Toward A Public Domain UMLS Patient Database

L.S. Fu, O. Bouhaddou, S.M. Huff, D.K. Sorenson, H.R. Warner

Department of Medical Informatics
University of Utah, School of Medicine
Salt Lake City, Utah

ABSTRACT

This paper describes a unified structure and with an associated vocabulary to represent and store patient cases derived from different computerized patient databases. The unified structure is based on the concept of event definitions which are generic templates for representing clinical data in a patient database. An implementation of this structure has been evaluated using patient cases from two expert systems (Iliad [1] and QMR [2]) and a hospital information system (HELP [3]). The primary focus of the UMLS patient database is to accumulate patient information from different sources and provide enhanced statistical estimates of clinically important variables. Inter-communication and navigation among medical information systems are other potential benefits of this unified computerized medical record system.

GOAL

Since 1985, the UMLS project has been a major National Library of Medicine (NLM) initiative designed to facilitate the retrieval and integration of information from many machine-readable information sources [4]. Although a current major focus of the UMLS project is to build a Metathesaurus encompassing terms from several controlled vocabularies and classifications, access to a source of patient data is also important for UMLS experimentation [4, 5].

In this project, we present a unified patient record structure and vocabulary component designed to represent and store patient data recorded in different medical information systems. The goal is to create a resource that will allow people to share patient data from multiple institutions in order to expedite the collection of statistical estimates of clinically important parameters such as disease prevalence, sensitivities and specificities of disease findings, and outcome measures of treatment protocols. Other potential benefits of such a resource, not detailed in this paper, include mapping between electronic medical vocabularies for the purpose of linking clinical records to knowledge bases, as well as comparing alternative knowledge bases on the same patients.

Historical Background

Event definitions are descriptions or templates of the structure of clinical data as it resides in a database. The concept of event definitions has grown out of a desire to better define, organize and characterize clinical data in order to support clinical research and knowledge engineering. Historically and conceptually the current form of event definitions has evolved from the previous work of Huff *et. al.* [6] while drawing experience from the generic frame concept work of Miller and Masarie *et. al.* [7, 8] and Cimino *et. al.* [9]. Since event definitions are in many ways similar to generic frames, the two models will be briefly described and contrasted here. A main hypothesis underlying the event definition model is that there is some small set of definitions (less than 100) that can be used to record all relevant clinical findings. An example of an event definition is shown in Figure 1.

Event Identifier: 1
Description: Used for recording symptoms from multiple patient databases
Semantic Requirements:
Mandatory: A
Body:
(A) Concept (e.g., pain, cough, dyspnea)
(B) Quality (e.g., sharp, stabbing, tearing)
  (B.1) Multiple term relationship (e.g., and, or, and not)
(C) Severity (e.g., severe, intense, mild)
(D) Occurrence Pattern (e.g., chronic, seasonal, recurs)
  (D.1) Duration (e.g., lasting)
    (D.1.1) Comparator (e.g., more than, less than)
    (D.1.2) Value (e.g., one, two, three)
    (D.1.3) Time (e.g., minute, hour, day)
  (D.2) Frequency (e.g., often, daily, continuous)
    (D.2.1) Comparator (e.g., more than, less than)
    (D.2.2) Value (e.g., one, two, three)
    (D.2.3) Time (e.g., minute, hour, day)
(E) Trend (e.g., increase intensity, increase frequency,increase severity)
(F) Onset (e.g., gradual, acute, chronic, suddenly, onset, came on)
  (F.1) Comparator (e.g., more than, less than)
  (F.2) Value (e.g., one, two, three)
  (F.3) Time (e.g., minute, hour, day)
(G) Anatomic Site << Body Parts >> (e.g., chest, back)
  (G.1) Location (e.g., left, right)
  (G.2) Multiple term relationship (e.g., and, or, and not)
(H) Spatial Relationship (e.g., radiates to, migrating to, moved to, shifted to)
  (H.1) Anatomic Site << Body Parts >> (e.g., arm, shoulder, back)
    (H.1.1) Location (e.g., left, right)
(I) Associated with << Actions,Concept Names,States >> (e.g., exercise, coughing, stress)
  (I.1) Effect (e.g., aggravated, alleviate, relieved, improved, worsen)
  (I.2) Substance (e.g., food, drug)
  (I.3) Time Relationship (e.g., before, after, during)
  (I.4) Trend (e.g., increase intensity, increase frequency, less vigorous, deeply)
  (I.5) Comparator (e.g., more than, less than)
  (I.6) Value (e.g., one, two)
  (I.7) Time (e.g., minute, hour, day)
(J) Probability (e.g., present, absent, appears to, probable, seems to)

Figure 1: An Eventt Definition used to describe patient symptoms. Clauses are labelled alphabetically and sub-clauses have numerical indexes.

Generic frames were developed as part of a UMLS -related experiment by the Pittsburgh group to help capture and organize information about clinical observations [10]. The common aspect of event definitions and generic frames is that both methodologies are using frames that contain slots to describe clinical observations. However, there are also major

0195-4210/90/0000/0170$01.00 © 1990 SCAMC, Inc.

differences. The first major difference is that there are many more generic frames (e.g. there are 750 generic frames in the symptoms category alone) than there are event definitions (less than 100 in total) because of the granularity of the concepts that are modeled. For instance, there is one event definition for symptoms with the main slot being the "concept name" which would take values such as "pain," "cough" or "shortness of breath." Additional slots in the event definition describe the location of the pain and associated characteristics. However, different generic frames exist for each type of pain such as "chest pain," "abdominal pain," "back pain," and "headache." A second difference is that the slots in generic frames point to lists of items while slots in event definitions point to nodes in existing hierarchies of vocabulary terms. The value to a user of the event definition is that inferences can be made based on the vocabulary hierarchy. Thus, since "thumb" is subordinate to "hand" which is a child of "arm" in the vocabulary hierarchy, if a clinical researcher is interested in arm injuries, the system can infer that an injury to the thumb is an injury to the arm.

IMPLEMENTATION ISSUES

In addition to the Event Definitions (EDs) structures, the building components of the UMLS patient database are the Master Object Index (MOI) file , the ED instances and the actual patient records. The MOI is the repository of all the coded medical concepts as they have been collected from each specific source of patient records while the EDs describe a structure for medical information as it exists in the clinical world. An ED instance is the mapping of a specific vocabulary term to an ED structure where the value of the ED slots have been filled to reflect the meaning of the specific term. Each source included in the UMLS patient database contributes its unique medical terms to the MOI file and has its dictionary terms mapped to a set of ED instances. The actual unified patient file stores the identity of the patient, the id of the database where the patient data comes from, the id of the ED instance and the time the event occurred. Figure 2 is a schematic representation of these components.

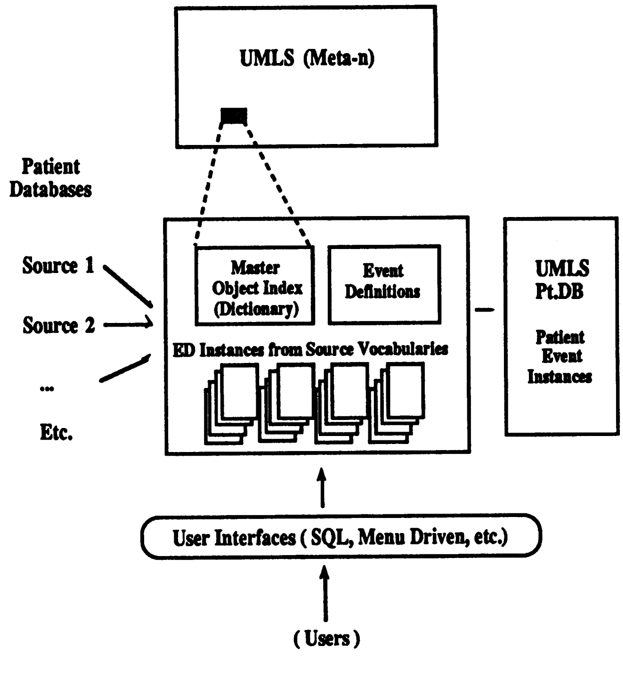


Figure 2: Schematic diagram describing the components of the UMLS patient database

The Master Object Index (MOI) file

The MOI file is built by accumulating the unique terms of each dictionary source. At present, the MOI file has been populated with terms from HELP, Iliad, QMR and MeSH. In the future, terms of the MOI could be obtained from the Meta-1 thesaurus developed separately by the UMLS participants [5], given this resource includes clinical databases information. Below is a portion from the symptoms hierarchy of the current MOI file where each row represents a different record:

alias	obj	N/L	text	hierarchical codes	level
1	13	N	Terms		0
1	13	N	symptom names	10	1
2	13	N	findings	10	1
1	13	N	cough	10.4	2
1	13	L	hemoptysis	10.4.2	3
1	13	L	dyspnea	10.6	2
2	13	L	shortness of breath	10.6	2
1	13	L	faintness	10.8	2
2	13	L	syncope	10.8	2
1	13	L	nausea	10.10	2
2	13	L	vomiting	10.10	2
1	13	L	thirst	10.12	2
1	13	L	fever	10.14	2
1	13	N	chills	10.16	2
1	13	L	shaking	10.16.2	3
1	13	L	appetite	10.40	2
1	13	N	pain or discomfort	10.44	2
2	13	N	pain	10.44.2	3
1	13	N	ache	10.44.2.2	4
1	13	L	headache	10.44.2.2.2	5

Legend:

alias -- indicated whether the text portion represents an alias for the record instance with the same code.

obj -- indicates the object type of the item. Type 13 refers to text terms. Other types could be frames, fields or disease hierarchies.

N/L -- indicates if the term is a Node or Leaf in the hierarchy.

text -- the textual description of the item.

hierarchical code -- represents a hierarchical position relative to other items in the MOI.

level -- indicates the exact level or depth in the hierarchy structure of the MOI.

Event Definitions (EDs)

As previously described, Event Definitions (EDs) are frame-based structures to describe medical events. The information is partitioned into slots and slot fillers (values). An example of an event definition for a symptom description event is given in Figure 1. Additional event definitions have been developed for physical examination observations, chemistry and hematology laboratory tests and diagnoses.

The implementation of an ED is composed of a header and a body part. The header defines pointers to the MOI file: the ED id, the ED description and the permissible values that the ED can take (e.g., symptom names hierarchy). Similarly, the body part is a set of MOI pointers describing the various attributes of the ED (e.g., symptom location, time pattern, onset, duration... etc.).

ED instantiations

ED instantiations share all the same structure (that of EDs) and represent different examples of the ED. Both the MOI and ED files use hierarchical vocabularies to instantiate source terms. The advantages of a hierarchical representation can be found in the literature [6]. However, alternative dictionary structures have been used to code patient information (e.g., a flat list of terms combined with properties describing interdependence between terms as in the QMR program [11]). In the UMLS patient database it is possible to combine different implementation schemes (hierarchical vs. flat list) into one common representation. Unique diagnostic concepts from each vocabulary source are added to the MOI hierarchical structure, and a set of ED instantiations is constructed automatically to describe the dictionary terms of each source. For example, the three items below

(QMR) CHEST PAIN SUBSTERNAL EXERTIONAL (833)
(Iliad) CHEST PAIN (1.10.14.0.0.0)
WITH EXERCISE (1.10.14.18.0.0)
(HELP) DO YOU GET CHEST PAIN WHEN YOU EXERCISE?
(7.1.120.2.1.28.0.0)

were mapped into two event definition instances (since the Iliad and HELP terms describe the same information but not the QMR term) as shown below:

map_id	ED_ID	Instances
3482	1 (symptom)	A:Concept: PAIN G:Body Parts: CHEST I:Actions: EXERCISE
3480	1 (symptom)	A:Concept: PAIN G:Body Parts: SUBSTERNAL I:Actions: EXERCISE

The UMLS patient database structure

A relational database model with highly normalized records was selected for the implementation of the UMLS patient records. The main tables, illustrated in Figure 3, encode the patient demographic data and the clinical information.

Pt_Demg: (Patient demographic data)

s_db	pt_num	name	sex	birthday	admit_date
1	46758	Alb	M	19480102	19850203
2	34215	Blo	F	19390403	19860714
4	83942	Cli	M	19400708	19880705

Pt_DB: (Patient data records)

s_db	pt_num	time_stamp	map_id	value
1	46758	56748	100	1.000
1	46758	57686	200	5.000
2	34215	58767	400	1.000
2	34215	74833	300	5.000
4	83942	58393	700	5.000
4	83942	93773	800	1.000

Figure 3: Organization of the demographic and clinical information in the UMLS patient database where s_db refers to the source database (1=Iliad, 2=QMR, 3=HELP), pt_num is the patient identification number and map_id is the assigned instantiation id number.

EXPERIENCE WITH A PROTOTYPE

To build a prototype of the UMLS patient database, we have selected two medical expert systems (QMR and Iliad) and a hospital information system (HELP). Each system includes a database of patient cases. Subsets of cases from each system were selected as follow:

HELP All cases admitted to the hospital are in the HELP database. Most patients do not have history and physical exam data, but all have lab, X-ray, pharmacy and discharge diagnosis.

Iliad All cases entered by junior medical students during their clerkship on medicine at each of the three teaching hospitals used by the University of Utah, medical school.

QMR Arrangements were made with Dr. Randy Miller at UMLS meeting for QMR cases from New England Journal of Medicine CPC's to be included in the UMLS database.

Distinct controlled medical vocabularies are used in each system to represent patient information. Also, the patient record and database structure in each system is different. The patient record structure in Iliad and QMR is simply a list of dictionary codes and associated values, since both systems do not deal with temporal data, whereas the structure of the clinical database and patient record in HELP is far more complex [3].

For each system a utility program was written to convert patient information into relational format and then to transfer it into a separate text files while patient identification was scrambled for privacy purposes. Also, for the purpose of this prototype, only a subset of data for each patient was extracted (i.e., present history items, physical exam observations, chemistry and hematology laboratory results and discharge diagnoses.) The MOI file information and the ED instantiations were built from these text files.

Special attention was given to how these three systems represent numerical values. Iliad and HELP store laboratory test results as a dictionary code and the actual numerical test result (continuous variables), where as QMR only expresses the range in which the value falls (e.g., "WBC 14000 TO 30000") (discrete variables). This discrepancy is handled by assigning to the QMR term a two valued instantiation. Similarly, a user search for patients with "WBC greater than 16000" is instantiated in the same way and records from the UMLS patient database will be retrieved by matching the test name instances and then evaluating if the values satisfy the search criteria.

A prototype system was built based on a set of patients (total 279) who were diagnosed with chronic bronchitis, pulmonary embolism, emphysema, asthma or pneumococcal pneumonia. The relevant data (i.e., historical and physical exam, chemistry and hematology labs and diagnoses) were extracted from the source databases and transferred into the appropriate UMLS data structure.

Automatic instantiations

Once the EDs and MOI file have been built to reflect the medical information contained in each source dictionary, ED instances are built automatically for each dictionary term. Creation of the ED instances involves several steps. First, source dictionary terms are manually selected by corresponding ED. Secondly, each sentence is broken into words and the words are matched against slot values (i.e., hierarchy of terms) which are retrieved from the MOI file. Finally, matches are determined on the basis of the MOI hierarchical code and therefore only exact matches are accepted. This automated algorithm has proven efficient with 84% success as measured on the experimental prototype. The reasons for failure of automatic instantiations are distributed in three categories as follow:

1. Incomplete ED database (3%). The sentence described is an instance of an ED that has yet to be constructed (e.g., exposure history, family history, medication history)
2. Incomplete set of aliases in the MOI file (9%). (e.g., Shortness of Breath, SOB, Dyspnea)
3. The natural language barrier (4%). (e.g., "Are you aware of a tumor or growth?" (HELP))

Growth of the MOI file with new sources

Another experiment was conducted to evaluate the amount of overlap in the MOI entries between Iliad, QMR and HELP within the limited domain of this prototype. The goal of this is to get a sense for the rate of growth of the MOI file which, theoretically should level out as new dictionary sources are added. Figure 4 illustrates the number of common concepts between the three systems for the specific vocabulary subdomain considered (i.e., 17 cases of asthma and pneumonia).

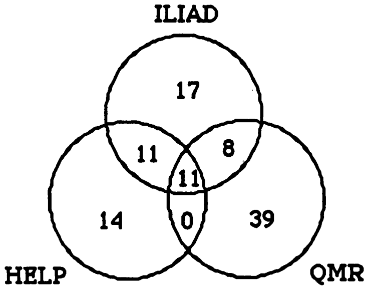


Figure 4: Degree of overlap between Iliad, QMR and HELP dictionaries based on 17 cases of pneumonia and asthma. The values shown represent the percentages of dictionary terms in each subset.

Figure 5 offers another perspective at the growth of the MOI file. The figure delineates the decrease in percentage as new terms from new vocabulary scources are added to the MOI file.

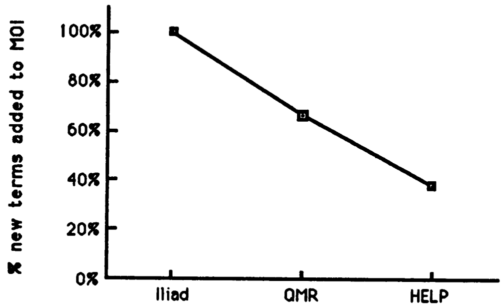


Figure 5: Percent new terms contributed to the MOI file as new dictionary sources are added to the system.

Sample searches from the UMLS patient database

Users' queries are treated as another dictionary source to be instantiated dynamically by the system. Some sample searches from the prototype system of 279 patients are detailed below.

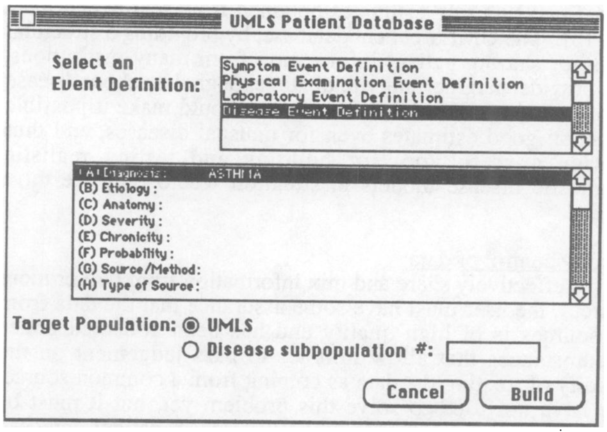
1. Specify search using Event Definitions

To perform a search the user selects an event definition from a menu (see Figure 6). Slots in the ED frame are filled by selecting the slot and then choosing from a list of eligible items of the desired value. In Figure 6, the user wishes to generate a subpopulation of patients with the disease Asthma from the entire UMLS patient database. To accomplish this the selected event definition is then displayed in the bottom window. In this example, the disease slot is highlighted and the user types in "asthma", which is used as a key to select the diagnosis "asthma" from the terms in the diagnosis hierarchy. The user then selects the database to be searched. Other slots may be filled to specify modifiers and narrow the search criteria.

Clicking on "Build" will start the search. When an instantiated event definition matches the search definition, that patient number will be saved in a list as a subpopulation file for future use and the user will be asked to assign a name and a number to this new subpopulation.

2. Searching to obtain an estimate of disease prevalence

To perform this function, the user first specifies the name of the subpopulation (Asthma) which is to be searched. The program displays the total number of patients found as well as the number found coming from each of the sources (Iliad, QMR, and HELP), the total number of patients searched in each source (the denominator), and the prevalence of the disease estimated from these numbers (see Figure 7). Of course, the prevalence may be quite different, depending upon the conditions under which the population was sampled and on the incidence of the disease in different areas.



Select an Event Definition:
Symptom Event Definition
Physical Examination Event Definition
Laboratory Event Definition
Disease Event Definition

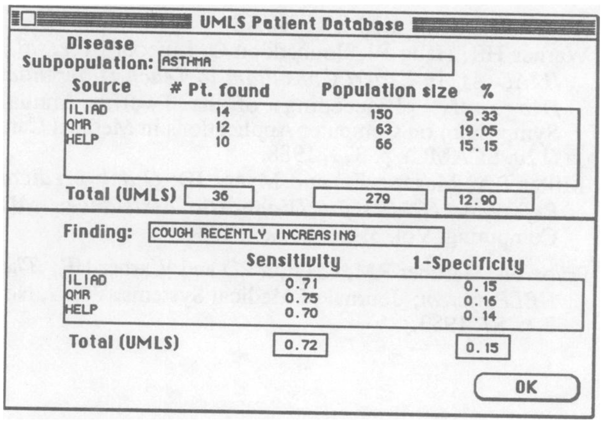
(A) Diagnosis: ASTHMA
(B) Etiology :
(C) Anatomy :
(D) Severity :
(E) Chronicity :
(F) Probability :
(G) Source/Method:
(H) Type of Source :

Target Population: ● UMLS
○ Disease subpopulation #: □

Figure 6: A sample of a UMLS patient database query window

3. Estimating sensitivity and specificity of a finding

In the bottom half of Figure 7, the user has specified a finding by selecting the symptom event definition and entering "cough" as the symptom name and "recently increasing" as a modifier. The sensitivity and 1-specificity of this finding in patients from each of the source files with asthma is displayed for the user. If these statistics are similar from each source, the user may use the cumulative values across the whole UMLS patient database with added confidence.



Disease Subpopulation: ASTHMA

Source	# Pt. found	Population size	%
ILIAD	14	150	9.33
QMR	12	63	19.05
HELP	10	66	15.15
Total (UMLS)	36	279	12.90

Finding: COUGH RECENTLY INCREASING

	Sensitivity	1-Specificity
ILIAD	0.71	0.15
QMR	0.75	0.16
HELP	0.70	0.14
Total (UMLS)	0.72	0.15

Figure 7: A sample of a UMLS patient database search summary window

It took about five seconds to complete this particular search on this small patient database. As patient cases are added to the system, search time would be expected to increase proportionally.

POTENTIAL BENEFITS AND PROBLEMS

Knowledge engineering support

We have shown that statistical estimates such as these can be used to effectively model the diagnostic decision process of expert clinicians, and that such a model can be used as both a teaching tool and a consultant [1]. Our knowledge engineering efforts to build such a model of diagnosis begin by asking domain experts for subjective estimates of sensitivities, specificities, and *a priori* probabilities. We have shown that the performance of these models improves as these estimates are refined by using estimates obtained from real patient data [12, 13]. The UMLS patient database, by providing a structural basis for sharing patient information from many institutions, will provide an opportunity to obtain enough data about disease manifestations. This UMLS resource should make it possible to obtain good estimates even for unusual diseases, and thus provide a useful tool for building and testing realistic diagnostic disease models in situation where they are most needed.

Quality control of data

To effectively share and mix information from two or more sources, the user must have some assurance that the data from all sources is of high quality and has been collected under circumstances that allow him/her to pass judgement on the validity of treating the data as coming from a common source. We have not tried to solve this problem yet, but it must be approached in the future to make the UMLS patient database the kind of useful resource we hope it will become.

There exists other potential benefits of this approach to a unified medical record system which will be discussed in future papers. These include 1) coupling real patient records to medical expert systems, 2) "free text" input of patient cases in medical information systems (e.g., QMR, HELP), 3) "free text" query of HELP patient database or of MeSH indexed literature database. These benefits should provide the appropriate stimulus for interested groups to share their patient information resources.

ACKNOLEDGEMENT

This work is supported in part by contract number N01-LM-8-3515 from the National Library of Medicine.

REFERENCES

[1] Warner HR, Haug PJ, Bouhaddou O, Lincoln M, *et. al.*; *ILIAD As An Expert Consultant to Teach Differential Diagnosis*; Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care (12th SCAMC), p. 371, 1988.

[2] Miller RA, Masarie FE and Myers JD; *Quick Medical Reference (QMR) for Diagnostic Assistance*; MD Computing, Vol. 3, No. 5, p.34, 1986.

[3] Pryor TA, Gardner RM, Clayton PD and Warner HR; *The HELP system*; Journal of Medical Systems, Vol. 7, No. 2, p. 87, 1983.

[4] Humphreys BL and Lindberg DAB; *Building the Unified Medical Language System*; Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care (13th SCAMC), p. 475, 1989.

[5] Tuttle M, Sherertz D, Erlbaum M, Olson N and Nelson S; *Implementing Meta-1: The First Version on the UMLS Metathesaurus*; Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care (13th SCAMC), p. 483, 1989.

[6] Huff SM, Craig RB, Gould BL, Castagno DL and Smilan RE; *Medical Data Dictionary for Decision Support Applications*; Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care (11th SCAMC), p. 310, 1987.

[7] Miller RA, Masarie FE, Claudon CH, Giuse NB, Warner HR and Bouhaddou O; *Mapping of Medical Knowledge Representations: INTERNIST-I, HELP, and MeSH*; Final Task Report, Task 4, Unified Medical Language System, Contract No. N01-LM-6-3522,October 9, 1987.

[8] Masarie FE, Cimino JJ, Giuse NB and Miller RA; *Mapping Between Controlled Vocabularies: QMR and DXplain*; Report of Results of Task 5, Subtask 6, N01-LM-6-3522, April 14, 1988.

[9] Cimino JJ and Barnett GO; *Automated translation between medical terminologies using semantic definitions*; Proceedings of the AAMSI Congress, 1989.

[10] Masarie FE, Miller RA, Bouhaddou O, Giuse NB and Warner HR; *Creating an Interlinga for Electronic Interchange of Medical Information: Frames for Mapping Between clinical Vocabularies;* (in preparation).

[11] Masarie FE, Miller RA, Myers JD; *INTERNIST-1 Presenting Common Sense and Good Medical Practice in a Computerized Medical Knowledge Base";* Computer and Biomedical Research, vol. 18, p.458, 1985

[12] Haug PJ, Clayton PD, Shelton P, Rich T,Tocino I, Frederick PR, Crapo RO and Morrison WJ; *Revision of diagnostic Logic using a Clinical Database*; Proceedings of the AAMSI Congress, p. 238, 1987.

[13] Bouhaddou O, Haug PJ and Warner HR; *Use of the HELP Clinical Database to Build and Test Medical Knowledge*; Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care (11th SCAMC), p. 64, 1987.