, a total of 1,003 unique CCs were selected for the training set.

EMT-P Revisions

We customized EMT-P for acute respiratory illness surveillance by using methods from the early EMT-P development,[4] and focusing on the most common patterns of respiratory (e.g., *cough*, *SOB*) and constitutional (e.g., *fvr*, *chills*) terms from ED CCs.

Initially, the training set was processed with EMT-P v.2.1. A team of three ED clinicians then manually examined the processed CCs to determine the accuracy of EMT-P segmenting and mapping to UMLS concepts. First they evaluated the accuracy of EMT-P for segmenting CCs into separate concepts. For example, many CC entries (e.g., *fever/shortness of breath*) contain two or more concepts and are partitioned by the system into separate CC segments (e.g., segment on the slash to create two separate segments: 1- *fever* and 2- *shortness of breath*). Next, the clinician reviewers evaluated the accuracy of EMT-P for mapping cleaned CCs to standard concepts in the UMLS. Using a rating system from our previous study[5] that was based on the NLM's Large Scale Vocabulary Test, the experts rated the EMT-P mapping as: equivalent, more general, more specific, associated, non-match or incorrect match. The results of this training set analysis were included in our initiative to revise EMT-P.

We also analyzed additional data to identify areas for improvement of EMT-P, including all frequent non-matching CCs in the NC DETECT dataset for 2005-06, and controlled CC lists from those hospitals that use them to document CC for ED patients.

EMT-P Evaluation for Respiratory Surveillance

To evaluate EMT-P as a chief complaint pre-processor for syndromic surveillance, we compared the standard NC DETECT syndrome classification method (keyword queries of raw CC text) to queries of CUIs generated during pre-processing with EMT-P. The current keyword queries are written in standard query language (SQL) and search raw CC entries from ED visit records in NC DETECT. The CUI queries are also written in SQL and search EMT-P output (CUIs corresponding to cleaned CCs). Both queries identify whether the ED visit record meets the acute respiratory illness case definition.

The testing set of 3,699 NC DETECT records was used for the EMT-P evaluation. The final sample was queried twice for syndromic classification, once using the keyword query on the raw CC entries, and once using the CUI query on the CC entries after pre-processing with EMT-P. We compared the results of each query to the gold standard ratings. SAS version 9.1 (Cary, NC) was used to generate kappa statistics, sensitivity (Se), specificity (Sp), positive predictive value (PPV) and negative predictive value (NPV).

Results

EMT-P Revisions

981 (98%) of the 1,003 CCs were judged to be accurately segmented by EMT-P v.2.1 and 315 (76%) of 418 segments were accurately mapped to UMLS concepts. 18 CC segments were incorrectly mapped to the UMLS and 85 (20%) were not mapped to the UMLS. All inaccurately segmented and mapped CCs were examined to identify areas for improvement to EMT-P. The team also analyzed frequent non-matching respiratory and constitutional terms from NC DETECT chief complaints in the 2006 dataset. Based on these analyses, we revised the system to version 2.2. Examples are shown in Table 2.

Table 2- Revisions to EMT-P for Version 2.2

Update	Examples- CCs (replacements)
Add acronyms, abbreviations, truncation for expansion	v d (vomiting diarrhea)
	code 500 (assault)
	shob (shortness of breath)
Add misspellings for correction	absecss (abscess)
	pneomonia (pneumonia)
	vomioting (vomiting)
New module to map non-matching terms from controlled CC lists to terms matching UMLS concepts	pain all over (generalized aches and pain- C0281856)
	fussy/pediatric (fussy infant- C0849993)
	respiratory complaint (signs & symptoms, respiratory- C0037090)
Correct inaccurate mappings to UMLS concepts	arrest (map to C0600228- cardiopulmonary arrest instead of C0392351- law enforcement arrest)

EMT-P Evaluation for Respiratory Surveillance

The results of the raw and pre-processed queries are shown in Table 3, with statistics shown in Table 4.

Table 3- Raw versus Pre-processed CC Queries*

N=956,015- weighted
N= 3,699- actual
		Gold standard Ratings	

		Resp +	**Resp -**
Raw CC *(keyword query)*	**Resp +**	16,736 (180)	12,511 (143)
	Resp -	55,398 (325)	871,370 (3,051)
Pre-processed CC *(CUI query)*	**Resp +**	17,266 (182)	12,590 (144)
	Resp -	54,868 (323)	871,291 (3,050)

**Frequencies weighted (actual number of records reviewed are listed in parentheses)*

Table 4- Raw versus Pre-processed CC Queries

	Se (%)	**Sp (%)**	**PPV (%)**	**NPV (%)**
Raw CC *(keyword query)*	23.2	98.6	57.2	94.0
Pre-processed *CC (CUI query)*	23.9	98.6	57.8	94.1

Se= sensitivity, Sp= Specificity, PPV= positive predictive value, NPV= negative predictive value

Discussion

We found that the previous version of EMT-P accurately segmented most raw CCs. We revised it to improve mappings to UMLS concepts, and address respiratory and constitutional terms. We also added a new module to address controlled terms from some hospitals that did not match UMLS records.

We found that pre-processing with EMT-P v.2.2 performed as well as our standard method (raw CC query) for acute respiratory illness detection. In addition to correctly identifying all true positives that the text query identified, pre-processing with the EMT-P method also led to the identification of 2 cases missed by the text query due to lexical variation in the CC data. We analyzed the 323 true positive cases that were missed by the CUI query and identified areas for improvement. These include additional acronyms, abbreviations, truncated words and misspellings to add to EMT-P, as well as additional CUIs to add to the CUI query. We plan to make these additions and expect they will improve the sensitivity of the CUI queries

The raw CC queries, much like EMT-P, have been customized to address abbreviations and misspellings. The raw CC queries utilize a keyword search approach as opposed to the standard (CUI) identifier searching used with EMT-P processed CCs. We were pleased to find that EMT-P pre-processed CCs and CUI based syndrome queries work as well as the detailed and complex text queries of raw CC data. EMT-P has been developed by a group of information scientists and with experience in natural language processing. The raw text queries have been developed by a group of surveillance professionals with a great deal of expertise and experience and have taken years of development, evaluation, and revisions to reach their current level of performance. However, the raw CC queries are time-consuming to develop and maintain. There are also limitations of keyword searching (e.g., searching for *flu* can mistakenly identify records with the term *fluid*).

With any natural language data such as chief complaint, there will always be new terms for new clinical conditions (e.g., *avian flu*) and additional misspellings, acronyms, abbreviations (e.g., *diff. br* for *difficulty breathing*) and other local variations in text. EMT-P takes care of many of the text issues currently handled as part of the raw text queries. By pre-processing the data with EMT-P, the syndrome processing is much faster, resulting in better overall performance of the system. Use of a pre-processor such as EMT-P has the potential to streamline maintenance of existing, and development of new, syndrome queries based on CUIs rather than free text. For example, Table 6 shows the differences in raw text and CUI queries for *fever* and *dyspnea* terms.

Table 6: Query Differences

Text Query	**CUI Query**
CC like ‘%fev% OR CC like ‘%febrile%’ or CC like ‘%fvr%’	CUI like ‘C0015967’
CC like ‘%diff br%’ or CC like ‘%diff. br%’ or CC like ‘%difficulty br%’ or CC like ‘%diff. bthg%’ or CC like ‘%dyspnea%’	CUI like ‘C0013404’

It is well documented that biosurveillance is hampered by the variability of CC data.[1,2,13] CC data are often entered in free text form and contain abbreviations, acronyms, misspellings and other lexical and semantic variants. Even when hospitals utilize controlled CC lists, the terms may be non-standard and don’t necessarily match UMLS concepts. We have found that other ED data elements, including clinical notes and temperature, improve the sensitivity of biosurveillance.[14] However, these elements are available for less than 25% of all NC DETECT visits. The chief complaint continues to be the most widely used data element for biosurveillance and is universally available from NC emergency departments. In light of this, we plan to use the results of this study to improve EMT-P and then implement it for pre-processing of CC data for biosurveillance with NC DETECT. We also plan to customize EMT-P for additional syndromes using the method developed for this study.

Even with improvements to EMT-P, the level of sensitivity for acute respiratory surveillance may continue to be less than optimal. We are conducting a related study of the NC DETECT syndrome definitions which is yielding promising results.[15] The current, more stringent ARI definition requires the presence of both a respiratory (e.g., *cough*, *SOB*) and a constitutional (e.g., *fvr*, *chills*) symptom. We are exploring the use of a less stringent ARI definition that requires only a respiratory symptom. Preliminary results indicate that this approach yields a moderate loss of specificity with a fairly significant improvement in sensitivity. We plan to explore methods for using EMT-P with the less stringent syndrome definitions and study the impact of these changes on syndromic classification.

The limitations of this study include the small sample size of the training set. It is likely that we did not identify all of the important natural language patterns in the 1,003 raw CC records that were manually reviewed. However, funding constraints limited our ability to review additional records.

Conclusions

Pre-processing with EMT-P v.2.2 performed as well as our standard query method for detecting acute respiratory illness. Syndrome queries that search EMT-P output are easier to develop and maintain than our standard text query method. Additional improvements to the syndrome case definitions and queries are needed to improve sensitivity.

Acknowledgements

Funding by the CDC for grant #-R01-PH000038-01. Thanks to John Crouch, Jennifer MacFarquhar, Meichun Li, Scott Wetterhall and George Ghneim.

References

1. Shapiro AR. Taming variability in free text: Application to health surveillance. MMWR Morb Mortal Wkly Rep. 2003; 53 (Suppl): 95-100.

2. Husk G, Akhtar S. Chief complaints, emergency department clinical documentation systems, and the challenge of dealing with the patient's own words. Acad Emerg Med. 2007;14(1):69-73.
3. Heffernan R., Mostashari F, Das D, Besculides M, Rodriguez C, Greenko J, et al. New York City syndromic surveillance systems, MMWR. 2004; 24(53): 23-7.
4. Hripscak G, Bamberger A & Friedman C. Fever detection in clinic visit notes using a general purpose processor. Adv Dis Surv. (2007); 2:14.
5. Komatsu K, Trujillo L, Lu H, Zeng D & Chen H. Ontology-based automatic chief complaint classification for syndromic surveillance. Adv Dis Surv. 2007; 2:17.
6. Travers, DA, Haas SW. Using nurses' natural language entries to build a concept-oriented terminology for patient's chief complaints in the emergency department. J Biomed Inform. 2003; 36:260-270.
7. Travers DA, Haas SW. Evaluation of Emergency Medical Text Processor, a system for cleaning chief complaint data. Acad Emerg Med. 2004; 11(11): 1170-1176
8. Li M, Ising A, Waller A, Falls D, Eubanks T, Kipp A. North Carolina bioterrorism and emerging infection prevention system. Adv Dis Surv. 2006; (1):80.
9. Centers for Disease Control and Prevention (October 23, 2003). Syndrome definitions for diseases associated with critical bioterrorism-associated agents. Retrieved from http://www.bt.cdc.gov/surveillance/syndromedef/index.asp. Accessed June 24, 2005.
10. Travers D, Kipp A, MacFarquhar J, Waller A. Evaluation of EMT-P for pre-processing chief complaint data for syndromic surveillance. Adv Dis Surv. 2006; (1):71.
11. Chromy JR. Design optimization with multiple objectives. Proc Survey Research Methods Section, Amer Stats Assoc. 1987; 194-199.
12. Ghneim GS, Wu S, Westlake M, Scholer MJ, Travers DA, Waller AE, Wetterhall SF. Defining and applying a method for establishing gold standard sets of ED visit data. Adv Dis Surv. 2:9.
13. Dara J, Chapman W. Evaluation of preprocessing techniques for chief complaint classification. Adv Dis Surv. 2006; (1):19.
14. Ising AI, Travers AD, MacFarquhar J, Kipp A, Waller, AE. Triage note in ED-based syndromic surveillance. Adv Dis Surv. 2006; (1):34.
15. Scholer MJ, Ghneim GS, Wu, SW, Westlake, M, Travers DA, Waller, AE, McCalla A, Wetterhall, SF. Defining and applying a method for improving the sensitivity and specificity of an ED early detection system. Proc 2007 AMIA Fall Symposium, accepted.