SRI International

NATURAL LANGUAGE ACCESS TO A MELANOMA DATA BASE

Technical Note 171

September 1978

By: Martin N. Epstein
Division of Computer Research and Technology
National Institutes of Health
Bethesda, Maryland

Donald E. Walker
Artificial Intelligence Center
SRI International
Menlo Park, California

SRI Project 676D32

To appear in Proceedings of the Second Annual Symposium on Computer Applications in Medical Care
Washington, D.C., 5-8 November 1978.

The work reported herein was supported in part by the SRI Internal Research and Development Program.

SRI International®
333 Ravenswood Ave. • Menlo Park, California 94025
(415) 326-6200 • Cable: SRI INTL MPK • TWX: 910-373-1246

NATURAL LANGUAGE ACCESS TO A MELANOMA DATA BASE

Martin N. Epstein
National Institutes of Health
Division of Computer Research and Technology
Bethesda, Maryland 20014

Donald E. Walker
SRI International
Menlo Park, California 94025

Summary

This paper describes ongoing research towards developing a system that will allow physicians personal access to patient medical data through natural language queries to support both patient management and clinical research. A prototype system has been implemented for a small data base on malignant melanoma. The physician can input queries in English that retrieve specified data for particular patients or for groups of patients satisfying certain characteristics, that perform simple calculations, that allow browsing through the data base, and that assist in identifying relations among attributes. The system supports dialogue interactions; that is, the user can follow a line of inquiry to test a particular hypothesis by entering a sequence of queries that depend on each other. Classes of questions that can be processed are described and examples using the system are given.

Introduction

Medical groups are collecting data in machine-readable form that consist of information on patients with specific diseases. To make effective use of this available data, clinicians need to be able to examine that data rapidly and easily, and compare and contrast patients according to any subset of those data in a highly flexible manner. Systems providing these capabilities would encourage and support efforts at studying and modeling the disease process, assessing prognosis, and evaluating therapeutic protocols [1].

A limitation of current medical data systems is that they are difficult to use for people who are not familiar with the specialized retrieval operation required, regardless of their simplicity [2]. This is particularly true of physicians or other clinical investigators who are usually unwilling or unable to spend the time required to learn special query languages or conventions. Therefore an intermediary must act as translator between the individual who wants to know and the data system that may have the answers. This requirement inhibits intellectual access to the data and prevents the majority of physicians, even those who have collected the data, from having free and informal interaction with their data.

One approach to reducing this dependency on an intermediary is to structure the system so that the physician is presented successively with a series of alternatives which, based on the user's selection, drive the interrogation process down a particular predetermined path. This type of organization makes it unneccessary for a user to know much about the operation of the system, but since the system controls the inquiry process, the user is limited in the operations that he can initiate and in the order in which options can be considered. This approach has been characterized as a 'menu tree' or 'branching logic' model and has found application in many areas of medicine for data acquisition as well as data retrieval [3,4].

The objective of our research is to develop a clinical data base system that will allow clinicians personal access to data on specific diseases using queries formulated in English, study the use of the system by physicians and to evaluate the effects of using the system to aid the physician in patient management and in studying the course of a disease more precisely. This paper describes the first part of this task, the implementation of a prototype natural language question answering facility to provide a direct and convenient interface with a data base on patients with malignant melanoma. Thus, the goal is not to handle all of natural language but only that portion of language required to access a restricted domain of medicine. Several research groups in areas outside of medicine are working on English language access to data bases [5,6,7,8].

Development of a Prototype System

For physicians to make use of a computer facility for support of patient management or clinical research, there must be (1) a convenient and powerful interface that allows them to interact directly with a data base, and (2) the data available must support the purposes for which the facility is being used.

To meet the first requirement, we have made use of a language processing package, called LIFER (for Language Interface Facility with Ellipsis and Recursion) developed at SRI International. LIFER has been developed as a practical system for creating English language interfaces to a variety of different kinds of computer software. The system has demonstrated its utility in allowing people who

are not computer specialists to interact easily with a variety of data management systems [5].

The selection of a data base for the prototype system was influenced by the fact that the University of California, San Francisco (UCSF) has been a participant in a cooperative group of four institutions which has been collecting data on patients with malignant melanoma. One of the purposes of the project was to create a comprehensive data base in machine-readable form. A subset of those data, those patients who were seen at UCSF, was selected for inclusion in the data base for the prototype system.

The following sections will describe in greater detail the data base, the LIFER concept for creating the language interface, the initial access capability, and experience with the prototype system. The prototype system described is implemented on a DEC KL-10 computer.

Melanoma Data Base

The data base for the prototype system contains information about 156 attributes for 130 patients. The starting point in creating the data base was data collected at UCSF as part of the Melanoma Clinical Cooperative Group from 1972, when the first patients entered the study, through December, 1976. These data include personal data, patient background, patient physical examination, history of the primary lesion, physical examination of the primary, family history of skin disorders, pathology data, lymph node procedures, and follow up results.

The criteria used to eliminate items from the original study were: (1) information missing in a high percentage of cases; (2) information consistently recorded incorrectly; (3) attributes considered irrelevant on the basis of discussions with melanoma experts; (4) attributes with the same value in greater than 95% of the patients. In spite of this selection of items, the melanoma data base was found to be incomplete and to contain numerous errors, especially in the pathology data and follow up results. Therefore, an extensive effort was required to re-extract the pathology and follow up data elements from the primary records, to edit the attributes and values, and then to create the data base. To ensure that minimal criteria for completeness and accuracy were met, selected medical records were checked and corrections made to the data base where necessary. The data base can be updated as new data are collected or corrections made.

Use of LIFER to Create A Language Interface

LIFER is composed of two basic parts: a set of interactive language specification functions and a parser. The language interface builder uses the language specification functions to define an application language, a subset of English appropriate for interacting with the data base on patients with melanoma. The LIFER functions permit specification of patterns, fixed phrases, sets of words associated with word classes, and predicates. Other LIFER facilities include an automatic facility for handling ellipsis (incomplete sentences), a spelling corrector, a grammar editor, and a paraphrase mechanism to extend the language interface by defining new constructs at the phrase or sentence level. Several of these features derive in part from its implementation in INTERLISP [9], an interactive list processing programming language designed for working with complex non-numeric data. A complete manual is available describing the LIFER facility [10].

The following simplified discussion illustrates the types of information that need to be defined, compiled and stored in LIFER to respond to queries about the melanoma data base:

(1) A lexicon of words and phrases containing all the vocabulary items used in the queries. It lists all terms that may be assigned to a category. Thus, there are entries such as LIST, WHAT ARE, SHOW ME, and DISPLAY assigned to the category <WH-ARE/LIST>; TUMOR THICKNESS, LEVEL, and LEVEL OF INVASION are specified as <NUMERIC-ATTRIBUTES>; SSM and SUPERFICIAL SPREADING MELANOMA are assigned to <TYPE>; ARM and HAND to <SITE>; ELECTIVE, ELND, and ELECTIVE LND to <DISS-V>, and AXILLARY, CERVICAL, and INGUINAL to <LND-SITE>. Currently the lexicon contains over 750 words, excluding abbreviations and numbers, in 170 categories.

(2) A grammar containing rules that establish meaningful relations among vocabulary terms. Substantial amounts of semantic as well as syntactic information are embedded in these rules. Rules consist of two parts; a pattern to be recognized, and an associated response expression indicating what action should be taken. For example, [(<WH-ARE/LIST> <NUMERIC-ATTRIBUTES> <OF-FOR> <PATIENTS> <TABULATE-CLAUSE>) (F0041)] is a rule consisting of a pattern with F0041 as a response expression. The category <WH-ARE/LIST> is used to interpret a variety of English questions that ask for some kind of output display as a response. <TABULATE-CLAUSE> includes the specification of constraints on the output display, and the category <PATIENTS> contains rules (called subgrammars) relating <TYPE>, <SITE>, <DISS-V> and <LND-SITE>. The above rule is used to interpret the request "List the tumor thickness and level of invasion for people with ssm of the upper extremities who had Elective lymph node dissections ordered by site of the LNP". Matching the pattern with the above request would create the following query of the data base: (LEVEL ? THICKNESS ? TYPE SSM SITE (FMEMB * (HAND ARM)) LYMPHADENECTOMY.TYPE ELND LYMPHADENECTOMY.SITE ?).

(3) Special functions developed to evaluate particular classes of questions by performing the specified operations on the data base and then formatting the output to provide an appropriate response. In this example, using the response function F0041, the data base access program is called, the query matched against the data base, patient data satisfying the query constraints are returned, and an appropriate output display is presented.

Classes of Questions

The classes of questions implemented to date are summarized with examples as follows:

1. Information about the data base: the system can respond to requests about the attributes associated with an information class in the data base.
For example:

List the classes of attributes in the melanoma data base.
What items are associated with follow-up results?
Show me the data items associated with the history of the primary.
Display the values in the study for mitotic rate.
How is few mitoses defined?
List the pathology results for patient S-72-002.

2. Counts: the data base is searched and the number of instances satisfying a particular combination of attribute-values is displayed.
For example:

How many cases were there characterized by an increase in the size of the lesion for 6 months prior to initial therapy?
Count the number of individuals who had lymph node procedures.
What is the total number of lymph node procedures performed?
In how many patients was level 5 disease seen?
How many men with melanoma of the superficial spreading type had many mitoses, thick primaries and histologic recurrence 2 years following initial therapy?

3. Relations and Distributions: the relation between several attributes is determined and several output displays are provided. The types of displays include tabulated groupings, tabulated counts, and listings organized by specific items in the data base.
For example:

Tabulate patients with ssm by depth of involvement of the primary.
List the site of the primary, site of lymph node procedures, and number of positive nodes found for patients who had an elective lymph node dissection, ssm, and tumor thickness less than 1.5 mm.
How many patients had ssm invasive to level 4 ordered by site of the primary lesion?
Stratified by site of the primary lesion, how many folks had ssm followed by histologic recurrence within 1 year after initial therapy?
Which patients with high risk primaries had histologic recurrence within the first year?

4. Calculations: Computations are performed on selected sets of data and the results displayed. Calculations that may be performed include proportions, percentages, average, median, maximum, and minimum.
For example:

What is the minimum tumor thickness for patients with level 3 disease?
What is the average tumor thickness and age of patients with ssm organized by level of invasion?
Determine the percentage of people with ssm and thick tumors who had regression of the primary lesion.
What proportion of patients with ssm 3 had the primary in a pre-existing mole?

5. Yes-no: response to a question provides a yes-no answer and a count of the number of individuals satisfying a particular constraint.
For example:

Did any patients with ssm 5 have the following: regression of the primary, few mitoses, and histologic recurrence in the first year of follow up?

6. References to prior questions: request makes use of things mentioned in a previous question to provide limited recognition of pronoun references.
For example:

How many patients with ssm?
How many of them had level 5 disease?
List the age, sex, and tumor thickness for these people.
How many of these people had histologic recurrence within 24 months?
Who are they?

Thus, questions in classes 1-5 above may be used to designate a subset of the data base upon which subsequent questions may be posed.

Examples of Questions

Figure 1 contains a sample transcript of interactions with the system. Each question is input as shown. The system responds with 'ANALYZED' when the request has been processed and the data base is accessed to return an appropriate answer. These examples illustrate some of the sentences the system can process together with the kinds of responses that are currently provided. It also shows something of the process of inquiry itself, in particular, the way that successive queries can build directly on preceding ones.

Items 1-5 illustrate queries that elicit information about the data base, with an example of the spelling correction feature given in question 5. Time oriented requests are exemplified by questions 7 and 8. The use of pronouns is shown in question 9, where the pronoun 'these' refers to the subset of patients considered in the previous request.

FIGURE 1. SAMPLE INTERACTIONS WITH MELANOMA DATA BASE

```
1_how many patients in the melanoma data base?
ANALYZED

130

2_list the classes of attributes
ANALYZED

PERSONAL DATA
FAMILY HISTORY OF DISEASE
PATIENT BACKGROUND
PATIENT PHYSICAL EXAMINATION
HISTORY OF PRIMARY
PHYSICAL EXAMINATION OF PRIMARY
PATHOLOGY DATA
LYMPH NODE PROCEDURES
FOLLOW UP DATA

3_what are the pathology attributes?
ANALYZED

SURGICAL PATH NUMBER
TYPE
LEVEL
SITE
THICKNESS (MM.)
MITOSES
CELL TYPE
RISK FACTOR
CLONISM
REGRESSION
NUMBER OF LYMPHOCYTES
NUMBER OF MACROPHAGES
ANGIONEOGENESIS
PLASMA CELLS
NEVUS

4_show me the values for mitotic rate.
ANALYZED

MITOSES = MODERATE, FEW, MANY, NS, ABSENT, NIL

5_how is many mitoses defined?
        MANY <-spelling
ANALYZED

Many mitoses is defined as more than one mitoses per high power
field, or greater than 8 mitoses per square mm.

6_how many individuals are there in the study tabulated by type of
 melanoma?
ANALYZED


MM.TYPE = LMM
Number of items in this category = 2

MM.TYPE = NM
Number of items in this category = 23

MM.TYPE = SSM
Number of items in this category = 101

MM.TYPE = UNCLASSIFIED
Number of items in this category = 3

MM.TYPE = UNDETERMINED
Number of items in this category = 1


7_how many people with ssm had many mitoses and were followed for
24 months?

ANALYZED

4

8_12 months?
Trying Ellipsis:
          (HOW MANY) PEOPLE WITH SSM HAD MANY MITOSES AND WERE
FOLLOWED FOR 12 MONTHS

8

9_display the age, sex, and level for these patients ordered by
tumor thickness.

ANALYZED

THICKNESS = (  0.76 .   1.25)

        THICKNESS AGE
ID      (MM.)     (AT DIAGNOSIS) SEX    LEVEL
S-74-048   1.00   46             FEMALE 4
S-73-016   1.00   52             FEMALE 3
----------
THICKNESS = (  1.26 .   1.50)

        THICKNESS AGE
ID      (MM.)     (AT DIAGNOSIS) SEX    LEVEL
S-75-016   1.35   27             FEMALE 4
----------
THICKNESS = (  1.51 . 20)

        THICKNESS AGE
ID      (MM.)     (AT DIAGNOSIS) SEX    LEVEL
S-75-024   2.20   57             MALE   4
S-75-017   2.00   66             FEMALE 4
S-75-007   2.55   57             FEMALE 4
S-74-061   4.50   39             MALE   5
----------
THICKNESS = *OTHER-VALUES*

        THICKNESS AGE
ID      (MM.)     (AT DIAGNOSIS) SEX  LEVEL
S-73-027 NAVAIL   28             MALE 4
----------


10_how many of them had histologic recurrence?
ANALYZED

2

11_list the surgical pathology number for these patients.
ANALYZED

        SURGICAL
        PATH
ID      NUMBER
S-74-061 CS74-3609
S-75-007 CS75-2238

12_what proportion of patients with ssm and few mitoses who were
seen for 12 months had histologic recurrence?
ANALYZED

39 cases were accepted
0  cases were rejected due to missing or unspecified values.
PROPORTION IS   0.13

13_list those patients with histologic recurrence.
ANALYZED

ID = S-73-005, S-73-010, S-73-033, S-75-010, S-75-018

14_what is the average tumor thickness for people with ssm
organized by level of invasion?

ANALYZED

----------
MM.LEVEL = 2

THICKNESS
25 items accepted.
The following items were rejected in the averaging process: NS
NAVAIL NAVAIL NAVAIL NAVAIL NAVAIL NAVAIL NAVAIL NAVAIL NS NAVAIL
NAVAIL NAVAIL NS NAVAIL
Average is   0.48

----------
MM.LEVEL = 3

THICKNESS
26 items accepted.
The following items were rejected in the averaging process: NAVAIL
NAVAIL NIL NAVAIL NIL
Average is   1.06
```

FIGURE 1. (continued)

```
MM.LEVEL = 4

THICKNESS
15 items accepted.
The following items were rejected in the averaging process: NAVAIL
NAVAIL NAVAIL NAVAIL NAVAIL NAVAIL NAVAIL
Average is   1.90

----------
MM.LEVEL = 5

THICKNESS
5 items accepted.
The following items were rejected in the averaging process: NAVAIL
Average is   4.30

----------
MM.LEVEL = UNCLASSIFIED

THICKNESS
2 items accepted.
Average is   1.28


15_nodular melanoma
Trying Ellipsis:
                    WHAT IS THE AVERAGE (TUMOR THICKNESS) FOR
PEOPLE WITH (NODULAR MELANOMA) ORGANIZED BY (LEVEL OF INVASION)

----------
                    .
                    .
                [output omitted]
                    .
                    .

16_how many individuals had lymph node procedures?
ANALYZED

60

17_how many lymph node procedures?
ANALYZED

69


18_let (hrp1)
...be defined by (tumor thickness > 1.5 mm. and many mitoses]

ANALYZED


  TUMOR THICKNESS >   1.50 MM AND MANY MITOSES
(HRP1)

19_how many individuals with hrp1 were seen for 2 years?
  (HOW MANY INDIVIDUALS WITH TUMOR THICKNESS >   1.50 MM AND MANY
MITOSES WERE SEEN FOR 2 YEARS)

ANALYZED

2

20_one year
Trying Ellipsis:
          (HOW MANY) INDIVIDUALS WITH (TUMOR THICKNESS) >   1.50
MM AND MANY MITOSES WERE SEEN FOR ONE YEAR

15

21_what percentage of these people had histologic recurrence?
ANALYZED

15 cases were accepted
0  cases were rejected due to missing or unspecified values.
PERCENTAGE IS  60.00

22_how many of them died of disease?
ANALYZED

2
```

The processing of incomplete sentences (ellipsis) is shown in question 8, where the request is analyzed by matching the pattern '12 months' to a similar context of the previous request (also see question 15).

Items 4-13 constitute a possible dialogue that considers the hypothesis that 'the number of mitoses affects the outcome for patients with superficial spreading melanoma'. Similarly, questions 14 and 15 can be used to verify the relationship between level of invasion and tumor thickness for classification of patients with ssm and nodular melanoma.

Questions 16 and 17 show the capability of the system to distinguish between number of individuals with lymph node procedures and number of lymph node procedures.

Item 18 illustrates the use of the paraphrase facility to define the concept 'hrp1' (high risk primary), and to immediately test whether it is a good predictor variable of histologic recurrence (questions 19-22). Such definitions can be tailored to each individual investigator using the system and stored and accessed in subsequent sessions.

Note that in these examples punctuation as well as abbreviations are handled. Rules also can be written to process sentences that are not grammatical, such as, 'list people ssm 5 no regression'. A more detailed description of the language interface to the melanoma data base including extensive examples and a discussion of use of the system to aid in studying prognosis is in preparation [11]. The time to process a typical request, including access to the data base, is approximately 1-2 seconds of CPU time.

Experience with the System

The system has been demonstrated to physicians who are specialists in melanoma, other cancers, and a variety of other diseases. The reactions have been uniformly positive, even though the existing data base is not large enough to provide significant results. The ability to formulate queries in ordinary English and to have the results displayed immediately is extremely appealing. The facility for correcting spelling errors received special commendation.

The availabilty to physicians of on-line English language access has encouraged more active interest and participation by physicians in the data acquisition process, in completing missing values in the data base, and in actively perusing the information in the data base.

Future Plans

Future plans include stepwise enhancement of the access capabilities of the system as well as an expanded data base on melanoma and other cancers. It also is important to incorporate a knowledge base in the system to reflect physician defined higher order concepts and thus reflect the current state of medical understanding of the disease. A knowledge base also would contain rules for understanding the context of requests and developing appropriate prompts and clarification dialogues that show a greater degree of understanding of the medical domain. Such dialogues could be used for rephrasing questions to ensure that the question has been properly analyzed or to interrogate the user about the intent of the question being asked [6].

As the size of the data base increases, more complex statistical operations certainly are required. Statistical tests such as chi-square or t test could be incorporated into the system by interfacing to a statistical package such as the BMDP series [12].

English language access provides a way to study the use of the system by physicians. By analyzing and evaluating how interactions are performed, it should increase our understanding of how medical decisions are made both for patient management and clinical research. At the same time, changes to the system will be determined and modifications made. In particular, it is hoped that experience with such a system will aid in obtaining a better understanding of the variables that may serve as predictors of outcome for patients with melanoma.

Conclusions

The methods applied in this research provide an iterative approach for determining the classes of question that physicians desire to ask of a clinical data base. The classes of questions described are a core set of questions that can be augmented over time. It provides a convenient access to medical data bases to allow physicians to browse through their data. The system supports a large number of questions involving recall and hypotheses of interest can be studied. A requestor can follow a line of inquiry to study a particular hypothesis by entering a sequence of requests that depend on each other. Thus, groups of attributes can be aggregated and tested to determine whether they are good prognostic indicators.

LIFER provides a facility that allows a language interface builder to specify rules required to process English language inquiries. These rules are used to create and modify the grammar, lexicon and language interface functions. The fact that these interactions can be accomplished on-line and the results of new rules immediately tested is one of the attractions of this approach.

Acknowledgements

The assistance and support of G. G. Hendrix and J. Slocum (computer scientists) and M. S. Blois and R. W. Sagebiel (physicians) is greatly appreciated.

References

[1] Fries, J., A data bank for clinicians, N Engl J Med, vol 294, no. 25, June 17, 1976, pp. 1400-1402.

[2] Zuckerman, A.E., and Stenn, H.M., A general purpose report generator for the computer stored ambulatory record: A tool for use by physicians to monitor their practice, In Proc. First Annual Symposium on Computer Applications in Medical Care, IEEE Computer Society, 1977, pp 165-167.

[3] Feinstein, A.R., Rubenstein J.F., and Ramshaw W.A., Estimating prognosis with the aid of a conversational-mode computer program, Annals of Internal Medicine, vol. 76, no. 6, June 1972, pp. 911-921.

[4] Rosati, R.A., McNeer, J.F., Starmer, C.F. et al, A new information system for medical practice, Arch Intern Med, vol. 135, August 1975, pp 1017-1024.

[5] Hendrix, G.G., Sacerdoti E.D., Sagalowicz, D. and Slocum, J., Developing a natural language interface to complex data, ACM Transactions on Database Systems, vol. 3, no. 2, June, 1978, pp. 105-147.

[6] Codd, E.F., How about recently? (English dialog with relational data bases using RENDEZVOUS version 1), in Databases: Improving Usability and Responsiveness, B. Shneiderman(ed.), Academic Press, New York, 1978, pp 3-28.

[7] Petrick, S.R., On natural language based computer systems. IBM J. Res. Develop. vol. 20 no. 4, July 1976, pp. 314-325.

[8] Waltz, D.L., An English language question answering system for a large relational database. Comm. ACM, vol 21, no. 7, July 1978, pp 526-539.

[9] Teitelman, W., INTERLISP reference manual. Xerox PARC, Palo Alto, Calif. 1975.

[10] Hendrix, G.G., The LIFER Manual: A guide to building practical natural language interfaces Tech. Note 138, SRI Artificial Intelligence Center, Menlo Park, Calif., 1977.

[11] Epstein, M.N., Natural Language Access to Clinical Data Bases, Ph D. dissertation (in preparation), University of California, 1978.

[12] Dixon, W.J., and Brown, M.B. (eds.), BMDP-77: Biomedical Computer Programs, University of California Press, Berkeley, California, 1977.