

Computerized Language Processing: Implications for Health Care Evaluation

By NAOMI SAGER, Ph.D.
and MARGARET LYMAN, MD

The following is a talk delivered June 14, 1977 at the Eighth Annual Multidisciplinary Conference on Health Records sponsored by the Association for Health Records.

At New York University we are developing a computer technique for treating medical narrative. The computer program accepts as input the text of a medical document, without editing or coding, and converts the information in the document into a completely structured format, such as one might obtain from a detailed checklist covering the same material.

In this form the information in the documents can be searched and tabulated by computer. For example, narrative discharge summaries which have been processed by this program can be screened automatically for compliance with health care audit criteria, such as those written following the Performance Evaluation Procedure (PEP) of the Joint Committee on Hospital Accreditation. An example of this application is described below.

The assumption in this work is that it is going to be increasingly feasible, and possibly necessary, to keep some portion of the medical record in computer readable form. For health care evaluation and other purposes, the hospital discharge summary is a likely candidate and so the programs are being tested on this type of document. In addition,

the narrative in a discharge summary is sufficiently complex so that if the technique works on this type of document it should also readily apply to simpler material, such as reports of clinic visits.

Although it is more common to enter data into the computer in coded form, based on a checklist, there are several reasons why it is desirable to treat medical information in its narrative form.

First: Unless the checklist is prohibitively long, pre-coding inevitably loses some of the information, whereas the technique described here preserves virtually all the information in the document. Thus, *no information is lost*.

Second: By capturing all the information rather than a selection needed for a particular purpose, the information is available for a variety of purposes: health care proper, health care evaluation, clinical research, and teaching. That is, *the information can be re-used*.

Third: By not forcing the physician to cram his notations into preset categories we not only avoid distortion and reduction of the information, we avoid turning the physician into a high class coder. In other words, *the physician does what he is trained to do—observe, evaluate, treat—not code*.

The purpose of the program, then, is to convert the narrative of the hospital document into a structure which is equivalent to a pre-coded format for the same information, and to do this automatically, without imposing the structure on the person who is doing the medical reporting. That person should be able to report in a completely natural way, in English, without even knowing the detailed structure into which the computer will arrange the information.

But how is this possible? Where does the structure come from? And how can a document be automatically converted to a structured form without distorting the information? The answer is that the information structure is really already in the narrative, and our task is only to expose it, to strip away the stylistic and grammatical variations, and to align into columns the words which are carrying



NAOMI SAGER, Ph.D., is Senior Research Scientist, Adjunct Professor of Linguistics, and Principal Investigator of the National Library of Medicine grant on Computer Structuring of Medical Narrative at New York University.

MARGARET LYMAN, MD, is an Associate Professor of Clinical Pediatrics at New York University.

the same types of information.

The information in medical documents is actually of limited types, very similar in kind (though not in factual detail) from document to document. It appears in different linguistic forms due to the many alternative ways of expressing the same content, which language provides.

Corresponding to the limited content, the vocabulary is limited, and only certain combinations of the words occur in particular grammatical relations. For example, one may find "The patient was

admitted into Bellevue," but never "Bellevue was admitted into the patient."

On the basis of a study of which nouns occur as the subject or object of which verbs we discover word classes specific to the medical material. Thus *patient* is in a different noun class from *Bellevue*, and *admitted* is in a verb class that connects words like *Bellevue* (the medical institution class) to patient-words.

In terms of these classes we find patterns of word-class co-occurrence in the sentences of the

Figure 1

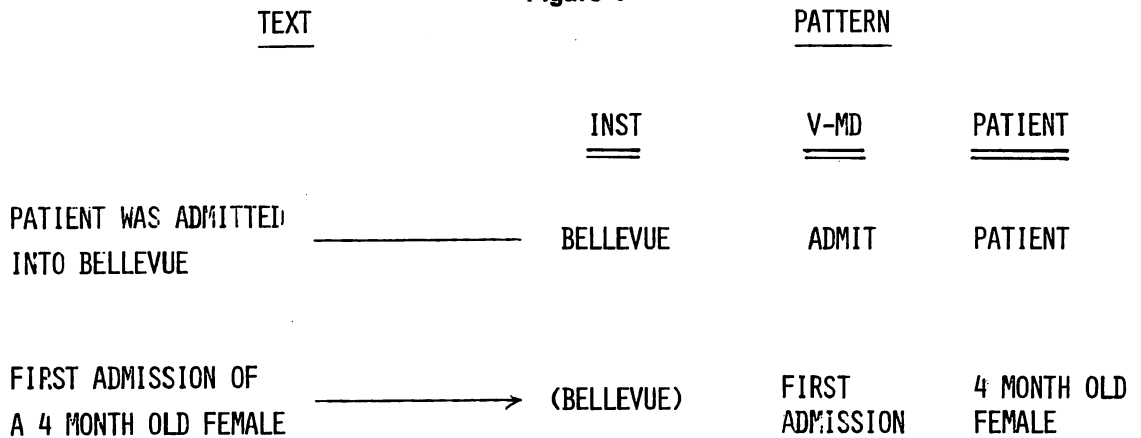


Figure 2

(REASON FOR ADMISSION) - COLD; HIGH FEVER; VOMITING.

(PERTINENT HISTORY)

PRESENT ILLNESS - FIRST ADMISSION OF A 4 MONTH OLD FEMALE. 5 DAYS BEFORE ADMISSION PATIENT DEVELOPED MILD COLD. ONE DAY BEFORE ADMISSION HIGH FEVER WAS FIRST NOTED. FEVER HAS PERSISTED.

FORMAT

PARAGRAPH	DATA												
	PATIENT	TREATMENT V-MD	T-P CON	PATIENT STATUS					TIME			MODS EVIDENTIAL	
				V-PT	FINDING				EVENT-TIME	TENSE	ASPECT		
				BODY MEAS.	NORM- ALCY	QUANT	QUAL S/S						
1	REASON FOR ADMISSION							COLD	(ADMISSION)				
2								HIGH FEVER	(ADMISSION)				
3								VOMITING	(ADMISSION)				
4	PERTINENT HISTORY	4 MONTH OLD FEMALE	FIRST ADMISSION										
5		PATIENT		DEVELOP				MILD COLD	5 DAYS BEFORE ADMISSION	PAST			
6								HIGH FEVER	1 DAY BEFORE ADMISSION	PAST PASS- IVE			FIRST NOTE
7								FEVER		PRES. PERF.	PER- SIST		

documents, which correlate with the information in the sentences. Thus, as illustrated in Figure 1, "The patient was admitted into Bellevue" and "First admission of a 4-month-old female" are both instances of a particular word-class co-occurrence pattern.

This pattern consists of the noun class INST, which contains words referring to medical institutional personnel, followed by the verb class V-MD, which contains the medical action verbs, followed by the noun class PATIENT, containing words which refer to the patient.

Note in Figure 1 that some words are present implicitly and can be filled in. Thus, in a Bellevue document the omitted INST word is *Bellevue*, shown in parentheses. Also, different forms of the same root word (*admit*, *admission*) are treated alike.

The overall pattern for sentences of the documents is called an "information format." Figure 2 shows the computer output for the first few sentences of a typical pediatrics discharge summary.

The FORMAT has a column for the PARAGRAPH heading and many columns and subcolumns, of which only a few are shown here, for the DATA. There is a column for references to the PATIENT, and under TREATMENT a column for medical action verbs V-MD. Also under TREATMENT there is a column (not shown) for MEDICATION, with slots for dosage, etc. There is place for a connective between TREATMENT and PATIENT STATUS, as in "admitted for meningitis."

Under PATIENT STATUS, there is a column for a patient verb, such as *complain of*, *have*, and a number of columns for FINDINGS. These include a column for BODY MEASUREMENTS, such as temperature, pulse, etc., a column called NORMALCY for the word *normal* and words or word-parts indicating a return to the normal state; QUANT for numbers and words of quantity; QUAL for qualitative findings such as SIGNS OR SYMPTOMS, written S/S.

Every medical fact or event has a TIME associated with it, primarily an EVENT TIME but sometimes also aspectual information, like duration or frequency. Language also allows for modifiers (MODS), for example, EVIDENTIAL, modifiers stating the amount of certainty, or the source of the information (e.g., "noted by mother"), and importantly, whether the finding is negated or not—for example, instead of "infection": "no sign of

infection."

The first sentence of the discharge summary was formatted by the computer as shown in line one of the format in Figure 2. REASON FOR ADMISSION went into the PARAGRAPH column, *cold* into the SIGN/SYMPTOM column, *fever* also into the SIGN/SYMPTOM column, with *high* going into the QUANT column, and *vomiting* into the SIGN/SYMPTOM column. The EVENT TIME for these entries was *admission* which is shown in parentheses because it was filled in by the computer at a later stage of regularization, in this case by reference to the paragraph heading REASON FOR ADMISSION.

Continuing the narrative of this discharge summary, shown at the top of Figure 2, we read in the paragraph PERTINENT HISTORY—PRESENT ILLNESS: "First admission of a 4-month-old female." The phrase "4-month-old female" was put under PATIENT, and "first admission" under V-MD. The next sentence is "5 days before admission patient developed mild cold." The computer program put *patient* into the PATIENT Column, *develop* into the patient verb column, (V-PT), *mild* into the QUANT column, *cold* into the SIGN/SYMPTOM column, and *5 days before admission* into the EVENT TIME column.

The next sentence: "One day before admission high fever was first noted" was formatted by the computer (line 6 in Figure 2) in a similar manner: *1 day before admission* is the EVENT TIME, the SIGN/SYMPTOM entry is *fever*, with QUANT equal to *high*. Here there is an EVIDENTIAL modifier, *noted*, the past, passive form of *note*. We retain tense information because it is sometimes needed in calculating the EVENT-TIME if a date or time is not specified.

In the last sentence in Figure 2, "Fever has persisted," the verb *persist* appears under TIME:ASPECT indicating that it constitutes descriptive time information about the symptom or medical event on the same format line.

The computer program for information-formatting is applied to each successive sentence of the input document. To date, the program is being applied on an experimental basis to a selection of narrative documents which are produced in machine readable form on a routine basis in the Department of Pediatrics of Bellevue Hospital.

In conjunction with a comprehensive health care project for children and youth at Bellevue Hospital, a computer-based narrative medical record

Figure 3

(PDS)

IDENTIFICATION NO -
BELLEVUE HOSPITAL - PEDIATRIC DISCHARGE SUMMARY
NAME - SEX - F
DATE OF ADMISSION - 05/29/71 DATE OF DISCHARGE - 06/10/71
LOCATION - 8-WEST
BIRTH DATE - 01/03/71

DOCUMENT NUMBER 393569

(REFERRING PHYSICIAN) - NONE GIVEN.

(REASON FOR ADMISSION) - COLD; HIGH FEVER; VOMITING.

(PERTINENT HISTORY)

PRESENT ILLNESS - FIRST ADMISSION OF A 4 MONTH OLD FEMALE. 5 DAYS BEFORE ADMISSION PATIENT DEVELOPED MILD COLD. ONE DAY BEFORE ADMISSION HIGH FEVER WAS FIRST NOTED. FEVER HAS PERSISTED. SLEEPINESS NOTED BY MOTHER. SEEN IN PES AND ADMISSION ADVISED. SIGNIFICANT PAST HISTORY - BORN AT ROOSEVELT HOSPITAL. PATIENT IS THE PRODUCT OF A GRAVIDA 10, PARA 9, NORMAL UNCOMPLICATED PREGNANCY. DELIVERY WAS SPONTANEOUS. NO NEONATAL COMPLICATIONS. BIRTH WEIGHT WAS 8 LBS 9 OUNCES. INITIAL FAMILY HISTORY WAS NEGATIVE. DURING THIS ADMISSION IT WAS LEARNED THAT 1 SON HAD SICKLE CELL ENEMIA AND REQUIRED FREQUENT TRANSFUSIONS. HE IS NOW DEAD FROM CAR ACCIDENT. GROWTH AND DEVELOPMENT OF PATIENT WAS NORMAL.

(EXAMINATION ON ADMISSION) - TMP 102, PU 175, RR 75, WEIGHT 15 LBS. ANTERIOR FONTANELLE WAS SLIGHTLY BULGING. THROAT WAS HYPEREMIC; NO EXUDATE. NECK WAS RIGID AND STIFF. CHEST SHOWED SLIGHT INTERCOSTAL RETRACTIONS. LUNGS REVEALED BILATERAL RHONCHI; NO RALES HEARD. ABDOMEN SHOWED NO ORGANOMEGALY; UMBILICAL HERNIA PRESENT. EXTREMITIES WERE SLIGHTLY HYPERTONIC. NEUROLOGICAL EXAM SHOWED POSITIVE KERNIGS, POSITIVE BRUDZINSKI'S, AND POSITIVE BABINSKI'S SIGNS; DTR'S WERE NORMAL.

(IMPRESSION ON ADMISSION) - MENINGITIS.

(COURSE IN HOSPITAL) - IMPRESSION OF MENINGITIS WAS CONFIRMED BY FINDING CLOUDY CEREBROSPINAL FLUID. TREATMENT WAS AMPICILLIN 200 MG/KG I.V. WAS GIVEN. A TRANSFUSION OF 100 CC PACKED RED BLOOD CELLS WAS GIVEN FOR EXTREME ANEMIA. STUDY OF PRE-TRANSFUSION SPECIMEN REVEALED POSITIVE SICKLE CELL PREPARATION. AFTER COMPLETION OF 4 HOUR TRANSFUSION, PATIENT LAPSED INTO COMA. SHE DEVELOPED SEIZURES AND WAS GIVEN PHENOBARBITAL AND VALIUM. GOOD CONTROL OF SEIZURES WAS OBTAINED. THERE WAS MARKED CLINICAL IMPROVEMENT. PERSISTENT NECK STIFFNESS AND LOW GRADE TEMP FINALLY CLEARED. HEMOGLOBIN REMAINED STABLE AFTER THE FIRST TRANSFUSION.

(STATUS AT DISCHARGE) - EATING WELL. AFEBRILE. EXAMINATION NORMAL.

(LABORATORY DATA)

HEMATOLOGY - AT ADMISSION WBC 17,000, WITH 31% POLYS, 63% LYMPHS. HCT 16, HG 5. POST TRANSFUSION TO DISCHARGE, HCT 27 TO 34.
URINE - AT ADMISSION, PROTEIN 1+ AND CASTS PRESENT. AFTER HYDRATION, URINE ALWAYS NORMAL.

Figure 4

**MENINGITIS &/or SEPTICEMIA
IN SICKLE CELL DISEASE**

- ✓ = found in PDS
- X = Found in complete record
- O = item not present in patient
- ∅ = item not in PDS or in record

AUDIT CRITERIA DATA

ELEMENTS

STD % MEETS VAR ND
100.0 E₁ E₂ CM VAR CM

Numbers = Elements
Letters = Exceptions

- Diagnosis of Meningitis
- 1. Positive CSF culture (or A + B) 100%
 - A. Admission history contains all of ff.:
 - (1) Fever
 - (2) Stiff neck
 - (3) Vomiting or headache
 - B. First CSF shows 2 of following:
 - (1) Positive smear
 - (2) WBC greater than 10/cmm.
 - (3) Glucose less than 30 mg%
 - (4) Protein greater than 40 mg%
- Diagnosis of Septicemia
- 2. Positive blood culture 100%
- Diagnosis of Sickle Cell Disease
- 3. One of following: 100%
 - (1) Positive sickle cell preparation
 - (2) Hemoglobin electrophoresis = HgS'
 - (3) Statement in history "known sickler" or equivalent
- Special Procedures
- 4. Cultures of CSF and blood at admission 100%
- 5. Tuberculin test if not done in past year "
- 6. Monitoring of hematocrit or hemoglobins or "
- A. Hematocrit or hemoglobin stable
- 7. Chest X-ray taken 100%
- Discharge Status
- 8. Afebrile 100%
- 9. Amelioration of symptoms at admission 100%
- 10. Plan for on-going care 100%
- 11. Length of Stay - 10-20 days 100%
 - A. Age at admission less than 4 weeks
 - B. More than 1 transfusion required
 - C. Transferred or signed out or died
 - D. Complications of meningitis
- 12. Mortality 0%
- Complications of Meningitis
- 13. Subdural effusion 0%
 - CM = Aspiration or drainage
- 14. Persistent Seizures (lasting over time) 0%
 - CM1 = EEG ordered
 - CM2 = Anticonvulsant therapy

CM = Critical Management.

system was developed which allows capture, storage and retrieval of all types of health care data. A wide variety of uses, both for patient care and administrative purposes have been made of the data.

Among the many different types of medical reports captured for daily patient care is the Discharge Summary. The discharging physician writes or dictates the summary in accordance with a standard outline which provides paragraph headings for the narrative summary. A medical typist then keys-in the written or dictated report on an interactive terminal with computer-prompting of the paragraph headings.

Processing and retrieval make the printed discharge summary available for future outpatient health care, notification to referring physicians, and other purposes, including analysis by computer techniques, as described here. The first page of a typical discharge summary in this system is shown in Figure 3. The output-format for the first few sentences of this document were shown in Figure 2. This document will serve to illustrate how a computer program, operating on computer-formatted narrative discharge summaries, can screen cases for compliance with health care audit criteria.

Consider the criteria for meningitis/septicemia in sickle cell disease, specified on a PEP Audit Criteria Data Sheet, as shown in Figure 4. The computer program applies these criteria one by one to the formatted discharge summary and determines whether an element is present, and if not, whether the specified variations were documented. Each criterion is realized as a small subprogram which searches the appropriate format columns for the required type of information and performs calculations where needed.

To illustrate, consider just the items mentioning fever in the PEP data sheet of Figure 4. Once the narrative document is formatted, applying analysis procedures by a computer program is straightforward. This can be demonstrated with reference to the output lines (Figure 5), obtained by the formatting program for the sentences that pertained to fever in the above discharge summary. (In Fig. 5, the TREATMENT area does not appear because the fever items only involve PATIENT STATUS and the program does not print the column heading when the column entry is empty. The column EVENT-TIME is shown in greater detail than previously.)

First, to explain the format lines in Figure 5. The first three lines were already seen in Figure

Figure 5

FEVER - FORMATTED NARRATIVE

(REASON FOR ADMISSION) - ... HIGH FEVER;
 (PERTINENT HISTORY) - ... ONE DAY BEFORE ADMISSION HIGH FEVER WAS FIRST NOTED. FEVER HAS PERSISTED
 (EXAMINATION ON ADMISSION) - TMP 102, ...
 (COURSE IN HOSPITAL) - ... LOW GRADE TEMP FINALLY CLEARED, ...
 (STATUS AT DISCHARGE) - ... AFEBRILE. ...

FORMAT

PARAGRAPH	DATA				TIME					V-TENSE	CHANGE	ASPECT	MODS EVIDENTIAL
	FINDING				EVENT-TIME								
	BODY MEAS	NORM-ALCY	QUANT	QUAL S/S	NUM	TIME UNIT	T-PREP	REF POINT					
1	REASON FOR ADMISSION			HIGH	FEVER					(ADMISSION)			
2	PERTINENT HISTORY			HIGH	FEVER	ONE	DAY	BEFORE		(ADMISSION)	PAST PASSIVE		FIRST NOTE
3					FEVER						PRESENT PERFECT		PERSIST
4	EXAMINATION ON ADMISSION	TMP		102						(ADMISSION)			
5	COURSE IN HOSPITAL	TEMP	[GD]	LOW GRADE				FINALLY			PAST	CLEAR [GD]	
6	STATUS AT DISCHARGE				FEBRILE					(DISCHARGE)			[NEG-PREFIX]

2. The fourth line is straightforward. With regard to line 5, the sentence "Low grade temperature finally cleared," though only five words, is actually a rather complicated sentence for automatic formatting.

Low grade temperature formats easily with *temperature* going to BODY-MEASURE and *low grade* to QUANT under FINDING. However, *finally* is a time adverb which refers both to a previous time and also to the present but leaves both times unstated. *Cleared* is a verb formed from the adjective *clear*. As a verb it indicates positive change, and as such it is formatted into two columns: the word *clear* under CHANGE, and a marker (GD) under NORMALCY.

Format line 6, for *afebrile* in the paragraph STATUS AT DISCHARGE, is interesting because the word *afebrile* is analyzed into two parts: its negative prefix *a-* (represented by a marker in the MODS column) and the remainder-*febrile*, appearing in the SIGN/SYMPTOM column. Thus, *afebrile* is treated similarly to *no fever*.

The EVENT-TIME in line 6 is DISCHARGE, given by the paragraph heading and filled-in during the stage of computer-regularization of the format entries.

Fever appears in two places in the PEP data sheet of Figure 4, under DIAGNOSIS OF MENINGITIS and under DISCHARGE STATUS. The audit criterion for a diagnosis of meningitis is a positive CSF culture. However, exceptions A + B are allowed. Suppose the patient had had antibiotics the day previous and the culture grew out nothing. Then in applying the audit criteria the program would have to find in the admission history, evidence of fever, as well as the other symptoms noted under A and B.

Procedurally this means that the words *fever*, or an exact synonym, must appear in the document or else a temperature greater or equal to 100.2 (for Pediatrics); and that the time given for that observation should be at admission or within a certain time up to 48 hours of admission. Similarly, *Afebrile at Discharge* would be established by finding the negation of these findings associated with the time of discharge.

The program logic for establishing the PEP criterion 1.A(1) of Figure 4, *Fever in Admission History*, can be summarized as follows:

PEP 1.A(1) *Fever in Admission History*

In DATA columns

Either S/S = *fever/febrile*; MODS \neq negation

Or both BODY-MEASURE = *Temperature/Temp/TMP*
And either QUANT = $n \geq 100.2$
or QUANT = *word* (e.g. *high, low grade*).

In TIME Columns

Both REF-POINT = *Admission* or *a* (Date of Admission)

And Either EVENT-TIME (other columns) = \emptyset
Or calculate Event Time *t*; *t* should be within 48 hours of *a*.

Here, "MODS \neq negation" stands for a subprogram which checks for absence of negation.

As applied to the output lines shown in Figure 5, the program FIRST scans the S/S column for *fever* or *febrile*. It finds such an entry in line 1: *fever*. Then it checks by a subprogram that the MODS do not imply negation, e.g. *No fever, no evidence of fever, fever not observed*. Line 1 has no MODS and therefore no MODS equal to negation, so the first test is passed and the program skips to the TIME entries in format line 1.

To meet the criterion, the reference point should be the word *admission* or an actual date which checks with the date of admission. Line 1 has the word *admission* as its REF-POINT entry. But the program must also check that *admission* is not occurring as a reference point for some much earlier time, e.g., "1 week before admission patient had a high fever but it cleared." Since line 1 has no modifiers of *admission*, this test is passed and it is established that *Fever in Admission History* has been documented.

The fact that the information is present in several places enables us to check for consistency. This is the case for lines 2 and 4 of Figure 5, which also pass the test for *Fever in Admission History*. With regard to line 2, the S/S column contains *fever*; the MODS Column is not empty as in line 1, but its contents have no *negative* marker. Under TIME, the REF-POINT is *admission* and a calculation by a subprogram establishes that "1 day before admission" is within 48 hours of *admission*.

Line 4 is another verification of fever at admission. Here the SIGN/SYMPTOM column is empty but BODY MEASURE contains *TMP*, and QUANT contains a number $102 \geq 100.2$. Under TIME, the REF-POINT is *admission* and the other columns are empty. Thus, line 4 also passes the test for fever at admission. The calculations for *afebrile at discharge* are very similar.

With the development of these programs it is possible to envision completely automated screen-

Figure 6

CHECKLIST FOR RECORDING ADMISSION HISTORY — FEVER

ADMISSION HISTORY

FEVER

PRIOR TO ADMISSION DAY: YES ___ NO ___

- IF YES: A) NUMBER OF DAYS _____
- B) HIGHEST TEMPERATURE IF RECORDED _____
OR ESTIMATE OF FEVER: HIGH ___
MODERATE ___ LOW GRADE ___
- C) DID FEVER PERSIST: YES ___ NO ___

ON ADMISSION DAY: YES ___ NO ___

- IF YES: A) HIGHEST TEMPERATURE IF RECORDED _____
OR ESTIMATE OF FEVER: HIGH ___
MODERATE ___ LOW GRADE ___

ing of narrative discharge summaries for conformity to health care audit criteria. The documents would first be formatted by the language programs. Then the audit criteria would be automatically applied. This would save human time for cases which do not meet the criteria in a straightforward fashion. This would also allow health care patterns to be established by summarization over many evaluations.

Perhaps just as important is a point mentioned earlier. Because the language-based formatting preserves all the information, the computerized data can be used for a variety of purposes. This is not the case when the input to the computer is based on checklists, because to get complete information the checklists would simply be too long.

For example, consider the checklist (Figure 6) that would be required just to get the equivalent of the first four lines regarding fever in the computer-formatted Discharge Summary. All this coded information was obtained automatically from the free narrative.

From the formats, in addition to screening for health care evaluation, statistical summaries and

comparisons can be generated. Suppose as part of the evaluation and education activities of a hospital, it is desired to review how the hospital handled all cases of bacterial meningitis.

Questions could be answered by computer from the formats—such as: Of the cases that stayed longer than two weeks, what were their problems? (DIAGNOSIS and SIGN/SYMPTOM columns)? How many of all meningitis patients were on antibiotics? How many days on IV vs. oral? [MEDICATION column under TREATMENT (not shown in Figures)].

The hospital's performance could then be compared with the performance at other institutions via such sources as *Medical Clinics of North America*. This technique could help to identify patient management problems within an institution.

It also could be used in clinical research, making it possible to review old data in light of new criteria, and to keep abreast of incoming data by having reports monitored for completeness while the data sources are still available to supply missing information. ■