CHARTLINE: Providing bibliographic references relevant to patient charts using the UMLS Metathesaurus Knowledge Sources*

Randolph A. Miller, Filip M. Gieszczykiewicz, John K. Vries, and Gregory F. Cooper
University of Pittsburgh School of Medicine, Pittsburgh, Pa.
*Supported by UMLS Contract N01-LM-1-3535 from the National Library of Medicine.

ABSTRACT

A successful medical informatics program helps its users to match their information needs as closely and efficiently as possible to the capabilities of the system. CHARTLINE is a computer program whose input is a free text, "natural language" patient chart in ASCII format. Using the UMLS Metathesaurus Knowledge Sources, CHARTLINE can suggest bibliographic references relevant to the patient case described in the chart. The program does not attempt to "understand" the natural language content of the chart. CHARTLINE only recognizes UMLS Metathesaurus Main Concept terms (or their synonyms) as they occur in the medical text, since those terms represent the tokens used to index the literature. The program depends on user feedback to determine which topics of a large number of potentially relevant subjects are of interest to the user.

INTRODUCTION

The goal of CHARTLINE is to produce clinically relevant bibliographic references from input consisting of a textual (ASCII) patient record. We describe the system at a relatively early stage in its development as an example of how useful applications can be based on the National Library of Medicine's (NLM's) Unified Medical Language System (UMLS) Metathesaurus Knowledge Sources. The version of the UMLS Metathesaurus used to develop CHARTLINE is Meta-1.1 (October 1991 release). While CHARTLINE is under development at the University of Pittsburgh, all components described in this document exist as running prototypes. CHARTLINE will have significant impact in our Medical Center in the near future, and has the potential to improve patient related bibliographic access at remote settings in the intermediate term future.

DESCRIPTION OF CHARTLINE

The three main components of the current version of CHARTLINE consist of a clinical workstation, textual patient records from our hospital information system (HIS), and a MEDLINE bibliographic reference system. For the current version of CHARTLINE, a TCP/IP ethernet network connects the three components of CHARTLINE - the workstation (an IBM RS/6000® UNIX® System in the Section of Medical Informatics), the electronic chart, and the bibliographic database. The electronic patient records reside in the University of Pittsburgh's Medical ARchival System (MARS), a hospital information system developed by Vries and Yount [1] locally. MARS is housed in the Office of Biomedical Informatics at a site distant from the Section of Medical Informatics. Through the Medical ARchival System (MARS) we have available, on-line, over 200,000 patient records. These records contain history and physical examination and discharge summary notes; over 400,000 radiology reports; 188,000 pathology reports; 9,300 cytology reports; 1,000 autopsy reports; 25,000 dictated outpatient clinic notes; 11,000 referral letters; 4,700 cardiac catheterization reports; and a variety of other reports. The MARS system is both expanding in scope (new forms of reports are being added to those available) and content (new patient records are added to the system each time patients are seen in the hospital or outpatient clinics). The current bibliographic retrieval system is a local MEDLINE implementation that resides on the same machines as the MARS system.

The CHARTLINE System has been constructed generically. Future versions of CHARTLINE can function in a physician's office, using an IBM-compatible PC running Microsoft Windows® as both the workstation and MEDLINE bibliographic reference system (via CD-ROM). In such an office setting, the textual document might be provided by a dictated history and physical exam note that was transcribed on a word processor (or alternatively a typed or legibly printed paper document that was scanned into an ASCII file using an OCR system).

METHODS

At present, a CHARTLINE user identifies the patient record of interest by submitting a search query to the MARS system. A MARS query could be as simple as "jones joseph", which would retrieve all information (H&PEs, discharge summaries, radiology reports, lab results, etc.) about the patient named "joseph jones". The retrieved patient record is next processed on the RS/6000 Workstation (in several steps) to produce pointers to the medical literature. The steps are carried out by programs written in the "C" programming language.

0195-4210/92/$5.00

To create CHARTLINE data structures, we process the Meta-1.1 CD-ROM data files to create more compact and more rapidly accessible disk files on the RS/6000. The first CHARTLINE data file ("UMLS words file") contains all unique words obtained from terms in the UMLS "MRMC" (main concepts) distribution file. We create Key Word in Context (KWIC) lists to relate entries from the UMLS words file (via their unique word identifier numbers) to the UMLS main concept terms (via their Metathesaurus unique identification numbers).

Figure 1 shows a sample patient record that has been parsed to identify UMLS words that appear in the chart. In Figure 1, words from a sample chart appear on the left-hand margin of the column (see full text, Figure 2), and the manner in which the CHARTLINE parser matched these words to Meta-1.1 words is indicated on the right hand side of Figure 1. The words in a given Meta-1.1 term may appear arbitrarily in their singular, plural, or possessive forms (e.g., Wilson's Disease, Wilsons Disease, or Wilson Disease). Similarly, the words in a patient record may be arbitrarily in singular, plural, or possessive form. It is therefore necessary to match any word appearing in a chart with the singular, plural, or possessive variants it may have in Meta-1.1. For the purpose of handling singular, plural, and possessive variants of Meta-1.1 words, we wrote a program to convert a word into its potential English, Latin or Greek singular, possessive or plural form. The program is used to generate the alternative forms for chart words that are then matched against words from the Metathesaurus. In Figure 1, "complaint" from the chart matches "complaints" in the Metathesaurus word index; "history" matches both "history" and "histories", and so on.

A "stop list" of common English words is used to eliminate false positive matches. We set a cutoff word length (4 characters), below which chart words are ignored. As seen in Figure 1, words of three characters or less are STOPWORDs due to their length, and generate "NO MATCH" automatically. On the other hand, "with" and other common English words were specifically listed as STOPWORDs. It is our experience that the majority of words in noun phrases found in medical charts are "medical words", in that they are words that participate in terms of the UMLS Metathesaurus.

The next step in processing a patient record is to retrieve the KWIC lists for all words recognized in the chart (i.e., words identified in Figure 1) and to use them to identify potential matches with single or multi-word Meta-1.1 terms. A serial sliding-frame methodology is used to accumulate potential matches. For each word in the patient record that is recognized as a Meta-1.1 word, the word's KWIC list is logically ORed with its singular, plural, or possessive variant's KWIC lists (if such variants are indeed Meta-1.1 words) to create a word-KWIC-set. The first "combined" word-KWIC-set from Figure 1 would be created by combining (ORing) the KWIC lists of the words "history" and "histories", because the words "complaint" and "shortness" only match one singular or plural form -- their KWIC lists directly form their word-KWIC-sets.

```
[ chief   NO MATCH ]
[ complaint ..... complaints    #6499 ]
[ shortness ..... shortness     #24281 ]
[STOPWORD of      NO MATCH ]
[ breath ..... breath       #4394 ]
[STOPWORD and     NO MATCH ]
[ swelling ..... swelling     #25549 ]
[ feet ..... feet         #10736 ]
[STOPWORD      NO MATCH ]
[STOPWORD      NO MATCH ]
[ history ..... histories     #13064
   ..... history      #13066 ]
[STOPWORD of      NO MATCH ]
[ present   NO MATCH ]
[ illness ..... illness       #13858 ]
[STOPWORD mr      NO MATCH ]
[ jones NO MATCH ]
[STOPWORD is      NO MATCH ]
[STOPWORD a     NO MATCH ]
[STOPWORD 76      NO MATCH ]
[STOPWORD year     NO MATCH ]
[STOPWORD old     NO MATCH ]
[ white ..... white         #28296
   ..... whites        #28298 ]
[ male ..... male          #16183 ]
[STOPWORD with     NO MATCH ]
[STOPWORD a     NO MATCH ]
[ history ..... histories     #13064
   ..... history      #13066 ]
[STOPWORD of      NO MATCH ]
[ myasthenia ..... myasthenia    #17893 ]
[ gravis ..... gravis        #12286 ]
[ coronary ..... coronary      #6769 ]
[ artery ..... arteries      #2998
```

Figure 1: CHARTLINE Identification of Meta-1.1 Words in Chart

To identify potential matches of phrases in the medical chart with single or multi-word Meta-1.1 terms, the word-KWIC-sets for successive words recognized from the chart are intersected (logically ANDed) together serially. In Figure 1, the first such AND operation would intersect the KWIC list of "complaints" with the KWIC list of "shortness". When the intersection of a series of KWIC lists process produces a "NULL" result, (i.e., when the resultant set is not consistent with any Meta-1.1 main concept name or synonym), the program backtracks to retrieve the most recent set produced by KWIC list intersection that was not the null set.

CHIEF COMPLAINT: **SHORTNESS OF BREATH AND SWELLING FEET** HISTORY OF PRESENT ILLNESS: Mr. Jones is a 76-year old **white** male with a history of **myasthenia gravis, coronary artery disease,** status post **myocardial infarction** in approximately July 1978 and **seizure** disorder who presents now with a 2 to 4 week history of increasing **dyspnea** and peripheral **edema. His cardiac** history began in July 1978 when he was evaluated at an **outpatient clinic** and found to have new EKG changes consistent with a recent anterior wall **myocardial infarction.** The patient states approximately 2 to 3 weeks before he had been seen in the clinic, he had an episode of substernal squeezing-like **chest pain** associated with **dyspnea** which also radiated down both **arms.** At the time of evaluation at the **outpatient clinic,** it was decided that the patient had already had his event and that further evaluation was not necessary. He was therefore sent home on **Diltiazem** and **nitrates** for the anginal-type **chest pain** that he was still having. Since that time, he has done well up until about one month prior to admission when started noting the **paroxysmal nocturnal dyspnea** and orthopnea. On the day of **admission, the patient** presented to the **Geriatric** Center for evaluation of his worsening **dyspnea** and **edema.** He denies any episodes of **chest pain, palpitations, nausea, vomiting,** diaphoresis, **cough,** decreasing **urination** or other **symptoms.** He says his **symptoms** have been slow on onset and have been stable over the past 2 to 3 weeks. From the clinic, he was sent directly to the **University hospital** emergency department and was found to be in A-fib **flutter.**	M0037002 MC0013404 Shortness of breath 66 M0038999 MC0038999 Swelling 100 M0043157 MC0043157 Whites 100 M0026896 MC0026896 Myasthenia Gravis 100 M0010063 MC0010068 Coronary Artery Disease 100 M0027051 MC0027051 Myocardial Infarction 100 M0036572 MC0036572 Seizures 100 M0013404 MC0013404 Dyspnea 100 M0013608 MC0013608 Edema, Cardiac 100 M0029916 MC0029916 Outpatient Clinics, Hospital 66 M0027051 MC0027051 Myocardial Infarction 100 M0008031 MC0008031 Chest Pain 100 M0013404 MC0013404 Dyspnea 100 M0003792 MC0003792 Arm 100 M0029916 MC0029916 Outpatient Clinics, Hospital 66 M0012373 MC0012373 Diltiazem 100 M0028122 MC0028124 Nitrate 100 > M0028125 MC0028125 Nitrates 100 M0008031 MC0008031 Chest Pain 100 M0013406 MC0013406 Dyspnea, paroxysmal nocturnal 100 M0030673 MC0030673 Patient Admission 100 M0017469 MC0017469 Geriatrics 100 M0013404 MC0013404 Dyspnea 100 M0013604 MC0013604 Edema 100 M0008031 MC0008031 Chest Pain 100 M0030252 MC0030252 Palpitations 100 M0027498 MC0027498 Nausea and vomiting 66 M0010200 MC0010200 Cough 100 M0042034 MC0042034 Urination 100 M0001726 MC0001726 Affective Symptoms 100 > M0004941 MC0004941 Behavioral Symptoms 100

Figure 2: Matching Text From Sample Chart to Meta-1.1 Terms Using CHARTLINE

This "non-null intersected KWIC list" corresponds to the longest current string of Meta-1.1 words from the chart which participate in common Meta-1.1 term names. In Figure 1, the word "complaints" would generate a non-null intersected KWIC list. Since no Meta-1.1 terms contain the words "complaints" and "shortness", the phrase "complaints" is not extended. The two words "shortness" and "breath" would generate the next non-null intersected KWIC list. The number of Meta-1.1 terms to which a non-null intersected KWIC list points can vary from one term to hundreds of terms.

A set of heuristic rules is next applied to see if an appropriate match exists between the words in the chart and one (or a few) Meta-1.1 term(s). In effect, we must determine if it is reasonable to match a series of "recognized" words from the chart to the set of Meta-1.1 terms corresponding to those words' non-null intersected KWIC lists. The first heuristic employed counts the number of chart words that actually appear in each candidate Meta-1.1 term "matched". In the right hand column of Figure 2, the numbers to the right of the terms matched indicate the percentage of words in each Meta-1.1 term that appeared in the chart. If less than 51% of the words in the candidate Meta-1.1 term appeared in the chart, the term is rejected (and does not appear in the right hand column in Figure 2). In addition, only the candidate Meta-1.1 term with the highest percentage of matched words from the chart is retained (others do not appear in Figure 2). For example, if a phrase found in a chart is "insulin dependent diabetes", then 75% of the words in the Meta-1.1 term "Diabetes Mellitus, Insulin Dependent" are matched, but only 60% of the words in the term "Diabetes Mellitus, Non Insulin Dependent" are matched, so the latter term is dropped. If, after applying these heuristic rules, five or fewer Meta-1.1 terms remain as candidate matches for a phrase in the chart, (i.e., there are five or fewer Meta-1.1 survivors from the phrase's non-null intersected KWIC list), then the chart phrase and terms are considered "possibly matched". Figure 2 shows how the words from a sample chart (left hand column) are matched by this algorithm to Meta-1.1 terms (right hand column).

Next, the Meta-1.1 terms identified from the chart are processed using the co-occurrence of terms data from the UMLS Metathesaurus Knowledge Sources (MRCOT files). The MRCOT file links individual Meta-1.1 terms with other Meta-1.1 terms, whenever both terms were used concurrently as "main MeSH headings" in indexing a given article (or set of articles) from the literature (for MEDLINE from 1983-91). For example, there might have been 254 articles (in the literature indexed by MEDLINE between 1983 and 1991) which concurrently discuss "Coronary Artery Disease" along with "Chest

Pain". The fact that 254 articles refer to both terms can be retrieved from the MRCOT file. Accessing the MRCOT file for a given Meta-1.1 term produces a list of all other Meta-1.1 terms that appeared with that term as main concept headings in literature articles, as well as the number of such co-occurrences for each pair of terms.

The next step in processing a patient record is to determine potentially interesting MEDLINE searches that CHARTLINE might suggest for its users. We must, for each Meta-1.1 term identified from the chart ("focus" term), retrieve the list of co-occurring terms from the literature (via the MRCOT file). We then intersect the list of all Meta-1.1 terms that co-occur in the literature with the "focus" term (as determined by the MRCOT file) with the list of all Meta-1.1 terms recognized from the chart. The resultant list is the set of Meta-1.1 terms that both appear in the chart and are related to the focus term via one or more MEDLINE-indexed literature articles. These pairs of terms are guaranteed to produce non-null retrievals when submitted as conjunctive searches to MEDLINE. The lists are currently displayed to the user sorted by descending frequency of term co-occurrence, since terms that co-occur frequently are more likely to be related in a medically meaningful way. Figure 3 shows the result of processing the patient chart from Figure 2 using "Atrial Flutter" and then "Cellulitis" as the focus terms.

At this point, CHARTLINE has identified many bibliographic searches (in theory, all possible pairwise searches) that can relate two separate terms from the patient chart to the medical literature (i.e., the literature indexed by MEDLINE). The final step in CHARTLINE is for the user to identify which of the potential searches (that might be submitted to MEDLINE) are of actual interest to the user. At present, this is done manually. Imagine that the user identifies, from an output similar to that of Figure 3, that he or she is interested in references that relate the terms Procainamide and Myasthenia Gravis. Using MARS MEDLINE, the user conducts the search, and obtains the reference "Procainamide-induced myasthenic crisis" (Godley PJ et al; Ther Drug Monit 1990; 12:411-414) which indicates that procainamide may induce respiratory failure in patients with myasthenia gravis. Thus, CHARTLINE might be of clinical value in managing a patient with myasthenia gravis and atrial flutter, in whom physicians mentioned procainamide as a possible therapy in their admission history and physical examination note.

An Alternative Strategy for Text Recognition

Author GFC and his colleagues are investigating a probabilistic method for identifying a set of MeSH Headings that correspond to the concepts expressed in the text of a patient chart. The method generates the probability of each possible MeSH Heading given a text phrase. The text phrases are taken from the text in titles and abstracts of MEDLINE articles and are associated probabilistically with the MeSH headings already assigned to the article in which the text phrase appears. The calculations are made for text phrases of one to four words in length.

Processing 4239 Atrial Flutter	**Processing 7642 Cellulitis**
Co-occurs with (11 times) Quinidine 34414	Co-occurs with (19 times) Septicemia 36690
Co-occurs with (8 times) Digoxin 12265	Co-occurs with (16 times) Neck 27530
Co-occurs with (7 times) Procainamide <1> 33216	Co-occurs with (8 times) Cholecystitis 8325
Co-occurs with (4 times) Heart Failure, Congestive 18802	Co-occurs with (6 times) Leg 23216
Co-occurs with (4 times) Diltiazem 12373	Co-occurs with (5 times) Antibiotics 3232
Co-occurs with (3 times) Myocardial Infarction 27051	Co-occurs with (5 times) Edema 13604
Co-occurs with (1 times) Coronary Disease 10068	Co-occurs with (3 times) Blindness 5752
Co-occurs with (1 times) Cardiopulmonary Bypass 7202	Co-occurs with (2 times) Arm 3792
Co-occurs with (1 times) Heart Enlargement 18800	Co-occurs with (2 times) Venous Insufficiency 42485
Co-occurs with (1 times) Heart Diseases 18799	Co-occurs with (1 times) Groin 18246
Co-occurs with (1 times) Myasthenia Gravis 26896	Co-occurs with (1 times) Hydronephrosis 20295
Co-occurs with (1 times) Substance Withdrawal Syndrome 38587	Co-occurs with (1 times) Coronary Disease 10068
Co-occurs with (1 times) Venous Insufficiency 42485	Co-occurs with (1 times) Cholecystectomy 8320
	Co-occurs with (1 times) Myocardial Infarction 27051

Figure 3: CHARTLINE co-occurrence output for two sample foci, using patient chart of Figure 2

One assumption made in computing such probabilities is that the meaning of a phrase in a MEDLINE abstract is the same (or almost the same) as its meaning in a clinical chart. An advantage of the probabilistic approach is that MEDLINE provides an enormous training set with millions of indexed MEDLINE articles and abstracts. Preliminary results have been encouraging when using only a small training set of about 20,000 MEDLINE articles. Examples of the results obtained with this training set are (for brevity, we list here only the probabilities above 0.25): "chronic pancreatitis" (text phrase) --> MeSH P(Pancreatitis | chronic pancreatitis) = 0.60; P(Chronic Disease | chronic pancreatitis) = 0.40; P(Pancreas | chronic pancreatitis) = 0.27; "varicella zoster" (text phrase) --> MeSH P(Varicella-Zoster Virus | varicella zoster) = 0.56; P(Acyclovir | varicella zoster) = 0.44; P(Herpes Zoster | varicella zoster) = 0.44; and P(Chickenpox | varicella zoster) = 0.33. While the these examples suggest the potential for the probabilistic

approach to yield useful results when applied to clinical text, formal testing is needed.

DISCUSSION

It is desirable to interconnect medical records and medical decision support systems with the relevant medical literature (at least in the form of bibliographic references if full text is not available). A number of previous efforts have developed systems for such purposes. As part of the UMLS Project in the late 1980s, Dr. James J. Cimino carried out initial studies of the utility of using co-occurences of terms for determining medical causality. Powsner et al described PsychTopix, a system which allowed users to "underline" phrases of interest in an electronically displayed in a psychiatric consultation note. The PsychTopix knowledge base stored a series of canonical text phrases identifying "potential topics of interest", such as DSM-IIIR categories. Each topic in the knowledge base also had "canned MeSH logic" indicating an optimized search strategy for obtaining literature relevant to the canonical term. User-underlined topics of interest were then matched with known canonical, "searchable" topics, allowing users to obtain pertinent references [2]. Chris Cimino et al, in the Intelligent Query Workstation (IQW), are also using the UMLS Metathesaurus and UMLS Information Sources Map to link text from a patient record to a number of information sources, including MEDLINE [3]. Hersh et al, in developing the Sapphire system, used the Shoval algorithm (a semantic network expansion technique that identifies concepts relevant to input terms) to link patient records to auto-indexed reference material on AIDS, including a local subset of literature references on AIDS [4]. Hammond et al developed a MEDLINE interface for the TMR medical record system [5]. A number of medical decision support systems, including AI/RHEUM [6], Iliad [7], and QMR[8], have "hooks" to bibliographic search engines. These systems allow their users to conduct searches related to topics the users identify during the course of interacting with the medical decision support systems.

The simple approach embodied in CHARTLINE is both a strength of the system and a weakness. The program merely identifies the words in the chart that exist in any Meta-1.1 term, since only those words match Meta-1.1 terms. The recognized words are used to match Meta-1.1 terms to the contents of charts, using heuristic methods we have described. We take advantage of the term-level synonymy provided in the Meta-1.1 to avoid complex computationally expensive linguistic parsing techniques and thesaurus expansion techniques. Our heuristic strategy allows CHARTLINE to present lists of potentially relevant searches to the user, in a manner that constrains the user to selecting searches that will retrieve references in MEDLINE. CHARTLINE then allows the user to determine what is most relevant or of greatest interest. Combining the lexical and probabilistic methods for term recognition will ultimately produce better results than either alone.

While the present version of CHARTLINE only relates components of a single text document to other portions of the same document, it would not be difficult to extend the methodology to relate, for example, the text of a radiology report to the contents of the patient's history and physical examination, or an autopsy report to the corresponding patient discharge summary. CHARTLINE could also be applied to text documents that are not patient records. For example, if a medical student wanted to obtain bibliographic references relevant to his or her lecture notes, if the lecture notes were available as an ASCII text file, CHARTLINE could be used to accomplish the desired linkages.

REFERENCES

1. Yount RJ, Vries JK, Council CD. The Medical Archival System: An information retrieval system based on distributed parallel processing. *Information Processing & Management.* 1991; 27:379-391.
2. Powsner SM, Miller PL. Linking Bibliographic Retrieval to Clinical Reports: Psych Topix. *Proc SCAMC 13.* Wash,D.C. IEEE Press. 1989; 431-435.
3. Cimino C, Barnett GO. Standardizing Access to Computer-Based Medical Resources. *Proc SCAMC 14.* Wash, D.C. IEEE Press. 1990; 33-37.
4. Hersh WR, Pattison-Gordon E, Evans DA, Greenes RA. Adaptation of Meta-1 for SAPPHIRE, A General Purpose Information Retrieval System. *Proc SCAMC 14.* Wash, D.C. IEEE Press. 1990; 156-160.
5. Hammond JE, Hammond WE, Stead WW. Information Integration through Distributed Resources: the TMR-NLM Connection. *Proc SCAMC 14.* Wash, D.C. IEEE Press. 1990; 719-723.
6. Kingsland LC III, Rosenberg KM, Cheh ML. CTX: The NLM Criteria Engine. *Demo Digest, SCAMC 12.* Wash, D.C. IEEE Press. 1988; 23-24.
7. Fan CL, et al. Odyssey: A Program to Access Medical Knowledge. *Proc SCAMC 11.* Wash,D.C. IEEE Press. 1987; 488-491.
8. Miller RA, Jamnback L, Giuse NB, Masarie FE Jr. Extending the Capabilities of Diagnostic Decision Support Programs through Links to Bibliographic Searching: Addition of "Canned MeSH Logic" to the Quick Medical Reference (QMR)* Program for use with Grateful Med*. *Proc SCAMC 15.* Wash, D.C., McGraw-Hill, 1991.