Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 41 (2008) 613–623

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Evaluation of preprocessing techniques for chief complaint classification

Jagan Dara [a], John N. Dowling [a], Debbie Travers [b],
Gregory F. Cooper [a], Wendy W. Chapman [a,*]

[a] Department of Biomedical Informatics, University of Pittsburgh, 200 Meyran Avenue, VALE M-183, Pittsburgh, PA 15260, USA
[b] School of Nursing, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Received 27 June 2007
Available online 18 December 2007

Abstract

Objective: To determine whether preprocessing chief complaints before automatically classifying them into syndromic categories improves classification performance.

Methods: We preprocessed chief complaints using two preprocessors (CCP and EMT-P) and evaluated whether classification performance increased for a probabilistic classifier (CoCo) or for a keyword-based classifier (modification of the NYC Department of Health and Mental Hygiene chief complaint coder (KC)).

Results: CCP exhibited high accuracy (85%) in preprocessing chief complaints but only slightly improved CoCo's classification performance for a few syndromes. EMT-P, which splits chief complaints into multiple problems, substantially increased CoCo's sensitivity for all syndromes. Preprocessing with CCP or EMT-P only improved KC's sensitivity for the Constitutional syndrome.

Conclusion: Evaluation of preprocessing systems should not be limited to accuracy of the preprocessor but should include the effect of preprocessing on syndromic classification. Splitting chief complaints into multiple problems before classification is important for CoCo, but other preprocessing steps only slightly improved classification performance for CoCo and a keyword-based classifier.

Keywords: Public health surveillance; Surveillance; Infection control; Syndromic surveillance; Syndrome; Chief complaints; Natural language processing

1. Introduction

The threat of bioterrorism attacks has led to the development of early warning systems focused on timeliness of detection [1]. These systems often use data collected routinely for other purposes, resulting in the collection and analysis of data earlier when compared to conventional public health surveillance methods. Examples of such data include sales of over the counter drugs [2,3], telephone triage data [4], discharge diagnoses, and web access logs [5,6].

Healthcare registrations that include patient chief complaints recorded on admission are another data source with high potential for biosurveillance and early detection of outbreaks. The chief complaints are recorded in coded or free-text form and can be automatically classified into syndromic categories. Various classifiers exist for categorizing free-text chief complaints into syndromic groups, and the classifiers have shown high specificity and moderate sensitivity at detecting patients with syndromes relating to outbreaks [7–11].

Real-time Outbreak and Disease Surveillance (RODS) [12] is an automated biosurveillance system that monitors chief complaint data routinely collected during an emergency department admission. RODS uses a naïve Bayesian classifier called CoCo [13] to classify every chief complaint into one of seven syndromic categories relevant to public health or bioterroristic outbreaks: Gastrointestinal, Constitutional, Respiratory, Rash, Hemorrhagic, Botulinic, and Neurological. The eighth syndrome, Other, is a catch-all for everything not relevant to syndromic surveillance. Univariate and multivariate statistical detection algorithms are then used to detect anomalous patterns and alert users to abnormal syndrome counts.

[*] Corresponding author. Fax: +1 412 647 7190.
E-mail address: chapman@cbmi.pitt.edu (W.W. Chapman).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.11.004

Free-text chief complaints are challenging to work with due to substantial word variation [14,15]. There is no standard terminology for expressing a chief complaint, resulting in differences idiosyncratic to a specific area or hospital. Chief complaints are recorded in busy medical settings, increasing the occurrence of concatenations, such as *flusxs—flu symptoms*, and misspelled or mistyped words, such as *nausa—nausea*. Another complication with chief complaints is the use of abbreviations, such as *ha—headache,* and acronyms, such as *n/v—nausea and vomiting*. Some hospitals' electronic interface limits the number of characters that can be entered, resulting in truncations, such as *diar—diarrhea.* Symptoms patients present with can be defined in multiple ways by using synonyms like *shortness of breath* and *dyspnea*, paraphrases like *grandmother sts pt c/o having a flu,* and different parts of speech, such as *coughs* and *coughing*. Moreover, some chief complaints describe multiple medical problems that could be classified into more than one syndromic category, such as *headache/vomiting*, which denotes neurological and gastrointestinal syndromes.

We hypothesized that preprocessing chief complaints by standardizing word variations, correcting misspellings, and splitting a complaint into separate problems before classification would result in more accurate syndromic classifications, potentially increasing sensitivity and specificity of detection. Our objective for this study was to compare classification performance from chief complaints with and without preprocessing.

2. Methods

We measured classification performance of two chief complaint classifiers with and without preprocessing to determine whether preprocessing improves sensitivity and specificity of classification. Both classifiers were used to classify a set of chief complaints into any of eight syndromic categories currently used by the RODS system: Respiratory (congestion, shortness of breath, cough, etc.); Gastrointestinal (nausea, vomiting, abdominal pain, etc.); Rash (most rashes); Hemorrhagic (bleeding from most sites); Constitutional (fever, malaise, body aches, etc.), Botulinic, Neurological (non-psychiatric neurological symptoms, such as headache or seizure), and Other (genitourinary complaints, trauma, etc.). Chief complaints were preprocessed using combinations of two preprocessors. Below, we describe the preprocessors and chief complaint classifiers used in this study, along with the evaluation we performed.

2.1. Preprocessors

We applied two preprocessors to chief complaints and measured whether preprocessing improved syndromic classification performance. We developed the first preprocessor (CCP) to decrease word variation, expand abbreviations, and correct misspellings. The second preprocessor (EMT-P)[17] was developed by Travers and colleagues for standardizing chief complaints. In addition to decreasing word variation and expanding abbreviations, EMT-P also splits chief complaints into multiple problems based on syntactic and semantic properties. We applied CCP, EMT-P, CCP combined with EMT-P, and CCP combined with only the splitting module of EMT-P (without normalization modules such as synonym replacement) to determine whether preprocessing improves classification performance. We describe CCP and EMT-P next.

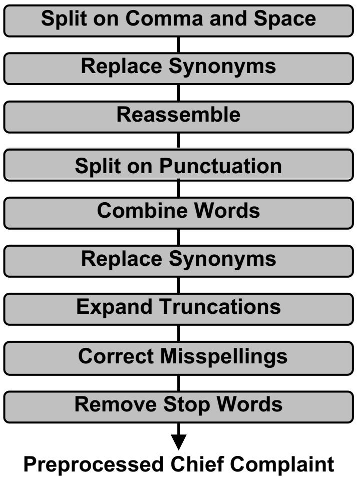

Fig. 1. The CCP modules for preprocessing chief complaints.

2.1.1. Chief complaint processor (CCP)

CCP is a chief complaint preprocessing algorithm that (i) standardizes chief complaints by replacing synonyms, (ii) replaces abbreviations, acronyms and truncations with expanded forms, (iii) corrects misspellings and typographical errors, and (iv) removes words that do not have clinical meaning. The modules involved are shown in Fig. 1. A description of each module follows.

2.1.1.1. Split chief complaint on comma and space

abd cramps, n/v	→	*abd* *cramps* *n/v*

In spite of their brevity, chief complaints often contain punctuation. A comma or space is often used for word separation. The first CCP module splits the chief complaint *abd cramps, n/v* into three separate words. The first is *abd,* second is *cramps,* and the third is *n/v.*

2.1.1.2. Replace synonyms

shortness of breath	→	dyspnea
gx	→	ground transportation
n/v	→	nausea/vomiting

Synonyms, acronyms and abbreviations occur frequently in chief complaints. We used a local dictionary of 3036 synonyms to replace acronyms and abbreviations with their full forms (Appendix 1). In addition, we used a local list of 10 context sensitive synonyms (Appendix 2) that are dependant on the co-occurring words in the chief complaints. In the initial stage, chief complaints are split on comma and space and checked for synonyms. But the chief complaints may contain a multitude of punctuations such as semicolons, hyphens etc. For this reason, the synonym replacement module is applied a second time after all punctuation has been removed.

2.1.1.3. Combine words

hea dache fever	→	*headache fever*

A common cause for word error in chief complaints is the introduction of a blank space before the completion of the word. Since the chief complaints are typed in busy medical settings, this could very well be a typographical error or a systematic error while storing or transferring chief complaints. CCP checks each word for a misspelling by using the spell checking module (described below). Misspelled words are combined with the word immediately following them. The combination is retained if approved by the spell checker.

2.1.1.4. Expand truncations

crying diar	→	*crying diarrhea*

In some cases, words in the chief complaints are truncated. A truncation could be the result of a nurse creating unique abbreviations on the go or of words being terminated due to a computer-system space constraint. If the spell checker module suggests that a word needs correction, the truncation module checks for truncations in a local library of 636 words commonly occurring in chief complaints (Appendix 3). The expansion is retained on the spell checker's approval.

2.1.1.5. Correct misspellings

dizziness nausa	→	*dizziness nausea*

The spell-checking module uses the Java API from the National Library of Medicine's (NLM's) GSpell[16]—a spelling suggestion tool that uses a mix of algorithms to retrieve close neighbors. The API was used to create a suggestion tool that uses a hierarchy of two dictionaries: (1) NLM's 2003 Specialist Lexicon term list, which contains 292,979 words, and (2) the local dictionary of 636 words (Appendix 3). The tool uses the 2003 lexicon to search for a suggestion. If a suggestion exists, indicating a word error, the local dictionary is used to retrieve a suggestion. If the local dictionary fails to yield a suggestion, the tool retains the suggestion from the 2003 lexicon.

2.1.1.6. Remove stop words

flu symptoms	→	*flu*

Frequently occurring words that are unlikely to help classification are considered stop words, which, in our experience, can cause classification errors. The stop word removal module removes any word in a chief complaint that occurs in a local list of 303 stop words (Appendix 4).

2.1.2. Emergency medical text processor (EMT-P)

EMT-P [17] was developed at the University of North Carolina at Chapel Hill. EMT-P is a set of natural language processing modules that clean chief complaint text in order to extract standard clinical terms. The terms are then mapped to concepts in the Unified Medical Language System® (UMLS®). The individual EMT-P modules are written in Perl with a controller program written in JAVA. The goal of EMT-P is to minimize processing of the raw chief complaints (CCs) while facilitating a match with a standard UMLS term. The standard terms produced by EMT-P can then be aggregated for secondary uses such as biosurveillance, tracking the major reasons why patients visit the ED, or clinical research. The system has been validated for general clinical purposes [17].

The EMT-P Perl modules are organized in rounds that range from least aggressive to most aggressive, and each CC is processed only until a UMLS match is made. A basic cleaning module is also run at the start of EMT-P and in each round; the cleaning module performs basic processing such as replacing multiple spaces with single spaces, converting CC's to lower case, and eliminating most non-alpha-numeric characters. Fig. 2 shows the three rounds of EMT-P.

2.1.2.1. Round 1: replace and correct

Dirreah	→	*diarrhea*

In Round 1, EMT-P replaces acronyms, abbreviations, truncations and other synonyms, and misspellings with standard terms. At the conclusion of Round 1, all CC's are compared to the UMLS and those that match standard UMLS concepts are not processed further. The remaining, non-matching terms are processed in the next round.

2.1.2.2. Round 2: punctuation and segmentation

chest/abd pain	→	*chest pain and abdominal pain*

In Round 2, EMT-P addresses punctuation in several modules. First, chief complaints with punctuation marks are expanded (e.g., h/a → ha). Then coordinate structures are processed using context-sensitive rules for the most com-

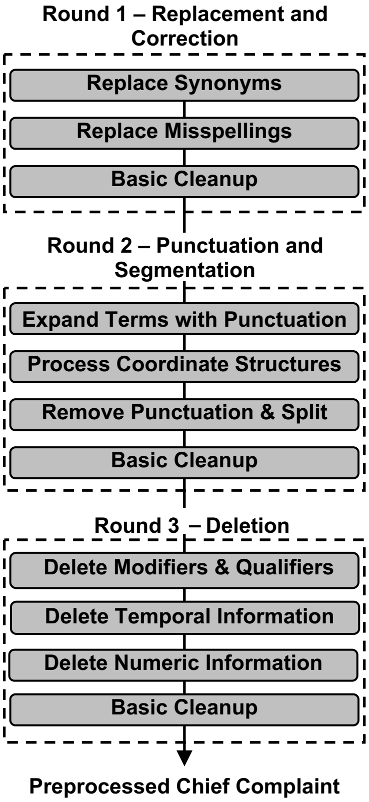
Round 1 – Replacement and Correction
Replace Synonyms
Replace Misspellings
Basic Cleanup
Round 2 – Punctuation and Segmentation
Expand Terms with Punctuation
Process Coordinate Structures
Remove Punctuation & Split
Basic Cleanup
Round 3 – Deletion
Delete Modifiers & Qualifiers
Delete Temporal Information
Delete Numeric Information
Basic Cleanup
Preprocessed Chief Complaint

Fig. 2. The EMT-P modules for preprocessing chief complaints.

mon coordinate structures found in ED CC data. Semantic type information from the UMLS is used to split coordinate structures with body parts on either side of a slash or conjunction, into separate CCs. For example, *chest/ abd pain* is split into 2 records, *chest pain* and *abdominal pain*, but *rash on chest/fever* is not split with the coordinate structures module. After the coordinate structures module, remaining CCs with slashes and other abbreviation (comma, semi-colon) are split on the punctuation (e.g., *rash on chest/fever* is split into 2 records, *rash on chest* and *fever*).

2.1.2.3. Round 3: deletion

chest pain since 3pm	→	*chest pain*

Round 3 is the most aggressive round, and CCs are only processed in this round if they fail to match a standard UMLS concept after previous rounds. Modifiers, qualifiers, numbers and temporal information are deleted in this round. While modifiers, qualifiers and numbers are important data for clinical and other purposes, the goal of EMT-P is to distill CCs into easily aggregated categories.

We evaluated classification performance of two classifiers after preprocessing with CCP and EMT-P alone, with CCP and EMT-P combined, and with CCP and EMT-P's splitting module. Combining CCP and EMT-P involved applying EMT-P after the spell correction module of CCP, as shown in Fig. 3.

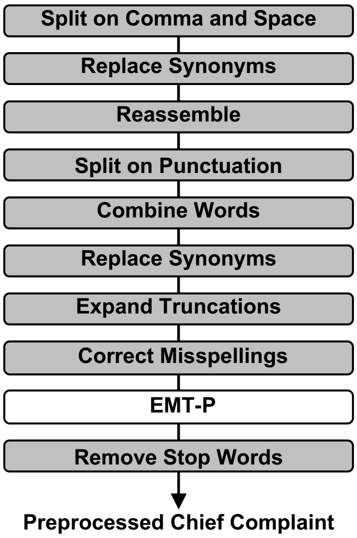

Fig. 3. Process for combining CCP and EMT-P modules.

3. Classifiers

We compared classification performance of unprocessed chief complaints against that of preprocessed chief complaints using two different classifiers—a statistical classifier (CoCo) and a rule-based classifier (Keyword Classifier, or KC).

CoCo 3.0—The RODS system uses CoCo to automatically classify free-text chief complaints into syndromic categories [12,13]. CoCo is a naïve Bayesian classifier that assigns one syndrome to each chief complaint. A training set of 28,990 chief complaints was used to estimate the prior probabilities of unique words for each syndrome. Given a chief complaint, CoCo calculates a joint probability estimate over the eight possible classifications. In its current implementation, CoCo assigns the chief complaint the syndromic category with the highest posterior probability.

Keyword Classifier (KC)—We implemented a keyword classifier based on keywords contained in the New York City Syndromic Macros—a SAS algorithm that scans chief complaints for character strings and assigns matching chief complaints to a syndrome category [18]. For example, if the characters *diar* are found in a chief complaint, the complaint is assigned the category Gastrointestinal. The New York City Syndromic Macros include character strings representing misspellings, abbreviations, acronyms, and truncations. As a keyword search, the algorithm can assign multiple syndromic categories to a chief complaint. We converted the SAS version of the algorithm into Java code. To compare KC's syndromic classifications against CoCo's classifications, we modified the algorithm to assign the syndromes monitored by RODS instead of the syndromes monitored by the New York City Department of Health and Mental Hygeine. To do this, physician author JND mapped the keywords in the SAS code to RODS' syndromic categories. For example, the respiratory keywords *cough* and *sob* also mapped to RODS' Respiratory category, whereas the respiratory keyword *earache* mapped to RODS' Other category. In practice, the New York City Department of Health and Mental Hygeine algorithm uses a hierarchy of syndromes to select the most important syndromic classification when the algorithm assigns more than one syndrome to a single chief complaint. For our Keyword Classifier, we did not employ the hierarchy and allowed KC to assign multiple syndromic categories based on the view that there may be more than one valid classification for a single chief complaint.

3.1. Data Sets

We used three data sets in this study.

1. CoCo's Training set—CoCo's training set consists of 28,990 chief complaints from Utah. The training set was manually annotated by a single physician board-certified in internal medicine and infectious diseases with 30 years of clinical experience (author JND). We also used a subset of the training set to validate the accuracy of our implementation of the KC by comparing KC's classifications against the manual classifications.
2. Development set—The development set consisted of 20,293 chief complaints collected over seven months from the RODS alerts of 2003–2004 from Utah and Pennsylvania. We used the development set to tune the parameters of CCP (e.g., find new synonyms to add, evaluate the performance of the spell-checking module using different dictionaries, etc.) and to informally evaluate our translation of the NYC algorithm to RODS' syndromes.
3. Test set—The test set comprised 10,161 chief complaints from Utah and Pennsylvania. Reference standard classifications for the test set were generated by consensus of a physician (JND) and two emergency department ICD diagnosis coders, who used case definitions we developed to classify the chief complaints into any of eight syndromes. To generate consensus classifications, we used a modification of the Delphi method in which the physician—who is an infectious disease practitioner and is the author of the syndromic definitions—determined the final classification with feedback from the coders. The ICD coders independently classified the chief complaints then came to consensus on their disagreements. We compared the ICD coders' consensus classifications with those made by the physician and presented the disagreements to the physician. The physician reviewed each disagreement and decided whether to change his classification to match that of the ICD coders or retain his original classification. We counted the number of changes made by the physician after viewing the ICD coders' classifications and used Cohen's kappa to measure agreement between the physician and the coders.

3.2. Evaluation

We evaluated classification performance of two chief complaint classifiers (CoCo and KC) with and without preprocessing. Because CoCo is a statistical system trained on manual annotations, both the test and the training set were preprocessed for evaluating the effects of preprocessing. The overall design for evaluation is shown in Fig. 4.

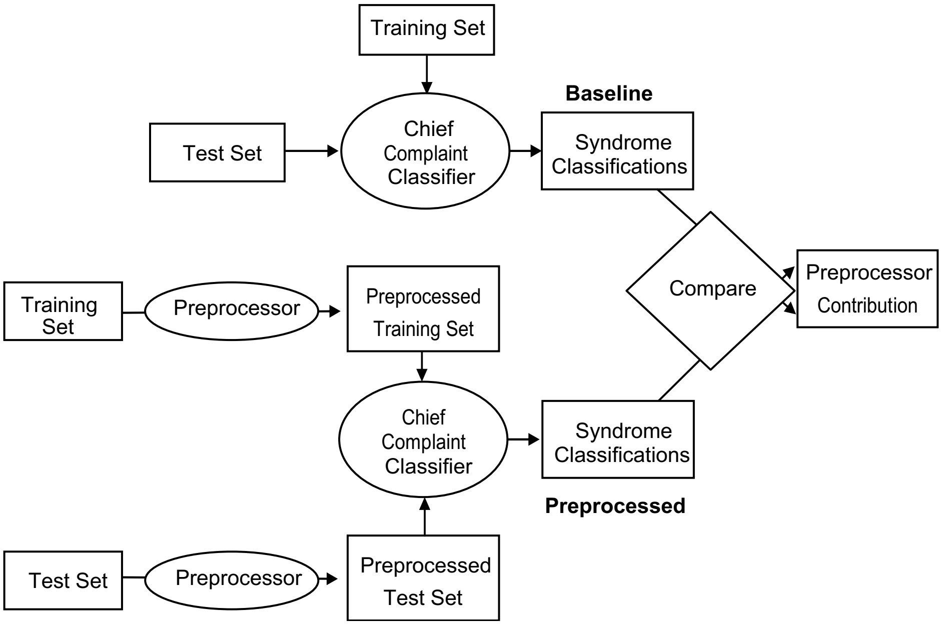

Fig. 4. Overall evaluation design comparing syndromic classifications with and without preprocessing. Only CoCo used a training set.

3.3. Outcome measures

We calculated the proportion of the 10,161 test chief complaints in the test set modified in any way by CCP and EMT-P. We analyzed CCP's mistakes by examining false positive and false negative classifications. We estimated the proportion of false positives on a set of 559 randomly selected test chief complaints changed by CCP. Author JND used domain knowledge to evaluate two linguistic changes made by CCP: (i) spelling correction and (ii) synonym replacement and abbreviation expansion. JND tabulated the number of times a change made by CCP resulted in an incorrect meaning. We estimated the proportion of false negatives on a set of 500 randomly selected chief complaints that were not changed by CCP. JND categorized a chief complaint as a false negative if the chief complaint contained a clinically relevant abbreviation or misspelling that would have been expanded by a perfect preprocessor.

We measured classification performance of both classifiers (CoCo and KC) with and without preprocessing, using the following combinations of the two preprocessors: CCP, EMT-P, CCP + EMT-P, and CCP + EMT-P (splitting module). We measured classification performance of the classifiers by calculating outcome measures for each of seven syndromes—we did not calculate outcome measures for the syndrome Other. To measure the effect of preprocessing on statistical (CoCo) and non-statistical (KC) syndrome coding systems when compared to the reference standard classification, we calculated sensitivity, specificity, and their 95% confidence intervals for every syndrome:

- *Sensitivity:* The number of correct positive classifications divided by reference standard positive classifications for a given syndrome.
- *Specificity:* The number of correct negative classifications divided by reference standard negative classifications for a given syndrome.

We also calculated the number of 10,161 chief complaints whose classification with CoCo or EMT-P changed from being incorrect to being correct or from being correct to being incorrect. Additionally, for each syndrome we calculated CoCo's sensitivity and specificity of classification after each module of CCP to determine the effect of adding that module.

Table 1
Average agreement on classifications between physician and two ICD coders on 10,161 chief complaints

Syndrome	Kappa	95% Confidence interval
Gastrointestinal	0.94	0.93–0.95
Constitutional	0.82	0.80–0.84
Respiratory	0.92	0.91–0.94
Rash	0.88	0.84–0.92
Hemmorhagic	0.87	0.84–0.90
Botulinic	0.73	0.65–0.80
Neurological	0.89	0.88–0.90
Other	0.90	0.89–0.90

4. Results

A physician and two ICD coders generated the reference standard classifications for the test set. Agreement between the physician and the coders' consensus classifications averaged 0.87 over all syndromes. Table 1 shows Cohen's kappa values for each syndrome. Of the 10,161 test cases, the physician changed the classification for 470 (4.6%) chief complaints after viewing the ICD coders' classifications, resulting in an average kappa of 0.94 between the physician's original classifications and his revised classifications.

The majority (57%) of changes the physician made were to correct mistakes in his initial classifications (270/470). Annotating text is tedious and complex, and our experience has shown that no matter how well trained an annotator is,

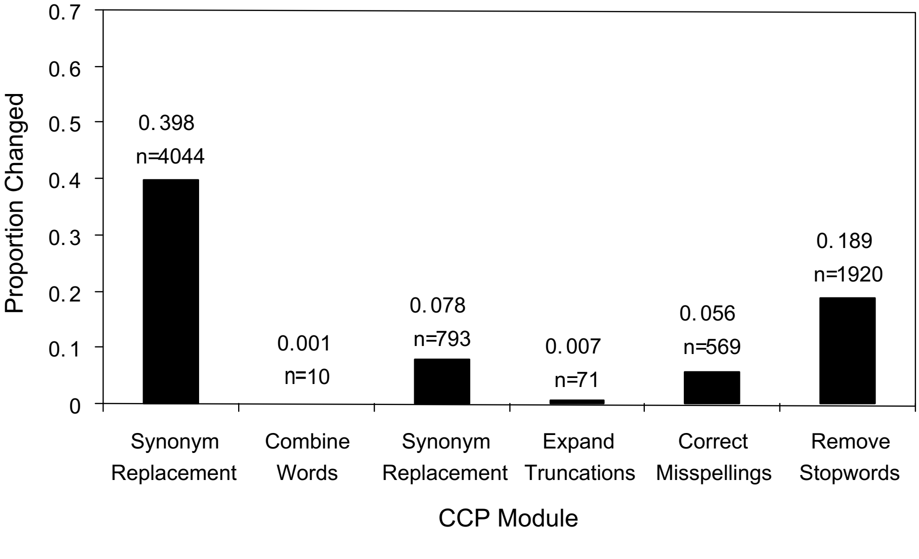

Fig. 5. Proportion of 10,161 chief complaints changed after each module of CCP.

he or she will make mistakes, which is an important reason for having multiple annotators involved in generating a reference standard classification. The physician changed his classification in 104 cases (22%) because he felt the coders' classification was more correct than his. For example, he originally classified the chief complaint "weak/confusion" as Constitutional, but the coders classified it as Constitutional and Neurological, and he agreed with their assignment of two syndromes. The remaining changes he made (20% or 96/470) were due to a change we made in our annotation protocol between the time he and the coders classified the complaints.

Of the 10,161 chief complaints in the test set, CCP changed 55% and EMT-P changed 82%. EMT-P split 10,161 complaints into 17,463 based on more than one concept appearing in the complaint. Preprocessing CoCo's training set with CCP decreased the number of unique words in the training set from 2775 to 2309. Fig. 5 shows the proportion of chief complaints changed after each module of CCP. The majority of the changes occurred in the synonym replacement and stop word removal modules. The two synonym replacement modules accounted for changes in 48% of test chief complaints, whereas stop word removal changed 19%.

Physician coauthor JND evaluated two randomly selected subsets of chief complaints. The first subset contained chief complaints changed by CCP. Of 559 complaints changed by CCP, 473 were correct changes (true positive rate of 85%) and 126 were incorrect (false positive rate of 15%). JND scored changes such as *abd pain/back pain* to *abdominal pain back pain* as correct and changes like *A FIB ANXIETY* to *a fibula anxiety* as incorrect. He ignored deletion of stop words, e.g., *follow-up* or *possible*. However, he marked word deletions that changed the meaning of the chief complaints as incorrect, e.g., changing *unable to urinate* to *urinate.* The second subset he examined contained chief complaints not changed by CCP. Of 500 chief complaints not changed, he considered 27 to be false negatives (5%). Sixteen of the 27 false negatives (59%) resulted from the single abbreviation *lac*. CCP's synonym replacement file contains 17 expansions to *laceration* but did not include *lac*. According to JND, nine other words in the 500 complaints should have been expanded: *k, g tube, od, mri, sub, injs, iv,* and *staph*. We examined CoCo's syndromic classifications of the 27 false negatives and found that 25/27 were correctly classified as Other, one was correctly classified as Neurological, and one was incorrectly classified as GI (should have been Other). These results suggest that although CCP did not change the chief complaint and should have, CoCo still classified the chief complaints correctly. We manually changed the 27 false negatives the way a perfect preprocessor would have (e.g., expanding *lac* to *laceration*) and ran the revised chief complaint through CoCo. None of CoCo's classifications differed after modifying the chief complaints.

Although CoCo outputs a probability distribution over the eight syndromic categories, in production, CoCo selects

Table 2
CoCo: sensitivity and specificity of classification for each syndrome, before preprocessing, with CCP, with EMT-P, and with a combination of CCP and EMTP

Syndrome	Sensitivity					Specificity				
	Before	CCP	EMTP	CCP + EMTP	CCP + EMTP (splitting modules)	Before	CCP	EMTP	CCP + EMTP	CCP + EMTP (splitting modules)
Botulinic	55.3	50.6	**76.5**	74.1	62.4	99.9	99.9	99.9	99.9	99.9
	44.7–65.4	40.2–61.0	**66.4–84.2**	63.9–82.2	51.7–74.9	99.9–1.0	99.8–1.0	99.8–99.9	99.8–99.9	99.8–99.9
Constitutional	50.0	53.1	**84.3**	**89.6**	**90.0**	98.9	98.8	96.9	96.7	96.7
	44.6–53.4	49.7–56.4	**81.7–86.6**	**87.4–91.5**	**87.7–91.8**	98.7–99.1	98.6–99.0	96.5–97.2	96.4–97.1	96.4–97.1
Gastrointestinal	67.2	67.1	**94.9**	**95.2**	**95.3**	99.0	99.0	98.7	98.8	98.9
	64.6–69.8	64.4–69.7	**93.6–96.1**	**93.8–96.2**	**93.9–96.3**	98.8–99.2	98.7–99.2	98.5–98.9	98.6–99.0	98.7–99.1
Hemorrhagic	60.9	66.2	**77.2**	**76.3**	**78.8**	99.7	99.7	99.3	99.3	99.2
	55.5–66.1	60.9–71.1	**72.4–81.5**	**71.4–80.6**	**74.0–82.9**	99.5–99.8	99.5–99.8	99.2–99.5	99.1–99.5	99.0–99.3
Neurological	53.9	52.4	**61.7**	**62.5**	**62.9**	98.8	98.6	98.2	98.9	98.8
	51.1–56.7	49.6–55.1	**59.0–64.4**	**59.8–65.2**	**60.2–65.6**	98.5–99.0	98.0–99.2	97.9–98.4	98.7–99.1	98.6–99.1
Rash	75.4	78.5	**90.0**	**90.0**	**90.8**	99.8	99.8	99.7	99.7	99.7
	67.3–82.0	70.6–84.7	**83.6–94.1**	**83.6–94.1**	**84.6–94.6**	99.7–99.9	99.6–99.8	99.6–99.8	99.5–99.8	99.5–99.8
Respiratory	75.5	77.9	**91.8**	**92.2**	**93.6**	99.0	99.2	98.4	98.7	98.7
	73.0–77.8	75.5–80.2	**90.1–93.2**	**90.6–93.6**	**92.0–94.8**	98.7–99.2	99.0–99.4	98.1–98.6	98.4–98.9	98.4–98.9

Bolded values do not have overlapping confidence intervals with performance before preprocessing.

Table 3
Sensitivity of CoCo after each module of CCP

Syndrome	Sensitivity						
	Before	Module 1	Module 2	Module 3	Module 4	Module 5	Module 6
Botulinic	55.3	58.8	58.8	57.6	57.6	55.3	50.6
Constitutional	50.0	54.5	54.5	52.1	52.1	52.8	53.1
Gastrointestinal	67.2	64.8	64.8	67.5	67.5	67.2	67.1
Hemorrhagic	60.9	65.2	65.2	64.6	64.6	65.2	66.2
Neurological	53.9	51.4	51.4	52.8	52.8	52.9	52.4
Other	96.9	96.6	96.6	97.0	96.9	97.0	96.8
Rash	75.4	81.5	81.5	80.0	80.0	80.0	78.5
Respiratory	75.5	76.4	76.4	78.6	78.5	78.9	77.9

Before: Split cc on space and comma.
Module 1: Synonym replacement split cc on all punctuation.
Module 2: Combine words.
Module 3: Synonym replacement.
Module 4: Truncations.
Module 5: Spell checking.
Module 6: Stop word removal—the final output of the preprocessor.

the syndrome with the highest probability as the single classification for the patient. Table 2 shows the sensitivity and specificity of CoCo's classification for each syndrome prior to and after preprocessing, along with the 95% confidence intervals. Specificity did not significantly change after preprocessing, regardless of the preprocessor. Sensitivity did not increase significantly after preprocessing with CCP. However, sensitivity increased significantly for each of the seven syndromes after applying EMT-P, showing increases between seven (Neurological syndrome) and 34 (Constitutional) percentage points. Combining CCP and EMT-P showed similar performance to EMT-P alone. Combining CCP with EMT-P's splitting module showed similar performance for most syndromes as applying EMT-P alone except for Botulinic syndrome—whose sensitivity did not increase significantly—and Constitutional syndrome—whose sensitivity was higher using CCP + EMT-P splitting module than it was with any other preprocessing combination.

To learn how each of CCP's modules affected classification performance, we measured CoCo's performance after each of CCP's modules made modifications to the chief

Table 4
Number of chief complaints whose classification outcomes from CoCo changed after preprocessing by CCP (a) and by EMT-P (b)

Syndrome	(TP –>FN)	(TN –>FP)	(FP –>TN)	(FN –>TP)	Net Gain
(a) After applying CCP					
Gastrointestinal	40	31	28	38	−5
Constitutional	37	33	27	63	20
Respiratory	46	14	34	75	49
Rash	3	11	5	7	−2
Hemorrhagic	3	9	8	20	16
Botulinic	6	4	2	2	−6
Neurological	61	18	41	42	4
Subtotal	196	120	145	247	76
Total	316		392		76
	3.1% decrease		3.9% increase		0.8% gain
(b) After applying EMT-P					
Gastrointestinal	13	68	43	352	314
Constitutional	22	214	25	312	101
Respiratory	19	88	34	214	141
Rash	0	21	8	19	6
Hemorrhagic	8	42	11	61	22
Botulinic	2	9	2	20	11
Neurological	105	109	57	201	44
Subtotal	169	551	180	1179	639
Total	720		1359		639
	7.1% decrease		13.4% increase		6.3% gain

TP, true positive; FN, false negative; FP, false positive; TN, true negative. Net change indicates the number of complaints with better outcomes after preprocessing.

Table 5
KC: sensitivity and specificity of classification for each syndrome before and after preprocessing with CCP, EMT-P, and CCP combined with EMT-P

Syndrome	Sensitivity				Specificity			
	Before	CCP	EMTP	CCP + EMTP	Before	CCP	EMTP	CCP + EMTP
Botulinic	41.1	43.5	41.2	38.8	99.9	99.9	99.9	100.0
	31.3–51.8	33.5–54.1	31.3–51.8	29.1–49.5	99.9–100	99.9–100	99.9–100	99.9–100
Constitutional	80.0	77.0	**85.3**	84.9	96.5	96.5	96.2	96.2
	77.2–82.6	72.5–78.2	**83.9–88.5**	83.4–88.1	95.8–96.6	95.0–95.8	95.8–96.6	95.8–96.6
Gastrointestinal	86.0	88.7	88.4	88.1	99.5	99.5	99.4	99.4
	83.5–87.4	86.8–90.4	84.5–90.1	86.1–89.8	99.3–99.6	99.3–99.6	99.3–99.6	99.3–99.6
Hemorrhagic	86.5	88.9	89.2	88.0	99.1	99.1	99.1	99.1
	82.3–89.8	85.1–91.9	85.4–92.6	84.0–91.1	98.9–99.3	99.0–99.3	98.8–99.2	98.9–99.3
Neurological	58.5	58.2	60.1	58.4	94.8	95.1	94.8	95.0
	55.8–61.3	55.4–60.9	57.3–62.8	57.3–62.8	94.3–95.3	94.6–95.5	93.9–94.9	94.5–95.4
Rash	78.5	78.5	80.0	80.0	99.9	99.9	99.9	99.9
	91.1–94.1	70.6–84.7	72.3–86.6	72.3–86.0	99.9–100	99.9–100	99.9–100	99.9–100
Respiratory	92.7	93.3	93.6	93.5	94.2	94.7	94.7	94.3
	91.1–94.1	91.8–94.6	92.1–94.9	91.9–94.7	93.8–94.7	94.3–95.2	94.2–95.1	93.8–94.8

Bolded values indicate a significant difference from before preprocessing.

complaints, as shown in Table 3. By measuring performance after each module, we hoped to identify modules that were critical for successful classification or modules that decreased classification performance. Unfortunately, we did not find any generalizable patterns for CCP's modules. For instance, synonym replacement (module 1) increased performance for most syndromes but decreased performance for Gastrointestinal and Neurological syndromes. Spell checking (module 5) slightly increased performance for Constitutional, Hemorrhagic, and Respiratory syndromes but decreased performance for Botulinic syndrome. None of the modules increased sensitivity by more than 4.5 percentage points, and each module caused sensitivity to decrease at least once for every syndrome.

Table 4 details the changes in CoCo's classification outcomes for the seven syndromes after preprocessing by CCP and EMT-P. The first columns (TP→FN and TN→FP) show the classification errors that occurred after preprocessing. The last columns (FP→TN and FN→TP) show the number of chief complaints whose classification improved after preprocessing. Summing across syndromes, preprocessing showed an overall classification improvement with only a small net gain of 76 (76/10161 = 0.7%) for CCP and a larger net gain (6.3%) for EMT-P. The largest gain from preprocessing is due to the number of FN classifications that became TP classifications after preprocessing with EMT-P and splitting chief complaints that have multiple problems. Table 4 also points out a surprising number of erroneous classification changes caused by preprocessing.

Table 5 shows the effect of preprocessing on the KC algorithm over the test set of 10,161 chief complaints. Based on non-overlapping confidence intervals, preprocessing significantly increased classification performance for only one syndrome-preprocessor combination: Sensitivity in classifying Constitutional syndrome increased from 80 to 85 percent when preprocessing with EMT-P.

5. Discussion

Preprocessing with CCP, which involved replacing synonyms, expanding truncations, and correcting spelling errors, did not improve classification performance for either CoCo or KC. CCP changed 55% of the chief complaints in the test set, and an analysis of a subset of the changes showed that CCP made mostly correct changes, with a true positive rate of 85%. An analysis of a subset of chief complaints not changed by CCP showed a false negative rate of 5%. However, our evaluation was more demanding than simply measuring whether the chief complaint was preprocessed correctly (e.g., corrected the misspelling or replaced the synonym correctly), because we measured whether the changes resulted in better classification performance and found no significant improvement after preprocessing.

A partial explanation for why CoCo's classification performance did not improve after preprocessing by CCP is that many of the gains CCP achieved were offset by erroneous classification changes. In fact, although CCP changed more than half of the 10,161 chief complaints, CCP only provided a net gain of 76 (Table 4). To understand the source of erroneous classifications made after processing by CCP, we examined chief complaints for Neurological and Gastrointestinal syndromes that were correct before preprocessing and incorrect after. We found that very few of the errors (5/73) were directly due to the preprocessing (e.g., changing the complaint "MS" to "mental status" when it should have been changed to "multiple sclerosis"). On the surface, CCP worked quite well. The majority (82% (60/73)) of mistakes in classification (true positives/negatives that became false negatives/positives after preprocessing) occurred for chief complaints that described multiple complaints. Given a chief complaint, CoCo selects the syndrome with the highest probability, and it appears that the relative probabilities for the co-occurring problems changed when re-training CoCo on the preprocessed training set. For instance, "blurred vision/dizziness" was classified by the reference standard as Botulinic and Neurological. Before preprocessing, CoCo classified this chief complaint as Neurological (a TP for the analysis of Neurological and a FN for the analysis of Botulinic). After preprocessing, even though the chief complaint was not altered by CCP, the complaint was classified as Botulinic (a FN for the analysis of Neurological and a TP for the analysis of Botulinic). We suspect that many of the changes in performance after preprocessing, such as the change described above, were due to changes in relative posterior probabilities after training CoCo on preprocessed chief complaints.

Preprocessing with EMT-P significantly improved CoCo's classification performance. We believe the improvement was mainly due to EMT-P's ability to split chief complaints into multiple problems. CoCo is a naïve Bayesian classifier that selects the syndrome with highest posterior probability. Assigning only a single syndrome is a limitation when chief complaints actually have multiple classifications (e.g., "headache/nausea" refers to a Neurological and GI complaint).

As a solution to the multiple classification problem, we looked at a natural language processing approach to splitting the chief complaints based on punctuations and syntax by preprocessing with EMT-P. Table 1 shows that preprocessing with EMT-P resulted in an increase in sensitivity for all syndromes. This increase was not only statistically significant but also quite large, showing a 68% increase for Constitutional syndrome (sensitivity increased from 0.50 to 0.84). CoCo's performance increased whether using the full version of EMT-P or just the splitting module, suggesting that improvement was mainly due to splitting complaints where appropriate.

CoCo's significant and substantial improvement after preprocessing with EMT-P is contingent on our choice to allow the reference standard annotators to assign multiple syndromic categories to a single chief complaint—if the reference standard classifications comprised a single classification, CoCo would have performed very well without EMT-P's splitting module. We chose to allow multiple reference standard classifications, because in reality many chief complaints represent more than one syndromic presentation. For instance, the complaint "cough/nausea" represents both respiratory and gastrointestinal complaints. If we required the annotators to select a single syndrome, they would either need to arbitrarily select a classification or select one based on some sort of hierarchical preference for a particular syndrome over another, which would have been an artificial distinction. Preprocessing with EMT-P allows CoCo to more closely represent the reality of what is presented in the chief complaint.

However, applying a preprocessing module that splits chief complaints into multiple problems could potentially be a limitation for a Bayesian classifier, because altering the words in the chief complaint (e.g., splitting the complaint "headache and cough" into two individual complaints "headache" and "cough") can reduce the evidence the classifier could potentially leverage from the full chief complaint. Therefore, preprocessing with EMT-P may reduce the potential advantage a Bayesian classifier has over a keyword classifier. We could plausibly modify the way we train CoCo to avoid losing information from splitting chief complaints. For example we could alter the values of each word in the training set from binary values of "present" or "absent" to "primary," "secondary," or "absent." In spite of the possibility of losing information by splitting chief complaints into parts, when we do so with EMT-P the net effect appears to be more helpful than harmful, as evidenced by significant improvements in sensitivity for all syndromic categories without any significant decline in specificity.

Preprocessing did not improve classification performance of the KC algorithm indicating that replacing synonyms, correcting misspellings, expanding truncations, and removing stop words may not be required to achieve good classification performance for a keyword-based system, which already includes abbreviations and word stems in its list of keywords. The KC algorithm was designed to assign multiple syndromes to a single chief complaint. Therefore, splitting chief complaints into multiple problems did not improve KC's classification performance, as indicated by the results of EMT-P in Table 3. Constitutional was the only syndrome that showed a statistically significant increase in sensitivity (based on non-overlapping 95% confidence intervals) after preprocessing.

Preprocessing with EMT-P is an important step to CoCo's achieving high sensitivity. Comparing performance and confidence intervals in Table 2 and Table 5, KC performed with higher sensitivity than CoCo before preprocessing on all syndromes except Botulinic and Neurological, which showed overlapping confidence intervals. After preprocessing with EMT-P, CoCo did not statistically differ from KC on five of the syndromes, and CoCo showed significantly higher sensitivity on Botulinic and Gastrointestinal syndromes.

6. Limitations

There are several limitations to our study. First, because we designed CCP, we focused our error analysis on CCP and did not investigatoe EMT-P's errors as extensively. Second, because the NYC algorithm maps to different syndromes than CoCo does, we adapted the algorithm and generated our own keyword classifier for comparing performance. How similar KC is to the original New York City algorithm is difficult to measure. We could not measure our JAVA version against the original SAS version, because the output is a different set of categories. We used the development set to informally evaluate the performance by looking at the output and ensuring that the program was working correctly. It would have been useful to perform a more formal analysis of our modification of the New York City algorithm on a subset of the development set by comparing the JAVA code output against reference standard output and analyzing the errors. In spite of the limitations, the KC algorithm we adapted from the NYC Syndromic Macros performed very well on our test set, suggesting a successful keyword modification to classifying into RODS' syndromes.

7. Conclusion

We applied a strict but meaningful assessment methodology for evaluating the usefulness of preprocessing chief complaints by measuring changes not only in the resulting chief complaints but also in resulting classification performance. Classification performance improved slightly with CCP for some syndromes but dropped for others, because many correct classifications became incorrect. An error analysis showed that the misclassifications were not generally directly due to the performance of the preprocessor. Chief complaints with multiple classifications represent a problem for CoCo. Errors due to selecting a single syndrome for complaints with multiple syndromes were largely corrected by using EMT-P to split relevant chief complaints into multiple complaints before classification. KC allows multiple classifications and was therefore not affected by EMT-P's splitting module. After preprocessing with EMT-P, CoCo and KC performed similarly for most syndromes, with KC achieving significantly higher sensitivity than CoCo on Hemorrhagic and CoCo achieving significantly higher sensitivity on Botulinic and Gastrointestinal.

Evaluation of preprocessing systems should not be limited to technical accuracy of the preprocessor but should include the effect of preprocessing on syndromic classification. Our results suggest that splitting chief complaints into multiple problems is important for CoCo and that other preprocessing steps only slightly improve classification performance for both CoCo and a keyword-based classifier.

Acknowledgments

This work was supported by Pennsylvania Department of Health Grant ME-01-737, NLM Grant K22 LM008301 "Natural language processing for respiratory surveillance," and NSF Grant 0325581.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2007.11.004.

References

[1] Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, Caruana RA, et al. The emerging science of very early detection of disease outbreaks. J Public Health Manag Pract 2001;7(6):51–9.

[2] Wagner MM, Robinson JM, Tsui FC, Espino JU, Hogan WR. Design of a National retail data monitor for public health surveillance. J Am Med Inform Assoc 2003;10:409–18.

[3] Hogan et al. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. JAMIA 10/6 (Nov/Dec): 555–562, 2003.

[4] Espino JU, Hogan WR, Wagner MM. Telephone triage: a timely data source for surveillance of influenza-like diseases. Annual Fall Symposium of the American Medical Informatics Association (for publication in J Am Med Inform Assoc, special supplement symposium issue), 2003.

[5] Zeng X, Wagner MM. Modeling the effects of epidemics on routinely collected data. J Am Med Inform Assoc 2002;9:S17–22.

[6] Johnson HA, Wagner MM, Hogan WR, Chapman WW, Olszewski RT, Dowling JN, Barnas G. Analysis of web access logs for surveillance of influenza. MEDINFO 2004. November 2004.

[7] Espino JU, Wagner MM. Accuracy of ICD-9-coded chief complaints and diagnoses for the detection of acute respiratory illness. Proc AMIA Symp 2001:164–8.

[8] Ivanov O, Wagner MM, Chapman WW, Olszewski RT. Accuracy of three classifies of acute gastrointestinal syndrome for syndromic surveillance. Proc AMIA Symp 2002:345–9.

[9] Tsui et al. Value of ICD-9-Coded chief complaints for detection of epidemics. JAMIA, Special Issue on Enabling Patient Safety Through Informatics 9/6 (Nov/Dec): S41–S47, 2002.

[10] Chapman WW, Dowling JN, Wagner MM. Classification of emergency department chief complaints into seven syndromes: a retrospective analysis of 527,228 patients. Ann Emerg Med 2005;46(5):445–55.

[11] Beitel AJ, Olson KL, Reis BY, Mandl KD. Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. Pediatr Emerg Care 2004;20(6):355–60.

[12] Tsui et al. Technical description of RODS: a real-time public health surveillance system. JAMIA 10/5 (Sept/Oct) 399-408, 2003.

[13] Olszewski RT. Bayesian classification of triage diagnoses for the early detection of epidemics. In: Recent advances in artificial intelligence: Proceedings of the Sixteenth International FLAIRS Conference; 2003: AAAI Press; 2003. p. 412–16.

[14] 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;36:260–70.

[15] Shapiro AR. Taming variability in free test: application to health surveillance. MMWR 2004;53:95–100.

[16] The NLM's GSpell: http://specialist.nlm.nih.gov/nls/GSpell_web/index.html.

[17] Travers DA, Haas SW. Evaluation of emergency medical text processor, a system for cleaning chief complaint data. Acad Emerg Med 2004;11(11):1170–6.

[18] Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. Emerg Infect Dis. 2004 May;10(5):858–64. Erratum in: Emerg Infect Dis. 2006 September;12(9):1472.