ISSN 0026-1270
E 3146 F

Methods of Information in Medicine

Special Issue on Concepts, Knowledge, and Language in Health-Care Information Systems (IMIA)
Editors: A.T. McCray, J.-R. Scherrer, C. Safran, C.G. Chute

1/2-95 Vol. 34

1 **Editorial**
Concepts, Knowledge, and Language in Health-Care Information Systems
A.T. McCray et al.

5 **On the Heuristic Nature of Medical Decision-Support Systems**
C.F. Aliferis, R.A. Miller

15 **Issues in the Structuring and Acquisition of an Ontology for Medical Language Understanding**
P. Zweigenbaum et al.

25 **Knowledge-Acquisition Tools for Medical Knowledge-Based Systems**
G. Lanzola et al.

40 **Objects and Domains for Managing Medical Data and Knowledge**
G. Wiederhold

47 **Cognitive Models of Clinical Reasoning and Conceptual Representation**
V.L. Patel, J.F. Arocha

57 **Naming Notes: Transitions from Free Text to Structured Entry**
J. Gregory et al.

68 **Natural Language Processing, Lexicon and Semantics**
E. Wehrli, R. Clark

75 **Sharing of Worldwide Spread Knowledge Using Hypermedia Facilities & Fast Communication Protocols (Mosaic and World Wide Web): The Example of ExPASy**
D.F. Hochstrasser et al.

79 **Searching for Answers on a Clinical Information System**
C. Safran

85 **Development of a Controlled Medical Terminology: Knowledge Acquisition and Knowledge Representation**
M.A. Musen et al.

96 **Knowledge-Based Medical Image Analysis for Integrating Content Definition with the Radiological Report**
C.A. Kulikowski et al.

104 **An Overview of Statistical Methods for the Classification and Retrieval of Patient Events**
C.G. Chute, Y. Yang

111 **Uncertainty and Decisions in Medical Informatics**
P. Szolovits

122 **The Learning Process in the Epistemology of Medical Information**
W.J. Clancey

131 **The Separation of Reviewing Knowledge from Medical Knowledge**
J. van der Lei, M.A. Musen

140 **Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding**
N. Sager et al.

147 **A Terminology Server for Medical Language and Information Systems**
A.L. Rector et al.

158 **Use of the Unified Medical Language System in Patient Care at the Columbia-Presbyterian Medical Center**
J.J. Cimino

165 **Conceptual Analysis for Knowledge-Base Design**
J.F. Sowa

172 **The ICD Family of Classifications**
M. Gersenovic

176 **Representing Clinical Narratives Using Conceptual Graphs**
R.H. Baud et al.

187 **Read Codes Version 3: A User Led Terminology**
M. O'Neil et al.

193 **The Representation of Meaning in the UMLS**
A.T. McCray, S.J. Nelson

202 **Decision Support Systems from the Standpoint of Knowledge Representation**
P. Degoulet et al.

209 **SNOMED-Based Knowledge Representation**
D.J. Rothwell

214 **Navigating to Knowledge**
M.S. Tuttle et al.

Schattauer

IMIA

Meth. Inform. Med., Vol. 34

This journal is regularly listed in Current Contents

Methods of Information in Medicine

N. Sager, M. Lyman, N. T. Nhàn, L. J. Tick

Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding

Courant Institute of Mathematical Sciences and School of Medicine, New York University, New York, USA

Abstract: A linguistic approach is presented to develop a representation of patient data. Semantic categories developed for computer processing of narrative clinical reports are shown to be similar to the Medical Concepts used manually to extract data from narrative in Exercises of the Computer-based Patient Record Institute. Clinical statement types composed of these categories are used in the Linguistic String Project (LSP) medical language processing (MLP) system to convert narrative information into relational database tables of patient information. A procedure for mapping the output of the LSP MLP system into SNOMED International codes was developed. Preliminary results and further requirements are discussed.

Keywords: Natural Language Processing, Automatic Encoding, SNOMED, Clinical Data Representation

1. Introduction

While the move toward a standardized computer representation of patient data has been gathering momentum over the past few years, fortunately so have developments that can support this objective: affordable powerful computers; experience with, and maturing designs for, medical information systems [1–8]; better user interfaces to stored medical data [9]; increased attention paid to lexical issues, including expanded terminologies [10, 11], and to methods of maintaining controlled vocabularies [12]; efforts to integrate or make reusable well-developed medical knowledge bases [13, 14]; and finally techniques of medical language processing, the subject of this paper.

2. Language and the Presentation of Information

Whether information is captured by an individual's filling in the slots of a prefigured template or by the processing of free-text input to map it into an equivalent template, the underlying predicative structure of language (and thereby also of information) is inescapable [15]. Information is to say something about something, e. g., about a bladder that it is intact (on examination after an accident).

Predicates build on predicates, producing modifiers of various types. Typical in medical statement types are modifiers that describe the source of the information (e. g., "cystogram revealed intact bladder"), the amount (where measurable), the time of the event (if appropriate), etc. Variations and specializations of medical statement types are numerous but far from infinite.

Medical vocabulary is very large; accordingly the predicative combinations, if unrestrained, would be prohibitively large for meaningful communication. The vocabulary, however, falls into a relatively small number of categories (anatomy, diagnostic procedures, surgical procedures, names of diseases, symptoms, to name a few). And the relations among these categories, seen in the large, are relatively few in number (e. g., a symptom is often associated with an anatomical site, a treatment is given for a pathological condition). Thus, while grossly simplifying the representational issue, it is nonetheless true that there is a set of categories and relations appearing in characteristic linguistic combinations, that provide the underpinnings to a representation of patient information.

3. Categories of Information

3.1 Medical Concepts

There are several ways of arriving at the categories of information that comprise the elements of an informational representation. One way is to "tap the expert". Let a panel of experts determine a set of categories ("medical concepts") and extract from patient reports the words and phrases that instantiate these concepts in each case. These attribute-values pairs (<medical concept> <*word string instantiation*>) constitute the essential elements of a representation of the patient data in the report. It can further be investigated how many of these elements can be mapped to a standard set of elements, a "code".

This was the activity undertaken by the Computer-based Patient Record Institute (CPRI) in two Exercises, each involving 10 case reports [16]. Medical Concepts and their instantiations were manually determined for each case from such documents as discharge sum-

Table 1 Retrieval of negative findings from sentence 17-CRPI Exercise:
014A.1.17 He had no associated headache, loss of consciousness, head injury, chest pain, or shortness of breath.

SID	ROW	DIAG-SGN_SYM	Bodypart-Bodyfunc	QUANT	TXPROC	TIMEWDS	Negation
014A.1.17	1	associated headache					No
014A.1.17	2	associated # loss of consciousness					No
014A.1.17	3	associated # injury #	head				No
014A.1.17	4	associated # pain #	chest				No
014A.1.17	5	associated # short-ness of breath	chest				No

Medical Concepts other than Physical Finding	
Primary Diagnosis	*melanoma, neurocardiogenic syncope, vasodepressor syncope, adenocarcinoma, tumor*
Diagnosis	*cancer, hypertension, aortic regurgitation, anemia, rash, renal insufficiency, coronary artery disease, ulcer*
Symptoms	*light-headedness, nausea, diaphoresis, spotting, cramping, discomfort, red, difficulty with vision, tearing*
Qualitative	*tiny shotty, asymmetric, soft, uneventful, standard, within normal limits, unremarkable*
Anatomy	*thigh, groin, hip, acetabulum, femoral head, greater trochanter, endometrial, ovary*
Topology	*left, superficial, right, midline, right side, lower, proximal, bilaterally, anterior, lower left anterior*
Functional Status	*appetite, oral intake, usual state of health, about to fall, risk of recurrent cancer, appetite, normally active*
Severity	*severe, mild, excessive, markedly, increased in frequency, moderately high, worsened, moderate, acute*
Stage	*stage IC, benign, 2+*
Grade	*grade 2–3, grade 2, high grade, benign*
Extent	*Clark's level 2, maximum depth two-thirds of, normal endocervix, 0/3 lymph nodes*
Quantitative	*0.84 mm [depth of invasion], 1.5 cm long [incision length], 5–8 [number of occurrences], multiple, 2 units [amount transfused]*
Diagnostic Exam	*Holter monitor, tilt table test, Ortolani test, Barlow test, chest x-ray, echocardiogram, hip x-rays, scanogram, ECG*
Laboratory Name	*CBC, creatinine, ejection fraction, hemoglobin, ferritin, iron binding saturation, TIBC, folate, iron, B12*
Therapeutic Procedures	*wide local excision, adductor tenotomy, open reduction, pinning, D & C - [Dilatation & Currettage], hysterectomy, arthrotomy*
Treatments	*propranolol, hydralazine, ibuprofen, darvocet, gold salt therapy, Pavlik harness, Freijka-type splint, irradiation*
Disposition	*observation, poor prognosis, do nothing, refer for evaluation, patient wishes to proceed*
Negation	*do not recommend, no, negative, without evidence of, no associated, denies, no ... abnormality*
Reliability	*question of, rule out, approximately, very suggestive, was nearly, essentially, possible*
Chronicity	*history of, recurrent, past medical history, chronic, recurrences*
Time	*recent, nocturnal, usual, present for 2 years duration, 04/25/91 [date of diagnosis], four years [duration of diagnosis]*

Fig. 1 a Medical concepts and associated text, representative examples.

Physical Finding	
B1. Physiological function or anatomic site examined + Result	*motion limited, percussion tenderness, positive pivot shifts, skin folds asymmetric, visceromegaly, healing of vaginal cuff, neck supple, bleeding at biopsy site*
B2. Physiological function or anatomic site examined	*internal rotation, respirations, heart rate, BP, hip motion, abduction external rotation, flexion, range of motion, extremities*
B3. Result	*prominent, nodes, incision, lymphadenopathy, less developed, masses, incision, infection, retained sutures, lymphadenopathy, healthy appearing, looks younger than stated age, polyp, development [joint development], scar, laxity*

Fig. 1 b Medical concepts and associated text, representative examples.

maries, nurses notes, and some other natural language sources. The results were then the basis for manual coding into a number of standard terminologies.

The major goal of the CPRI Exercises was to test the coverage of medical concepts by existing medical terminologies. At the same time, inadvertently perhaps, the Exercises provide some insights into the mind-set of the physician, functioning as an abstractor of information from the patient record. One observes the types of information (Medical Concepts) that were considered most important to extract, and the words and phrases that were seen to constitute that information in patient documents. Figure 1 shows a list of the Medical Concepts appearing in the CPRI Exercise 1 data, with examples of word values that appeared as their instantiations. For brevity, Medical Concepts with less than 3 occurrences in the Exercise 1 data are not included in Fig. 1, or Fig. 2; the words fitted reasonably into other Medical Concepts used in the Exercises.

In the Exercise 1 data, instances of the Medical Concept <Physical Finding> appeared to cover different amounts of the finding information. This depended, in part, on which other Medical Concepts were represented in the text. In Fig. 1 b, occurrences have been subclassified here accordingly. In category B1 are examples of those occurrences that comprised a subject and a predicate, i. e. they named a physiological function or anatomic site and stated an observation about it *(motion limited)*. In B2 of Fig. 1 b are occurrences that named a physiological function or anatomic site but did not include the finding proper, which may have been noted in another Medical Concept associated with the same sentence (*internal rotation* occurring along with <Severity> *excessive*). In B3

are occurrences that stated only the finding proper. It might be one that applied to the patient as a whole *(looks younger than stated age)* or one *(prominent)* occurring along with another Medical Concept (<Anatomy> *greater trochanter*) that completes the statement of the finding. We see in Fig. 1 b the beginning of a definition of a statement type for physical findings.

3.2 Linguistic Categories

Another approach to determining major medical concepts and the words associated with them is by linguistic methods: treating the texts of physicians like the myths of native peoples: i. e., to determine from the regularities observed in word cooccurrences in many utterances the categories these words fall into and their characteristic combinations. The LSP group of New York University has adopted this approach to determining the semantic categories required for processing narrative clinical documents. A comparison was made between the Medical Concepts seen in the Exercises data and the LSP semantic categories used in text processing. It was quite striking to observe that the semantic categories that emerged from applying linguistic methods were similar to those that the reviewing physicians of the CPRI Exercises developed on an informed intuitive basis.

Figure 2 shows the correspondence between the Medical Concepts of the CPRI Exercises and the linguistically based categories of the LSP system. The latter, in Fig. 2, are either lexical categories from the LSP medical dictionary; or categories computed during LSP processing (starred in Fig. 2), e. g., a number with a unit-word as the value of QUANT; or a combination of lexical classes and/or syntactic units, as in the LSP equivalent of the Medical Concept <Physical Finding> (Fig. 1 b), which will be seen below as part of a prototype clinical statement format used in text processing.

Figure 3 contains a list of LSP lexical categories that appeared to be comparable to CPRI Exercise Medical Concepts, along with examples of the categories from the LSP dictionary. These categories appear in Fig. 3 in the same order as in Fig. 2. A more complete list of LSP lexical categories, arranged by broad semantic group, appears in [17], and detailed definitions of the categories (some name changes) are given in Appendix A of [18]. The examples of LSP categories in Fig. 3 were chosen quite arbitrarily and do not show which particular words of the Exercise data appeared in the LSP dictionary. In some cases a particular word was classified differently, e. g., *rash* as Symptom or Diagnosis. This is a well known grey area of classification. What was of major interest was to see that elements were distinguished as being essential in the information (Medical Concepts), and to see that these were virtually the same as the categories needed to process the information in its free-text form.

Medical concepts (CPRI Exercises)	Corresponding linguistic categories (LSP dictionary classes, or LSP-computed if starred)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Primary Diagnosis	DIAG						
Diagnosis	DIAG						
Symptom	INDIC						
Qualitative	DESCR						
Qualitative	NORMAL						
Anatomy		PTPART					
Topology		PTAREA					
Functional status		PTFUNC					
Severity			AMT				
Degree			QUANT*				
Stage			QUANT*				
Grade			QUANT*				
Extent			QUANT*				
Quantitative			QUANT*				
Diagnostic Exam				TXPROC			
Laboratory Name				TXVAR			
Therapeutic Procedures					TTCHIR		
Treatments					TTMED		
Treatments					TTCOMP		
Disposition					TTGEN		
Chronicity						TMPER	
Time						TMLOC	
Time						TIME*	
Negation							NEG
Physical Finding*							MODAL

* The LSP system recognizes quantitative phrases syntactically, and similarly for time expressions involving quantities ("3 days post op"). The notion of Physical Finding in the LSP representation is computed from a number of the above elements to comprise a *clinical statement type* as described in the text.

Fig. 2 Medical concepts and corresponding linguistic categories.

4. Representational Structures for Clinical Data

At present a major concern in the development of a representation of patient data for a Computer-based Patient Record CPR (or Electronic Medical Record EMR) is the determination of a standard medical terminology. This is indeed a Herculean task, and fortunately, considerable progress has been made, as the report of the CPRI Exercise demonstrated [16]. Indeed, in the beginning there is the word. But soon thereafter is the concern with the relation among words. Negated findings are an obvious case. Then, uncertainty, degree, time order and temporal aspect influence the significance of a finding. Finally, there is the composition of the finding itself as a unit of information, and the relations among the units (e. g., temporal).

In Figure 2, the arrangement of linguistic categories into numbered columns is meant to illustrate that these categories (equivalently the Medical Concepts listed at the left) can be read as a prototype clinical statement type:

The finding (1), in the physiological function or anatomic site of the patient (2), to the amount (3), found by procedure or laboratory test (4), was treated by (5);

and one may state with regard to (1), (4), or (5) that

it occurred at time (6) and/or with temporal features (6);

and also with regard to (1), (4), or (5) that

it did not occur or there was some doubt (7).

Figure 4 shows this prototype clinical statement type broken into 3 less cumbersome subtypes: Patient State, Laboratory Finding, and Treatment. This reflects the fact that in patient documents, these substatements often occur as separate sentences in different parts of the document.

The statement types seen in Fig. 4 serve in the LSP system as the basis for data structures into which processed text is mapped. They might equally well serve as a guide to the design of graphical user interfaces (GUI) for the capture of patient data on-line. Natural language-like formal languages for data capture have pitfalls. Nevertheless, the knowledge of how physicians and other healthcare workers formulate their observations and report their actions when NOT constrained by a computer program should be of help in designing software for clinical data capture. It should be reassuring to realize that the information is reported in relatively regular patterns of medical concept occurrences.

DIAG	*adenoma, AIDS, beriberi, cancer, diabetes, fibroelastosis, histoplasmosis, leukemia, malaria, papilloma, tumor*
INDIC	*abnormal, bacteremia, bilious, cramping, discomfort, nausea, puffiness, rales, rash, swollen, tachycardia, warmth, watery, pallor*
DESCR	*acinar, classic, false, inactive, laminated, neutral, non-specific, open, primitive, radical, shotty, symmetrical, villous*
NORMAL	*nondistended, orientated, pure, rational, satisfactory, successfully, uncomplicated, undistressed, uneventful, well, WNL*
PTPART	*decidua, face, gallbladder, ileum, jaw, labial, malar, nail, obturator, palatal, ramus, sacral, stapes, talus, vagina, wrist*
PTAREA	*adjacent, edge, field, inferior, lateral, left, opposite, quadrant, right, tail, terminal, unilateral, ventral, wall, zone*
PTFUNC	*air-entry, ambulate, blood pressure, eat, fall asleep, gait, heal, libido, pulse, respiration rate, secretion, uptake, vision, wake*
AMT	*impressive, largely, many, mild, moderate, numerous, partially, rare, scant, severe, significant, tolerable, waxing and waning*
TXPROC	*xray, analyzer, billroth II, biopsy, echocardiogram, holter, mammogram, paracentesis, radiographic, scan, ultrasound*
TXVAR	*acetone, basophil, calcium, DNA, electrical axis, electrolyte, FEP, gamma globulin, hemoglobin, immunoglobulin, ion, WBC, VDRL, zinc*
TTCHIR	*ablation, bunionectomy, laminectomy, neurosurgical, operate, re-excision, shortened, shunt, tenotomy, transected, valvotomy, wired*
TTMED	*butazolidine, ranitidine, calcichew, depomedrol, ecotrin, flagyl, gamma globulin, heparin, immunization, lasix, mannitol*
TTCOMP	*appliance, bandage, cane, defibrillator, enema, foley catheter, hemodialysis, lavage, mask, opsonize, pacemaker, splint, radiation*
TTGEN	*admission, care, follow up, processed, reassess, refer, schedule, transfer, visit, watch, workup*
TMPER	*briefly, ongoing, on occasion, period, permanent, persist, short term, sustain, transient, trend, usual*
TMLOC	*childhood, earlier, finally, hence, immediate, infancy, interim, juncture, juvenile, last, lately, midstage*
NEG	*absence, deny, didn't, fail, neither, never, no, not, no evidence of, unable, won't*
MODAL	*estimate, feel, hope, likely, maybe, necessary, need, offer, oppose, pending, perhaps, proposed, questionable, recommended*

Fig. 3 Medical linguistic categories appearing in Figure 2, examples from LSP dictionary.

A. Patient State					
A finding (1)	in patient part or function (2)	of amount (3)	found by (TXPROC) (4)	at time/ chronicity (6)	with doubt or negation (7)
B. Laboratory Finding					
A finding (1)	in patient specimen (2)	of amount (3)	of item (TXVAR) (4)	at time/ chronicity (6)	with doubt or negation (7)
C. Treatment					
----------{A} or {B}----------			treated by (5)	at time/ chronicity (6)	with doubt or negation (7)

Fig. 4 Prototype medical statement types based on columns of Figure 2.

5. Text Processing for Applications

The possibility of automatically extracting relevant clinical information from free-text patient documents has been a challenge for a small number of research groups over the years. The language processing problems are daunting but hopefully not insurmountable. Some recently reported applications of MLP systems include two devoted to healthcare monitoring and follow-up [17, 19], one to healthcare support [20], and one to automatic encoding [21]. The data used in [19, 20] are radiology reports; the data in [17, 21] are discharge summaries.

The importance of structure has been recognized by the various groups concerned with MLP. Parsers specifically adapted for the representation of natural language clinical data have been and continue to be developed [22, 23]. Parsing is followed by, or combined with, a mapping to the medical semantic representation which is used in the application. While systems differ with regard to the type and amount of syntactic analysis employed in the processing, some baseline syntactic requirements are generally recognized; for one, the ability to separate positive from negative findings.

After processing via the LSP MLP system, using the statement types illustrated in Fig. 4 as the basic informational units, the CPRI analyzed text was mapped into a relational database. Retrieval programs could then generate a table of positive findings and a table of negative

014A.1.17 He had no associated headache, loss of consciousness, head injury, chest pain, or shortness of breath.

A. CPRI Exercise 1 Data

Medical Concept

<Diagnosis>	headache
<Diagnosis>	loss of consciousness
<Diagnosis>	head injury
<Diagnosis>	chest pain
<Diagnosis>	shortness of breath
<Negation>	not associated

Manual SNOMED Coding from Medical Concepts

F-A2700	01	Headache, NOS
F-A5570	02	Loss of consciousness, NOS
DD-21000	02	Head injury, NOS, without skull fracture
F-37000	01	Chest pain
F-20040	02	Shortness of breath
G-A201	01	Negative

B. LSP Experimental Automatic Encoding from MLP Output

1. *Results*

Row 1:	G-A201	01Negative
	F-A2700	01Headache, NOS I 784.0
Row 2:	G-A201	01Negative
	F-A5570	02Loss of consciousness, NOS
Row 3:	M-10000	01Injury, NOS I 959.–
	T-D1100	01Head, NOS I C76.0
	G-A201	01Negative
Row 4:	F-A2600	01Pain, NOS
	T-D3000	02Chest I C76.1
	G-A201	01Negative
Row 5:	F-20040	02Shortness of breath I 786.09
	T-D3000	02Chest I C76.1
	G-A201	01Negative

2. *Type of Search*

Row 1: Single word search
Row 2: Text flow break search
Row 3: Text flow break search
Row 4: Text flow break search
Row 5: Text flow break search

Fig. 5 Example of automatic encoding from text.

findings (e. g., as seen in Table 1). In the database tables, the symbol # marks a "breakpoint" between words in a given field that were separated in the sentence.

Over and above minimal requirements, such as positive vs. negative findings, MLP systems are influenced by the type of document to be processed (e. g. discharge summary vs. pathology report) and the requirements of the application (e. g., decision support vs. epidemiology).

6. Automatic Encoding – Different Approaches

One approach to the automatic encoding of clinical data from free-text sources is to view it as a task of translation. The document is written in language L1 (natural language); the target language L2 is the set of word strings comprising the textual entries of the coding system. A correct translation of the document would consist of groups of textual code entries in L2 conveying the same information as the corresponding sentences or sentence parts of the document in L1. The field of machine translation has faced many of the same problems (e. g., natural language ambiguity) as other language processing efforts in addition to the special problems of translation. There is much to be learned from that experience [24].

Strategies of machine translation are broadly characterized as translation between language pairs ("transfer rules approach") or translation that maps the starting text first to an underlying representation and hence to any target language ("Interlingua approach") [25]. Applying this distinction to the translation of clinical free-text into clinical codes, one would consider the transfer rules approach to be one that tries to match parts of the (possibly preprocessed) clinical document with word strings of the code. The Interlingua approach would seek a decomposition of both text sentences and word strings of the code into elementary units that could be individually matched. This would be followed by the use of an algorithm to find the *best match* where a given stretch of text had a match, or partial match, with more than one word string of the code. The matching algorithm completes the "translation".

In his extensive work on medical vocabulary and automatic encoding, Wingert adopted the Interlingua approach [26]. Wingert's algorithm for automatic indexing into SNOMED provided for a preliminary decomposition of both input text strings and SNOMED word strings into their elementary morphosemantic units (by analogy, the Interlingua). The comparison then took place at the level of these units and the algorithm contained a means of choosing the best match (briefly: maximal matched units; minimal number of codes to cover matched units). Since medical vocabulary contains many Latin-based terms that combine several semantic elements into one word, it was necessary to provide a morphological decomposition of these words into their semantic elements [27].

The attractiveness of the Wingert approach lies in its principled methodology and, possibly in the long run, greater practicality with regard to the maintenance and further updating of the encoding system and the code itself. As Wingert recognized, it requires, in addition to morphological analysis of composite words, also ways of treating word relations, including synonomy, hypernomy, hyponomy, and those relations carried by syntax. While the details of Wingert's algorithm may not be applicable to all types of clinical text, it is clear that for automatic encoding a common ground has to be found between text and code. The task is not all

on the text processing side. The size and complexity of SNOMED, for example, forces one to seek within it a reduction to elements that can be compared with text elements, just as the text of a document requires analysis into smaller units so they can be sought in word strings of the code.

7. Automatic Encoding – Issues and Examples

7.1 An Encoding Algorithm

In the CPRI Exercises the path from patient document to code was via Medical Concepts:

Text → Medical Concepts → Coded Text.

The similarity of LSP semantic categories to Medical Concepts suggested a similar path for automatic encoding:

Text → LSP MLP System → Coded Text.

To explore this possibility, an algorithm was devised (still quite preliminary) to map the output of the LSP MLP system into SNOMED International. We used texts from the CPRI Exercise 1 as input so that the results obtained by the LSP procedure could be compared with the SNOMED codes that had been assigned manually to the texts as part of the Exercise.

Figure 5 contains results of the automatic encoding experiment for one sentence from the CPRI Exercise. Part A contains the CPRI Exercise 1 data: Medical Concepts and SNOMED codes; and Part B shows the results of the LSP automatic encoding, row by row. B 1 displays the SNOMED codes obtained by the (still quite preliminary) LSP search algorithm; B 2 lists the particular procedures of the search algorithm that contributed to the result.

The unit of text that is input to the encoding algorithm consists of the words in a row of the relational database table generated by LSP-processing of the document. Table 1 shows an example. Each row corresponds to one of the Statement Types shown in Fig. 4. For this experiment, entries in the TIME field were not included in the search.

The Search Algorithm proceeds row by row. Four types of search are employed. "Full search" constructs a text string of all the words in the row in their order of occurrence in the sentence, and searches SNOMED strings for a complete or partial match. (Partial match has limits on how partial and how many.) "Field search" does the same for each nonempty field in the row. "Text flow break search" does the same for the substrings within a field (strings set off by break marks "#"). Finally, as a last resort, "Single word search" treats every word as a text string to be matched against SNOMED strings. The final result is the set of SNOMED codes which covers the most words in the least number of codes (similar to Wingert strategy).

7.2 Preliminary Results

Some of the lessons learned from this experiment can be illustrated with reference to the results in Fig. 5.

Figure 5 illustrates the value of making explicit the syntactic relation of negative words to their arguments, i. e. to the words the negation applies to. The language processing has "expanded the conjunction" so that *no* applies individually to each negated finding.

Ongoing work on SNOMED has shown it to have extensive coverage of medical vocabulary, and by its very richness also to contain redundancy, i. e., more than one way to code certain word strings [28, 29]. Figure 5 contains an example. In Part A, *chest pain* has been manually coded as a single unit (F-37000 01). In Part B, the search procedure found the two words separately (F-A2600 01 Pain, NOS; T-D3000 02 Chest). [The search procedure would find both codings by searching the combined Anatomy and Symptom fields.] Some redundancy in a code is probably inevitable. *Chest* must be a single code entry as it appears in many contexts, and similarly for *pain.* Yet, *chest pain* as a medical concept also has a claim to a code. Without arguing the particulars of *chest pain,* it is clear that many codable concepts will contain words that occur often in other combinations and therefore require their own code. It seems inevitable that a record of multiple codings be maintained. Given automatic encoding, one could treat SNOMED as a text to be coded into SNOMED as a way of discovering redundancy.

One can note also in Part B of Fig. 5 a (semantically incorrect) syntactic association of *chest* with *shortness of breath* due to the MLP procedure that expands conjunctions. Here the procedure applied in a situation of syntactic ambiguity. Thus, *No chest pain or shortness of breath → No chest pain or [chest] shortness of breath* in the same way that *No joint swelling or redness → No joint swelling or [joint] redness.* Most ambiguities can be decided correctly with sufficiently detailed semantic rules or large stocks of correct word collocations. The issue is cost in relation to purpose.

Procedures for morphosemantic analysis of medical vocabulary to obtain the semantic elements in composite (Latin-based) words would be an essential prepocessing procedure, as envisioned by Wingert and others. In addition, there are many issues around term equivalence that require attention. General classifier terms like *problems, difficulty* and the like occur in documents, whereas the terms in a code are more precise. For example, in the manually coded Exercise 1 data, the Medical Concept <Symptom> *difficulty with vision* was assigned the SNOMED code DA-74900 02 Decreased vision, NOS.

The synonym list is consulted as part of every search. "Synonomy" in this context means an application-dependent term equivalence, not identity of meaning. Affixes impart meaning (e. g. the difference between singular and plural) but for coding purposes, only the medical semantic content is important. A full fledged automatic encoder would employ linguistic transformations of grammatical form in a preprocessor.

The search procedure at present is severely limited by having to work with database rows instead of linguistic trees. Syntactic relations and lexical attributes in the MLP output are not conveniently accommodated in a relational database. We would like, for example, to "peel off" adjectives in a noun phrase tree according to their position in the tree and their semantic class, e. g. *associated* in Sentence 17. It should be understood that the search procedure reported here represents the first of several strategies that are being investigated.

8. Conclusion

How to code an uncodified world? No one has the answer. But Medicine has been obliged to attempt it. Codes have evolved with increasing coverage and systematization. The challenge now is to provide structures for the medical concepts embodied in codes so that together, the structure and codes carry clinical information.

Can traditional ways of reporting be preserved but made to conform? The traditional way of reporting is by language. Finding the data structures that are implicit in the use of language in clinical documents may be helpful in the design of user interfaces for the capture of patient data. Processing natural language clinical documents may make it possible to continue traditional reporting, with automatic encoding to convert the free form to the established code.

REFERENCES

1. Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP System. J Med Systems 1983; 7: 2.
2. Bakker AR. The development of an integrated and co-operative hospital information system. Med Inf 1984; 9: 135–42.
3. Bleich HL, Beckley RF, Horowitz GL et al. Clinical computing in a teaching hospital. N Eng J Med 1985; 312: 756–64.
4. Diogène Staff. *The DIOGENE Hospital Information System.* Division Informatique, Hôpital Cantonal Universitaire de Genève, Switzerland, 1986.
5. Medical Archival System (MARS): Large scale medical data archiving at the University of Pittsburgh. Office of Biomedical Informatics, University of Pittsburgh, 1990.
6. McDonald CJ, Tierney WM, Overhage JM, Martin DK, Wilson GA. The Regenstrief medical record system: 20 years of experience in hospitals, clinics, and neighborhood health centers. MD Comput 1992; 9: 206–17.
7. Clayton PD, Sideli RV, Sengupta S. Open architecture and integrated information at Columbia-Presbyterian Medical Center. MD Comput 1992; 9: 297–303.
8. Stead WW, Bird WP, Califf RM et al. The IAIMS at Duke University Medical Center: transition from model testing to implementation. MD Comput 1993; 10: 225–30.
9. Safran C, Porter D, Lightfoot J et al. ClinQuery: A system for online searching of data in a teaching hospital. Ann Int Med 1989; 111: 751–6.
10. Humphreys BL, Lindberg DAB. The Unified Medical Language System Project: a distributed experiment in improving access to biomedical information. In: *MEDINFO 92.* Amsterdam: North-Holland, 1992: 265–8.
11. Coté RA, Rothwell DJ, Beckette R, Palotay J eds. *SNOMED International.* Northfield, IL: College of American Pathologists, 1993.
12. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc 1994; 1: 35–50.
13. Musen MA. Dimensions of knowledge sharing and reuse. Comput Biomed Res 1992; 25: 435–67.
14. Masarie FE, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An Interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res 1991; 24: 379–400.
15. Harris Z. *A Theory of Language and Information: A Mathematical Approach.* New York: Oxford University Press, 1991.
16. Case histories and concept identification obtained from Computer-based Patient Record Institute (CPRI) exercises, presented at the Annual Symposium on Computer Applications in Medical Care, Washington DC, November 1993.
17. Sager N, Lyman M, Bucknall C, Nhàn N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994; 1: 142–60.
18. Sager F, Friedman C, Lyman MS. *Medical Language Processing: Computer Management of Narrative Data.* Reading MA: Addison-Wesley, 1987.
19. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comput Biomed Res 1993; 26: 467–81.
20. 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: 161–74.
21. Satomura Y, Do Amaral MB. Automated diagnostic indexing by natural language processing. Med Inf (Lond) 1992; 17: 149–63.
22. Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physical findings: Canonical Phrase Identification System (CAPIS). In: *Proc Annu Symp Computer Applications in Medical Care.* 1991; 15: 843–7.
23. Baud RH, Rassinoux A-M, Scherrer J-R. Natural language processing and semantical representation of medical texts. Meth Inform Med 1992; 31: 117–25.
24. King M. Are there any lessons to be learned from machine translation? In: Scherrer J-R, Coté RA, Mandil S, eds. *Computerized Natural Medical Processing for Knowledge Representation.* Amsterdam: Elsevier Science Publ 1989; 73–82.
25. Hutchings WJ, Somers HL. *An Introduction to Machine Translation.* London: Academic Press, 1992.
26. Wingert F. An Indexing System for SNOMED. Meth Inform Med 1986; 25: 22–30.
27. Wingert F. Morphologic analysis of compound words. Meth Inform Med 1985; 24: 155–62.
28. Campbell KE, Musen MA. *Representation of Clinical Data Using SNOMED III and Conceptual Graphs.* Stanford CA: Stanford University, Knowledge Systems Laboratory Report KSL-92-13, 1992.
29. Rothwell DJ, Coté RA, Cordeau JP, Boisvert MA. Developing a standard data structure for medical language – the SNOMED proposal. In: *Proc Annu Symp Computer Applications in Medical Care.* 1993; 17: 695–9.

Address of the authors:
Naomi Sager, Ph. D.,
Margaret Lyman, M. D.,
Ngo Thanh Nhan, Ph. D.,
Leo J Tick, Ph. D.,
Courant Institute of Mathematical Sciences,
New York University,
251 Mercer Street,
New York, NY 10012, USA