

Ontology-enhanced automatic chief complaint classification for syndromic surveillance [☆]

Hsin-Min Lu ^{a,*}, Daniel Zeng ^{a,b}, Lea Trujillo ^c, Ken Komatsu ^c, Hsinchun Chen ^a

^a Management Information Systems Department, The Eller College of Management, University of Arizona, 1130 E. Helen Street, Room 430, P.O. Box 210108, Tucson, AZ 85721-0108, USA

^b The Institute of Automation, The Chinese Academy of Sciences, No. 95, Zhongguanchun East Road, Beijing, China

^c Arizona Department of Health Services, Phoenix, AZ 85007, USA

Received 18 May 2007

Available online 6 September 2007

Abstract

Emergency department free-text chief complaints (CCs) are a major data source for syndromic surveillance. CCs need to be classified into syndromic categories for subsequent automatic analysis. However, the lack of a standard vocabulary and high-quality encodings of CCs hinder effective classification. This paper presents a new ontology-enhanced automatic CC classification approach. Exploiting semantic relations in a medical ontology, this approach is motivated to address the CC vocabulary variation problem in general and to meet the specific need for a classification approach capable of handling multiple sets of syndromic categories. We report an experimental study comparing our approach with two popular CC classification methods using a real-world dataset. This study indicates that our ontology-enhanced approach performs significantly better than the benchmark methods in terms of sensitivity, *F* measure, and *F2* measure.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Medical ontology; UMLS; Free-text chief complaints; Chief complaint classification; Syndromic surveillance; Bootstrapping; Statistical evaluation

1. Introduction

Syndromic surveillance aims to detect early signs of natural disease outbreaks, bioterrorism attacks, or changes in public health status in a timely manner [1]. Instead of monitoring confirmed cases or waiting for diagnostic data, syndromic surveillance focuses primarily on pre-diagnostic health-related information in an effort to significantly shorten the time needed to detect unusual events worth further investigation [2].

Emergency department (ED) triage free-text chief complaints (CCs) are short free-text phrases entered by triage

personnel describing reasons for patients' ED visits. Symptoms, diseases, mechanisms of injury, and other medical or non-medical concepts are commonly seen in CCs. ED CCs are a popular data source used by many syndromic surveillance systems because of their timeliness and availability [3–7]. CCs are among the first data elements collected for any ED visit and many hospitals increasingly have free-text CCs available in electronic form.

For automatic capture of syndromic surveillance data, free-text CC records need to be systematically classified into syndromic categories according to the symptom-related information they contain. Temporal analysis of classified results then can be used for outbreak detection. In the early stages, many diseases have similar non-specific symptoms. Grouping CCs into syndromic categories or syndromes instead of specific symptoms may provide more informative indication of potential outbreaks [5,8]. In effect, most existing syndromic surveillance systems accept

[☆] Some preliminary results of the reported research were reported in a conference paper which appeared in the Proceedings of 2006 IEEE International Conference on Systems, Man, and Cybernetics [19].

* Corresponding author.

E-mail address: hmlu@email.arizona.edu (H.-M. Lu).

this approach of classifying CCs into syndromic categories [9–11]. However, major technical challenges remain for automatic CC classification. CCs are often taken verbatim as patients describe their problems and are often individually typed as free-text entries, sometimes by trained health-care personnel and sometimes by hospital staff without medical training. This results in geographic, facility, and individual level differences in synonyms, acronyms, abbreviations, spelling, and truncations of the patients' CCs.

The issue concerning the lack of a standard vocabulary for ED CCs can be addressed in different ways. A supervised learning method which learns from the pairs of raw CCs and assigned labels can entirely bypass this problem but requires a large manually labeled training sample. Other approaches such as medical thesauri, spell checking algorithms, and manually created synonym lists have also been tried in the past [12–15]. The performance of these approaches, however, to a large degree depends on the CCs used to construct the synonym list or tune the system. These approaches may perform poorly with CCs that are different from those used in the system development and tuning. These existing approaches do not take advantage of the fact that medical terms appearing in CCs can be semantically related. We argue that by exploiting such semantic relations through the help of a medical ontology, the CC vocabulary problem [16,17] can be better handled and in turn a more effective CC syndrome classification approach can be developed.

The use of ontologies has been discussed in the context of syndromic surveillance [18]. The discussion has focused on the integration of different data sources into a unified problem-solving architecture as opposed to processing specific data sources such as free-text CCs. In this article, we propose an ontology-enhanced method to classify CCs into syndromic categories. At the core of this approach is a new grouping method based on Weighted Semantic Similarity Scores (WSSS) [19]. Utilizing the semantic relationships from a medical ontology, this method can be effectively applied to process CC terms not covered by syndrome mapping rules or past CC records with known syndromic category associations. The CC classification subsystems from two syndromic surveillance systems, Early Aberration Reporting System (EARS) and Real-time Outbreak Detection System (RODS), are chosen as the benchmarks for performance comparison.

A reference standard dataset consisting of CCs and validated classification results is of critical importance in assessing the performance of any CC classification system. In our study, such a reference standard dataset with 1000 records was constructed with help from three domain experts. This dataset was used to evaluate the performance of both our system and the benchmark systems.

The remainder of this article is organized as follows. Section 2 provides the background of the CC classification problem. The next section articulates the research opportunities and objectives. Section 4 presents the details of our technical approach. We report in Section 5 the experiments

designed to evaluate our approach. Section 6 highlights some issues about the reference standard dataset generation and the ontology-enhanced classification approach. Finally, Section 7 concludes the paper with a summary of our findings.

2. Research background

This section describes the input and output of a typical CC syndromic classification system. As part of the research background, we also briefly discuss CC coding schemes and survey various CC classification methods. Some of these methods were used as benchmarks to compare with our own approach.

2.1. Free-text chief complaints

CCs are the first records generated by triage personnel for ED patients. Examples of terms commonly seen in CCs are: nvd (nausea, vomiting, and diarrhea), fv (fever), fv w/c (fever with cough); sob (shortness of breath); so (ambiguous meaning); poss uti (possible urinary tract infection). The non-standard nature (misspellings, word variations, institution-specific use of expressions, etc.) of free-text CCs hinders their subsequent use in a syndromic surveillance system [17].

An obvious approach to deal with this significant problem is various spell checking algorithms that have been successfully applied in information retrieval research [20,21]. However, previous research reported mixed results for spell checking algorithms such as those based on edit distance or phonetic similarity. For instance, spell checking algorithms provided limited value in CC classification systems based on Bayesian networks [3]. On the other hand, combining edit distance and phonetic similarity was reported to increase sensitivity of a chief complaint classification system [12]. Since acronyms and idiosyncratic expressions are major sources of variations in free-text CCs, spell checking algorithms are only of limited value in CC processing.

2.2. Chief complaint coding schemes and medical ontologies

A coding scheme is a set of standardized terminologies into which chief complaints can be mapped. Coding schemes facilitate information retrieval, aggregation, and analysis. Two kinds of coding schemes are commonly used in public health surveillance research. The first kind is a general-purpose coding scheme, and encompasses examples such as ICD-9 CM, the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and the Unified Medical Language System (UMLS). General-purpose coding schemes usually include clinical terminology covering diseases, clinical findings and procedures. They are designed for consistently indexing, storing, and retrieving clinical data across medical practitioners and care sites.

Large collections of terminologies are usually included in these general-purpose coding schemes. For example, UMLS, developed and distributed by the US National Library of Medicine, contains 2.5 million English terms. Similarly, SNOMED CT and ICD-9 CM contain 750,000 and 20,000 terms, respectively. Terms with the same meaning are usually organized by concepts. Hierarchies are constructed to reveal the relations among concepts. A major component of the UMLS is its Metathesaurus, which combines selected coding schemes including both ICD-9 CM and SNOMED CT. Fig. 1 shows a subtree that exhibits the relations among “cramp stomach,” “upper abdominal pain,” and “epigastric pain” in the UMLS. The hierarchy in the UMLS is a valuable resource for medical information processing [22,23]. For instance, Leroy and Chen [24] demonstrated that the semantic relations among medical concepts can be used to help patients or medical experts find terms outside of their field of expertise.

The SPECIALIST lexicon, another component of the UMLS, is a general English lexicon that includes many biomedical terms. It can be used to normalize expressions such that the output text strings are in uninflected form without punctuation, genitive markers, and stop words. For example, “treating” and “treated” can be normalized to “treat” by the SPECIALIST lexicon. The SPECIALIST lexicon is a valuable tool for medical information processing. For example, Tolle and Chen [25] showed that the performance of noun phrasing improved with the addition of the SPECIALIST lexicon.

The other kind of coding scheme is more domain specific. Reason for Visit Classification (RVC) [26] provides such an example in an emergency department care setting. The National Ambulatory Medical Care Survey uses RVC to classify chief complaints into one of the more than 770 standardized entries [27]. The Canadian Emergency

Department Information System (CEDIS) workgroup proposed a coding scheme of 161 entries [28]. Similar research [14,15,29] created coding schemes that range from 20 to 228 entries.

It should be noted that a small set of standardized codes with proper synonyms/keywords can capture a majority of chief complaint records. For example, it was reported that 67% of chief complaints in testing samples can be processed using 208 keywords which correspond to 20 chief complaint groups [14]. However, moving beyond this level of performance requires a disproportional amount of keywords or synonyms. For instance, only 85.7% of training records can be processed using 2557 keywords which correspond to 228 chief complaint groups [15].

Choosing a proper coding scheme is crucial in building a flexible and effective chief complaint classification system. The coding scheme is the basic building block of CC classification systems. Coding schemes focusing on ED care settings are usually built by analyzing data collected from the field and thus can be applied relatively easily to process ED CCs. However, it is not clear how much external validity this kind of coding scheme has. General-purpose coding schemes could provide a lot of useful information, as once CCs are mapped to them, the existing semantic relations among the entries can be readily exploited to facilitate the syndromic category mapping process. In either case, there are difficulties in connecting the coding schemes and free-text CC records as none of these coding schemes can perfectly fit in CC records collected from different institutions. A possible solution involves using a combination of both types of coding schemes. For instance, a particular general-purpose coding scheme may be chosen and a customized synonym list may be built by analyzing the CCs collected from the EDs in order to bridge the gap between free-text CC and the coding schemes. The Emergency Medical Text Processor (EMT-P) system is an example of this [13,30]. It uses manually compiled synonym lists and the SPECIALIST lexicon tool to map expressions in CCs into a standardized form. The UMLS Metathesaurus then is used as a dictionary to map CCs to UMLS concepts.

2.3. Syndromic categories

There are two issues related to using syndromic categories in CC classification. First, there is no consensus about a common set of syndromic categories a system should provide [31]. Each syndromic surveillance system may have its own emphasis on the detection targets which determine the most appropriate syndrome groups and syndrome definitions. For instance, Electronic Surveillance System for Early Notification of Community-based Epidemics (ESSENCE) classifies CCs into eight syndromes: gastrointestinal, neurological, rash, respiratory, sepsis, unspecified, death, and others. RODS also has eight syndrome categories: gastrointestinal, constitutional, respiratory, rash, hemorrhagic, botulinic, neurological, and respiratory. But there is only partial overlap between the two systems' categories.

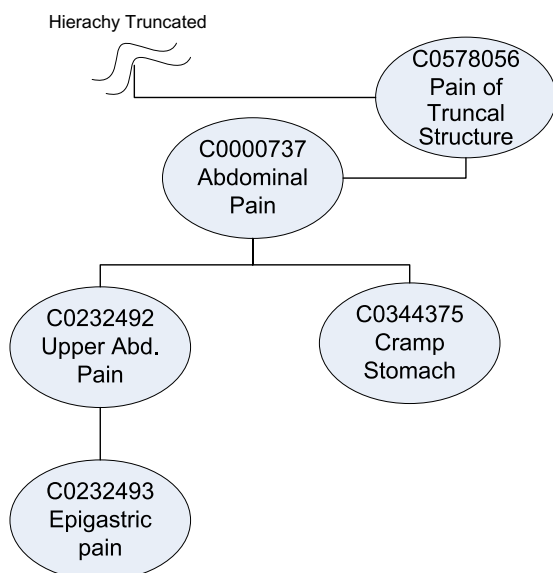


Fig. 1. An example of semantic hierarchy in the UMLS.

The syndrome coding systems of EARS and the New York City Department of Health and Mental Hygiene use 41 and 9 syndromic categories, respectively.

The existence of variations in syndromic categories implies that, if a CC classification system is designed to be widely used by many institutions, the system must be flexible, in that adding new syndromic categories or recoding from one set of syndrome definitions to another should be relatively straightforward. Most existing systems, however, have limited flexibility to support multiple sets of syndromic categories.

The second issue is related to the reliability of syndrome definitions. Syndrome assignments in the reference standard dataset are assumed to be accurate when calculating the performance of classifiers. However, syndrome assignments created using unreliable syndrome definitions may introduce errors into the reference standard dataset. As a result, additional variations are introduced into the performance measures.

It has been shown that human experts can generate a reliable reference standard dataset using broadly defined syndromic categories [5]. In the study by Chapman, Dowling, and Wagner [5], medical records were reviewed by multiple physicians and the level of agreement between physicians as measured by Cohen's kappa [32] is high for most syndromes. It should be noted, however, that the level of agreement can be low for some syndromes. The reliability of syndrome definitions, therefore, should be carefully examined before an evaluation study.

2.4. Existing automatic CC classification methods

There are two main approaches for automated CC syndrome classification: supervised learning and rule-based classification. A summary of selected syndromic surveillance systems that use CCs as one of their data sources and their underlying classification methods can be found in Table 1. The supervised learning methods require CC records to be labeled with syndromes before being used for model training. Naïve Bayesian [33,34] and Bayesian network [3] models are two examples of the supervised learning methods studied. Implementing the learning algorithms is straightforward; however, collecting training records is usually costly and time-consuming. For instance, 28,990 labeled records were used to train the RODS CoCo

naïve Bayesian classifier [10,33]. Bayesian network classifiers require fewer training records and can achieve better performance than the naïve Bayesian classifier. However, unlike most supervised learning methods, the training process was not fully automated. The system must interact with human experts to construct the semantic Bayesian network during the training process [3]. Another major disadvantage of supervised learning methods is the lack of flexibility and generalizability. Recoding for different syndromic definitions or implementing the CC classification system in an environment which is different from the one where the original labeled training data were collected could be costly.

Rule-based classification methods use a completely different approach and do not require labeled training data. Such methods typically have two stages. In the first stage, CC records are translated to an intermediate representation called “symptom groups” by either a symptom grouping table (SGT) lookup or keyword matching. For example, the ESSENCE system treats each CC as a document and each symptom group as a query. Symptom grouping, then, consists of running queries against CCs [35].

In the second stage, a set of rules is used to map the intermediate symptom groups to final syndromic categories. For instance, the standard EARS system uses 42 rules for such mappings.

A major advantage of rule-based classification methods is their simplicity. The syndrome classification rules and intermediate SGTs can be constructed using a top-down approach. The “white box” nature of these methods makes system maintenance and fine tuning easy for system designers and users. In addition, these methods are flexible: adding new syndromic categories or changing syndromic definitions can be achieved relatively easily by switching the inference rules.

A major problem with the rule-based classification methods is that they cannot handle symptoms that are not included in the SGTs. For example, a rule-based system may have a SGT containing the symptoms “abdominal pain” and “stomach ache” which belong to the symptom group “abd_pain”. This system will not be able to handle “epigastric pain” even though “epigastric pain” is closely related to “abdominal pain”. Our research is designed to address this vocabulary problem using an ontology-enhanced approach.

Table 1
Major CC classification methods

Methods	Systems	Related research
Rule-based method		
Keyword match, synonym list, mapping rules	DOHMH syndrome coding system	Mikosz et al. [51]
Same as above	EARS	Hutwagner et al. [52]
Weighted keyword match (vector cosine method), mapping rule	ESSENCE	Sniegowski [35]
Supervised learning		
Naïve Bayesian	RODS	Olszewski [33] and Espino et al. [34]
Bayesian network	N/A	Chapman et al. [3]

3. Research opportunities and objectives

Our review of existing CC classification methods reveals several research opportunities. First, the UMLS contains meaningful relations between symptoms that could be potentially leveraged in a CC classification system. Knowledge captured in existing CC classification methods, either learned through training samples or acquired directly from human experts, could be enhanced by these relations. However, most existing research ignores such semantic information. Second, the lack of one common standard for syndromic categories calls for an architecture which can support flexible syndromic categories. Finally, the existing term processing and syndromic classification research has resulted in concrete findings and system components that should be leveraged and reused when developing new approaches.

Based on these observations, our research is aimed at developing a novel free-text CC classification approach that can leverage a medical ontology to improve classification effectiveness.

4. An ontology-enhanced chief complaint classification approach

This section reports a new ontology-enhanced CC classification approach that meets the research objective discussed in the previous section. We first discuss its basic design and then discuss its major technical components.

4.1. A rule-based design

Our approach largely follows a rule-based design as opposed to a supervised learning method. As argued before, a rule-based method requires less training data and is flexible in incorporating new syndromic categories. Our approach will address its key weakness, i.e. the diffi-

culty associated with handling symptoms not previously encountered, by making use of semantic information contained in the UMLS ontology.

As depicted in Fig. 2, our syndromic classification approach can be divided into three major stages: CC standardization, symptom grouping, and syndrome classification. Central to our approach is the Weighted Semantic Similarity Score (WSSS)-based grouping component that automatically expands the coverage of the symptom grouping table by exploiting the semantic relations between symptoms. In the remainder of this section, we first introduce the symptom grouping table and then discuss the three major stages of our approach in turn.

4.2. The symptom grouping table (SGT)

A symptom grouping table records the mapping relations from symptoms to symptom groups. Symptoms to be classified in the same syndromic category are grouped together in a symptom group. For instance, nausea, vomiting, and sickness all point to the same gastrointestinal syndrome and thus are grouped together. Note that the granularity of symptom grouping depends on the final syndrome definitions. For example, if we are interested in respiratory syndrome only, the symptoms apnea, difficulty breathing, gasping, and hemoptysis can all be grouped together. However, if we also consider the hemorrhagic syndrome in addition to the respiratory syndrome, then the original symptom group must be broken down into two: one containing apnea, difficulty breathing, and gasping; and the other containing hemoptysis. Syndrome mapping rules can then be constructed so that the first group is mapped into the respiratory syndrome and the latter into both the hemorrhagic and respiratory syndrome.

Ideally, each and every symptom can only be mapped into one symptom group. Example entries in a SGT can be found in Table 2. The symptoms in the SGT are stored

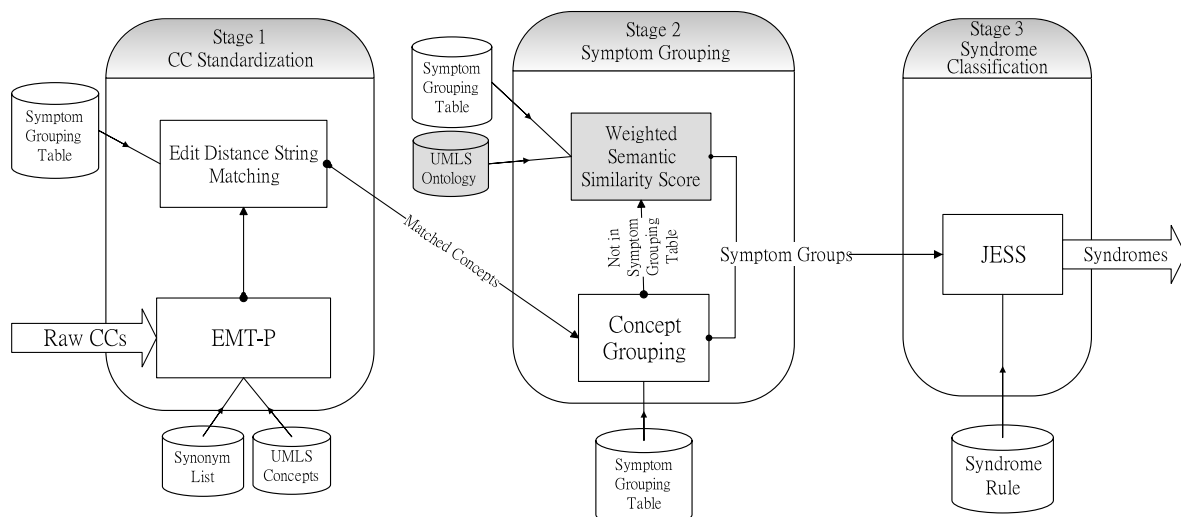


Fig. 2. System design for an ontology-enhanced chief complaint classification approach.

Table 2
Selected records in a symptom grouping table

Symptom group	Concept unique ID	Symptom name
Bleeding	C0019080	Bleeding
	C0017565	Bleeding gums
nvd ^a	C0151594	Bloody diarrhea
	C0011991	Diarrhea
	C0027497	Nausea
	C0027498	Nausea vomit
	C0221423	Sickness
	C0277525	Stomach flu

^a nvd stands for “nausea, vomiting, and diarrhea.”

in their standardized form following the underlying coding scheme, in our case, UMLS. For example, the second row in Table 2 indicates that the symptom “bleeding gums,” with a unique id C0017565 in UMLS, belongs to the group “bleeding.”

The SGT used in this study contains 61 groups and 392 symptoms (more discussion about the construction of the SGT can be found in “System Benchmarks”). In our study, it is implemented as a relational table with three fields: concept name, concept unique ID, and symptom group. The symptom grouping process identifies symptom groups that match the concept unique IDs from a CC.

4.3. Stage One: chief complaint standardization

In Stage One, the acronyms, abbreviations, and truncations in CC records are expanded and normalized using synonym lists, the SPECIALIST lexicon tool, and edit distance string matching. Then the standardized symptoms extracted from CCs are mapped to UMLS concepts. The EMT-P module is capable of expanding acronyms and truncation using synonym lists and the SPECILAIIST lexicon tool and is reused as a plug-in module in our system.

The EMT-P, nevertheless, has two shortcomings in this application. First, it cannot handle a simple typographical error such as “sore thorat” or word concatenation such as “sorethroat.” Second, the EMT-P does not consider symptoms in the current SGT as more relevant to the CC classification task and may decide to cut CCs into one or more concepts in its own way.

The edit distance string matching module is designed to address the first shortcoming. For each string that cannot be processed by the EMT-P, the similarity between terms in the SGT and the unrecognized string is calculated. The unrecognized string is deemed as similar to a term in the SGT if each word in the term can find a counterpart in the unrecognized string within a “small” distance and these words appear in the same order as they do in the SGT. Edit distance is considered small if: (a) the distance is zero; or (b) the word (in the SGT term) has more than five characters and the edit distance is one; or (c) two words have the same length, contain more than five characters, and have an edit distance of two. For example, for the unknown

string “sore thorat”, the term “sore throat” in the SGT is similar to it because “sore” and “throat” can find their counterpart “sore” and “thorat” in the unknown string in the same order as they appear in the SGT (that is, “thorat sore” would not be considered similar to “sore throat”).

EMT-P fails to process CC records with concatenated words, those formed by a group of words without any dividing signposts such as spaces or hyphens. For example, EMT-P maps “sore throat” to a UMLS concept successfully but fails to map “sorethroat” to the same concept. We use a simple approach to correct this problem for matching purposes: for each term in the SGT, we produce a concatenated word by linking all words of the term. This concatenated word is then used to match unknown strings.

Finally, as the terms in current SGT were created by domain experts familiar with target CCs of the classification system, the terms in the SGT should have higher priority over those found in the UMLS. (EMT-P does not treat the terms in the SGT differently from the UMLS concepts when determining how an expression should be divided into concepts). In our approach, we added another step searching the EMT-P *output* for terms in the SGT. The benefit of this step is that once part of a CC can be mapped to a term in the SGT, the grouping and subsequent syndrome classification can be done routinely. As such, the chance of finding any potential match to the current SGT is maximized. For instance, “arm injury” was mapped to one single symptom by EMT-P, as it prefers longer symptoms. In our approach, however, since “injury” appears in the SGT, the same record is standardized into both “arm injury” and “injury”.

Given that multiple methods are used to extract concepts in CCs, it is possible that some concepts come from overlapped terms. If multiple matches are found and terms from one match are contained in terms from the other match, these embedded shorter terms are dropped. For instance, if both “blood” and “blood pressure” are matched to “increased blood pressure sweat”, then “blood” is dropped because it is part of the term “blood pressure”.

To further illustrate the procedures used in Stage One, we discuss several additional examples. The first example is the free-text CC “DIARRHEA ABD CRAMPING”. The EMT-P component is first invoked and identifies this CC as two concepts in the UMLS: abdominal cramp (C0000729) and diarrhea (C0011991). The text strings in parentheses are the unique concept IDs in the UMLS. The entire free-text CC is successfully mapped to the UMLS concept. The edit distance search and word concatenation search are thus skipped. Terms in the SGT are used to search in “abdominal cramp” and “diarrhea” but no new concepts are found. The final output of step one is two concepts: abdominal cramp and diarrhea.

The second example is the free-text CC “STIFF NECK, UPPER SPINE PAIN”. EMT-P identified “stiff neck” (C0151315) as one UMLS concept but marked “upper

spine pain” as unidentified. The edit distance and word concatenation searches are invoked for the latter text strings. It turns out that no additional concept is identified. Finally, terms in the SGT are used to search new concepts in “stiff neck” and “upper spine pain” but no new concept is found. The final output has only one concept: stiff neck.

The third example is “SORE THORAT”. EMT-P failed to identify any UMLS concept from the input string. The edit distance identifies the string as similar to the concept “sore throat” (C0242429) in the SGT according to the rules described above. The word concatenation search does not find additional concepts. Finally, terms in the SGT are used to search the unmatched EMT-P output “sore thorat” but without a match. The final output in this case is “sore throat.”

4.4. Stage two: symptom grouping

In the second stage, each symptom extracted from the previous stage is mapped into an appropriate symptom group. As discussed before, symptom groups are intermediate representations that can enable system modularity, extensibility, and flexibility.

Our system uses the SGT to match symptoms to groups. If a corresponding group is located in the SGT, the grouping process terminates. However, it is likely that some symptoms do not directly appear in the SGT. In a traditional rule-based system design, the system simply ignores the unmatched symptoms.

The main technical innovation of our research is the development of an ontology-enhanced approach to process these unmatched symptoms. The basic intuition behind our approach is as follows. Unmatched symptoms may be semantically related to some symptoms in the SGT. If the semantic relations between these symptoms can be exploited, the system will be able to process the unmatched symptoms using the original SGT. In other words, the ontology can help expand the coverage of SGTs automatically.

At the center of our approach is the Weighted Semantic Similarity Score (WSSS). This score is based on the semantic distance between two concepts, defined as the path distance between them in the UMLS hierarchy. The specific definition of path distance and related computation are described below.

The path distance calculation involves four steps. Since each symptom may have more than one synonym, the first step is to identify all synonyms of the two symptoms between which semantic similarity is to be determined. Next, all ancestor nodes of identified synonyms are located. As the UMLS stores the concept hierarchy in a relational table, locating these ancestor nodes takes only one query which is efficient to execute. Third, the distance between a pair of terms is calculated by comparing their ancestor nodes. After calculating the distances of all possible pairs of synonyms from the two symptoms, the shortest distance is returned as the distance between the two symptoms.

Fig. 3 provides pseudocode for calculating the path distance between two concepts in the UMLS. For example, “swelling” and “abd. swelling” has a parent-and-child relation in the UMLS; the distance between these two symptoms is one. “Dysphagia” and “bloating” have a common parent “symptoms involving digestive system”; the distance between them is two.

Given a symptom C1 that is not in the SGT, the semantic distances from C1 to all symptoms in the SGT can be calculated. All distances are sorted in ascending order and distances larger than a threshold Z are discarded. The retained distances are then grouped together based on symptom groups. The WSSS measuring the “fitness” between C1 and all candidate symptom groups is then calculated by adding the reciprocals of semantic distance.

Formally, we define d_{ij} as the distance between C1 and symptom j in group i , and S_{zi} as the set of distances that satisfies threshold Z and belong to group i . Then the WSSS for group i of order Z is defined as

$$w_{zi} = \sum_{d_{ij} \in S_{zi}} \frac{1}{d_{ij}}$$

The threshold Z starts at one. The symptom group with the highest score is chosen as the resulting group for the unmatched symptom if only one group meets the threshold requirement. If two or more groups have the same WSSS, Z is increased by one. This process repeats itself until Z is too large to reveal a meaningful relation between the unrecognized symptom and groups in the SGT. Preliminary experiments showed that two symptoms with distances larger than four are usually related in a very weak manner. Thus the above process is repeated until Z is larger than four. It is possible that the WSSS will not result in a match if none of the groups can meet the threshold distance requirement.

We now use several real examples to illustrate the WSSS calculation process and how it is used to determine symptom group assignment. For example, the symptom “gall bladder pain” does not have a direct match in the SGT. By calculating semantic distances, we find that “abdominal pain” is the closest symptom in the SGT. The symptoms “cramp stomach” and “bladder pain” are the next closest. Table 3 lists the top 10 closest symptoms to “gall bladder pain”. Clearly, the symptom group “gi” (gastrointestinal) has the highest score with threshold Z equaling one. As a result, “gall bladder pain” is assigned to group “gi”. Another example can be found in Table 4. The unknown concept “groin swelling” cannot be matched with any symptom group in the SGT with the distance threshold set to 1; therefore the threshold is extended to two. Seven concepts satisfy the new threshold. The group “swelling” has three concepts with a distance equaling two. Thus the WSSS is $1/2 + 1/2 + 1/2 = 1.5$. Group “gi” has two concepts with a distance of two, and has a WSSS of $1/2 + 1/2 = 1$. The third group, “limbs_pain”, has one concept with a distance of two and a WSSS value of 0.5. The last

Semantic_Distance(C1, C2)

```

SET shortest_distance = a_large_number

SET syn_set1 = all synonyms of C1

SET syn_set2 = all synonyms of C2

FOR each syn1 in syn_set1

    FOR each syn2 in syn_set2

        CALCULATE all ancestors of syn1 RETURNING ancestor1

        CALCULATE all ancestors of syn2 RETURNING ancestor2

        CALCULATE the distance of syn1 and syn2 with ancestor1 and ancestor2

        RETURNING distance

        IF distance < shortest_distance THEN

            SET shortest_distance = distance

        ENDIF

    ENDFOR

ENDFOR

RETURN shortest_distance
    
```

Fig. 3. Pseudocode for calculating the semantic distance between two concepts C1 and C2 in the UMLS.

Table 3
Top 10 SGT symptoms closest to “gall bladder pain”

Distance	Symptom	Group
1	Abdominal pain	gi
2	Bladder pain	gi
2	Cramp stomach	gi
2	Left sided abdominal pain	gi
2	Lower abdominal pain	gi
2	Rectal pain	gi
2	Right sided abdominal pain	gi
2	Stomach ache	gi
2	Upper abdominal pain	gi
2	Groin pain	limbs_pain
(List truncated)		

Table 4
Top eight SGT symptoms closest to “groin swelling”

Distance	Symptom	Group
2	Arm swell	Swell
2	Groin lump	Swell
2	Leg swelling	Swell
2	Abdominal pain	gi
2	Abdominal swelling	gi
2	Leg pain	limbs_pain
2	Bradycardia	Tachycardia
3	Abscess	Cellulitis leg
(List truncated)		

group, “tachycardia”, also has a WSSS value of 0.5. As the group “swelling” has the highest WSSS, the unknown concept “groin swelling” is assigned to the group “swelling”.

One might argue that the closest symptom based on the semantic distance calculation could well serve as the best

match for the unmatched symptom under investigation. Based on our computational experience, however, this simplistic design can lead to rather arbitrary results, as typically the unmatched symptom could have multiple closely related SGT symptoms which suggest different groups. Our WSSS-based approach is designed to mitigate such

ambiguous situations. It can be viewed as a weighted voting scheme to determine the best group.

Though the method of grouping un-encountered symptoms using the WSSS is relatively new, it is conceptually similar to the widely used nearest-neighborhood methods (see, for example, [36]). Given a point X to be classified, the nearest-neighborhood method searches k points which are closest to the point X from the training dataset and assigns the classification result by the majority class of the k points. In our approach, we map the expressions in CCs to a concept space constructed by the UMLS and define the distance metric based on the UMLS. The SGT then serves as the training dataset in the nearest-neighborhood method to determine the classification result for symptoms that the system has not encountered before.

4.5. Stage three: syndrome classification

In the last stage, the system decides to which syndrome the CC belongs. This is done by mapping the symptom groups obtained from Stage Two to predefined syndromic categories using mapping rules. In our implementation a rule inference engine, JESS (<http://herzberg.ca.sandia.gov/jess/>), is employed.

As an example, Rule 1 in Fig. 4 dictates that CCs belonging to symptom groups gastrointestinal (gi); gastrointestinal bleeding (gi_bleed); nausea, vomiting, and diarrhea (nvd); constipation; or jaundice, but not caused by motor vehicle accident (mva), are assigned to gastrointestinal syndrome (GI_CAT). Similarly, Rule 2 in Fig. 4 dictates that CCs that belong to rash or hemorrhagic rash are classified into rash syndrome.

Note that all symptom groups from the same CC are combined to determine the syndrome classification results. For example, the raw CC “SOB AND NAUSEA” is standardized into “shortness of breath” and “nausea”. In Stage Two, “shortness of breath” is grouped into “respiratory”

and “nausea” is grouped into “nvd”. In the final stage the two groups, “respiratory” and “nvd”, are considered simultaneously and the CC is classified into two syndromic categories: respiratory syndrome and gastrointestinal syndrome.

In our implementation, the rule set is stored in a plain text file. This rule set consists of two parts. Rules in the first part encode the main logic behind the mapping from symptom groups to syndromes. Examples from this part can be found in Fig. 4. There are 17 such rules in total. Rules in the second part dictate the priority of syndrome assignments. For example, one rule in this part states that the “other” syndrome will be dropped if it is not the only syndrome identified. There are eight rules in the second part.

The rule set can be changed and updated easily and independently when new syndromic categories are needed. For instance, if a new syndrome, “febrile gastrointestinal”, needs to be added to existing syndromic categories, the user only needs to add one more rule to the rule file that combines the symptom groups involving the gastrointestinal syndrome and fever symptoms using an “and” operation. This design provides flexibility and extensibility for the system to meet the changing needs of syndromic surveillance.

In a full-fledged system, this simple approach of capturing rules in a plain text file could lead to scalability and maintenance issues as the rule set grows. A more formal, structured approach in dealing with such rules might be needed. However, since the rule set is built upon the symptom groups instead of individual symptoms, the number of rules is not likely to be large. This is another advantage of our symptom group-based approach.

5. An experimental study

In this section, we report on an experimental study conducted to evaluate the ontology-enhanced CC classification approach with respect to a human generated reference

```

Rule 1: (defrule s_gi "Gastrointestinal" (and (or (gi) (gi_bleed) (nvd) (constipation)
(jaundice)) (not (mva))) => (record GI_CAT))

Rule 2: (defrule s_rash "Rash" (or (rash) (hemorrhagic_rash)) => (record RASH_CAT))

Rule 3: (defrule s_botu "Botulism-like" (or (blurred_vision) (dysphagia) (paralysis)) =>
(record BOTU_CAT))

Rule 4: (defrule s_hemo "Hemorrhagic" (and (or (bleeding) (hematemesis) (hemoptysis)
(hemorrhagic_rash) (gi_bleed)) (and (not (chronic)))) => (record HEMO_CAT))

```

Fig. 4. Selected syndrome mapping rules.

standard dataset [37]. The evaluation focuses on the usefulness of the WSSS component and the performance difference between the ontology-enhanced system and the supervised learning system. The CC classification subsystems from two syndromic surveillance systems were chosen as benchmarks: Early Aberration Reporting System (EARS) and Real-time Outbreak Detection System (RODS). EARS is a syndromic surveillance system developed by the CDC after the terrorist attacks of September 11, 2001. Major data sources monitored by EARS include free-text CCs, 911 calls, school and business absenteeism, and OTC drug sales. RODS, developed by the University of Pittsburgh, was designed for detection and assessment of disease outbreaks. Similarly, data sources such as free-text CCs and OTC drug sales are monitored. RODS is used by more than 12 states in the US.

We first discuss the performance measures employed in our study and the statistical procedure used to test the performance differences between our approach and the two benchmarks. We then discuss the research test bed, reference standard dataset, and syndromic definitions used in this study. The last subsection reports our experimental findings.

5.1. Performance criteria

Sensitivity, specificity, and positive predictive value (PPV) have all been used extensively in previous research [3,4,33]. In addition we also consider the F measure and $F2$ measure [38,39]. The F and $F2$ measures, commonly used in the information retrieval literature, combine PPV and sensitivity to provide a single integrated measure to evaluate the overall performance of a given approach. The goal of a syndromic surveillance system is to detect disease outbreaks while minimizing the false alarm rate. This corresponds to high level of PPV and specificity. Excessive false alarms could happen if these two measures do not reach the desired levels. Sensitivity summarizes the portion of positive cases that can be captured by the classification system and can be linked to higher detection power. However, higher sensitivity could lead to lowered PPV and specificity and increases the false alarm rate. The F measure family is one way of characterizing the trade-off between detection power and the false alarm rate. In this family of measure, the F measure is the harmonic mean of sensitivity and PPV and thus can be interpreted as a measure that considers sensitivity and PPV equally important. The $F2$ measure assigns sensitivity twice as much weight as PPV and can be interpreted as a measure that is biased toward sensitivity. It should be noted that specificity is not included in the F measure and $F2$ measure calculations.

McNemar's test [40–42] is useful in determining whether two systems have the same level of accuracy. When considering only the positive cases of the same syndrome in the reference standard dataset, McNemar's test can also provide a statistical test for sensitivity comparison. A similar technique applies for specificity.

Unfortunately, this technique cannot be applied to PPV. Unlike sensitivity, the denominators of PPV, which are the positively classified CCs from the two systems, only partially overlap in most cases. Moreover, paired or independent comparisons are not applicable due to violated assumptions. The F measure and $F2$ measure, which encompass PPV, also lack proper statistical tests.

To overcome these difficulties, we apply the bootstrapping method for statistical inference on PPV, sensitivity, specificity, F measure, and $F2$ measure. Bootstrapping [43] is a general-purpose re-sampling technique for assessing statistical accuracy. The basic idea behind bootstrapping is to use the empirical distribution function obtained through the sample on hand to generate bootstrapping samples that in turn provide the sampling distribution of the statistics of interest.

Before proceeding with the detailed bootstrapping procedure used in this study, a clarification is in order. In the statistical learning literature, bootstrapping usually involves both system training and testing (see, for example, [44]). For instance, the bootstrapping method developed by Efron [45] estimates the system error rate by the weighted average of the training error using bootstrap sample and the testing error using instances not in the training sample. However, in a setting where training a benchmark system is not practical or where the training process is not fully automated, bootstrapping for both training and testing would be inappropriate.

In this study, our system and the benchmark systems are evaluated using the same reference standard dataset. The point estimator of various performance criteria can be calculated. A confidence interval of performance difference is required for statistical inference. The general bootstrapping procedure [43,46,47] can produce the confidence intervals for all performance criteria of interest. More specifically, we are interested in testing the null hypothesis $H_0: g_{\text{BioPortal}} - g_{\text{Benchmark}} \leq 0$ against the alternative hypothesis $H_1: g_{\text{BioPortal}} - g_{\text{Benchmark}} > 0$, where $g_{\text{BioPortal}}$ is the performance of our system (referred to as BioPortal) under a particular criterion, and $g_{\text{Benchmark}}$ is the performance of one of the benchmark systems under the same criterion.

This problem is equivalent to testing whether the performance difference $d = g_{\text{BioPortal}} - g_{\text{Benchmark}}$ is smaller than or equal to zero. Since this is a one-sided test, the hypothesis is rejected at $1 - \alpha$ confidence level if the $1 - 2\alpha$ level confidence interval is all positive. We define a bootstrap sample as a random sample with replacements from the original reference standard dataset of the same sample size as the original reference standard dataset. Then for bootstrap sample i , $i = 1, 2, \dots, B$, we calculate the performance difference d_i . The $1 - 2\alpha$ level confidence interval of d is then the interval covering the α percentile and $1 - \alpha$ percentile of $\{d_i\}$, $i = 1, 2, \dots, B$. A step by step procedure can be found in Fig. 5.

An important control parameter of our bootstrapping procedure is the total number of bootstrap samples, B .

1. Set the counter $i = 1$ and the total number of bootstrap samples $B=2500^*$.
 2. From the testing dataset of size n , draw a random sample with replacement of size n .
 3. Calculate the performance of our system $g_{BioPortal,i}$ and the benchmark system $g_{Benchmark,i}$ using the sample from the previous step.
 4. Calculate the difference $d_i = g_{BioPortal,i} - g_{Benchmark,i}$.
 5. Increase i by one.
 6. If $i \leq B$, repeat Steps 2-5.
 7. The $1-2\alpha$ level confidence interval is the interval covering the α percentile and $1-\alpha$ percentile of $\{d_i\}$.
 8. The null hypothesis $H_0 : g_{BioPortal} - g_{Benchmark} \leq 0$ is rejected at confidence level $1-\alpha$ if the $1-2\alpha$ level confidence interval is all positive.
- * See the discussion in the Performance Criteria section for the guideline of choosing B .

Fig. 5. Bootstrapping procedure for performance comparison.

When building confidence intervals using the bootstrapping method, B is typically set at larger than one thousand [46]. Through computational experiments, we also observed that the evaluation results become stable when B is set to a number larger than one thousand. In our study, we have chosen a conservative setting for B , 2500. Since the bootstrapping method as discussed above is generally applicable to evaluating system performance along all performance criteria used in our study, the subsequent analyses are mainly based on bootstrapping.

The bootstrapping method discussed above has not been used widely in previous public health surveillance studies. As a comparison framework, it can be applied to any similar research involving performance comparison between two systems. To ensure correct inference, the only assumption needed is the independence of the records in the reference standard dataset. It is our intended contribution to advocate this type of method for more rigorous studies of performance comparison between different surveillance methods.

5.2. Research test bed

The CC records used in this study were provided by the Phoenix Metropolitan Hospital through the Arizona Department of Health Services. The training dataset con-

tains 2256 CC records covering an interval of 11 days. The string length of records varies from 1 to 32 characters. The testing dataset is a random sample of one thousand records from July 2005 to November 2005, excluding the time interval when training data was collected. As the focus of this study is on improving the effectiveness of a CC classifier using a medical ontology, we are more interested in how the performance differs on distinct records as opposed to providing an unbiased estimation of classification performance. Therefore, duplicated chief complaint strings were removed before performing the random sampling.

The training dataset was used during the system development process and also for system tuning. The testing dataset was used to generate the reference standard dataset for system performance evaluation.

5.3. Syndromic definitions and reference standard dataset

Eleven syndromes were chosen by the Arizona Department of Health Service for evaluation: botulism, constitutional, gastrointestinal, hemorrhagic, neurological, rash, respiratory, upper respiratory, lower respiratory, fever, and other (the syndromic definitions used in this study can be found in Appendix). “Other” is a miscellaneous category for CCs that do not fit into any of the other

syndromes. One chief complaint could be assigned to more than one syndrome. If upper respiratory or lower respiratory is assigned, it automatically implies respiratory syndrome (but the reverse is not true).

To ensure that the syndrome definitions used were comparable to the benchmarks, text descriptions for each syndrome were compiled based on those used by the RODS Laboratory [3]. A mapping table of syndrome assigned by EARS, another benchmark system, to syndromes used by our system was constructed based on the descriptions. All syndromes except constitutional were successfully linked to the EARS syndromes. Table 5 lists the mapping from the EARS and RODS syndromes to those used in this study. More detailed discussion about the benchmark systems can be found in the next section.

To the best of our knowledge, there are no publicly available CC datasets labeled with syndrome definitions. Thus, the reference standard dataset for the system evaluation had to be constructed for this study. Three experts, including two physicians and one nurse in Phoenix, Arizona, were given the syndrome definitions and the testing set of one thousand chief complaints. They were asked to assign CCs to syndromes independently. After collecting the assignments from the experts, majority voting was used to determine the final syndrome assignment of each CC.

Out of these 1000 records, majority voting could not determine the syndrome assignment for 18 CCs that all three experts labeled differently. In these cases, a fourth expert, an emergency department physician, helped determine the final assignment. Another 85 CCs that had syndrome assignments mixed with “other” were also reviewed by this fourth expert. The final reference standard dataset contains 148 CCs that have been assigned to more than one syndrome. On average, one CC is assigned to 1.18 syndromes.

The prevalence of the 11 syndromes can be found in Fig. 6. Similar to previous research [3,33], the “other” category has the highest prevalence. Syndromes such as respiratory, gastrointestinal, and neurological have a prevalence of about 10%. Botulism has the lowest prevalence, at 0.6%.

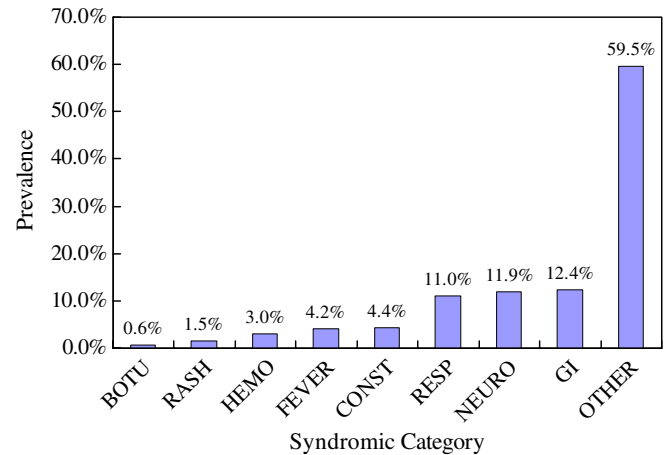


Fig. 6. Syndrome prevalence.

The kappa statistic is calculated using the assignment from the first three experts. The overall agreement is good (kappa = 0.71). Table 6 summarizes the kappa statistic of each syndromic category. Some syndromic categories such as botulism, constitutional, and lower respiratory syndromes have low agreement, while the fever and neurological syndromes have moderate agreement (according to the standard proposed by [48, p. 218]). The low kappa value for botulism syndrome may be due to its low prevalence. It is very difficult to have a large kappa value for a rare syndrome because a few disagreements can strongly influence the kappa value. In order to have reliable estimation of system performance, only syndromes with excellent agreement (kappa higher than 0.75) were used in our evaluation study.

5.4. System benchmarks

The CC classification subsystems of RODS and EARS serve as the benchmarks to compare against our ontology-enhanced approach. RODS uses its own CoCo naïve Bayesian classifier and is treated as a black-box CC classification method for the evaluation [33]. It is referred to as the CoCo naïve Bayesian classifier (CoCoNBC) in subsequent discussion. The CC classification subsystem of

Table 5
Syndrome mapping between the BioPortal system and the benchmark systems

BioPortal	EARS	RODS
Botulism-like	s_botulism	Botulism-like
Constitutional	N/A	Constitutional
Gastrointestinal	s_gastrointestinal, s_gicat	Gastrointestinal
Hemorrhagic	s_hemorrhagic	Hemorrhagic
Neurological	s_neurons, s_neurological	Neurological
Rash	s_rashcat	Rash
Respiratory	Upper respiratory, lower respiratory	Respiratory
Upper respiratory	s_upperresp, s_sb_upper_respiratory	N/A
Lower respiratory	s_lowerresp, s_sb_lower_respiratory	N/A
Fever	s_fever, s_febrile	N/A

Table 6
Kappa statistics of each syndromic category

Syndrome	Kappa
Botulism-like	0.22
Constitutional	0.24
Lower respiratory	0.38
Fever	0.46
Neurological	0.64
Other	0.74
Upper respiratory	0.77
Respiratory	0.80
Hemorrhagic	0.81
Rash	0.82
Gastrointestinal	0.85
Overall	0.71

EARS, on the other hand, is a rule-based classification system that shares some common architectural design elements with ours. It is referred to as EARS CC classification subsystem (ECCCS). ECCCS uses a symptom table to map raw chief complaints into groups and a set of rules to assign syndromes. In effect, the symptom table from ECCCS was used to construct the initial SGT of our system as follows. For symptoms listed in the ECCCS symptom table, EMT-P was used to standardize them into UMLS concepts. We took care to merge any redundant groups. For example, symptoms in “Poisoning” are very similar to those in “CO Poisoning”. In another example, there is no clear distinction between “Death” and “Unexplained Death” and they were merged together. The final SGT in our approach contains 61 groups and 392 symptoms. To ensure a fair comparison, the rule set from ECCCS was amended based on our syndrome definitions to construct the initial rule set for our system.

The setting using the SGT and rule set adapted from ECCCS is referred to as ECCCS in BioPortal. The BioPortal project is an infectious disease informatics project with funding support from the National Science Foundation and other federal state agencies. The reported research is part of this project [11,49,50]. ECCCS in BioPortal and ECCCS share the common symptom grouping table and compatible syndrome rules. As such, we can examine the effect of the WSSS component in isolation and fairness. Comparing ECCCS in BioPortal to CoCoNBC can help us evaluate whether an ontology-enhanced approach can achieve performance comparable to that of the naïve Bayesian method.

5.5. Performance comparisons

Table 7 summarizes the results of the comparison between ECCCS in BioPortal and ECCCS by syndromic categories. The second to the fifth columns of Table 7 list the true positive (TP) cases, false negative (FN) cases, true

negative (TN) cases, and false positive (FP) cases in each syndromic category. The sixth through the eighth columns list the PPV, sensitivity, and specificity measures. Comparing the TP and FN cases across syndromes, we find that the WSSS component raises the number of TP cases and reduces the number of FN cases. For example, the WSSS component increased the TP cases by 15 for the respiratory syndrome. At the same time, the FN cases decreased by the same amount. Raising the TP cases comes at the cost of an increased number of the FP cases. The increase of FP cases is different across syndromes. For instance, the respiratory syndrome only has three additional FP cases while the number of TP cases was increased by 15. On the other hand, for gastrointestinal syndrome, there are 20 additional FP cases while the number of TP cases is increased by 14.

From the discussion above, it is not surprising to observe that the WSSS component has opposite effects on PPV and sensitivity. The PPV of ECCCS in BioPortal is lower in most syndromic categories except in the upper respiratory syndrome. The difference is significant in the gastrointestinal syndrome (p -value < 1%), the hemorrhagic syndrome (p -value < 10%), and the rash syndrome (p -value < 10%). On the other hand, the sensitivity of ECCCS in BioPortal is significantly higher in all syndromes under consideration. The p -values are less than 1% in the gastrointestinal syndrome, the hemorrhagic syndrome, the respiratory syndrome, and the upper respiratory syndrome; and are less than 5% in the rash syndrome. ECCCS in BioPortal also has lower specificity. But since specificities in both systems are very high (all larger than 97.03%), the difference is not substantial. When considering PPV and sensitivity together, ECCCS in BioPortal has higher F measures and $F2$ measures in all syndromes. The differences are significant in the hemorrhagic syndrome, the respiratory syndrome, and the upper respiratory syndrome for both the F measure and $F2$ measure (p -value < 5%), and significant in the gastrointestinal

Table 7
Performance comparison between ECCCS in BioPortal and ECCCS

Syndrome	TP	FN	TN	FP	PPV	Sensitivity	Specificity	F	$F2$
<i>ECCCS in BioPortal</i>									
GI	104	20	850	26	0.8000	0.8387***	0.9703	0.8189	0.8254*
HEMO	19	11	967	3	0.8636	0.6333***	0.9969	0.7308**	0.6951***
RASH	10	5	976	9	0.5263	0.6667**	0.9909	0.5882	0.6122
RESP	90	20	879	11	0.8911	0.8182***	0.9876	0.8531***	0.8411***
URESP	36	7	935	22	0.6207	0.8372***	0.977	0.7129**	0.7500***
<i>ECCCS</i>									
GI	90	34	870	6	0.9375***	0.7258	0.9932***	0.8182	0.7849
HEMO	10	20	970	0	1.0000*	0.3333	1.0000*	0.5000	0.4286
RASH	7	8	982	3	0.7000*	0.4667	0.9970***	0.5600	0.5250
RESP	75	35	882	8	0.9036	0.6818	0.9910	0.7772	0.7426
URESP	27	16	938	19	0.5870	0.6279	0.9801	0.6067	0.6136

Statistical testing is based on 2500 bootstrappings.

* p -Value < 0.1.

** p -Value < 0.05.

*** p -Value < 0.01.

syndrome for the $F2$ measure (p -value $< 10\%$). Comparing the significance level of the F measure and the $F2$ measure, we find that the $F2$ measure is significant in gastrointestinal syndrome (p -value $< 10\%$) while the F measure is not significant. Similarly, the $F2$ measure is significant in the upper respiratory syndrome at the 1% level while the F measure is only significant at the 5% level. The differences in statistical significance levels reflect the fact that the $F2$ measure emphasizes sensitivity over PPV. To summarize, ECCCS in BioPortal achieves higher sensitivity but lower PPV. But in terms of the F measure and $F2$ measure, ECCCS in BioPortal outperforms ECCCS. Since the major difference between ECCCS and ECCCS in BioPortal is whether the WSSS grouping is used, we conclude that adding the WSSS component to a rule-based system increases its sensitivity and F and $F2$ measures at the expense of lowered PPV.

Table 8 summarizes the comparison between ECCCS in BioPortal and CoCoNBC. ECCCS in BioPortal has more TP and FP cases in most syndromes. The only exception is the hemorrhagic syndrome. CoCoNBC has one more TP case than that of ECCCS in BioPortal. ECCCS in BioPortal has lower PPV in three out of four syndromes. The difference, however, is only significant for the gastrointestinal syndrome. ECCCS in BioPortal has significantly higher sensitivity in most syndromes including the gastrointestinal syndrome, the rash syndrome, and the respiratory syndrome. CoCoNBC delivers higher sensitivity in the hemorrhagic syndrome. Given that the numbers of TP and FP cases in the hemorrhagic syndrome have only small differences between these two classifiers, it is not surprising that the statistical tests find no significant difference. We also observe that both systems have fairly high specificity. ECCCS in BioPortal has higher F measure and $F2$ measure in the gastrointestinal syndrome, the rash syndrome, and the respiratory syndrome but not in the hemorrhagic syndrome. The differences are significant in the gastrointestinal syndrome (p -value $< 5\%$) and the respiratory syndrome

(p -value $< 1\%$). Note that the $F2$ measure is significant at the 1% level in the gastrointestinal syndrome but the F measure is only significant at the 5% level. This difference, again, reflects the fact that the $F2$ measure puts more weight on sensitivity. Although the differences are significant only for half of the syndromes, these syndromes cover more than 80% of the CCs under consideration.

6. Discussion

This section discusses issues related to the reference standard dataset generation and the WSSS component. We then summarize the significance and limitations of our study and point out future work.

6.1. Discrepant kappas among syndromic categories

The kappa statistics of the 11 syndromic categories vary substantially. As briefly discussed before, rare syndromes such as botulism may have difficulty achieving a high kappa. However, the rash syndrome has moderate prevalence (1.5%) but a high kappa (0.82). Compared to the study of Chapman et al. [3], we find similarities and differences. The respiratory, hemorrhagic, and gastrointestinal syndromes share similar high levels of agreement in both studies. The rash syndrome has an excellent level of agreement in our results (kappa = 0.82), but the lowest level of agreement in [3] (kappa = 0.23). The botulism and constitutional syndromes have low levels of agreement in our research but have moderate levels of agreement in [3]. Fever has a moderate level of agreement in our results but a high level of agreement in [3]. Our kappa statistic for the neurological syndrome is about 15% lower than that reported in [3].

This comparison shows that it is not uncommon to have different agreement levels across different syndrome categories. One possible explanation is that the experts' different work experiences or specialty concentrations may lead to

Table 8
Performance comparison between ECCCS in BioPortal and CoCoNBC

Syndrome	TP	FN	TN	FP	PPV	Sensitivity	Specificity	F	$F2$
<i>ECCCS in BioPortal</i>									
GI	104	20	850	26	0.8000	0.8387 ^{***}	0.9703	0.8189 ^{**}	0.8254 ^{***}
HEMO	19	11	967	3	0.8636	0.6333	0.9969	0.7308	0.6951
RASH	10	5	976	9	0.5263	0.6667 [*]	0.9909	0.5882	0.6122
RESP	90	20	879	11	0.8911	0.8182 ^{***}	0.9876	0.8531 ^{***}	0.8411 ^{***}
URESP	36	7	935	22	0.6207	0.8372	0.977	0.7129	0.7500
<i>CoCoNBC</i>									
GI	80	44	867	9	0.8989 ^{**}	0.6452	0.9897 ^{***}	0.7512	0.7122
HEMO	20	10	968	2	0.9091	0.6667	0.9979	0.7692	0.7317
RASH	7	8	980	5	0.5833	0.4667	0.9949 [*]	0.5185	0.5000
RESP	65	45	881	9	0.8784	0.5909	0.9899	0.7065	0.6633
URESP	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Statistical testing is based on 2500 bootstrappings.

^{*} p -Value < 0.1 .

^{**} p -Value < 0.05 .

^{***} p -Value < 0.01 .

different interpretations of the chief complaints (and other information in ED reports in the study of [3]) and thus result in different syndrome assignments.

Although a detailed analysis about why syndromic categories such as botulism and constitutional have low levels of agreement and whether the syndromic definitions can generate a reliable reference standard dataset is beyond the scope of this article, a few examples may shed some light on this important topic. The chief complaint “NAUSEA WEAKNESS NOT EATING” was assigned to the constitutional and gastrointestinal syndromes by expert one, gastrointestinal by expert two, and constitutional by expert three. “HA VOMITING” was assigned to the neurological and gastrointestinal syndromes by expert one, neurological by expert two, and constitutional by expert three. The above examples indicate some possible reasons for low levels of agreement. First, the definition of some syndromes may not be clear. In some cases, even if the definitions are clear, the experts may have a difficult time fully understanding and consistently following these definitions. More research is needed to understand this issue further.

6.2. The effect of the WSSS component

As summarized in the previous section, the WSSS component is able to increase the number of TP and FP cases simultaneously. The resulting sensitivity is higher while PPV and specificity decrease. The increase in sensitivity means that the classification system can single out the desired signal better. At the same time, additional noise is introduced into the classification results because of higher PPV. Practically, for syndromes with low prevalence such as the rash syndrome, improving sensitivity should be the first priority to avoid delay in outbreak detection. Alternatively, if the syndrome has moderate or high prevalence, then the trade-off between sensitivity and PPV becomes less clear-cut. In cases where the classification system has moderate sensitivity but very high PPV, increased sensitivity and decreased PPV may benefit the subsequent statistical detection task by increasing the signal level higher than the noise it introduced. However, it is possible that this kind of adjustment make the detection task more difficult. If the relative importance between the detection ability of a surveillance system and the cost of having a false alarm can be determined, a weighting scheme which reflects the relative importance can be used to customize the measure from the F measure family. This measure then can be used to determine whether the trade-off between sensitivity and PPV is beneficial for the surveillance system. It should be noted that the decision to adapt the WSSS method is determined on a syndrome-by-syndrome basis. That is, the method may be applied to only some syndromes in a CC classifier while other syndromes are classified using the original method.

As noted in the “Research Test Bed,” the reference standard dataset contains only distinct CC strings. Evaluating the BioPortal CC classification system with this reference

standard dataset can tell us how the WSSS component extends the knowledge of a CC classifier. It is also interesting to know the performance impact of the WSSS component on the reference standard dataset that contains duplicated records (i.e. a random sample without duplicated records removed). We recalculated the performance of the ECCCS in BioPortal and ECCCS using the new reference standard dataset that contains duplicated records. The basic pattern of increased sensitivity and decreased PPV are the same. However, ECCCS in BioPortal has significantly lower F measure and $F2$ measure in the gastrointestinal syndrome. Looking into individual records, we find that the WSSS misclassified high frequency CCS “flank pain”, “left flank pain”, “right flank pain”, and “kidney pain” into the gastrointestinal syndrome. These false positives substantially reduced PPV of ECCCS in BioPortal. These cases, as a result, should receive higher priority in error analysis.

The gastrointestinal syndrome does not have good performance using the original reference standard dataset either. Among all syndromes, the gastrointestinal syndrome had the largest increment in FP cases using ECCCS as a benchmark. The number of FP cases (26) was also substantially higher than that of CoCoNBC (9). We thus select the gastrointestinal syndrome as the focus of error analysis.

The WSSS component utilizes semantic information from a medical ontology for symptoms grouping purposes. While the WSSS grouping results coincided with the assignments of human experts most of the time, it is possible that the WSSS assigns the wrong group to an unseen symptom. For example, “left flank pain” was assigned to group “gi” and subsequently classified to the gastrointestinal syndrome because it is very close to the symptom “abdominal pain” in the UMLS (distance = 2). But “left flank pain” was not considered part of the gastrointestinal syndrome in the reference standard dataset. “Vaginal pain” was also classified into group “gi” because its closest neighbor in the UMLS is “abdominal pain” (distance = 1). The right mapping for “Vaginal pain” was actually “other” in the reference standard dataset.

The above examples indicate that, in certain cases, the UMLS ontology is not suitable for the purpose of syndromic surveillance. Detailed error analysis should be able to provide a more complete picture about the potential factors that affect the performance of the ontology method and shed light on the direction of future performance improvement.

6.3. Significance and limitations of this work

This work proposed an approach that can potentially improve the effectiveness of a CC classification system. This approach is based on the use of a medical ontology in the CC classification process. As shown through an experimental study, semantic information captured in medical ontologies can be effectively leveraged to expand the coverage of the symptom grouping table automatically

without extra knowledge acquisition efforts. The specific technical approach developed, the WSSS component, can be seen as a booster for an existing rule-based CC classification system.

There are several limitations associated with this study. First, because of the low levels of agreement in some syndromic categories, we were forced to drop these categories from subsequent analyses. As a result, performance evaluation for syndromes such as botulism, constitutional, fever, neurological, and lower respiratory remains unknown. Second, for some syndromic categories such as rash, only a small number of valid cases are available. Thus no definitive conclusions can be drawn from these syndromes.

6.4. Future work

Besides obvious future work concerning additional data collection effort and testing to be performed to further evaluate our approach, several interesting research directions remain.

First, the National Ambulatory Medical Care Survey (NAMCS) provides datasets that contain CCs with standardized coding [27]. These datasets may provide new resources for future CC classification research. Second, CCs are often available in languages other than English in international contexts. How to develop a working CC classification system in a multi-lingual environment poses interesting technical challenges, such as a US/Mexico cross border syndromic surveillance system.

Finally, other uses of a medical ontology in the CC classification process may be worth exploring. For instance, in the current process of producing the symptom grouping table, the experts are completely on their own in coming up with the terms. One interesting extension is to use medical ontologies to help experts construct this table in an iterative manner by suggesting terms and groupings interactively.

7. Concluding remarks

In this paper, we developed and evaluated an ontology-enhanced approach to classify free-text chief complaints into syndromic categories. This approach can cope with multiple sets of syndrome definitions. At the core of this approach is the UMLS-based Weighted Semantic Similarity Score (WSSS) grouping method that is capable of automatically assigning previously un-encountered symptoms to appropriate symptom groups. An evaluation study shows that this approach can achieve a higher sensitivity, F measure, and $F2$ measure, when compared to the CC classification subsystem of EARS that has the same symptom grouping table and syndrome rules. This approach also outperforms RODS' CoCo naïve Bayesian classifier for syndrome categories that cover most CCs under consideration. As a side result, we also applied a bootstrapping-based statistical testing procedure to compare the performance of different methods. This procedure can be

applied to compare sensitivity, specificity, positive predictive value, F measure, and $F2$ measure as long as the systems under consideration share a common reference standard dataset in which the independent assumption among records is reasonable.

Acknowledgments

This work was supported in part by the US National Science Foundation through Grant No. IIS-0428241 and by the Arizona Department of Health Services. The authors thank Dr. Peter Kelly, Dr. Ayesha Bashir, Dr. Rebecca Sunenshine, and Ms. Leah Chinnaswamy for their significant help establishing the reference standard dataset and syndrome definitions used in this study. The second author wishes to acknowledge support from a Research Grant (60573078) from the National Natural Science Foundation of China, an International Collaboration Grant (2F05N01) from the Chinese Academy of Sciences, a National Basic Research Program of China (973) Grant (2006CB705500) from the Ministry of Science and Technology, and an Innovative Research Group Grant (60621001) from the National Science Foundation of China. We also appreciate valuable comments and suggestions for improvement from anonymous reviewers.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.jbi.2007.08.009](https://doi.org/10.1016/j.jbi.2007.08.009).

References

- [1] Lewis M, Pavlin J, Mansfield J, O'Brien S, Boomsma L, Elbert Y, et al. Disease outbreak detection system using syndromic data in the greater Washington DC area. *Am J Prev Med* 2002;23(3): 180–6.
- [2] Mandl KD, Overhage M, Wagner M, Lober W, Sebastiani P, Mostashari F, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. *J Am Med Inform Assoc* 2004;11(2):141–50.
- [3] Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artif Intell Med* 2005;33(1):31–40.
- [4] Ivanov O, Wagner MM, Chapman WW, Olszewski RT. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. *Proc AMIA Symp* 2002:345–9.
- [5] Chapman WW, Dowling JN, Wagner MM. Generating a reliable reference standard set for syndromic case classification. *J Am Med Inform Assoc* 2005;12(6):618–29.
- [6] Espino JU, Wagner MM. The accuracy of ICD-9 coded chief complaints for detection of acute respiratory illness. *Proc AMIA Symp* 2001:164–8.
- [7] Chapman WW, Dowling JN, Wagner MM. Fever detection from free-text clinical records for biosurveillance. *J Biomed Inform* 2004;37(2):120–7.
- [8] Lee N, Hui D, Wu A, Chan P, Cameron P, Joynt G, et al. A major outbreak of severe acute respiratory syndrome in Hong Kong. *N Engl J Med* 2003;348(20):1986–94.

- [9] Lombardo J, Burkom H, Elbert E, Magruder S, Lewis SH, Loschen W, et al. A system overview of the Electronic Surveillance System for Early Notification of Community-Based Epidemics (ESSENCE II). *J Urban Health* 2003;80(2):i32–42.
- [10] Tsui F-C, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM. Technical description of RODS: a real-time public health surveillance system. *J Am Med Inform Assoc* 2003;10(5):399–408.
- [11] Yan P, Chen H, Zeng D. Syndromic surveillance systems: public health and biodefense. *Ann Rev Inform Sci Technol* 2008;42.
- [12] Shapiro AR. Taming variability in free text: Application to health surveillance. *MMWR* 2004;53(suppl):95–100.
- [13] Travers DA, Haas SW. Evaluation of emergency medical text processor, a system for cleaning chief complaint textual data. *Acad Emerg Med* 2004;11(11):1170–6.
- [14] Day FC, Schriger DL, La M. Automated linking of free-text complaints to Reason-for-Visit categories and International Classification of Diseases diagnoses in emergency department patient record databases. *Ann Emerg Med* 2004;43(3):401–9.
- [15] Thompson DA, Eitel D, Fernandes CMB, Pines JM, Amsterdam J, Davidson SJ. Coded chief complaints—automated analysis of free-text complaints. *Acad Emerg Med* 2006;13(7):774–82.
- [16] Travers D. Identification of concepts from emergency department text using natural language processing techniques and the unified medical language system: University of North Carolina at Chapel Hill; 2003.
- [17] Chapman WW. Natural language processing for biosurveillance. In: Wagner MM, Moore AW, Aryel RM, editors. *Handbook of biosurveillance*. New York: Elsevier; 2006. p. 255–71.
- [18] Crubezy M, O'Connor M, Buckeridge DL, Pincus Z, Musen MA. Ontology-centered syndromic surveillance for bioterrorism. *IEEE Intell Syst* 2005;20(5):26–5.
- [19] Lu H-M, Zeng D, Chen H. Ontology-based automatic chief complaints classification for syndromic surveillance. In: *IEEE International Conference on Systems, Man, and Cybernetics*. Taipei, Taiwan; 2006.
- [20] Zobel J, Dart P. Phonetic string matching: lessons from information retrieval. *Proceedings of the 19th International Conferences on Research and Development in Information Retrieval*. Zurich, Switzerland; 1996.
- [21] Navarro G. A guided tour to approximate string matching. *ACM Comput Surv* 2001;33:31–88.
- [22] Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, Brinkley JF. Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base. *J Am Med Inform Assoc* 1998;5(1):17–40.
- [23] Achour SL, Dojat M, Rieux C, Bierling P, Lepage E. A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development. *J Am Med Inform Assoc* 2001;8(4):351–60.
- [24] Leroy G, Chen H. Meeting medical terminology needs—the ontology-enhanced medical concept mapper. *IEEE Trans Inform Technol Biomed* 2001;5(4):261–70.
- [25] Tolle KM, Chen H. Comparing noun phrasing techniques for use with Medical Digital Library Tools. *J Am Med Inform Assoc* 2000;51(4):352–70.
- [26] Schneider D, Appleton L, McLemore T. A reason for visit classification for ambulatory care. *Natl Center Health Stat Vital Health Stat* 1979;2:1–63.
- [27] McCaig LF, Nawar EW. National hospital ambulatory medical care survey: 2004 Emergency Department Summary. *Adv Data Vital Health Stat* 2006;372:1–32.
- [28] Grafstein E, Unger B, Bullard M, Innes G. Canadian Emergency Department Information System (CEDIS) presenting complaint list (Version 1.0). *Can J Emerg Med* 2003;5(1):27–34.
- [29] Aronsky D, Kendall D, Merkley K, James BC, Haug PJ. A comprehensive set of coded chief complaints for the emergency department. *Acad Emerg Med* 2001;8:980–9.
- [30] Travers DA, Haas SW. Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. *J Biomed Inform* 2003;36:260–70.
- [31] Graham J, Buckeridge D, Choy M, Musen M. Conceptual heterogeneity complicates automated syndromic surveillance for bioterrorism. *Proc AMIA Symp* 2002;1030.
- [32] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure* 1960;20(1):37–46.
- [33] Olszewski RT. Bayesian classification of triage diagnoses for the early detection of epidemics. *FLAIRS Conference* 2003; Menlo Park, California; 2003, p. 412–6.
- [34] Espino JU, Dowling J, Levander J, Sutovsky P, Wagner MM, Copper GF. SyCo: a probabilistic machine learning method for classifying chief complaints into symptom and syndrome categories. *Syndromic Surveillance Conference*. Baltimore, Maryland; 2006.
- [35] Sniegoski CA. Automated syndromic classification of chief complaint records. *Johns Hopkins APL Tech Digest* 2004;25(1):68–74.
- [36] Witten IH, Frank E. *Data mining: practical machine learning tools and techniques*. San Francisco, CA: Elsevier; 2005.
- [37] Hripisak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. *J Am Med Inform Assoc* 2002;9:1–15.
- [38] van Rijsbergen CJ. *Information retrieval*. London: Butterworth; 1979.
- [39] Pakhomov SVS, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *J Am Med Inform Assoc* 2006;13:516–25.
- [40] Agresti A. *Categorical data analysis*. Hoboken, New Jersey: John Wiley & Sons; 2002.
- [41] McNeman Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947;12:153–7.
- [42] Chapman WW, Haug PJ. Comparing expert systems for identifying chest X-ray reports that support pneumonia. *Proc AMIA Symp* 1999;216–20.
- [43] Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat* 1979;7:1–26.
- [44] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. New York, NY: Springer; 2001.
- [45] Efron B. Estimating the error rate of a prediction rule: some improvements on cross-validation. *J Am Stat Assoc* 1983;78:316–31.
- [46] Efron B, Tibshirani R. *Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy*. *Stat Sci* 1986;1:54–77.
- [47] Hinkley DW. Bootstrap methods. *J R Stat Soc Ser B* 1988;50:321–37.
- [48] Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York, NY: John Wiley & Sons; 1981.
- [49] Hu P, Zeng D, Chen H, Larson C, Chang W, Tseng C. Evaluating an infectious disease information sharing and analysis systems. In: Kantor P, Muresan G, Roberts F, Zeng D, Wang F-Y, Chen H, et al., editors. *IEEE International Conference on Intelligence and Security Informatics*. Atlanta, Georgia; 2005.
- [50] Chang W, Zeng D, Chen H. Prospective spatio-temporal data analysis for security informatics. *IEEE Conference on Intelligent Transportation Systems*. Vienna, Austria; 2005.
- [51] Mikosz CA, Silva J, Black S, Gibbs G, Gardenas I. Comparison of two major emergency department-based free-text chief-complaint coding systems. *MMWR*. 2004;53(suppl):101–5.
- [52] Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *J Urban Health*. 2003;80:i89–96.