Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 39 (2006) 196–208
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports

Wendy W. Chapman *, John N. Dowling

Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

Received 28 April 2005
Available online 22 August 2005

Abstract

Evaluating automated indexing applications requires comparing automatically indexed terms against manual reference standard annotations. However, there are no standard guidelines for determining which words from a textual document to include in manual annotations, and the vague task can result in substantial variation among manual indexers. We applied grounded theory to emergency department reports to create an annotation schema representing syntactic and semantic variables that could be annotated when indexing clinical conditions. We describe the annotation schema, which includes variables representing medical concepts (e.g., symptom, demographics), linguistic form (e.g., noun, adjective), and modifier types (e.g., anatomic location, severity). We measured the schema's quality and found: (1) the schema was comprehensive enough to be applied to 20 unseen reports without changes to the schema; (2) agreement between author annotators applying the schema was high, with an F measure of 93%; and (3) the authors made complementary errors when applying the schema, demonstrating that the schema incorporates both linguistic and medical expertise.

Keywords: Indexing; Natural language processing; Reference standard; Grounded theory; Annotations

1. Introduction

Automatically indexing clinical conditions from free-text medical sources, such as journal articles and dictated clinical reports, could be useful for many purposes, including discovering knowledge from the literature, identifying patients eligible for clinical trials, decision support, and outbreak detection. Researchers have developed automated indexing systems that use natural language processing techniques to automatically annotate relevant clinical information.

Evaluating automated indexing applications requires comparing automatically indexed terms against manual reference standard annotations. However, there are no standard guidelines for determining which words from a textual document to include in manual annotations, and the vague task can result in substantial variation among manual indexers. For example, in the sentence *Patient is experiencing severe left-sided chest pain radiating down her arm*, possible manual annotations of the clinical condition include *pain*, *chest pain*, *left-sided chest pain*, *severe left-sided chest pain*, and *severe left-sided chest pain radiating down her arm*.

Our objectives were (a) to induce from emergency department (ED) reports a standardized annotation schema for manually annotating clinical conditions that would integrate both medical and linguistic knowledge and (b) to measure the quality of the annotation schema, looking particularly for the schema's comprehensiveness, the ability to apply the schema with high agreement, and the contribution of both authors as an indication that the schema represents both linguistic and medical expertise.

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

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.06.004

2. Background

Several applications have been developed to automatically index clinical information from the literature and from clinical reports, including MetaMap [1], Saphire [2], IndexFinder [3], and those developed by Elkin et al. [4], Nadkarni [5], Berrios et al. [6], and Srinivasan et al. [7]. Indexing applications typically consist of two steps: locating clinical information in the text and mapping the text to clinical concepts in a standardized vocabulary, such as the UMLS.

Indexing usually focuses on identifying contiguous sections of text (generally noun phrases) and does not identify modifiers that are not contiguous to the topic being modified. For example, MetaMap successfully maps the phrase *enlarged heart* to the UMLS concept C0018800 Cardiomegaly but will not successfully map the entire concept if the modifier *enlarged* is not contained in the same phrase as *heart* (e.g., *the heart is enlarged*). Applications such as MedLEE [8], M+ [9], and applications created by Taira and Soderland [10], Baud et al. [11], and Hahn et al. [12] employ more sophisticated NLP techniques for grouping related information that may not be contiguous in the text. The annotation schema we developed and evaluated in this paper was designed for manually indexing clinical conditions from contiguous segments of text.

Evaluating the performance of an indexing application requires manual generation of reference standard annotations, and the first step is to manually identify the clinical concepts from the text. Pratt and Yetisgen-Yildiz [13] recruited six annotators to identify clinical concepts in titles from Medline articles. Friedman et al. [8] asked three physicians to identify relevant clinical terms in discharge summaries. In a paper by Chapman et al. [14], a physician indexed clinical concepts related to lower respiratory syndrome in ED reports. General guidelines for the task were given to the annotators in all of these studies. The two studies utilizing multiple annotators showed substantial variation among the annotators. The study involving a single annotator showed that the single physician made many indexing mistakes.

Our aim was to create a specific annotation schema for annotating clinical conditions from contiguous text that would reduce the vagueness inherent in the indexing task and would enable multiple annotators to perform the same task with high agreement. We focus on ED reports, because our research is ultimately concerned with detecting outbreaks from ED data.

Linguistic annotation is a key topic area in natural language processing research, and researchers have developed tools for creating annotations [15] and formal schemas to guide annotators [16]. The simplest type of annotation involves classification of text (documents, sentences, words, etc.) into a predefined set of values. The Penn Treebank Project has annotated every word in several large corpora with part-of-speech tags [15]. Annotators have classified chest radiograph reports into chest abnormalities [17,18], ED reports into bioterrorism-related syndromes [19], and utterances in tutoring dialogues into emotional states [20]. More complex annotation tasks involve not only classifying text into categories but also encoding more detailed characteristics of the annotations. For example, Green [21] developed a Bayesian network coding scheme for annotating information in letters to genetic counseling clients. In addition to classifying text segments into categories, such as history, genotype, symptom, test, etc., the annotators also represented probabilistic and causal relationships among the annotated concepts. Wiebe and colleagues [22,23] annotated text segments in newswire articles for expressions of opinion and emotion. For every identified opinion or emotion, annotators recorded additional information, including the source, the target, the intensity, and the attitude type. Evaluations of medical language processors that encode clinical conditions from text [18,24] require annotation not only of the condition itself but of the condition's presence or absence, severity, change over time, etc. The Genia corpus is a set of annotated abstracts taken from National Library of Medicine's MEDLINE database. The GENIA Corpus [25] contains annotations of a subset of the substances and biological locations involved in reactions of proteins and also collects part-of-speech, syntactic, and semantic information. Many annotation tasks require direct annotation of words in the text and rely on fairly consistent boundary segmentation by the annotators. The schema we developed focuses on which semantic categories to annotate and on which words to include in the annotation.

Annotation schemas act as knowledge representation tools involving semantic categories (e.g., types of information described in a genetic counseling letter) and specialized lexicons (e.g., names of symptoms or genotypes that are annotated). In some cases, such as annotating clinical concepts from patient reports, the annotations may be able to be mapped to terms from a controlled vocabulary [8], such as those contained in the Unified Medical Language System (UMLS). However, sometimes the annotation categories do not map directly to an existing lexicon. For instance, in our project, terms describing how to annotate the linguistic form of a clinical concept are not concepts contained in the UMLS Metathesaurus. Compiling a vocabulary for linguistic annotation that could be re-used by others would be a useful addition to biomedical text annotation, and we believe this project could contribute to such a compilation.

In this paper, we describe how we generated an annotation schema for manually annotating clinical conditions in ED reports. We present quantitative measures

of the schema's quality and discuss implications of our findings.

3. Methods

The objectives of this study were: (a) to integrate both medical and linguistic knowledge in establishing a standardized annotation schema for manually annotating clinical conditions from ED reports and (b) to measure the quality of the annotation schema. Both inducing the annotation schema and measuring its quality involved annotation of ED reports. We describe below the setting and selection of reports, our method for creating the annotation schema, and the outcome measures we applied to measure the quality of the schema.

3.1. Setting and selection of ED reports

The study was conducted on reports for patients presenting to the University of Pittsburgh Medical Center (UPMC) Presbyterian Hospital ED from February to May, 2004. Patients without an electronic ED report were excluded from the study, which was approved by the University of Pittsburgh's Institutional Review Board. We randomly selected 60 reports to be manually annotated by the authors for this study. We used 40 reports for creation of the annotation schema and 20 to validate the schema, as shown in Fig. 1.

3.2. Objective 1: To establish a standardized annotation schema for manually annotating clinical conditions from (ED) reports

Our aim was to create an annotation schema that would enable the two authors—one physician and one informatician—to individually annotate clinical conditions from ED reports in the same way. Both authors participated in creation of the schema, because we believed the task required both medical and linguistic expertise. We also believed the schema should reflect what actually occurs in the text, so we based our methodology for schema creation on the sociologic tradition of grounded theory [26,27], which refers to theory that is developed inductively from a corpus of data. Grounded theory has been applied to the biomedical domain for many studies, including development of theories of caregivers' and families' attitudes and experiences with patients [28,29]. Transcripts of interviews or social interactions are typical data sources for grounded theory studies. The grounded theory approach involves reading and re-reading text to discover or label variables and their interrelationships and involves a constant interplay between proposing and checking the theory. The end result of a grounded theory study is a theory, and, if performed well, the resulting theory will at least fit one dataset perfectly.

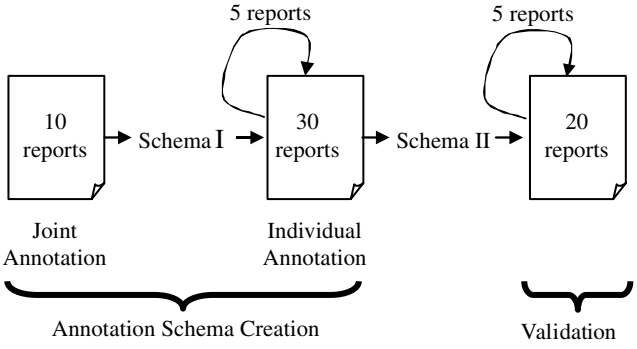

Fig. 1. We used 40 reports to induce the annotation schema and 20 reports to validate the resulting schema (Schema II). Schema creation involved iteratively annotating sets of five reports, discussing differences, and changing the existing schema. No changes were made to Schema II during the validation phase.

For this project, the theory being developed was an annotation schema to guide annotation of clinical conditions from text. The data source from which we induced the theory was dictated ED reports, and the variables to be discovered from text were syntactic and semantic elements used to describe the clinical conditions in the text, such as parts of speech and types of clinical conditions to be annotated. Examples of potential variables include symptoms, physical findings, nouns, and anatomic locations. As we induced the variables, we organized them into conceptual groups. The resulting annotation schema was meant to represent all potentially annotatable text describing clinical conditions in ED reports and could be applied to a particular annotation project by determining which of the variables should be annotated and which should not.

We developed the annotation schema in a multi-staged approach. First, we recorded a general, theoretical statement that declared in broad terms the annotation goal: to annotate the most specific, atomic clinical conditions in the text without annotating any modifiers related to time, uncertainty, or negation. Next, we iteratively applied grounded theory to sets of reports to enumerate specific variables to be considered when annotating clinical conditions. As an example, consider the sentence *This is a **AGE [in 40s]-year-old black woman who complains of 5 days of upper respiratory infection symptoms*. Five variables could be induced from this sentence. First, demographic information, such as the patient's age, race, and sex, could potentially be considered a clinical condition, so we could induce the variable (1) demographics. Second, the patient has some symptoms, and a symptom could be considered a clinical condition, so we could induce the variable (2) symptoms. In this example, the symptoms are specified by the phrase *5 days of upper respiratory infection*. *Respiratory* indicates an anatomic location of the infection symptoms, *upper* modifies the vertical location in the respiratory tract,

and *5 days* describes the temporal duration of the symptoms. Therefore, we could induce the following variables: (3) anatomic location; (4) verticality; and (5) numeric duration.

We iteratively read ED reports, defined variables occurring in the text, grouped the variables into related concepts, and determined whether the variables should be annotated or not for our annotation task. As can be seen in the above example, variables induced through this process remained faithful to the annotation goal while providing detailed information about how to realize that goal.

We applied this process to 40 reports (see Fig. 1). First, the authors jointly annotated ten reports. During annotation, we discussed every clinical condition described in the reports and jointly generated a list of variables representative of the examples we encountered (Schema I). Second, we generated Schema II as follows: Both authors applied Schema I while individually annotating 30 reports in increments of five. Fig. 2 shows an excerpt of an ED report with annotations of clinical conditions. After every five reports, we used a Python program to compare our annotations, and we discussed every discrepancy between us, making changes to the schema as needed. A change to the schema comprised either addition of a new variable to fit a previously unencountered example (e.g., medication names, such as *thorazine*) or addition of an exception to an existing variable (e.g., medication names that are part of a clinical condition, such as *allergic to thorazine*). We did not allow changes to whether or not a variable should be annotated, which would have required re-reading reports we had already annotated and undoing relevant annotations. We completed the schema creation process

CHIEF COMPLAINT:

Back pain

HISTORY OF PRESENT ILLNESS:

The patient is a **AGE[in 40s]-year-old man who has been having difficulty for approximately 48 hours prior to presentation He said that he has had intermittent **abdominal discomfort** which has prevented him from straightening fully upright It tends to wax and wane in intensity and, in fact, does go away completely at times At the time that I am seeing him, he has no **pain his abdomen** whatsoever He said that it was initially a **pain in his kidney** and points to an area in his right flank Initially, the **pain** was seen somewhat with urination, and the patient also complained of **some pain with defecation** He tells us that he very in touch with his body and that he knew that it was his kidney He said then he could feel it traveling in his leg to his left arm and shoulder where he had **some tingling** and **discomfort in his left arm** and hand He does not relate a history of true **numbness**, **weakness** or **clumsiness** He did not have any **difficulty using the hand** This symptom appeared to last for several hours and then resolve completely and spontaneously, that was yesterday There was no history of any associated **trauma**, **vision changes**, **lightheadedness**, **dizziness** or **diaphoresis** The patient has not had **chest pain** or **difficulty breathing** He denies any history of **heavy liftin** or other physical activity, which might have caused his **discomfort . . .**

Fig. 2. An excerpt from a de-identified report showing annotations of clinical conditions in bold. The schema takes into account the linguistic and medical context of the words. For example, the word *hand* is annotated in *difficulty using the hand*. However, *hand* is not annotated in *discomfort in his left arm and hand*, because *hand* is a modifier joined to another modifier (*arm*) by a conjunction. Similarly, *defecation* and *urination* are semantically similar, however *defecation* is annotated in *some pain with defecation*, but *urination* is not annotated in *the pain was seen somewhat with urination*, because *urination* is part of a verb phrase modifier of pain.

over a period of approximately three months, completing one or two cycles of annotations every week. Schema II can be downloaded from http://web.cbmi.pitt.edu/chapman/Annotation_Schema.doc.

For all individually annotated reports, we generated reference standard annotations based on consensus of the authors after discussing disagreements. The reference standard annotations will be used for a future study evaluating the performance of a negation algorithm against human negation of clinical conditions and were used in this study to help assess the quality of the annotation schema, as described below.

3.3. Objective 2: Measure the quality of the annotation schema

Our work represents initial research in creating a useful annotation schema for clinical conditions. To begin assessing the quality of Schema II, we measured the schema's completeness, the authors' ability to apply the schema with high agreement, and each author's contribution in creation of the schema.

3.3.1. Completeness of annotation Schema II

We calculated the number of variables and exceptions included in Schema I and compared that to the number of variables and exceptions in Schema II. To determine whether Schema II was comprehensive enough to apply to unseen reports, we froze Schema II and annotated 20 more reports, individually annotating reports in increments of five followed by discussion of disagreements (Fig. 1). We counted the number of new variables or exceptions that would have been required to successfully annotate the 20 new reports but did not alter Schema II.

3.3.2. Agreement when applying Schema II

To quantify agreement between annotators, we applied the *F* measure, as recommended by Hripcsak and Rothschild [30]. When manually annotating clinical conditions from text, the number of true negatives in the text (i.e., words that were correctly not annotated) is poorly defined, because the conditions may overlap or vary in length. For this reason, calculating standard agreement measures, such as *kappa*, is impossible. The *F* measure does not require the number of true negative annotations. Commonly used in information retrieval and information extraction evaluations, the *F* measure calculates the harmonic mean of recall (sensitivity) and precision (positive predictive value), as follows: $F = \frac{(1+\beta^2)\times\text{recall}\times\text{precision}}{(\beta^2\times\text{precision})+\text{recall}}$. The value of β can be used to weight recall or precision more heavily, and when $\beta = 1$, the two measures are weighted equally. In this case, the *F* measure is equal to positive specific agreement, which is the conditional probability that one rater will agree that a case is positive given that the other one rated it as positive, where the role of the two raters is selected randomly. The *F* measure approaches the value of *kappa* if the unknown number of true negatives is large [30].

Calculating recall and precision requires counting the number of true positive (TP), false positive (FP), and false negative (FN) annotations. To do this, we selected one annotator's answers to be the reference standard annotations and the other's to be the comparison annotations (which of the two annotators is the reference standard is irrelevant, because the *F* measure is equivalent for each annotator compared against the other). Comparing the comparison standard annotations to the reference standard annotations, we calculated the number of TP, FP, and FN annotations. If a comparison annotation overlapped but was not identical with a reference standard annotation, the proportion of overlapping words were used as the TP count, and the proportion of extraneous or missing words were used as the FP or FN count, respectively, as shown in Table 1. The TP, FP, and FN counts summed to one for each reference standard annotation.

For the 20 reports individually annotated with Schema II, we calculated agreement between the author annotators. For each annotator, we calculated the average number of annotations per report, the average length of the annotations, and the *F* measure.

3.3.3. Contribution of each author to annotation Schema

Author WWC has a B.A. in Linguistics and a Ph.D. in Biomedical Informatics. Author JND is a physician with 35 years of experience who received an M.S. in Biomedical Informatics late in his career. We believed the expertise of both authors was crucial to creation

Table 1
TP, FP, and FN counts for example annotations

Comparison annotation (position)	Reference annotation (position)	TP count	FP count	FN count
Fever (456–460)	—	0	1	0
—	Vision changes (992–1007)	0	0	1
Urinary tract infection (72–95)	Urinary tract infection (72–95)	1	0	0
Productive cough (1033–1051)	Cough (1047–1051)	1/2	1/2	0
Pain (619–622)	Severe pain in the abdomen (612–637)	1/5	0	4/5

Position is the character position in the ED report.

of an effective annotation schema and attempted to quantify the contribution of each author to the schema.

A potential weakness in a consensus-driven task is that one member will dominate the decisions by the group. In our case, it was possible that creation of the schema would be dominated by the physician, who possessed more medical knowledge, or by the informatician, who possessed more linguistic knowledge. If decisions about adding variables or exceptions were dominated by one author, we would expect that bias to be reflected in the similarity of the dominant individual's annotations when compared to the reference standard annotations, which reflect adherence to the annotation schema. We compared individual annotations against the consensus reference standard annotations for the 30 reports that were individually annotated when inducing Schema II (such a comparison was not possible for Schema I, because the authors jointly annotated the ten reports). We calculated the *F* measure, recall, and precision for the physician annotator vs. the reference standard and for the informatician annotator vs. the reference standard.

We also examined annotation errors made by each author when applying the schema. For each author, we calculated the proportion of disagreements due to missed annotations (FN), extraneous annotations (FP), or missing or extraneous modifiers (partial FN and FP). We also performed an error analysis for each annotator's disagreements with the reference standard for the 20 reports annotated with Schema II. The error analysis classified each error into one of three causal categories: lack of medical knowledge, incomplete understanding of the schema, or random error (i.e., we knew the rule but just did not follow it). We compared the distribution of disagreement categories for both annotators.

4. Results

4.1. Objective 1: To establish a standardized annotation schema for manually annotating clinical conditions from (ED) reports

We generated 1485 reference standard annotations for the 30 ED reports used to create Schema II, with an average of 45 clinical conditions per report. Schema I, which was created from jointly annotating 10 reports, contained 39 variables and two exceptions to existing variables. Schema II contained 45 variables and ten exceptions, which we grouped into three conceptual groups: Medical Concepts, Linguistic Form, and Modifier Type. Schema II is shown in Table 2. For every variable, Table 2 gives examples from the annotated reports and shows whether we annotated that particular variable or not for our annotation task. For example, for our task, we did not annotate demographics or medications. Other researchers could apply the variables in Table 2 with different annotation instructions suited to their particular task.

4.2. Objective 2: Measure the quality of the annotation schema

4.2.1. Completeness of annotation Schema II

Once we froze Schema II and annotated 20 additional reports, we did not encounter any clinical conditions requiring addition of a new variable or an exception.

4.2.2. Agreement when applying Schema II

Overall agreement between the annotators was high. For the 20 reports annotated with Schema II, author WWC individually generated 879 annotations (average of 44.0 per report), and author JND 891 (average of 44.6 per report). The average annotation length was identical for each author, with 14.5 characters and 2.1 words per clinical condition. The *F* measure between annotators was 93%.

4.2.3. Contribution of each author to annotation schema

When compared with reference standard annotations of the 20 reports, the *F* measure of physician annotations was 96.8% and of informatician annotations was 95.8%. Average recall was 96.1% for the physician and 94.5% for the informatician; average precision was 97.5% for the physician and 97.1% for the informatician.

Looking at individual annotation errors reveals differences between annotators that are not evident in the high *F* measure. Errors led to disagreements between annotators, and disagreements led to changes in the annotation schema. Therefore, it is reasonable that the types of errors made by the annotators reflect, in part, each author's contribution to the schema. For instance, both annotators averaged seven FP's per report, but the informatician averaged 4.8 more FN's per report (12.9 vs. 8.1), with a large proportion of the FN's resulting from two of the 20 reports describing orthopedic conditions the informatician was not familiar with. False negatives could be due to not annotating a relevant concept (e.g., not annotating *abdominal pain*) or annotating a relevant concept but leaving off a relevant modifier (e.g., annotating *pain* instead of *abdominal pain*). The informatician's FN's were more often due to not annotating a relevant concept than were the physician's (77% vs. 67%). Similarly, a larger proportion of the informatician's FP's were due to extraneous annotations than were the physician's (73% vs. 62%).

The error analysis reinforces this difference between the informatician and physician annotators. Table 3 shows the cause of FN's and FP's for each annotator on the 20 reports. Although the two annotators agreed well, their errors were complementary. About three-

Table 2
Summary of the annotation schema (Schema II)

Variables	Definition or Examples	Annotate	Annotation Example
Medical Concepts			
Atomic clinical condition*	Individual clinical conditions that may or may not be related	Y	*knee pain from osteoarthritis*
EXCEPTION: Multiple, inextricable clinical conditions	Multiple clinical conditions that share elements and cannot be extricated from each other	Y	*Normal PR, QRS, and QT intervals*
Historical findings	Findings that occurred previous to the current visit	Y	*Past history of pneumonia*
Symptoms*	Real or imagined symptoms	Y	*Patient presents with SOB*
Qualitative physical findings*	Findings that do not require a numeric value	Y	*Has good perfusion*
Quantitative physical findings*	Physical findings requiring a numeric value to be meaningful	N	*Temperature 38.6*
EXCEPTION: with qualitative modifier	Physical finding with a numeric value and a qualitative modifier	Y	*1+ pitting edema*
Qualitative radiological findings*	A radiological finding that does not require a numeric value	Y	*Left lower lobe opacity*
Qualitative diagnostic test results*	A lab test result with a non-numeric value – without the non-numeric value, the lab test would not be a clinical condition	Y	*low sodium*
Quantitative lab test results*	A diagnostic test result (e.g., lab test, radiological test, etc.) with a non-numeric value – without the non-numeric value, the lab test would not be a clinical condition	N	*calcium 9.3*
Diagnoses	Diseases or syndromes	Y	*Community Acquired Pneumonia*
Demographics*	Sex, race, age, or geographic location	N	*Sixty-year old gentleman*
Medications*	Medication names	N	*Currently taking thorazine*
EXCEPTION: Included in clinical condition	A clinical condition involving the medication	Y	*Patient is allergic to thorazine*
Non-specific clinical words*	Words indicating a clinical condition that do not give enough specific information to be clinically meaningful	N	No past medical history
EXCEPTION: with organ	A non-specific clinical word modified by an organ or system	Y	*cardiac medical history*
Trauma/Accident	An agent, force, or mechanism that causes trauma	N	*suffered a fall*
EXCEPTION: modifies a clinical concept being annotated	Trauma/accident provides more specific information about a clinical concept being annotated	Y	*shortness of breath after fall*
EXCEPTION: with organ system or body location	Trauma/accident to a specific organ system or body location	Y	*trauma to his joints*

Table 2 (*continued*)

Variables	Definition or Examples	Annotate	Annotation Example
Vague causes of a clinical condition	A cause for a clinical condition that is too vague to be clinically meaningful	N	*I do not see a cardiovascular cause for her findings*
Therapeutic interventions	Procedures that occurred at the visit	N	*A 6-French IJ-introducer was put into the right internal jugular vein*
Linguistic Form			
Noun phrase or verb phrase	A noun alone, preceded by an adjective, compounded with another noun, or followed by a prepositional phrase; a verb alone, modified by an adverb, followed by a prepositional phrase, or followed by a noun phrase	Y	*She vomited*
Clauses or verb phrases modifying a nominal clinical condition*	A verb phrase or clause modifying a clinical condition expressed as a noun phrase	N	*tenderness when I press on the sternum*
EXCEPTION: clinical condition alone is meaningless	Nominal clinical condition is meaningless without modifying clause or verb phrase	Y	*Difficulty when going from bed to bathroom*
EXCEPTION: verb phrase ends in *-ing*	Verb phrase modifying nominal clinical condition ends in *-ing*	Y	*Pain radiating down her anterior thigh*
Lexicalized expressions*	Lexicalized expressions are phrases or sentences that are treated as single words. Lexicalized expressions can be annotated in any linguistic form, regardless of other rules, as long as they belong to this short list	Y	*Tobacco, alcohol, drinks, smokes, drug, regular rate and rhythm, PERRLA, EOMI*
Section headings*	Headings that may contain information important for correct interpretation of the clinical condition	N	*Extremities: nontender*
Unambiguous adjective	An adjective that can have only one interpretation may stand alone without the noun it modifies, if another rule restricts annotation of both	Y	*Sclerae are icteric*
Ambiguous adjective*	An ambiguous adjective has multiple interpretations and, when another rule restricts annotation of both, can not stand alone without the noun it modifies	N	*Abdomen is soft*
Determiners used with nominal clinical condition*	A definite or indefinite article, possessive adjective, or demonstrative adjective used with a noun	N	*the pain*
EXCEPTION: determiner is internal to phrase*	Determiner is contained within the phrase being annotated	Y	*Pain in her lower back*
Coordinated modifiers*	Modifiers separated from the clinical condition by a coordinating conjunction	N	*Head or neck pain*

(*continued on next page*)

Table 2 (*continued*)

Variables	Definition or Examples	Annotate	Annotation Example
Modifier Types			
Integral to meaning*	A modifier that changes the meaning of the clinical condition and is necessary for correct interpretation	Y	*COPD exacerbation*
Void of clinical meaning*	A modifier that does not add any clinical meaning	N	*known drug allergies*
Anatomic location*	A modifier indicating the anatomic location of the condition	Y	chest *pain*
Sidedness*	A modifier indicating the sidedness of the condition	Y	*left-sided chest pain*
Verticality*	A modifier indicating the verticality of the condition	Y	*midline low back pain*
Quality*	A modifier indicating the quality of the condition	Y	*stabbing chest pain*
Non-numeric severity quantifiers*	A non-numeric quantifier indicating the severity of a condition	Y	*significant amount of bleeding*
Numeric quantifiers*	A numeric quantifier of a clinical condition	N	*vomited twice*
"otherwise" modifiers	Modifiers that indicate a condition in contrast to a previously mentioned condition by using a word like *other* or *otherwise*	Y	*other pneumonic process*
Causal/association*	An entity or action that caused or is associated with the clinical condition being annotated. Both the cause and effect are being described	Y	*tenderness to palpation*
Consequence modifiers without the causality*	A modifier indicating that the clinical condition is resulting or associated with a cause that is not described	N	*Associated palpations*
Uncertainty*	A modifier indicating uncertainty of the condition	N	*possible pneumonia*
Negation*	A modifier indicating the absence of the clinical condition	N	*no coughing*
Time of onset*	A modifier indicating the time of onset for the clinical condition	N	*chronic knee pain*
EXCEPTION: integral to meaning of phrase*	Time of onset modifier is not meant to indicate time of onset but is integral to the meaning of the clinical condition.	Y	*Acute myocardial infarction*
Change over time*	A modifier indicating a change over time for the clinical condition	N	*increase in peripheral edema*
EXCEPTION: required to be a clinical concept	A clinical condition would not exist without the change over time modifier – the change over time is the clinical concept	Y	*decrease in hearing*

Table 2 (*continued*)

Variables	Definition or Examples	Annotate	Annotation Example
Non-numeric duration*	A non-numeric modifier indicating the temporal duration of the clinical condition	N	*persistent chest pain*
Numeric duration*	A numeric modifier indicating the temporal duration of the clinical condition	N	*1 day of SOB*
Non-numeric Continuity/episodic*	A non-numeric modifier indicating that the clinical condition was episodic or continuous	N	*Occasional blurry vision*
Numeric Continuity/episodic*	A numeric modifier indicating the clinical condition was episodic or continuous	N	*diarrhea times 1*
Observer's perception*	A modifier indicating the observer's perception of the clinical condition	N	*Obvious distress*
Evidence*	A modifier indicating evidence or signs of a clinical condition	N	*Evidence of fracture*

Column 1 shows the three categories of variables included in Schema II. An asterisk (*) indicates that the variable also existed in Schema I. Column 2 gives definitions of the variables. Column 3 indicates whether we annotated (Y) or did not annotate (N) the variable for our annotation task. Column 4 provides a relevant example in which the text that should be annotated is underlined.

Table 3
Error analysis for physician and informatician listing three causes for error

	FP	TP-FP	FN	TP-FN	Total	Proportion
Physician						
Lack of medical knowledge	0	0	12	0	12	0.07
Misunderstanding of schema	15	10	6	22	53	0.76
Random	0	0	12	0	12	0.17
Informatician						
Lack of medical knowledge	8	20	25	16	69	0.74
Misunderstanding of schema	8	1	3	6	18	0.19
Random	0	0	6	1	7	0.07

TP-FP errors are annotations with extraneous modifiers, and TP-FN errors are annotations missing modifiers. The last column lists the proportion of all errors due to that cause.

fourths of the informatician's errors were due to a misunderstanding of what constitutes clinically meaningful information, e.g., extraneously annotating *sitting upright*, annotating the clinically irrelevant modifier *at home* in *having difficulty getting around at home*, not annotating *passive range of motion*, or not annotating *to command* in *moves all extremities to command*. Three-fourths of the physician's errors were due to not accurately applying the guidelines to clinical conditions in the text, e.g., extraneously annotating *pink* (which is an ambiguous adjective), extraneously annotating the modifier *use* instead of just the lexicalized expression *drug* in *drug use*, not annotating *slightly slurred* (which is an unambiguous adjective), and not annotating the modifier *relieved by nitroglycerin* in *chest discomfort relieved by nitroglycerin*. Both annotators had a fair number of FN's that were random errors, due simply to overlooking something they knew should be annotated.

5. Discussion

Our first objective was to create an annotation schema directly from ED reports. We hoped the schema would incorporate medical and linguistic knowledge and would allow annotators to exhibit high agreement when indexing clinical conditions from ED reports. The authors—one physician and one informatician—used a methodology based on grounded theory, involving repeating cycles of annotations from text combined with discussion of disagreements and formulation of relevant variables for annotation, to generate a fairly complete schema from the ED reports. The resulting schema (Schema II) contained six more variables than Schema I, suggesting that the bulk of the schema was derived from only ten reports. The main change to Schema II while annotating 30 more reports was the addition of eight new exceptions when we encountered new situations in the text. For example, in creating Annotation

Schema I, we decided not to annotate non-specific clinical words, such as *problem* or *history*. However, in the 30 reports, we encountered situations in which a non-specific clinical word was combined with an organ or system, such as *liver problem* or *cardiac medical history*. When combined with an organ or system, the condition became meaningful, and we added an exception to the variable.

Our second objective was to measure the quality of the annotation schema. A theory—in our case an annotation schema—resulting from grounded theory should fit a particular data set very well. After freezing Schema II and annotating 20 additional reports, we encountered no clinical conditions in the reports that required additional variables or exceptions, suggesting that the theory we induced from the training text was quite comprehensive for ED reports. However, every new report has the potential of introducing a variable we have not yet seen, and the annotation schema could feasibly require changes when applied in the future.

Applying the schema to a new annotation task could be carried out similarly to the methodology we used to induce the schema. First, a researcher could determine which of the variables should or should not be annotated. For example, although we did not annotate demographics for our annotation task, demographics may be a useful variable to annotate for another task. Next, the researcher could train expert annotators using the current schema, the annotators could apply the schema to a few pilot reports, and the researcher could measure their agreement. It is likely that the schema already includes the major syntactic and semantic variables consistently used when describing clinical conditions in ED reports. However, the researcher may want to make changes to the schema based on the experts' annotations and feedback. Types of changes most likely to be required when applying the schema to new reports include exceptions to existing variables and expansions of variables that were not required in our data set. For example, one of the variables is qualitative radiological finding. It would be logical to add a variable for quantitative radiological findings, but we did not encounter any in the reports we annotated. The experts could iteratively annotate pilot reports until their agreement surpasses a certain threshold. At that point, experts could annotate the test reports using the most current version of the schema. Because annotation would ultimately be done individually by each annotator, the number of annotators and their geographic location should not influence the ability to apply the schema to a new annotation task.

One measure of an annotation schema's quality is the ability for annotators to apply the schema consistently to text. We measured the agreement of the authors when applying Schema II to annotation of ED reports. Agreement between the annotators was high, with an *F* measure of 93%, which is equivalent to a positive specific agreement of 93%. If the unknown number of TN's in the text is large and unknown, the *F* measure is equivalent to *kappa*. A text annotation task potentially complies with the assumption of a large number of TN's, because the text contains numerous words and phrases that should not be annotated. In our case, there were 44.9 reference standard annotations per report that averaged 2.2 words long, comprising approximately 99 TP words per report. On average, a report contained 441 words, so we can estimate the number of TN single words per report as 342 (441–99). In addition to single TN words, the reports also contain multiple-word TN phrases that are not easily countable. So, although the number of TN's is unknown, it is likely to be quite large in our data set.

The reference standard annotations generated by consensus of the two annotators represented improvements to both annotators, indicating that the annotation schema was not strongly biased by one or the other of the authors. Both annotators consistently made mistakes, but their errors differed in a predictable way: most of the physician's errors were due to misapplication of the schema, whereas most of the informatician's errors were due to misunderstanding of what was clinically relevant. Complementary errors reinforce our belief that the linguistic expertise from the informatician was necessary in complementing the medical expertise of the physician so that the resulting schema would represent both linguistic and medical variables evident in descriptions of clinical conditions in ED reports.

The physician's recall was higher than the informatician's, and the informatician's false negative rate was higher than the physician's, probably reflecting the informatician's inability to match the physician's knowledge of what was clinically relevant in the text. Fewer mistakes by the physician is a positive result, because physicians will probably be the target users of the annotation schema. The fact that the physician author could apply the annotation schema better than the informatician suggests that other physicians could be trained to apply the annotation schema in spite of a lack of formal linguistic training. Lower performance by the informatician suggests that a non-physician annotator—who would be less expensive—could not perform equivalently to a physician annotator.

The majority of annotation errors were due to extraneous or missing concepts rather than to extraneous or missing modifiers. This result may indicate a couple of trends. First, individuals annotating text sometimes miss concepts they should be annotating, which is why it is important to include multiple annotators. Second, the schema's conceptual groups Medical Concepts and Linguistic Form were probably more difficult to implement than the group Modifier Types.

In spite of the fact that the annotators created the annotation schema, applied the evolving schema to 50 reports, and annotated over 2400 clinical conditions

over the course of this project, their agreement was still not perfect. In fact, 7–17% of the errors were not systematically related to the annotator's weaknesses but were random errors that may never be eliminated (see Table 3). Perfect agreement would not be expected by any reference standard generation task involving text, because subjective disagreement on what constitutes a valid clinical concept inevitably occurs when two or more people—even people with significant expertise in a given domain—are annotating clinical concepts. An *F* measure of 93% represents good agreement on a complicated task, leaving few disagreements to be decided on by consensus or majority vote if the annotations are to be used as reference standard annotations.

5.1. Limitations

The annotation schema generated for this project appears to be quite complete and useful. However, we manually developed the schema, which can limit the portability and utility of any data coded with this schema. Moreover, the schema was generated for ED reports and may not provide equivalent coverage for other types of dictated reports. Agreement when applying the schema was high, but the annotators were also the creators of the guidelines, and agreement by annotators not involved in creation of the guidelines may not be as high—the *F* measures reported in this paper may be the ceiling level that other annotators would hope to reach but not surpass. However, we believe that understanding and implementing the schema would not be much more difficult for other physicians and informaticians than it was for the creators of the guidelines: The fact that the two authors with such different backgrounds could agree on many variables outside of their expertise supports this claim.

Glaser suggests two main criteria for judging the adequacy of a theory emerging from grounded theory: that the theory fits the situation and that the theory works—that it helps the people in the situation to make sense of their experience and to manage the situation better [31]. This paper addressed the first criteria for judging the quality of the annotation schema we created. However, this paper does not address the second by answering the question of whether applying the schema improves agreement in manual annotations when compared to annotating with only the general annotation goal. A major limitation is that we did not measure agreement between the annotators before we began creating the schema. We are carrying out a study to determine whether annotators not involved in creation of the schema have higher agreement when applying the schema. We will measure change in agreement when seven annotators apply the general instructions without the schema (given only the theoretical annotation goal described in Section 3) and when they apply Schema II. We will also measure the size of the learning curve involved in applying the complicated schema.

6. Conclusion

We have applied a methodology based on grounded theory to create an annotation schema for assisting annotators in indexing clinical conditions from ED reports. An annotation schema narrows the semantic and syntactic domains from which annotations can be selected, which should naturally increase the level of agreement between annotators using the schema. The schema we created represents three conceptual classes of variables to consider when annotating textual descriptions of clinical conditions: Medical Concepts, Linguistic Form, and Modifier Types. Applying the annotation schema, the physician and the informatician who created the schema annotated almost 900 clinical conditions in 20 reports with an *F* measure of 93%. The schema sufficiently modeled clinical conditions and their modifiers in our sample of ED reports and could provide a starting point for creating an annotation schema for clinical conditions in other types of reports. Moreover, the iterative process we used of annotating text with the schema, discussing disagreements, and annotating more text could be useful in training annotators in order to achieve high annotation agreement. Resulting manual annotations could be used as a potentially reliable reference standard against which an automated indexing application could be compared.

Acknowledgments

This study was funded by NLM Grant No. 1K22LM008301-01. We thank Manoj Ramachandran for programming the interface we used to index the reports.

References

[1] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the metamap program. Proc AMIA Symp 2001:17–21.
[2] Hersh W, Hickam DH, Haynes RB, McKibbon KA. Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature. Proc Annu Symp Comput Appl Med Care 1991:808–12.
[3] Zou Q, Chu WW, Morioka C, Leazer GH, Kangarloo H. IndexFinder: a method of extracting key concepts from clinical texts for indexing. AMIA Annu Symp Proc 2003:763–7.
[4] Elkin PL, Tuttle M, Keck K, Campbell K, Atkin G, Chute CG. The role of compositionality in standardized problem list generation. Medinfo 1998;9(Pt 1):660–4.
[5] Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. J Am Med Inform Assoc 2001;8(1):80–91.

[6] Berrios DC, Kehler A, Fagan LM. Knowledge requirements for automated inference of medical textbook markup. Proc AMIA Symp 1999:676–80.
[7] Srinivasan S, Rindflesch TC, Hole WT, Aronson AR, Mork JG. Finding UMLS Metathesaurus concepts in MEDLINE. Proc AMIA Symp 2002:727–31.
[8] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392–402.
[9] Christensen L, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. Proc Workshop on Natural Language Processing in the Biomedical Domain 2002:29–36.
[10] Taira RK, Soderland SG. A statistical natural language processor for medical reports. Proc AMIA Symp 1999:970–4.
[11] Baud RH, Lovis C, Ruch P, Rassinoux AM. A light knowledge model for linguistic applications. Proc AMIA Symp 2001:37–41.
[12] Hahn U, Romacker M, Schulz S. MEDSYNDIKATE-a natural language system for the extraction of medical information from findings reports. Int J Med Inf 2002;67(1–3):63–74.
[13] Pratt W, Yetisgen-Yildiz M. A study of biomedical concept identification: MetaMap vs. people. Proc AMIA Symp 2003:529–33.
[14] Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using metamap. Medinfo 2004;2004:487–91.
[15] Linguistic Annotation. http://www.ldc.upenn.edu/annotation/. Accessed June 14, 2005.
[16] Bird S, Liberman M. A formal framework for linguistic annotation. Speech Commun 2001;33(1–2):23–60.
[17] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995;122(9):681–8.
[18] Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7(6):593–604.
[19] Chapman WW, Dowling JN, Wagner MW. Generating a reliable reference standard set for syndromic case classification. J Am Med Inform Assoc 2005 [in press].
[20] Litman DJ, Forbes-Riley K. Predicting student emotions in computer-human tutoring dialogues. In: Proc of 42nd annual meeting of the association for computational linguistics (ACL). Barcelona, Spain: The Association for Computational Linguistics, July 2004; p. 351–8.
[21] Green N. A Bayesian network coding scheme for annotating biomedical information presented to genetic counseling clients. J Biomed Inform 2005;38(2):130–44.
[22] Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in langauge. Lang Resour Evaluat 2005 [in press].
[23] Wilson T, Wiebe J. Annotating attributions and private states. Proc ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky 2005 [in press].
[24] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–74.
[25] Genia Corpus. Accessed June 14, 2005. http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/.
[26] Strauss A, Corbin J. Basics of qualitative research: grounded theory procedures and techniques. 2nd ed. Newbury Park, CA: Sage; 1998.
[27] Glaser BG, Strauss AL. The discovery of grounded theory; strategies for qualitative research. Chicago: Aldine; 1967.
[28] Morison M, Moir J. The role of computer software in the analysis of qualitative data: efficient clerk, research assistant or Trojan horse? J Adv Nurs 1998;28(1):106–16.
[29] Adams WL, McIlvain HE, Geske JA, Porter JL. Physicians' perspectives on caring for cognitively impaired elders. Gerontologist 2005;45(2):231–9.
[30] Hripcsak G, Rothschild AS. Agreement, the f measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(3):296–8.
[31] Glaser BG. Grounded theory: issues and discussions. Mill Valley, CA: Sociology Press; 1998.