Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 36 (2003) 260–270

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department

Debbie A. Travers* and Stephanie W. Haas

School of Information and Library Science, University of North Carolina at Chapel Hill, Campus Box 3360, Chapel Hill, NC 27599-3360, USA

Received 26 August 2003

Abstract

Information about the chief complaint (CC), also known as the patient's reason for seeking emergency care, is critical for patient prioritization for treatment and determination of patient flow through the emergency department (ED). Triage nurses document the CC at the start of the ED visit, and the data are increasingly available in electronic form. Despite the clinical and operational significance of the CC to the ED, there is no standard CC terminology. We propose the construction of concept-oriented nursing terminologies from the actual language used by experts. We use text analysis to extract CC concepts from triage nurses' natural language entries. Our methodology for building the nursing terminology utilizes natural language processing techniques and the Unified Medical Language System.

Keywords: Terminology; Nursing; Emergency department; Natural language processing; DEEDS; UMLS

1. Introduction

In the initial minutes of a patient's emergency department (ED) visit, the triage nurse determines the patient's chief complaint (CC), or reason for seeking care. The Emergency Nurses Association (ENA) Triage Curriculum emphasizes the importance of the CC in emergency nurses' decision-making; it is the first data element collected during the triage history and physical assessment [1]. The nature and severity of the CC directly influence many aspects of the patient's ED visit. The CC forms the basis of focused nursing and medical assessments, and is critical for patient prioritization for treatment and determination of patient flow through the ED. Despite the clinical and operational significance of the CC to the ED, and the increasing use of computers to document the CC, there is no standard terminology to describe this nursing data element. EDs currently document the CC in free text form, or using a variety of locally developed or adapted terminologies [2,3]. In the absence of a standard terminology, it is difficult to aggregate CC data. But there is growing interest in aggregated CC data for secondary uses, such as supporting clinical, health services, and epidemiologic research; public health surveillance and quality improvement activities [2,4–7].

The CC has been identified as a key data element in national efforts to develop terminology standards for the ED. Initial work by the ENA on an emergency nursing minimum data set was incorporated into a multi-disciplinary effort sponsored by the Centers for Disease Control and Prevention, which led to the release of Data Elements for Emergency Department systems (DEEDS) 1.0 in 1997 [2,8–10]. DEEDS data element 4.06 is "Chief Complaint." Since no standard vocabulary exists for documentation of CC, the DEEDS and ENA leaders recommended evaluation and adaptation of established terminologies as a solution to the need for an ED CC system, and identified several candidate systems including the International Classification of Primary Care and the Reason for Visit Classification and Coding Manual [11,12]. A standard CC terminology must

* Corresponding author. Fax: 1-919-962-8071 (beginning December 2003, the fax number will be: 1-919-966-3049).

E-mail address: dtravers@med.unc.edu (D.A. Travers).

1532-0464/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2003.09.007

represent the concepts that are used by triage nurses to describe the reasons why patients visit the ED. Therefore, before a CC terminology can be adapted or constructed, it is necessary to identify the concepts that comprise the domain of ED CC.

In this paper, we propose a method for building concept-oriented nursing terminologies grounded in the natural language used by domain experts. Then, we test the feasibility of that approach through a pilot study in which we identify concepts from triage nurses' CC entries using natural language processing techniques and the Unified Medical Language System.

2. Background

There is a lack of standardized vocabulary to express clinical findings, treatments and patient progress in electronic health record systems [13–15]. The nursing informatics community has addressed this issue on many fronts: by creating minimum nursing data sets, developing and evaluating standardized nursing terminologies for electronic systems, and more recently through information modeling and working toward a reference terminology as an international standard [16–23]. Controlled terminologies have been developed for specific nursing domains such as home health [24]. Although the ENA and DEEDS have delineated a set of essential emergency care data elements, specific terminologies have not been developed for the ED nursing domain [8–10].

2.1. Why CC terminology is needed

An ED CC terminology will benefit clinical care as well as enable secondary use of CC data. Direct clinical applications of a CC terminology include facilitating electronic health record systems, initiating and monitoring compliance with clinical guidelines, linking clinical information to bibliographic resources, facilitating development of decision support systems, and implementing complaint-specific history and physical exam prompts [8–10]. Secondary uses of CC data include insurance review and reimbursement for ED visits, research, quality improvement, and public health surveillance on the regional and national level [8]. For example, there are a growing number of systems that are exploring the use of ED CC information to facilitate symptom-driven surveillance for early detection of bioterrorism [3,25]. The systems utilize real-time clinical information including ED CC, diagnosis and laboratory data. These bioterrorism surveillance systems have the potential to alert authorities to possible geographic clusters of patients with similar symptoms that might indicate, for example, an attack with a biological agent that causes high fever and shortness of breath. Though existing public health surveillance and bioterrorism detection systems utilize a mixture of CC data in free text form or documented with various locally developed or established classification systems, surveillance and detection activities would be greatly enhanced by a standard CC terminology.

2.2. Build terminologies from the language of domain experts

Before a terminology can be built or adapted for CC, it is necessary to identify the concepts that comprise the domain of ED CC. Cimino [26] defines a concept as "an embodiment of a particular meaning" (p. 395). Examples of concepts for signs or symptoms are *chest pain* and *syncope* [27]. A useful approach to identifying the concepts in a domain is to map them from the terms used in the natural language of domain experts. Whereas concepts represent meaning, terms are the natural language phrases that represent and describe the concepts [13]. For example, the terms *fainted* and *fainting* are lexical variants that represent the same concept.

Medical informaticians have proposed methods for developing controlled terminologies using terms found in the clinical text or literature. Liu and Friedman [28] proposed a method for capturing clinical terms from pathology reports that included analysis of compositional information in the terms. McCray [29] described the process of building lexicons that reflect the common language shared by domain experts, by comparing it to literary warrant. Lexicon developers decide what terms and concepts should be included based on the frequency of use in the literature of the domain. Kreis and Gorman [30] created a structured data entry system for physical examinations by including the most frequent words used by trauma surgeons in dictated history and physical examination reports.

In the health care domain, clinical language has been described as a sublanguage, which is a restricted language used by a group of people in a specialized domain [31–36]. Features of sublanguage include specialized terminology and content, and patterns of occurrence and co-occurrence of words in text. Johnson and Gottfried [32] present a method for using sublanguage analysis as the basis for building controlled vocabularies in healthcare. Their sublanguage analysis methods involve collecting language (primarily from written text) used by domain experts, and then analyzing the data to ascertain the content and relationships of the sublanguage's terminology. The results of the analysis can then be incorporated into a model of the sublanguage that can be used to construct an information system for use by experts in the field.

Using natural language nursing text for building nursing vocabulary introduces several challenging

methodological issues. The fast-paced ED environment presents particular constraints on the data entry process, and leads to hurried, compressed, and potentially error-prone entries. The well-known characteristics of messy free text electronic data can be magnified: the non-standardized entries contain misspellings, data entry errors (e.g., hitting the 1 key instead of the q key), local expressions, abbreviations, and synonyms. These characteristics must be addressed in order to extract concepts from the free text entries. Concept extraction methods are more straightforward for edited text, such as automatic indexing of journal articles [37,38]. In contrast, methods for identifying concepts are more complicated for un-edited text like ED CC entries, reports and clinical progress notes. Issues with un-edited text that must be addressed, include ambiguous abbreviations, punctuation, and multiple word senses [39–42].

2.3. Natural language processing techniques

One approach to extracting concepts from free text data is to use natural language processing (NLP) techniques to clean and normalize the original data [31,43–45]. NLP encompasses a wide array of techniques for linguistic analysis of natural text. Those relevant to the current research project include normalization, segmentation, stemming, word sense disambiguation, word look-up, spelling correction, and abbreviation expansion. Normalization is the process of transforming data to eliminate minor differences such as upper and lower case, inflection and word order, and to remove stop words [46]. For example, an original clinical entry such as *injury to head* is normalized to *head injury*.

One of the earliest NLP developments was text segmentation, which is used for a myriad of NLP applications. There are two types of text segmentation: tokenization, which breaks up text into individual words, and sentence segmentation, which breaks up text into sentences or other phrase-like units [47]. Despite the fact that most written languages, including English, have white space boundaries between words, and punctuation to delineate sentences, there is no absolute definition of what constitutes a word or a sentence. For example, rules are needed to determine if characters surrounding hyphens, such as *e-mail* and *so-called*, should be considered as one or two words. Punctuation may not always indicate sentence or phrase boundaries, and has proven to be a challenging feature of the English language for sentence segmentation and tokenization applications. The frequency of use of periods to designate abbreviations versus ends of sentences varies, depending on the specific corpus. For example, in the Brown corpus, only 10% of the periods denoted abbreviations, as opposed to 47% in a *Wall Street Journal* corpus [47]. This information is useful to those developing segmentation applications for specific corpora. NLP system developers have used other contextual features to assist with punctuation processing in sentence segmentation. For example, case distinctions are useful for applications based on languages and corpora that consistently use upper- and lower-case letters (e.g., it's the end of a sentence if it's followed by a space or two and then a capital letter). And Palmer [47] found that parts of speech within three tokens (words or comparable text units) of a punctuation mark were useful in sentence segmentation.

An added feature of clinical text that must be addressed with clinical NLP systems is that punctuation is not used just to segment sentences. For example, punctuation is used in abbreviations such as *diarr*. for *diarrhea* and *h/a* for *headache*. The Unified Medical Language System (UMLS) lexical tool kit provides three different tools for processing punctuation [46]. Options include simple deletion of punctuation, replacement with spaces or replacement with spaces except where punctuation is between or just before numbers. In an earlier study, we found that a relatively rare punctuation mark, the slash (''/'') was used extensively in clinical entries [48]. The slash was used for many purposes including abbreviations and coordinate structures. Ammons [39] found that the slash (''/'') is used arbitrarily in scientific writing in psychology. He stated that this practice was ''producing jargon which hides rather than elucidates meaning'' (p. 418).

Once sentence and word boundaries are identified in textual data, the focus of NLP can shift to the word level. A foundational technique for word level analysis in NLP is stemming, which removes prefixes like *un-* and suffixes like *-ed, -ing, -ion, and -ions* through prefix- and suffix-stripping algorithms. Stemming is used in a wide variety of NLP applications to reduce morphological variants to the root form of the word. For example, stemming is useful for counting the words in a corpus to ascertain the most frequent words. A popular stemmer was developed by Porter [49], and is based on an algorithm with a limited number of suffixes. The UMLS lexical toolkit includes a stemmer that is useful for preparing text for comparison with UMLS records in the normalized string index [46].

Word sense disambiguation is an important NLP technique for applications that require some level of semantic processing. Words can be spelled the same but have different meanings or senses, such as *treat*, which can mean a special food (e.g., ''Trixie ate her *treat*'') or to give care (e.g., ''the nurses *treat* the patient with morphine''). Part of speech taggers are used to facilitate word sense disambiguation in cases where the different meanings occur in different parts of speech (e.g., *treat* as a noun and verb) [50]. The availability of electronic dictionaries has aided efforts to accomplish word sense disambiguation, and these techniques have proven useful in identifying domain-specific senses of words [51].

Corpus-based statistical methods for word sense disambiguation are costly and require large training sets, but have proven to be relatively accurate [50,51]. Many word sense disambiguation methods are context-sensitive, utilizing the words surrounding the target word to clarify meaning.

Many NLP applications also utilize word look-up programs to address synonyms, abbreviations, and misspellings. For example, Olszewski [52] developed a list of substitutions for misspellings and non-standard terms for an ED surveillance system that is used for early detection of disease outbreaks. The accuracy of the detection system improved with the domain and application-specific look-up and replacement program.

2.4. Using the Unified Medical Language System to build terminologies

The National Library of Medicine has encouraged the use of the UMLS and its source vocabularies as a tool to facilitate construction of new terminologies and thesauri [53]. The 2003 UMLS Knowledge Sources include the Metathesaurus, Semantic Network, and the SPECIALIST lexicon and lexical programs [46,54]. At the core of the UMLS is the concept-oriented Metathesaurus, which contains over 800,000 biomedical concepts from more than 100 source vocabularies. The concepts are organized in semantic categories (e.g., sign or symptom, body part, pathological function) with defined relationships (e.g., antibiotic is a pharmacologic substance) in the Semantic Network. The SPECIALIST programs include a suite of lexical processing tools to assist researchers with managing natural variation in biomedical language. The UMLS normalization tools abstract away case, inflection, and word order, as well as removing stop words and possessives, and replacing punctuation with spaces [55]. Normalized natural language terms can then be compared with the Metathesaurus string index to determine whether the terms correspond to a UMLS concept.

Concepts from the UMLS and its source vocabularies have formed the basis of many terminology applications in health care. Payne and Martin [56] used the UMLS to create a master problem list for a computer-based patient record system in a large cooperative of primary care facilities. Chute and Elkin [57] used the existing hierarchies of the UMLS to help structure the Mayo Clinic online problem list. Eisner [58] used terminology from the Metathesaurus to develop a core vocabulary for a dental school curriculum. Cooper and Miller [59] developed statistical and lexical methods for extracting controlled terms from clinical free text.

We propose the use of text analysis to build concept-oriented nursing terminologies that are grounded in the natural language of domain experts. In this paper, we specifically address the construction of an ED CC terminology from the language used by nurses in the context of triage. To test the feasibility of this method for creating concept-oriented nursing terminologies, we conducted a pilot study using electronic ED CC data entered by triage nurses. Though this study did not include a formal evaluation of emergency medical text as a possible sublanguage within the medial domain, we based this work on the assumption that triage nurses' CC entries represent the language of ED clinicians, who use a specialized set of terms, synonyms and concepts [32,42,46]. Our methodology included NLP routines directed at the specific characteristics of natural language found in the analysis of the nurses' text entries, and mapping the CC terms to the UMLS.

The goal of this study was to use text analysis in the construction of a nursing terminology. The analysis in this pilot study focused primarily on concept extraction and synonym identification, rather than on defining relationships. Our specific aims were to describe the characteristics of CC expressed in nurses' natural language that must be addressed in order to identify concepts, begin to develop NLP methods for processing the clinical text, map CC terms to the UMLS, and start to assemble the concepts that comprise the domain of ED CC.

3. Methods

A corpus of CC data was collected from three southeastern US EDs representing urban, rural and suburban academic medical centers. For the pilot project, the training corpus included all CC entries recorded for ED visits during January and August 2000. The IRBs at all three sites approved the study, and no patient identifiers were collected. The unit of analysis was the unique CC entry. Triage nurses entered the CCs directly into the hospital information system (HIS) upon patient arrival to the ED at all three sites. For this paper, we define a CC entry as exactly what was typed into the CC field(s) of the HIS. Some entries contain more than one CC term, such as *Fever/Throwing Up*. At two of the sites, CCs are entered only as free text; at the third site, nurses have the option of using a locally developed, controlled list of 238 terms, or entering the CC as free text. The nurses often supplement the controlled terms with additional information. For example, the controlled term *Chest pain/burning* is often augmented with modifiers such as *severe* or temporal information such as *for 2 weeks*.

We used the UMLS to identify ED CC concepts by employing a methodological approach developed in an earlier pilot study [48]. Our intent in using the UMLS was to map the CC entries to existing Metathesaurus concepts where possible, while acknowledging that some

CC concepts might not be represented in the UMLS. We began the experiment by mapping the unprocessed CC entries to the Metathesaurus, in order to identify corresponding concepts. We first evaluated how many CC entries exactly matched a UMLS concept. We then performed a normalized match on those entries that had not matched a UMLS concept exactly. After the normalized match, we again calculated the match rate with Metathesaurus concepts.

The entries that still did not match a UMLS concept were analyzed using a combination of automated and manual techniques to identify the characteristics of the written text and domain knowledge to interpret those characteristics. The most frequent non-matching entries in the corpus were identified through frequency counts, and the non-matching entries were also tokenized into words, which were then counted and sorted by frequency. The non-matching entries and words were examined by the investigator, a domain expert with 20 years' ED nursing experience and certifications in emergency and informatics nursing. A panel of four domain experts (a nurse and physician from two of the participating sites) were also consulted during this process and assisted in identifying language usage characteristics. The most common characteristics of the non-matching entries are summarized in Table 1.

Many of the non-matching CC entries were found to contain punctuation, the most frequently occurring (22%) of which was the slash. The slash was used most often for separating two or more CCs, but was also found in many coordinate structures (CS) and abbreviations. In the CS, words were dropped to create more compact expressions (known as ellipsis) with the slash replacing the word *and*. We also found twenty frequently used abbreviations that contained a slash; the panel of experts deemed all the abbreviations with slashes as unambiguous in the context of the CC entries. A small number of the non-matching CC entries contained a comma or semi-colon, the majority of which were used to separate two or more concepts.

Table 1
Characteristics of nurses' natural language chief complaint entries

Characteristic	Example
Slash: 2 or more separate concepts	Dizzy/fever
	Cough/diarrhea/congestion
Slash: coordinate structures	Hip/thigh/back pain
	Tingling feet/hands
	Testicle pain/redness
Slash: abbreviations	H/a
	B/p elevated
Comma, semi-colon: 2 or more concepts	Fall, rib pain
	Fever; cancer
Acronyms, abbreviations: not in the UMLS	FB
	MVC
Acronyms, abbreviations: ambiguous	Rx- reaction or prescription
	LOC- loss of consciousness or level of consciousness
Truncation	Diarr
	Pyelo
	Congest
Modifiers	*Right* leg injury
	Severe chest pain
Qualifiers	*History of* seizure
	Headache *since 5 am*

Other characteristics of the nurses' natural language CC entries were acronyms and abbreviations, many of which were ambiguous. For example, *rx* could mean *prescription* or *reaction*. The ED text also contained many truncated words that did not map to the UMLS. We also found that many non-matching entries were more specific than terms in the UMLS because they contained additional information indicating laterality, severity and temporality of the patients' chief complaints. We used the distinctions between modifiers and qualifiers that were set forth by Chute and Elkin [57]. They describe modifiers as words that alter the severity, location, or acuity of a clinical term, such as *acute* myocardial infarction. Qualifiers are words or phrases that qualify the meaning of a clinical term, such as *history of* a condition.

We first attempted to use the UMLS Metathesaurus and SPECIALIST lexical processing tools to address the punctuation and abbreviation issues described in Table 1, but these resources did not effectively process the unique language characteristics found in the ED text [46]. For example, the SPECIALIST punctuation processing routines include simple deletion of punctuation (the entry *dizzy/nausea* becomes *dizzynausea* instead of two terms, *dizzy* and *nausea*) or replacement with spaces (*h/a* becomes *h a* instead of *headache*), and the SPECIALIST abbreviation table expands the abbreviation *SI* to *systeme international d'unites* instead of *suicidal ideation*.

We then began to develop customized processing techniques to address the specific characteristics of the nurses' language. The NLP routines were written as Perl scripts [60]. Ongoing development of NLP techniques continues; pilot methods are described in this paper and focus on the most common characteristics of the triage nurses' language. Three groups of NLP routines were developed and applied in successive rounds, starting with simple techniques and proceeding to more aggressive techniques. For example, replacement was performed before modifiers were removed, so *h/a* was replaced with *headache* in an earlier round, and *severe* was removed from *severe chest pain* in a subsequent round. The goal of this processing was to follow the strategy described by Bodenreider [61] and maximize the match rate with existing Metathesaurus concepts, while minimizing the alteration of the original terms. After each round, the resulting CC terms were again com-

pared to the Metathesaurus to determine how many entries matched a UMLS concept through exact matching followed by normalized string matching. Smaller test corpora were utilized during the development of each processing step for evaluation of the accuracy of the programs and the impact of each program on the entry terms.

4. Results

There were 39,038 patient visits, and 13,494 unique CC entries recorded during the study period. We applied the NLP routines in three rounds, from least to most aggressive.

4.1. Punctuation processing

In Round 1, we addressed the commonly used punctuation patterns. We removed slashes, commas, and semi-colons and processed the entries as shown in Table 2.

First, the abbreviations containing slashes were replaced with the expansions identified by the domain experts. Next, we dealt with CS. Due to the challenges associated with processing CS in medical text, most NLP approaches have addressed a limited subset of coordinations [62–65]. Acknowledging the complexity of many of the CS in the ED CC corpus, we chose to focus a CS algorithm on a limited subset of CC entries. Semantic information from the UMLS was utilized to develop context-sensitive processing rules for the most common CS in our corpus. We excluded CS with commas and semi-colons from the CS processing, in keeping with our desire to apply the least aggressive alterations to the original terms. Though it is possible to have entries such as *hip, thigh*, *back pain*, and *hip/thigh/back pain*, in fact we found that of the 1175 entries in the corpus that contained a comma or semi-colon, only 37 were CS. The CS algorithm first identifies the semantic type of the words bordering the slash(es). In our manual analysis of entries with the slash, the semantic categories of body location, body part or spatial concept, were the most common types found on either side of the slash in CS, whereas the semantic categories of sign or symptom, disease or syndrome, or pathological function were most common in the entries with slash(es) separating two or more concepts. Thus, the CS algorithm processes only the entries in which word(s) bordering both sides of the slash(es) are semantic type body location, body part or spatial concept via look-up in the UMLS. For those entries, the algorithm distributes the other information in the entry to the words bordering the slash, and then splits the entry into two or more separate records. For example, for input CC 3 in Table 2, *hip*, *thigh*, and *back* are all body location, body part or spatial concepts, so the algorithm distributes the word *pain* to each of those words. But in input CC 8, *pain*, *bleeding* and *post-partum* do not belong to those semantic categories so the CS algorithm ignores that entry. The CS algorithm excludes the less common CS entries that have semantic types sign or symptom, disease or syndrome or pathological function bordering the slash, such as *testicle pain/redness* which is shown in Table 1.

The final two steps in the punctuation round dealt with unnecessary abbreviations, and then segmenting all remaining terms on the slash, comma and/or semi-colon. We identified two unnecessary abbreviations, as shown in the Table 2 examples. Since the data in this study were entered into a field specified for ED patients' chief

Table 2
Punctuation processing

Processing step	Input CC	Output CC
Replace	1. h/a	1. headache
	2. b/p elevated	2. blood pressure elevated
Expand coordinate structures, split into 2+ terms	3. hip/thigh/back pain	3a. hip pain
		3b. thigh pain
		3c. back pain
	4. tingling feet/hands	4a. tingling feet
		4b. tingling hands
	5. abdominal/inguinal rash	5a. abdominal rash
		5b. inguinal rash
Eliminate unnecessary abbreviations	6. c/o earache	6. earache
	7. nausea w/vomiting	7. nausea vomiting
Delete slash, comma, semi-colon, and split into 2 terms	8. abdominal pain/vaginal bleeding/post-partum	8a. abdominal pain
		8b. vaginal bleeding
		8c. post-partum
	9. dizzy;nausea	9a. dizzy
		9b. nausea
	10. fall, rib pain	10a. fall
		10b. rib pain

complaints, the abbreviation *c/o* for *complains of* is not needed to convey the meaning of the CC entry. Similarly, the abbreviation *w/* for *with* is also not necessary to convey the meaning of the CC information on either side of the word. Both abbreviations were eliminated from the CC entries. Finally, the remaining entries with slash, comma or semi-colon were split into two or more entries.

During the application of the punctuation rules, many entries were segmented into two (or more) CC terms. For example, the entry *fever;cough* was split into two separate terms, *fever* and *cough*. Some duplicate terms resulted from this process, which were then eliminated. Thus, the number of entries pre-Round 1 were not compared directly with the number of entries/terms post-Round 1. We continue refer to the units of language under study as CC entries for the remainder of this paper, acknowledging that some of the entries were split into distinct terms during Round 1.

4.2. Expansion of acronyms, abbreviations, and truncated words

In Round 2, we took the unmatched entries remaining from Round 1, and handled acronyms, abbreviations, and truncated words (AAT).

Given the restricted context of ED CC, we hypothesized that there would be fewer ambiguous AAT, as opposed to the larger domain of biomedicine, which is covered in the UMLS. We identified the frequently used AAT in the corpus by comparing the CC entries to the SPECIALIST lexicon acronym database, LRABR, and by manually reviewing the remaining unmatched CCs. The four domain experts reviewed the list of common AAT and identified one or more expansions for each AAT. Then, the list was compared to LRABR. We found that many of the ED AAT were missing from LRABR. Others mapped to more than one LRABR expansion and were thus ambiguous. Still, others were in LRABR but it did not include the sense most commonly used in the ED. For example, LRABR had eight expansions for the most common ED abbreviation, *cp*, but did not include the expansion, *chest pain*, identified as correct by the domain experts.

Since most of the AAT in the corpus were not present or matched more than one LRABR record, we created our own AAT dictionary. A consensus of the experts was used to determine the correct expansion for each AAT. Context-sensitive replacement has been used for word sense disambiguation in machine translation, information retrieval, and content and grammatical analysis [40,51,66]. We developed context-sensitive rules were developed for ambiguous AAT; for example, *AB* was expanded to *abortion* unless it preceded *pain* in which case it was expanded to *abdominal*. Examples of the expansions are shown in Table 3.

Table 3
Expansion of acronyms, abbreviations, and truncations

Processing step	Input CC	Output CC
Expand acronym	FB	Foreign body
Expand abbreviation	Rx	If after *all*, *allerg*, *allergic* then *reaction* Else *prescription*
Expand truncation	Diarr	Diarrhea

4.3. Deletion of qualifiers and modifiers

In the last round, we took the unmatched entries remaining from Round 2, and addressed modifiers and qualifiers. We tokenized the unmatched entries into individual words and performed word counts. The words were then compared to lists of modifiers and qualifiers identified in previous research [57,61,67]. Other authors may not differentiate between modifiers and qualifiers as Chute and Elkin [57] have done; in this study, we treated both types of words and phrases alike [32,61]. Previous researchers found that concept matching was improved when common modifiers and qualifiers were removed [57,61,67]. We identified those modifiers and qualifiers present in two or more CC entries, and deleted them. Examples of the altered entries are shown in Table 4. The modifiers and qualifiers were retained in a separate file for inclusion in the ED CC terminology that will be based upon this research (e.g., for pre- or post-combination).

4.4. Results summary

Table 5 shows a summary of the results from Rounds 1–3 of the study, as well as the results of comparing the raw data to the UMLS before any processing. Prior to Round 1, the sample of 13,494 unique CC entries was compared with UMLS concepts. 1137 of the entries exactly matched a UMLS concept, with no manipulation of the CC entries. After normalization, an additional 764 entries matched a UMLS concept, yielding 1901 (14%) matches for the pre-Round 1 phase.

In Round 1, we then applied the punctuation processing algorithms to the 11,593 non-matching entries. After application of the rules (which included segmenting some entries into more than one term) and elimination of duplicates, the modified sample included 10,553 unique CC entries. Of these, 733 exactly matched

Table 4
Deletion of modifiers and qualifiers

Processing step	Input CC	Output CC
Delete modifier	*Right* leg injury	Leg injury
	Severe chest pain	Chest pain
Delete qualifier	*History of* seizure	Seizure
	Headache *since 5 am*	Headache

Table 5
Summary of results

Round	CC entries compared to UMLS (N)	Matches (N, %) *NS = normalized string*	Non-matches (N, %)
Pre-round	13,494	Exact match—1137 NS match—764 Total—1901 (14%)	11,593[a] (86%)
Round 1—Punctuation	10,553[a]	Exact match—733 NS match—204 Total—937 (9%)	9616 (91%)
Round 2—Expansion	9616	Exact match—402 NS match—272 Total—674 (7%)	8942 (93%)
Round 3—Deletion	8942	Exact match—940 NS match—631 Total—1571 (18%)	7371 (82%)

[a]During the application of the punctuation rules, many entries were split into two CC terms. This resulted in some duplicate terms, which were then eliminated. Thus, the *N* for non-matching terms was 11,593 after the pre-Round but only 10,553 terms were compared to the UMLS for the Round 1 processing.

a UMLS concept. The remaining entries were again normalized and an additional 204 matched a UMLS concept, for a total of 937 (9%) for the punctuation phase of the study.

In the second round, we expanded acronyms, abbreviations, and truncated words. We found that 402 of the remaining 9616 entries exactly matched a UMLS concept. The non-matching entries were again normalized and 272 more matched a UMLS concept, for a total of 674 (7%) for the expansion phase of the study.

The final round involved deletion of 21 modifiers and qualifiers, after which 940 of the remaining entries exactly matched a UMLS concept. The non-matching entries were again normalized and 631 more matched a UMLS concept, for a total of 1571 (18%) for the deletion phase of the study.

In summary, in the course of Rounds 1–3 we identified a total of 5083 CC entries (or segments of entries) that matched one or more UMLS concepts, of which 2978 CC entry terms were unique. We found that 86% (4369) of the 5083 matched entries were identified with one UMLS concept only, and 14% were identified with two or more UMLS concepts. Another 7371 entries did not match a UMLS concept; further review is planned to identify any other patterns for which NLP routines can be developed.

The accuracy of the UMLS matches were evaluated in two ways. Smaller test corpora of 50–300 entry terms were utilized during the development of each processing step; corrections were made to the programs as needed to achieve the impact of each program on the entry terms. The automated concept matches from each round were also evaluated. First, the investigator took a random sample of 2% of the entries that matched only one UMLS concept, and manually examined the results for accuracy. Seventy-two of the 77 matches (92%) were deemed accurate. Of the 72 that matched, many did not match exactly but the match was semantically accurate. For example, the CC entry *arm laceration* matched UMLS concept C0432974, *laceration of upper limb*. A small number of matches (8%) were not accurate, for example, the CC entry *stepped on by sibling* matched UMLS concept C0337504, *step sibling*.

Those CC entries that matched more than one UMLS concept were evaluated using semantic information from the UMLS. A set of semantic groupings was developed by McCray et al. [68], to distill the 134 semantic types in the UMLS into 15 broader groups such as anatomy, disorders, and objects. In the current project, the semantic group was identified for each CC term and the matching UMLS concepts. Eighty-five percent of the CC entries matched at least one UMLS concept from the same semantic group.

5. Discussion

With this pilot study, we have demonstrated that text analysis is a useful approach for constructing a nursing terminology that is grounded in the language of domain experts. We began to build a CC terminology by extracting an initial set of concepts from triage nurses' CC entries. We accomplished this through identification of several characteristics of the natural language that nurses use in documenting CC, and then by developing NLP routines to address those patterns. We identified 5083 entries and corresponding UMLS concepts from the 13,494 unique CC entries, for potential inclusion in the CC terminology. The 5083 entry/concept matches identified in this pilot are at best a partial representation of the ED CC domain, and some may not be appropriate for inclusion in the final terminology. Additional review by domain experts and further NLP routines are needed to identify concepts for the 7371 non-matched ED CC entries. The routines will then be tested by applying them to a larger corpus including CC entries for

all ED visits to the three sites during a one year period, in order to provide a more complete representation of the ED CC domain by accounting for seasonality and less frequent CC entries. The remainder of the ED CC terminology will be built around this core set of concepts.

Authors have described the essential features of health care terminologies; key requirements for machine-readable controlled terminologies include concept-orientation and comprehensive content [69–71]. Though some ED CC terminologies have been locally developed at hospitals or by clinical systems vendors, there is no evidence that the systems contain key CC concepts based on a thorough analysis of the actual language used by nurses in describing patients' CCs [51–53]. These CC terminologies may also lack comprehensive vocabulary content for the domain represented by the system; most contain between 50 and 300 CC terms. The goal for the final terminology will be to follow the principle of warrant by including the most frequently used concepts from the natural language entries. While the optimum number of concepts necessary for the CC terminology has yet to be determined, our results show that it could be more than the 2978 concepts identified in this pilot study.

We found that some of the CC entries have a level of granularity that is finer than the standard vocabulary terms found in the UMLS. By deleting selected modifiers and qualifiers, we were able to broaden the entries and increase the concept match rate significantly. Since modifiers and qualifiers are frequently used in ED CC entries, they should be included in the final ED CC terminology, either through pre- or post-combination of modifier/qualifier-concept pairs. The final terminology will likely be hierarchical and allow for users to build applications that employ a smaller number of more general concepts or a larger number of more granular concepts. The terminology will also likely contain varying levels of granularity depending on the concept area. For example, in the emergency domain, more detail is needed about the concept of *chest pain* (*severe crushing* vs. *with coughing*) than about *rash* (*macular*, *vesicular*).

Abbreviations, acronyms, and truncated words were common in the ED CC entries, and required context-sensitive expansions in order to improve the match rate of entries containing them. Previous work has addressed abbreviations in both the biomedical literature and medical reports. Yu et al. [41] developed a software tool for identifying and extracting defined abbreviations in biomedical articles, and achieved an average 0.70 recall and 0.95 precision. Defined abbreviations are those in which the abbreviation and expanded form of the term occur together, such as: *abdominal pain* (*ABD*). They were also able to map 68% of the undefined abbreviations in their corpus to existing abbreviation databases. The researchers noted that ambiguous abbreviations were a problem in biomedical text. Stetson [45] found that abbreviations were common in three types of medical notes: signouts (end of shift notes to the next shift to care for the patient), ambulatory clinic notes and hospital discharge summaries. They also found that ambiguous abbreviations ranged from 8–18% of all abbreviations in the notes.

In the course of this pilot study, we developed a useful methodology for terminology construction. Using domain knowledge expressed in NLP algorithms, we were able to obtain a higher match rate with UMLS concepts than that obtained using the standard UMLS matching and normalization tools. Though our more aggressive approach introduces more risk for altering CC entries from their original representation, domain knowledge is essential in facilitating the appropriate processing of entries containing patterns such as punctuation and acronyms. In addition to supporting terminology construction, our NLP routines have other potential applications. For example, they may be useful for term and concept extraction from ED nurses' narrative notes, automatic classification (e.g., mapping text to NANDA or NIC), or linking the CC to patient outcomes [10].

Limitations of this pilot study include the relatively small corpus representing one region of the US. There is evidence that the nature of ED visits varies by season [72] and there may also be geographic variations. We plan to apply our methodology to a corpus of one year's CC entries from the three original hospitals, and in the future may expand to other regions of the country.

Another limitation is the lack of rigorous validation of the methods. Through this pilot work, we learned that there will be a need for manual review of the matched concepts as this project continues. The accuracy rates of 92% for single concept matches, and similar semantic groups for 85% of the multiple concept matches is encouraging. However, the final ED CC terminology will require a more accurate reflection of the language of the domain. While the accuracy judgments in this pilot were largely decided by one domain expert who was also the investigator, a more rigorous and independent review process will be needed to make decisions about concepts to include in the final terminology. The validation plan includes a formal review of the accuracy of the CC entry/UMLS concept matches by six domain experts, for all rounds of processing. One ED nurse and one physician from each of the participating study sites will participate in the formal validation, which will include a check that normalized entries are mapped correctly. For example, the entry *burn to chest* normalizes to the UMLS record, *chest burning*, which is very different than a burn injury to the chest.

The problem of ambiguous acronyms is significant and our limited AAT database may be inadequate to deal with the disambiguation necessary to address AAT in a larger corpus of ED text. Further context-sensitive rules may need to be developed. In addition, future work

is needed for CS processing. For example, a method is needed to address the subset of CS entries that have words of the semantic types sign or symptom, disease or syndrome, or pathological function bordering the slash, while not altering the entries with those semantic types bordering the slash that are in fact two or more separate concepts.

Future directions for our CC terminology work include identification of the UMLS source vocabulary that contains the most ED CC concepts, so we can follow the DEEDS recommendation to evaluate it for adaptation for ED CC. We also plan to collect emergency nursing concepts, terms and AAT that are not in the UMLS, and submit them to the National Library of Medicine, for consideration of inclusion in the national terminology system. We are also planning to compare and contrast the CC concepts and preferred terms from each of the three study sites, and have plans to expand our analysis to CC data from other regions of the country.

Another important step in the development or adaptation of a CC terminology for emergency nurses will be to more clearly define the CC [9]. While DEEDS and the ENA define the CC as representing as close to patients' words as possible [1,2], we have found that nurses' natural language CC's are often their interpretation of the patients' own words. For example, when a patient points to their left chest and states, ''I got a hurtin' right here,'' the nurse often records the CC as *chest pain*.

6. Conclusion

In this study, we tested the feasibility of text analysis as a tool for building concept-oriented nursing terminologies. We analyzed triage nurses' natural language entries and identified several characteristics of the CCs that needed to be addressed in order to identify concepts. We developed an initial set of NLP routines to address those characteristics, and increased the match rate with UMLS concepts. Text analysis is a useful approach for building a concept-oriented terminology for ED CC and should be further investigated in other nursing domains.

Acknowledgments

National Library of Medicine training grant number LM07071 supported Ms. Travers' work on this project. Funding was also provided by the North Carolina Office of Public Health Preparedness and Response in the Bioterrorism Branch of the Epidemiology Section of the Division of Public Health. The authors wish to thank Olivier Bodenreider, MD, PhD, for his mentorship during the development of the pilot methods and guidance on the focus of this manuscript.

References

[1] Emergency Nurses Association. Making the right decision: a triage curriculum (2nd ed.). Des Plaines, IL: Author 2001.
[2] National Center for Injury Prevention and Control. Data elements for emergency department systems, release 1.0. Atlanta, FA: Centers for Disease Control and Prevention; 1997.
[3] Lober WB, Karras BT, Wagner MM, et al. Roundtable on bioterrorism detection: information system-based surveillance. J Am Med Inform Assoc 2002;9(2):105–15.
[4] Yasnoff WA, Overhage JM, Humphreys BL, LaVenture M. A national agenda for public health informatics. J Am Med Inform Assoc 2001;8(6):535–45.
[5] Martinez R. Into the looking glass. Acad Emerg Med 1995;2:83–4.
[6] Garrison HG, Runyan CW, Tintinalli JE, et al. Emergency department surveillance: an examination of issues and a proposal for a national strategy. Ann Emerg Med 1994;24:849–56.
[7] Wears RL. Computer data base for ED visits. Ann Emerg Med 1992;21:67–8.
[8] Pollock DA, Adams DL, Bernardo LM, Bradley V, Brandt MD, Davis TE. Data elements for emergency department systems (DEEDS), release 1.0: a summary report. J Emerg Nurs 1998;24:35–44.
[9] Bradley V. Innovative informatics: development of an emergency data set: a worthwhile challenge. J Emerg Nurs 1996;22(3):238–40.
[10] Bradley V. Toward a common language: emergency nursing uniform data set (ENUDS. J Emerg Nurs 1995;21(3):248–50.
[11] World Organization of National Colleges (WONCA), Academies and Academic Associations of General Practitioners/Family Physicians. International classification of primary care (ICPC-2) (2nd ed.) New York: Oxford University Press; 1998.
[12] National Center for Health Statistics (NCHS). Reason for visit classification and coding manual. Hyattsville, Maryland: NCHS, Centers for Disease Control, US Department of Health and Human Services; 1997.
[13] Campbell JR, Carpenter P, Sneiderman C, Cohn SP, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, terminology, mapping, definitions, and clarity. J Am Med Inform Assoc 1997;4:238–50.
[14] McDonald CG. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc 1997;4:213–21.
[15] Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. J Am Med Inform Assoc 1998;5:503–10.
[16] Werley H, Lang N, editors. Identification of the nursing minimum data set. New York: Springer Publishing Company; 1988.
[17] North American Nursing Diagnosis Association (NANDA). Nursing diagnoses: definitions and classifications 1999–2000. Philadelphia: Author 1999.
[18] McCloskey JC, Bulechck GM, editors. Iowa intervention project nursing interventions classification (NIC. third ed. St. Louis: Mosby-Year Book; 2000.
[19] Bakken Henry S, Warren JJ, Lange L, Button P. A review of major nursing vocabularies and the extent to which they have the characteristics required for implementation in computer-based systems. J Am Med Inform Assoc 1998;5:321–8.
[20] Harris MR, Graves JR, Herrick LM, Elkin PL, Chute CG. The content coverage and organizational structure of terminologies: the example of postoperative pain. Proc AMIA Symp 2000:335–9.
[21] Bakken S, Warren JJ, Casey A, Konicek D, Lundberg C, Pooke M. Information model and terminology issues related to goals. Proc AMIA Symp 2002:17–21.
[22] Ozbolt J. Terminology standards for nursing: collaboration at the summit. J Am Med Inform Assoc 2000;7:517–22.

[23] Beyea SC. Data fields for intraoperative records using the Perioperative Nursing Data Set. AORN J 2001;73(5):952–4.
[24] Martin KS, Scheet NL. The Omaha system: applications for community health nursing. Philadelphia: W.B. Saunders; 1992.
[25] Guidotti TL. Bioterrorism and the public health response. Am J Prev Med 2000;18:178–80.
[26] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Meth Inform Med 1998;37:394–403.
[27] National Library of Medicine (NLM), Unified Medical Language System (UMLS). Available from: http://www.nlm.nih.gov/research/umls/.
[28] Liu H, Friedman C. A method for vocabulary development and visualization based on medical language processing and XML. Proc AMIA Symp 20000:502–6.
[29] McCray AT. The nature of lexical knowledge. Meth Inform Med 1998;37:353–60.
[30] Kreis C, Gorman P. Word frequency analysis of dictated clinical data: a user-centered approach to the design of a structure data entry interface. Proc AMIA Symp 1997:724–8.
[31] Spyns P. Natural language processing in medicine: an overview. Meth Inform Med 1996;35:285–301.
[32] Johnson SB, Gottfried M. Sublanguage analysis as a basis for a controlled medical vocabulary. Proc Ann Symp Comput App Med Care 1989:519–23.
[33] Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002;35(4):222–35.
[34] Harris Z. A theory of language and information: a mathematical approach. Oxford: Clarendon Press; 1991.
[35] Grishman R, Kittredge R, editors. Analyzing language in restricted domains: sublanguage description and processing. Hillsdale, NJ: Lawrence Erlbaum; 1986.
[36] Haas SW. Disciplinary variation in automatic sublanguage term identification. J Am Soc Info Sci 1997;28(1):67–79.
[37] Anderson JD, Perez-Carballo J. The nature of indexing: how humans and machines analyze messages and texts for retrieval part 2: machine indexing and the allocation of human versus machine effort. Info Proc Man 2001;27(2):255–77.
[38] Aronson AR, Bodenreider O, Chang HF, et al. The NLM indexing initiative. Proc AMIA Symp 2000:17–21.
[39] Ammons D. Psychology of the scientist: LXXIII. Miscommunication in technical writing. Percep Motor Skills 1998;86(2):418.
[40] Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care 1994:240–4.
[41] Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc 2002;9(3):262–72.
[42] Stetson PD, Johnson SB, Scotch M, Hripcsak G. The sublanguage of cross-coverage. Proc AMIA Symp 2002:742–6.
[43] Dale R, Moisl H, Somers H, editors. Handbook of natural language processing. New York: Marcel Dekker; 2000.
[44] Haas SW. Natural language processing: toward large-scale, robust systems. In: Williams ME, editor. Ann Rev Info Sci Tech, vol. 31. 1996. p. 83–119.
[45] Sager JC. A practical course in terminology processing. Amsterdam: John Benjamins Publishing Company; 1990.
[46] National Library of Medicine (NLM), Unified Medical Language System (UMLS) 2003AA User Documentation. Available from: http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML.
[47] Palmer DD. Tokenisation and sentence segmentation. In: Dale R, Moisl H, Somers H, editors. Handbook of natural language processing. New York: Marcel Dekker; 2000. p. 11–35.
[48] Travers DA, Bodenreider O. Identifying medical concepts in free text chief complaint data. Acad Emerg Med 2002;9:511 [abstract].
[49] Porter MF. An algorithm for suffix stripping. Program 1980;14:130–7.
[50] Yarowsky D. Word-sense disambiguation. In: Dale R, Moisl H, Somers H, editors. Handbook of natural language processing. New York: Marcel Dekker; 2000. p. 629–54.
[51] Ide N, Veronis J. Introduction to the special issue on word sense disambiguation: the state of the art. Comp Ling 1998;24(1):1–40.
[52] Olszewski RT. Bayesian classification of triage diagnoses for the early detection of epidemics. In: Recent Adv Artificial Intelligence: Proc 16th Int FLAIRS Conf 2003. Menlo Park, California: AAAI Press; 2003. p. 412–6.
[53] McCray AT. The nature of lexical knowledge. Meth Inform Med 1998;37:353–60.
[54] Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The unified medical language system: an informatics research collaboration. J Am Med Inform Assoc 1998;5:1–11.
[55] McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Ann Symp Comput App Med Care 1994:235–9.
[56] Payne TH, Martin DR. How useful is the UMLS Metathesaurus in developing a controlled vocabulary for an automated problem list? Proc Ann Symp Comput App Med Care 1994:199–203.
[57] Chute CG, Elkin PL. A clinically derived terminology: qualification to reduction. Proc AMIA Symp 1997:570–4.
[58] Eisner J. The developing electronic curriculum consortium. J Dent Ed 1990;54:598–9.
[59] Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. J Am Med Inform Assoc 1998;5:62–75.
[60] Christiansen T, Torkington N. Perl cookbook: solutions and examples for Perl programmers. Sebastopol, CA: O'Reilly; 1998.
[61] Bodenreider O. Using UMLS semantics for classification purposes. Proc AMIA Symp 2000:86–90.
[62] Schneiderman CA, Rindflesch TC, Bean CA. Identification of anatomical terminology in medical text. Proc AMIA Symp 19980:428–32.
[63] Huang X. Dealing with conjunctions in a machine translation environment. Proc Coding 1984:243–6.
[64] Hirschman L. Conjunction in meta-restriction grammar. J Logic Prog 1986;3:299–328.
[65] Rindflesch TR. Integrating natural language processing and biomedical domain knowledge for increased information retrieval effectiveness. Proc 5th Ann Dual-use Technol App Conf 1995:260–5.
[66] Dang HT, Palmer M. Combining contextual features for word sense disambiguation. Proc Word Sense Disambiguation: Recent Successes and Future Directions 2002:88–94.
[67] McCray AT, Brown AC. Discovering the modifiers in a terminology data set. Proc AMIA Symp 1998:780–4.
[68] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. MEDINFO 2001;10:216–20.
[69] Aronsky D, Kendall D, Merkely K, James BC, Haug PJ. A comprehensive set of coded chief complaints for the emergency department. Acad Emerg Med 2001;8(10):980–9.
[70] DiPasquale JT, Nichols JA, Garvey JL. Chief complaint coding in the emergency department. Acad Emerg Med 1995;2(5):442.
[71] Afilalo M, Unger B, Boivin JF, et al. Evaluation of a chief complaint classification system for comparing emergency department clienteles. Ann Emerg Med 2001;28(4):S82 [abstract].
[72] Haas SW, Travers DA, Waller AE, Hilligoss B, Cahill M, Pearce PF. Defining clinical similarity among ICD-9-CM diagnosis codes: diagnosis cluster schemes. In: Efthimiadis E, editor. Advances in Classification Research, vol. 12: Proceedings of the 12th ASIST SIG/CR Classification Research Workshop [in press].