Implementation of Emergency Medical Text Classifier for Syndromic Surveillance

Debbie Travers, PhD, RN[1,3], Stephanie W. Haas, PhD[2], Anna E. Waller, ScD[3], Todd A. Schwartz DrPH[1, 4], Javed Mostafa[2], Nakia C. Best, MSN, RN[1], John Crouch, BA[1,3]

Schools of [1]Nursing, [2]Information and Library Science, [3]Medicine/Emergency Medicine, [4]Public Health/Biostatistics
University of North Carolina, Chapel Hill, NC

Abstract
Public health officials use syndromic surveillance systems to facilitate early detection and response to infectious disease outbreaks. Emergency department clinical notes are becoming more available for surveillance but present the challenge of accurately extracting concepts from these text data. The purpose of this study was to implement a new system, Emergency Medical Text Classifier (EMT-C), into daily production for syndromic surveillance and evaluate system performance and user satisfaction. The system was designed to meet user preferences for a syndromic classifier that maximized positive predictive value and minimized false positives in order to provide a manageable workload. EMT-C performed better than the baseline system on all metrics and users were slightly more satisfied with it. It is vital to obtain user input and test new systems in the production environment.

Introduction
The purpose of this study was to implement a new system, Emergency Medical Text Classifier (EMT-C), into daily production for syndromic surveillance and evaluate system performance and user satisfaction. EMT-C was initially developed to extract infectious disease concepts from triage nurses' notes in emergency department visit records, and was previously tested in the laboratory setting and shown to improve classification accuracy over traditional information extraction methods (Mahalingam, Mostafa, Travers, Schwartz, Haas, Waller, submitted). In order to implement EMT-C into daily production, we modified it to extract concepts from multiple data fields (notes, chief complaints and temperature), classify visit data batched twice daily during uploads to the syndromic surveillance system, and address user workload issues.

Syndromic Surveillance Using ED Textual Data
Public health officials are increasingly reliant on syndromic surveillance systems to facilitate early detection and response to acute infectious disease outbreaks. These systems utilize routinely collected data from electronic health records that are timely, population based and electronically available. A commonly used source of such data is emergency department (ED) visits. A variety of ED visit data elements have been evaluated for use in syndromic surveillance (Table 1).

Table 1. Emergency department visit data elements used for syndromic surveillance

Data element	Description	Size	Availability	Timeliness
Diagnosis	Final diagnosis(es) assigned by provider at end of visit	Coded using ICD-9-CM	Widely available	Delayed
Temperature	Temperature measured in emergency department	Integers	Not widely available	Immediate
Chief complaint	Brief description of the patient's primary symptom(s)	Few terms	Captured electronically by most EDs	Immediate
Triage nurse's note	Expanded history of present illness	Paragraph	Becoming more available electronically	Immediate

Travers, Waller, Haas, Lober, Beard (2003)

Though accurate, diagnosis data are often not available for several days/weeks, so are typically not utilized in syndromic surveillance[2] (Travers, Barnett, Ising, Waller, 2006). The temperature measured in the ED is also accurate, but is not widely available in electronic form in syndromic surveillance systems and doesn't indicate the body system affected (respiratory, gastrointestinal). Chief complaint (CC) data are timely and widely available, and are utilized by most syndromic surveillance systems. However, the CC field may contain only one term and thus can lack the rich information available in the longer triage nurses' notes. Table 2 illustrates this with three syndromic surveillance records that all have a chief complaint of fever. When a triage note is added, the records meet three different syndrome definitions.

Table 2. Triage note improves sensitivity of syndromic surveillance over chief complaint data

Chief Complaint	Triage Note*	Syndrome
Fever	amb c/o yest **fever 102**, **n/v**.today just general **aches**	Gastrointestinal
Fever	27 yo male, c/o **cough** and **SOB** x1 wk. Denies wheez.Yellow **sputum**.	Respiratory
Fever	**Rash** on chest, spread to back, abd & neck. **Fever** today with back pain. **Aches**.	Fever Rash

Information from triage note that triggered positive syndrome classification is **in bold*

A key to effective use of ED symptom data for syndromic surveillance is the accurate classification of symptoms that are often recorded in textual form. However, standards are not in place for codification of these data, so syndromic surveillance systems require methods for processing clinical text. Triage notes are now becoming more available in electronic form, and adding triage notes increases the sensitivity of outbreak detection because of the additional symptom data. In a consensus recommendation published by the CDC, triage notes have been recommended as part of the minimum data set of electronic health record data to include in syndromic surveillance systems for support of meaningful use (CDC, 2012)**.** Though not all EDs have electronic triage notes available, we found in a previous study that a disproportionate percentage of visits that are flagged as positive for one or more syndromes in surveillance reports include a triage note (46%)[3] (Ising, Travers, MacFarquhar, Kipp, Waller, 2006). In another study, we found that the sensitivity for acute respiratory surveillance went from 13% without a triage note to 35% with a triage note (Scholer, MacFarquhar, Sickbert-Bennett, Kipp, Travers, Waller, 2006).

Even when triage notes are available, current syndromic surveillance methods generate suboptimal results because of lexical and syntactic variation in triage notes used for syndromic surveillance. For example, triage note search methods result in both false positives (e.g., no vomiting or fever generates a signal) and false negatives (the system misses variants such as V/D for vomiting and diarrhea, fvr. for fever, or misspelled terms like dierhea), limiting the utility of the data for syndromes such as infectious gastroenteritis. Tools are needed to address the variation in symptom terms in ED data in order to improve the accuracy of syndromic surveillance. Natural language processing tools have been used for concept extraction from semi-structured clinical data (e.g., dictated reports that contain headings, etc.) but there has been limited application of these techniques to unstructured ED triage notes.

Syndromic Surveillance System The North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT) provides clinical data for syndromic surveillance as part of a statewide system to detect and respond to infectious diseases outbreaks (Waller, Scholer, Ising, 2010). The system assists public health officials in early identification of the onset of disease outbreaks, and provides information about the geographic distribution and spread of disease and the demographics of infected persons, enabling public health officials to implement control measures earlier than with traditional disease surveillance methods. The sources of data in NC DETECT include emergency departments, the statewide poison center, ambulance services, the state Department of Public Instruction, and select urgent care centers.
Major users of NC DETECT are public health epidemiologists with the NC Statewide Program for Infection Control and Epidemiology, who are based at the 11 largest hospitals in NC and collectively monitor data for 65% of all ED visits in the state. Other users are epidemiologists with the North Carolina Division of Public Health (DPH)'s Communicable Disease Branch, who query the system daily to: 1) monitor infectious diseases and community-acquired infections of special interest, 2) perform syndromic surveillance, 3) initiate outbreak investigations, and 4) monitor public health threats, including disease outbreaks and disasters. Other users include epidemiologists in other branches of the NC Division of Public Health and divisions of the NC Department of Health and Human Services, as well as communicable disease nurses and epidemiologists in local health departments around the state. Hospital infection control professionals and ED clinicians and administrators also use the system to monitor patients at their hospital EDs. Additionally, the State Center for Health Statistics provides a link to the NC DETECT web portal and refers requests for ED data reports to NC DETECT.

The NC DETECT Web application includes syndromes for acute infectious diseases, chronic diseases and injuries (NC DETECT, 2013). The infectious disease syndromes include acute gastrointestinal illness, influenza-like illness and fever-rash illness, among others. Users can view daily trends for these syndromes and the system will flag any aberrations, i.e., higher than expected counts, based on comparison to a moving average from the very recent past. Aberrations are detected using a CUSUM algorithm developed by the CDC (Hutwagner, Thompson, Seeman, Treadwell, 2003).

Emergency Medical Text Processor (EMT-C)
For this project, we evaluated a new natural language processing system, Emergency Medical Text Classifier (EMT-C). EMT-C was designed to address the complexity of ED triage notes and extract key symptom concepts using a combination of heuristic and statistical natural language processing modules. Our research team developed and tested EMT-C v.1 in a laboratory setting previously (Mahalingam, Mostafa, Travers, Haas, Waller, 2012; Mahalingam et al., submitted).

In the laboratory setting, EMT-C v.1 was tested on a manually annotated set of NC DETECT records (Scholer et al., 2007) and achieved improved sensitivity (Se) (0.77) while maintaining acceptable specificity (Sp) (0.85) over the

standard NC DETECT query method ("baseline") (Se 0.28, Sp 0.97). However, we found that EMT-C v.1.1 was not as effective in identifying true positives and thus had worse positive predictive value (PPV) over baseline (EMT-C 0.18 versus baseline 0.28). EMT-C maintained a high negative predictive value (NPV) (EMT-C 0.99 versus baseline 0.97).

The ultimate evaluation of any system comes when it is actually used in production. In this project, our team sought to make adjustments to improve classification, such as incorporating term weighting, we also needed to adjust the design of EMT-C to better fit the production setting. These issues included adapting the system to handle the smaller daily batches of ED visits as they are uploaded to NC DETECT (as opposed to testing on a fixed corpus) and dealing with the variation in processing accuracy on records both with and without triage notes.

EMT-C is both a pre-processor and classifier, and incorporates both heuristic and statistical natural language processing (NLP) methods. The system utilizes heuristic tools from our previously-developed pre-processing system, Emergency Medical Text Processor (EMT-P) which is used in several surveillance systems (Lu,, Zeng, Trujillo, Komatsu, Chen, 2008; Travers & Haas, 2004) and from analysis of fever and GI terms in NC DETECT triage notes and chief complaints (Travers, Haas, Waller, Crouch, Mostafa, Schwartz, 2010). EMT-P is a pre-processor that employs linguistic methods to clean CC's with the goal of matching a standardized term from the Unified Medical Language System (Conway, Dowling & Chapman, 2013). The statistical NLP method used in EMT-C incorporates vector-space modeling. The vector-space model for classification requires generation of numeric arrays (vectors), whereby each element in the array represents the presence or absence of key terms relevant to the processed content. The key terms are pre-identified and placed in a dictionary (sometimes referred to master term list in the literature) and the content is matched against the dictionary to generate the vector. Numeric values can be either binary or weighted depending on the degree of precision desired.

EMT-C was designed so that the individual syndrome, e.g. GI, and the triage notes were represented as vectors of equal length and the Similarity Value between these vectors were computed using the cosine similarity metric. The master term list used by EMT-C was comprised of syndrome-specific and constitutional terms from the NC DETECT GI syndrome case definition and from our previously-developed pre-processor, EMT-P. Additional terms were found through a term frequency-inverse document frequency (tf-idf) analysis computed on syndrome-positive data. For each ED visit, a query vector was generated based on the presence or absence of each term. The use of EMT-P's misspellings and synonyms modules provided additional help in determining each term's presence by standardizing syndrome-related terms. Cosine similarity values were computed between the document vector and all query vectors. The similarity values were considered analogous to the likelihood of a visit being syndrome-positive, with the average of these values acting as a threshold for positive/negative classification. A high similarity threshold value made assignment to the syndrome category more stringent and rare, and alternatively, by using a low similarity threshold value assignment to the syndrome category is made easier and potentially more frequent. Varying the threshold value allowed us to tune the classifier for lower or higher Se, Sp, PPV, or NPV.

Methods

The goal of this project was to prepare EMT-C for production and then implement it as a user-centered system that could best meet the needs of the Public Health Epidemiologists who use NC DETECT on a daily basis. We deployed EMT-C for one syndrome, Acute Gastrointestinal Illness (GI) syndrome classification in parallel with the current (baseline) GI classifier in use for NC DETECT. The evaluation included analyses of performance data as well as user feedback in narrative form.

EMT-C Revision Phase

In order to prepare the system for implementation into the production environment for NC DETECT, we made iterative revisions to the design of EMT-C. We assessed each version by processing samples of production data that included ED records with and without triage notes. We then tested each version by comparing the GI classifications from EMT-C to those of the baseline system using a manually classified dataset of 3353 records containing gold standard ratings by experts (Scholer et al, 2007). As we prepared EMT-C for implementation into production, we encountered several issues that affected the design of the system. These included a shift in user preferences for system performance, the need for term weighting, variation in the size of data uploads, and different challenges in processing records with and without triage notes.

User Input: Based on pilot triage note research at NC DETECT (Ising et al., 2006), our initial focus with this project was on maximizing Sensitivity (Se), which involves True Positives (TP) and False Negatives (FN). Our first revision, EMT-C v.1.4a, achieved a Se of 0.60 but the PPV dropped to 0.17 (see Table 3). After seeking input from the public health epidemiologists who use NC DETECT for syndromic surveillance, we learned that they preferred minimization of False Positives (FP), as those are resource-intensive to investigate. That is, for GI, the users would prefer to be given a smaller number of records (True and False Positives combined), even with the risk of missing some relevant ones (False Negatives). Based on this user input, we modified EMT-C again, seeking to optimize PPV because it emphasizes True Positives (TP) and diminishes False Positives (FP). The resulting EMT-C v.1.4e

favored PPV over Se while maintaining excellent Sp and NPV (Table 3). Given these results, we chose to implement EMT-C v.1.4e.

Table 3. Performance of EMT-C v.1.4e compared with baseline

	BASELINE Syndrome Queries	EMT-C v.1.4a	EMT-C v.1.4e
Sensitivity:	**0.28**	**0.60**	**0.18**
Specificity:	**0.97**	**0.88**	**0.99**
Positive Predictive Value:	**0.28**	**0.17**	**0.41**
Negative Predictive Value:	**0.97**	**0.98**	**0.97**

n=3353 from BioSense Gold Standard Set (manually annotated), weighted to N=2,418,167

Term weighting: We incorporated term weighting into EMT-C to help improve performance. The baseline system is essentially a query written in Structured Query Language; the query terms representing the syndrome are either present or absent in the ED record. There must be at least one constitutional (e.g., fever, aches) and one syndrome-specific (e.g., vomiting, diarrhea) term present, but no term weights are used; all terms are considered to be equally informative. In contrast, EMT-C is based on the vector-space model, which allows term-weighting; terms that are considered to be highly indicative of the syndrome can receive a higher weight than those that are less so. For example, although *weak, lethargic,* and *nausea** are symptoms of GI, they are also associated with many other conditions and are not specific to the GI syndrome. In contrast, *diarrhea, vomiting, and dehydrat** are strong indicators of GI while introducing minimal noise from other illnesses. Domain experts from the research team assigned weights to each term in the syndrome vector. The effect of these weights is that only one or two highly-weighted terms are needed to pass the Similarity Value threshold and be classified as GI, but a record must contain a greater number of weaker terms to do so. This prevents a record containing only two less specific terms (e.g., *dizzy, weak)* from being classified as syndrome positive, while a record containing two highly weighted terms (e.g., *diarrhea, dehydrat*)* would be classified as syndrome positive.

Daily batches of data: In designing EMT-C v.1.4e, we also dealt with a problem related to batching of data as it is uploaded into NC DETECT. EMT-C computes threshold values based on the average Similarity Value of processed records. In the production system, ED visit data are processed in batches twice daily, and the batches can vary significantly in size. With a very small batch, or a batch that contains few syndrome-related records, the resulting low threshold value would open the possibility of flagging numerous false positives. To correct this, a minimum threshold (0.14) was set as added insurance against broadly flagging records positive. Furthermore, adding the requirement of a syndrome-positive visit needing both constitutional and syndrome-specific terms added supplementary protection.

Records with and without triage notes: Another design decision was also driven by the realities of working with ED records in the context of the production system of NC DETECT. EMT-C v.1 development was based only on ED records with triage notes (TNs), leveraging the additional terms they contain. In practice, however, only about 30% of ED records submitted to NC DETECT currently contain TNs; the chief complaint (CC) is the *only* text field in the remaining 70%. EMT-C performance declines for CC-only records. We tested a two-path system that used the baseline system for CC-only records and EMT-C for those records with TNs, but performance was no better than EMT-C alone. Thus, we decided to use the simpler, one-path design.

Evaluation Phase

Syndromic Surveillance Data Analyses

Number of records classified as GI positive: To evaluate the performance of EMT-C in the production environment, we compared the classification of the baseline system to EMT-C v.1.4e and calculated the number of daily GI syndrome-positive records.

Manually reviewed sample: We also calculated Se, Sp, PPV and NPV for a sample of records that were manually classified by two subject matter experts. For this review, experts saw the full NC DETECT record for each ED visit, which included diagnosis. The diagnoses are typically not available in real time to the epidemiologist users because of delays in data entry from participating EDs (Travers et al., 2006). The manually reviewed sample was composed of a random set of 500 records drawn from NC DETECT during the post-implementation phase of 1-10 to 1-30-2013. In an attempt to ensure inclusion of GI syndrome-positive records, we "pivoted" around a Similarity Value that is slightly near but under typical EMT-C thresholds (pivot value=0.20, while thresholds have typically been 0.21-0.22). Records with similarity values below the pivot point are considered unlikely to be positive, and those above considered likely to be positive. We then computed the total number of records below the pivot point, and total number of records at or above it. From the N=273,409 records under consideration, the final stratified random sample included 250 records from each category, for a total of 500 records. The ratio of the 250 records to the total number of records in each respective category provides the inverse of the weighted value for each record. In other

words, a record's weighted value is determined by whether or not it is in the "likely" or "unlikely" sets, and the weighted values were incorporated into the calculation of Se/Sp/PPV/NPV. For our sampling, there were N=123,779 records in the "low" category and N=149,630 in the "high" category. Hence, the weights were computed as follows:

w[high]= 149,630 / 250 = 598.52
w[low]= 123,779 / 250 = 495.12

User Evaluation

Survey: The user evaluation was comprised of an anonymous pre- and post-implementation survey sent to the public health epidemiologists (PHE) who use NC DETECT daily as part of their surveillance duties. Prior to each survey distribution, research staff met with the PHE users to discuss survey instructions and the method used to identify false positive GI classification during the study period. Users were also reminded of the case definition for the GI syndrome. Both surveys asked users for their opinion of 4 aspects of performance for the surveillance classification system they used (baseline in the pre-implementation survey and EMT-C v.1.4e in the post-implementation survey).

- Accuracy of classifications (1 = very inaccurate, 5= very accurate)
- Ease of interpretation of syndrome data (1 = very difficult, 5 = very easy)
- Omission of information from classification results (1 = never, 5 = always)
- Provision of needed information in classification results (1 = never, 5 = always)

These aspects were identified, from discussions with users, as being important in terms of contribution of surveillance data to their work-flow and decision-making. Users were also asked to rate their overall satisfaction with the system. In addition, the post-deployment survey asked users to compare performance of EMT-C with that of the baseline system along the same aspects.

- Accuracy (1 = much less accurate, 5=much more accurate)
- Ease of interpretation (1= much more difficult, 5 = much less difficult)
- Omission of information (1 = much less often, 5 = much more often)
- Provision of needed information (1 = much less often, 5 = much more often)

Users were invited to provide examples and other comments about the systems in both surveys. A link to the pre-implementation survey was sent via electronic mail to the 11 Public Health Epidemiologist users in February 2012, and the link to the post-implementation survey was sent to the 9 users from February 2013. Two PHE positions were vacant at the time of the second survey.

False positive records: In the post-EMT-C implementation period, we also collected user feedback on specific ED records classified by EMT-C and the baseline system. Users could flag and comment upon NC DETECT records classified as GI positive by either the baseline system, EMT-C, or both, that the users considered to be potential or actual false positives (FP). A record could be flagged as a definite FP ("yes"), a possible FP ("maybe") or one about which there was some other question or concern ("other").

Results

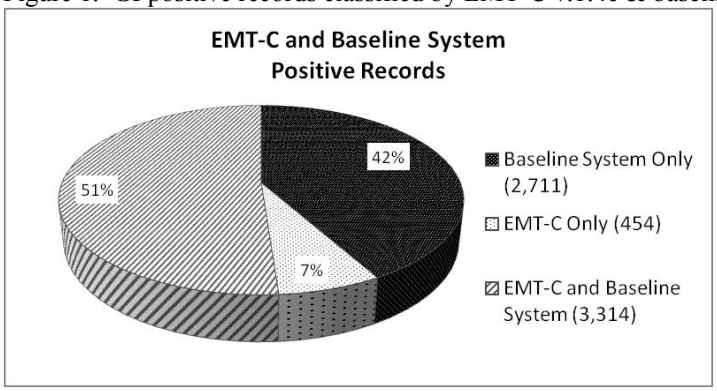
The EMT-C v.1.4e GI syndrome was implemented into production at NC DETECT in parallel with the baseline system GI syndrome on January 1, 2013. We collected post-implementation syndromic surveillance data from January 10-30, 2013. We evaluated user feedback on the baseline GI syndrome from February 2-23, 2012 and on the EMT-C v.1.4e GI syndrome from January 10-30, 2013.

Syndromic Surveillance Data Analysis

There were 268,987 ED visit records included in NC DETECT for the study period of 1/10/13-1/30/13. The EMT-C generated Similarity Values for those dates ranged from 0.01 to 0.60, with thresholds varying between 0.20 and 0.23.

Number of records classified as GI positive: Of all 268,987 records in NC DETECT during the post-EMT-C implementation study period, 6,479 (2.4%) records were classified as GI syndrome positive: 3,768 (1.4%) by the EMT-C system and 6,025 (2.2%) by the baseline system. 3,314 of those records were classified as positive by both EMT-C and the baseline system. These results are displayed graphically in Figure 1.

Figure 1. GI positive records classified by EMT-C v.1.4e & baseline



Total N=6479 ED records from 1/10/13-1/30/13 that were classified positive for GI Syndrome

Manually reviewed sample: In the manual classification of the 500 records by the two subject matter experts, we found 86.4 % agreement with a kappa of 0.55 (95% CI 0.47-0.64). The experts met and came to consensus on classifying the remaining records as true positive or true negative. Statistical analyses of the performance of the baseline and EMT-C systems on the 500 records reviewed are shown in Table 4.

Table 4. Performance of EMT-C v.1.4e compared with baseline

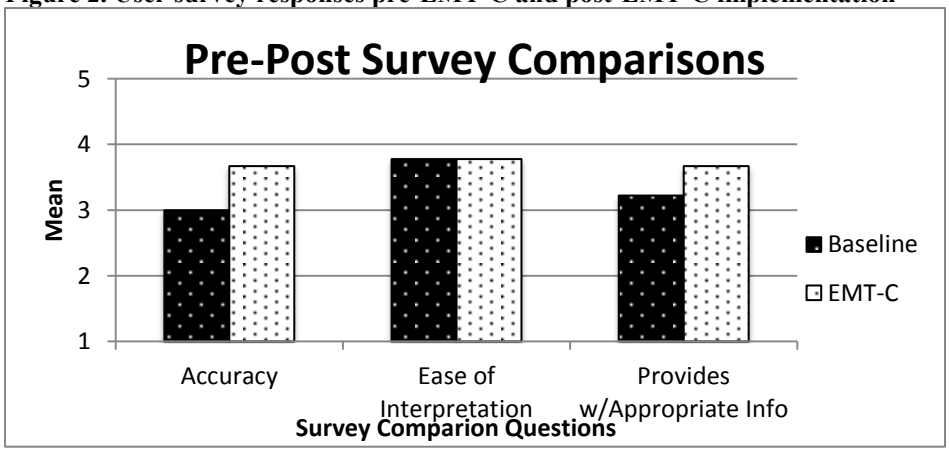
	EMT-C	Baseline System
True Positives:	70	61
False Positives:	62	75
False Negatives:	21	30
True Negatives:	347	334
Sensitivity:	0.79	0.69
Specificity:	0.83	0.80
Positive Predictive Value:	0.53	0.46
Negative Predictive Value:	0.94	0.91

n=500 records, weighted to N=273,409, reviewed by subject matter experts

User Evaluation

Survey: 11 users were invited to participate in the 2012 pre-deployment study and received email invitations and links to the pre-implementation survey. The response rate was 100%. At the time of the post-EMT-C implementation survey in 2013, all 9current users were invited to participate and the response rate was 100%. Responses to the pre-implementation survey questions indicated that they considered baseline system performance to be slightly positive, with users giving a mean overall satisfaction rating of 3.67 (1 = very dissatisfied, 5 = very satisfied).

Figure 2. User survey responses pre-EMT-C and post-EMT-C implementation



Compared to the baseline system, (Fig. 2) users perceived EMT-C as more accurate (baseline mean = 3.27 vs. EMT-C 3.67) and more capable of providing information needed to make decisions or take appropriate action (mean 3.22 vs. 3.67). Users perceived both systems as easily interpretable (mean 3.78). Users were slightly more satisfied with EMT-C than with the baseline system (mean = 3.44, where 1 = much less satisfied, 5 = much more satisfied).

False positive records: Nine users annotated 268 records during the post-EMT-C implementation data collection period (mean per user = 29.78 annotations); 251 of the records included comments (mean per user = 27.89). Content analysis of the associated comments resulted in 8 comment classes. Table 5 shows the number of Yes, Maybe, and Other flags for each comment class.

The most frequent class of comment provided some evidence as to *why* the user judged the record to be a false positive (FP), such as listing symptoms that were not consistent with the GI case definition (e.g., the visit was for an injury), or noting that the onset of the symptoms was not acute. All of these were flagged as definite or possible FPs. The next most frequent class noted that the record had been classified as GI by the baseline system but not by EMT-C. While true, the combination of the "other" flag and the comment did not indicate whether the user judged it to be FP.

Table 5. Frequency of comment classes associated with each false positive (FP) Flag type

	Frequency (%)- Flag			**Frequency (%)- Comments**
Comment Class	**Yes**	**Maybe**	**Other**	
Explanation of why record isn't GI (symptoms, onset characteristics, etc.)	90 (33.58)	6 (2.24)	0	96 (35.82)
Missed by/not in EMT-C TRUE	8 (2.99)	7 (2.61)	70 (26.12)	85 (31.72)
Missed by/not in EMT-C FALSE	0	0	20 (7.46)	20 (7.46)
Etiology not identified	19 (7.09)	1 (0.37)	0	20 (7.46)
No GI symptoms; doesn't meet case definition	14 (5.22)	0	0	14 (5.22)
Other comment	0	1 (0.37)	6 (2.24)	7 (2.61)
Not in baseline	0	0	6 (2.24)	6 (2.24)
Technically meets case definition but isn't GI	2 (0.75	0	0	2 (0.75)
No comment provided	*2 (0.75)*	*16 (5.97)*	*0*	*18 (6.72)*
Total	**135 (50.37)**	**31 (11.57)**	**102(38.06)**	

To further investigate records annotated by the users flagged as FPs ("Yes"), a member of the research team who is also a domain expert (DT) reviewed a sample of the records (N=129) and judged whether, in her opinion, they were FPs. The team expert agreed with users that 82.17% of the records were FP. Most of the agreed-upon FP records (N=85) were classified as GI by both EMT-C and the baseline system, and thus were FP. Of these 85, 69 were accompanied by comments explaining why the record was not GI, 11 had comments stating they did not meet the case definition, 4 had comments that the etiology doesn't fit, and 1 did not have an associated comment.

Table 7 shows examples of some of these FP records, which include the final diagnoses which were not in NC DETECT in real-time but were available later during the expert review. The examples illustrate the challenges for classification systems of dealing with sparse data and context issues**.**

Table 7. Examples: False positive records determined by both system users and expert review

CC	Temp	TN	Dx
N/V/D; FEVER	None listed	None listed	V58.69 - LONG-TERM (CURRENT) USE OF OTHER MEDICATIONS *-* 172.4 - MALIGNANT MELANOMA OF SKIN OF SCALP AND NECK *-* 787.01 - NAUSEA WITH VOMITING *-* 789.09 - ABDOMINAL PAIN OTHER SPECIFIED SITE *-* 197.7 - MALIGNANT NEOPLASM OF LIVER
FEVER/ HEADACH/ VOMITING	None listed	None listed	311 - DEPRESSIVE DISORDER NOT ELSEWHERE CLASSIFIED *-* 305.1 - NONDEPENDENT TOBACCO USE DISORDER *-* 616.4 - OTHER ABSCESS OF VULVA *-* V13.01 - PERSONAL HISTORY OF URINARY CALCULI
FEVER	**38**	2yo boy w/ broviac, Jtube significant **gastrointestinal** history who presents with **fever** for 3 days responsive to Ibuprofen and Broviac that has problems drawing and flushing for 3 days. pt is on chronic TPN. Has been sleeping more the last few days. Was recently admitted with broken broviac that was repaired 5 days previously.	Not available at time of manuscript submission

Discussion

In this study, we addressed the challenges of implementing into practice an NLP system that was previously trained and validated in a laboratory setting. Based on our sample of 500 records, EMT-C v.1.4e performed better than baseline on all metrics: Se, Sp, PPV, and NPV. Responses on the post-implementation survey suggested that users tended to view EMT-C's performance on all aspects slightly more positively than the baseline system. This is encouraging, especially given the drastic reduction in the number of positive records EMT-C identified, and that users thus had to review.

Our design decision to favor PPV over Se may be a better direction than striving to achieve high Se in real-world syndromic surveillance systems. As public health users have more and more data sources available for monitoring, it becomes even more important to minimize user workload associated with potential signal review. The focus of syndromic surveillance systems may be shifting from using syndromic surveillance systems to detect unusual events to identifying signals that require action. Samoff and colleagues (2012) evaluated public health surveillance signals in NC and found that a small proportion of signals from syndromic surveillance were actionable. They examined the integration of syndromic surveillance data in daily practice at local health departments (LHDs). Structured interviews were conducted with local health directors and communicable disease nursing staff from a stratified random sample of LHDs from May through September 2009 to capture information on direct access to the North Carolina syndromic surveillance system and on the use of syndromic surveillance information for outbreak management, program management, and creation of reports. They analyzed syndromic surveillance system data to assess the number of signals resulting in a public health response, and found that only a small proportion of signals (<25%) resulted in a public health response. They recommend that syndromic classifiers be designed to generate fewer signals, and increase the likelihood of signals likely to need action.

Evaluating a system in the laboratory against gold standard, curated data is an important step in system development. But design decisions must also be based in the reality of how the system will be used, and research project timelines do not always allow for this; indeed, it can be difficult to have access to a real-world setting. Preparing EMT-C for testing in actual use forced us to deal with small set sizes, missing and sparse data, and the knowledge that the users would be using it for real work and real decisions.

There will never be a perfect computer classification system. There will always be false positives and negatives. One factor contributing to the false positives problem in this application is the sparse data available in the records at the time they are classified by NC DETECT. The CC and TN may consist of only a few terms, e.g., "N/V/D; FEVER" (nausea, vomiting, diarrhea, fever), which result in a positive GI classification. But these fields do not

include relevant information such as the presence of a chronic disease that would exclude the record from the case definition. These types of false positives (FP) cannot be eliminated from a real-time classification system such as EMT-C. Other FP problems can more easily be addressed in future versions of EMT-C. For example, the addition of features to identify GI symptoms that describe the patient's medical history, not the reason for the current visit, could be incorporated as evidence against classifying a record as GI positive.
We chose to design EMT-C to allow for adjustments in Se, Sp, NPV and PPV. This allows for flexibility in adjusting EMT-C for local and temporal conditions and user preferences. The system also permits acquisition of term weights using both automated and manual approaches. Syndromes have different characteristics. Some, like GI and influenza-like illness, are commonly recognized, capturing thousands of records (including true and false positives) every week, most of which are not considered actionable by public health users. Others, such as meningococcal and anthrax, receive very few hits, but the consequences of missing a true positive are more serious. This suggests that the design decision to maximize PPV at the possible expense of Se may not be appropriate for all syndromes. User input on the comparative risks and benefits of maximizing PPV or Se is vital to the continued improvement**.**

Free text fields in medical records are, in some ways, a mixed blessing. There is no doubt that free text allows clinicians to express their observations, concerns, patients' own words, and other potentially useful information without being limited by a controlled vocabulary or a series of checkboxes. On the other hand, with free text comes non-standard language; unusual words or phrases, idiosyncratic abbreviations, misspelled words, and so on. The architecture of EMT-C provides a framework that accommodates a wide range of information representation, including structured data fields, standardized terminologies and text, and wildly creative language.

Limitations

The manually classified sample of 500 records was small but was all that our resources would allow. Since most of the thousands of records uploaded to NC DETECT daily are not positive for any acute infectious disease syndrome, it is very costly to adequately evaluate classification performance. Thus we rely on the use of stratified random samples.

Responses to both pre and post-implementation user surveys must be interpreted in light of the fact that there is no alternative surveillance system available to the users; judgments in the pre-implementation survey may reflect users' perspectives that, although the system is not perfect, it is better than nothing. Of course, no significance testing is possible with such a small group of users.

At present, triage notes are not widely available for syndromic surveillance so the utility of EMT-C is somewhat limited. While there is a state mandate to send chief complaint data to NC DETECT, the transmission of triage notes to NC DETECT is currently optional. However, the availability of triage notes in NC DETECT continues to increase, and our study results have motivated NC DETECT users to request that EDs in North Carolina be required to send triage notes to the state for syndromic surveillance. As electronic health record adoption increases in NC and nationally, the availability of real-time triage and other clinical ED notes is expected to increase.

A small number of epidemiologists participated in our user evaluation of EMT-C, so the results lack generalizability. Our findings were promising, however, and provide a fresh perspective on the issue of maximizing PPV over Se which should be evaluated with larger samples of syndromic surveillance users.

Future Directions

Future plans call for extension of EMT-C to the acute respiratory and fever rash illness syndromes, and to test the system on data from other syndromic surveillance systems.

Conclusion

Our new system EMT-C v.1.4e was successfully implemented into production with key input from users, and was adjusted to favor positive and potentially actionable signals over high sensitivity. Users were positive about the performance of the new system. Unstructured clinical notes can be a valuable source of syndromic surveillance data if the challenges of free text form can be managed.

Acknowledgements

We wish to thank Clifton Barnett, Lana Deyneka, the PHEs, and the NC Division of Public Health. Funding by National Library of Medicine #1G08LM009787-01A1. 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.

References

1. Biosurveillance Data Steering Group- Biosurveillance Workgroup, American Health Information Community, USDHHS. (2007). Retrieved June 29, 2008 from http://www.hhs.gov/healthit/ahic/materials/meeting10/ bio/BDSG_ Minimum_DataSet.doc.
2. Centers for Disease Control and Prevention. (2012). PHIN messaging guide for syndromic surveillance: Emergency department and urgent care data. Retrieved August 19, 2013 from http://www.cdc.gov/phin/library/guides/PHIN_MSG_Guide_for_SS_ED_and_UC_Data_v1_1.pdf.
3. Conway M, Dowling JN, Chapman WW. (2013). Using chief complaints for syndromic surveillance: A review of chief complaint based classifiers in North America. *Journal of Biomedical Informatics 46*: 734-743.
4. Goth G. Analyzing medical data. *Communications of the ACM*. 2012 June;55(6):13-15
5. Hutwagner, L., Thompson, W., Seeman, G.M., Treadwell, T. 2003. The Bioterrorism preparedness and response Early Aberration Reporting System (EARS). *JUrbanHealth*; 80: i89–96
6. Ising, A., Travers, D.A., MacFarquhar, J., Kipp, A. and Waller, A. 2006. Triage note in emergency department-based syndromic surveillance. *Adv Dis Surv*. (1): 34.
7. Li, M., Ising, A., Waller, A., Falls, D., Eubanks, T. and Kipp, A. 2006. North Carolina bioterrorism and emerging infection prevention system. *Adv Dis Surv*. (1):80.
8. Lu, H., Zeng, D., Trujillo, L., Komatsu, K. and Chen, H. 2008. Ontology-enhanced automatic chief complaint classification for syndromic surveillance. *Journal of Biomedical Informatics* (41): 340-356.
9. Mahalingam D, Mostafa J, Travers DA, Schwartz TA, Haas SW, Waller A. (submitted). Emergency Medical Text Classifier: A new system for processing and classifying triage notes.
10. Mahalingam, D, Mostafa, J., Travers, D., Haas, S.W. & Waller, A. (2012). Automated syndrome classification using early phase emergency department data. Paper presented at: *Proceedings of the American Computing Machinery Special Interest Group on Healthcare Information Technology*; 2012 Jan 28-30; Miami, Florida. New York: New York; 2012. P. 373-378.
11. NC DETECT. 2013. Retrieved March 10, 2013 from: http://www.ncdetect.org.
12. Samoff E, Waller AE, Fleischauer A, et al. Integration of syndromic surveillance data into public health practice at state and local levels. *Public Health Reports*. 2012 May-June;127(3):310-317.
13. Scholer, M.J., Ghneim, G.S., Wu, S., Westlake, M., Travers, D.A., Waller, A.E. et al. 2007. Defining and applying a method for improving sensitivity and specificity of an emergency department early even detection system. *Proceedings of the 2007 AMIA Symposium,* 651-655.
14. Scholer, M.J., MacFarquhar, J., Sickbert-Bennett, E., Kipp, A., Travers, D. & Waller, A. (2006). Reverse engineering of a syndrome definition for influenza. *Advances in Disease Surveillance, 1*, 64.
15. Travers, D. A., Barnett, C., Ising, A., & Waller A. 2006. Timeliness of emergency department diagnoses for syndromic surveillance. *Proceedings of the AMIA Symposium,* 769-773.
16. Travers D, Haas SW, Waller A, Crouch J, Mostafa J, Schwartz T. (2010). Identifying evidence of fever in emergency department text. Poster presented at: Proceedings of the *2010 American Medical Informatics Association*; 2010 Nov 13-17; Washington, District of Columbia. Bethesda: Maryland; 2010. P. 1280.
17. Travers, D.A. and Haas, S.W. 2004. Evaluation of Emergency Medical Text Processor: A system for cleaning chief complaint data. *Academic Emergency Medicine, 11*(11), 1170-1176.
18. Travers, D.A., Waller, A., Haas, S.W., Lober, W.B., Beard, C. 2003. Emergency department data for bioterrorism surveillance: electronic data availability, timeliness, sources and standards. *Proceedings of the AMIA Symposium, 664*, 8.
19. Waller, A. E., Scholer, M.J., Ising, A. I. (2010). Using emergency department data for biosurveillance: The North Carolina experience. In Castillo-Chavez C. et al. (Eds). *Infectious Disease Informatics and Biosurveillance: Research, Systems, and Case Studies.* New York: Springer Publishing Company; 46-66.