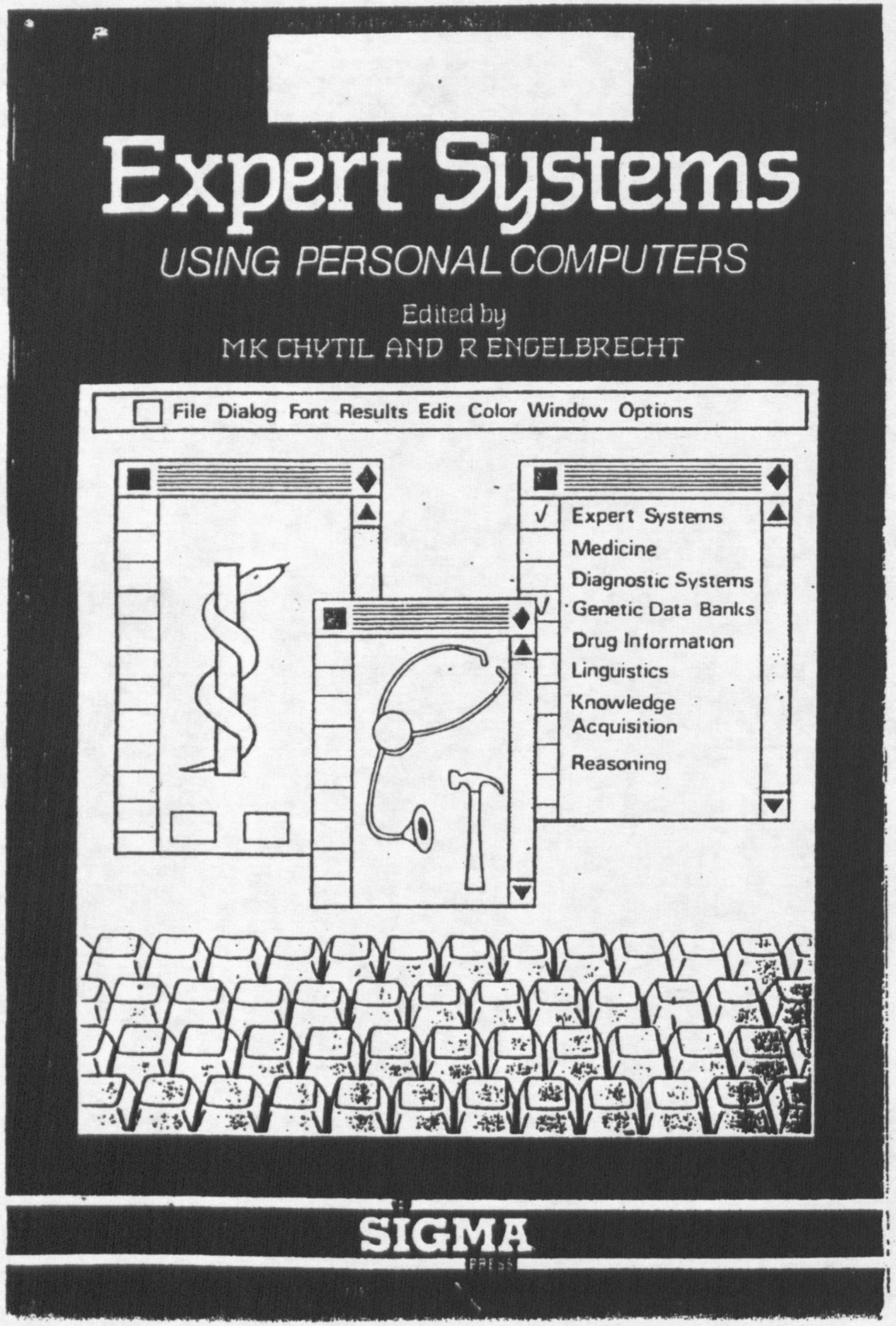

Medical Knowledge Acquisition for Expert Consultation Systems

PETER STANCHEV

ABSTRACT

This paper presents a method to acquire medical knowledge for an expert consultation system by means of a specially developed subsystem of a hospital information system. This knowledge acquisition subsystem captures information from clinical cases, confirmed to a different degree, in chosen domains.

A model of such a subsystem is described in the paper. Among its modules are the following:

- numerical data processing using statistical methods;
- text data processing operations of fuzzy values [1];
- graphical data processing using pattern recognition techniques.

The model is developed in Pascal for IBM PC XT.
An example illustrating how the model functions is given.

Keywords: Knowledge acquisition, knowledge data bases, diagnosis, health information systems, expert consultation systems, fuzzy set theory, statistics, pattern recognition, personal computer, PASCAL.

Introduction

The acquisition of initial knowledge is one of the most important tasks when developing systems that should attain an expert level performance in automated problem solving systems. It is seldom discussed in the literature concerning such systems. The main reason is that:

Knowledge acquisition is at a very early stage of development, where different experiences are still being gathered, and general principals have not emerged [2].

The idea we propose in this paper is the following: The acquisition of domain specific knowledge for a medical expert consultation system (ECS) results from the functioning of a special Knowledge Acquisition Subsystem of a Hospital Information System (KAS of HIS) during a sufficient period. The KAS is supposed to function simultaneously with other subsystems of HIS, such as the Patient Administration Subsystem, Clinical Laboratory Subsystem and other subsystems related to health care. The main goal of KAS is to capture and to present information from real clinical cases

With the help of such a subsystem knowledge bases could be created and the existing ones updated, but always by the decisive participation of a human expert Figure 1.

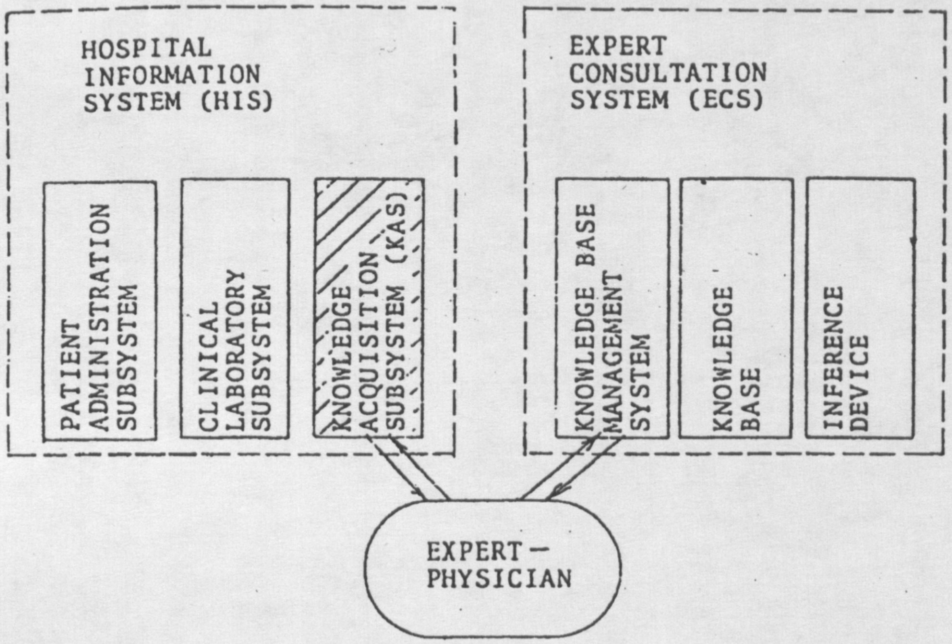

Figure 1. Knowledge acquisition for medical expert consultation system by KAS of HIS

Description of the KAS

The purpose of the system is to explore, capture and process the necessary medical data in an optional way. In individual applications this kernel can be combined with the corresponding information system which is currently used in the given medical centre.

The data in the model are of the following types:

a) fixed values - an integer or real numeric value, a linguistic or a graphic value;
b) interval values - a pair of fixed values (bounds of a closed interval);
c) vector values - a consequence of fixed values.

The data are distributed in two types of semantic records with an equal structure:

1) Records with patient's data - contain results from subjective and objective examination of patients, questioning, physical check-up, tests, messages about the progress of the disease or about the confirmed diagnosis, confidence degree, etc. All these data will be referred to as diagnostic parameters. Every diagnostic parameter is defined by name, type (the same as the type of its values - (a), (b) or (c)), coding scale

2) Records with the medical background knowledge - contain results accumulated from recently solved and confirmed (to a different degree) clinical cases which concern a certain disease. These records are identifiable by the name of the disease.

The files in the model include records from one type only, i.e. (1) or (2). The information in the files is structured by the matrices (patient matrix or knowledge matrix):

	dp_1	dp_2 ---	dp_m
obj_1	v_{11}	v_{12}	v_{1m}
obj_2	v_{21}	v_{22}	v_{2m}
|			
obj_n	v_{n1}	v_{n2}	v_{nm}

where:

$dp_1,...,dp_m$ means diagnostic parameters;
$obj_1,...,obj_n$ means names of patients or names of diseases;
$v_{ij}(i=1,...,n;\ j=1,...,m)$ means the value of the j-th diagnostic parameter for the i-th patient from the population or the value of the j-th diagnostic parameter for the i-th disease.

In the physical organization of the KAS the direct access method is used (every row of the matrix corresponds to a record). v_{ij} values occupy one or several fields with different length according to the parameter type. In the case of the graphical data, instead of a v_{ij} value, a label of the area where the graphic and its description are stored is written in the physical record.

The interaction with the KAS is performed by commands. Every command stimulates a set of questions with the help of which the parameters are entered.

The commands, their parameters and functions are as follows:

File operations

ADD	<file specification> <file specification> [<file specification> ...] ;	concatenates files
DELF	<file specification > ;	deletes a file

Data operations

CREATE	<file specification> <number diag param> <dp_1 name> <dp_1 type> [,<dp_2 name><dp_2 type>...] <max number obj>;	creates a file
DELDP	<file specification> dp_j name> ;	deletes the value of a diagnostic parameter

DELOBJ	<file specification> obj$_i$ name > ;	deletes a record
INSERT	<file specification> (<obj$_{i_1}$ name> <v$_{i_1}$ 1> [,<v$_{i_1}$2>...]) [(<obj$_{i_2}$ name> <v$_{i_2}$ 1> ,[v$_{i_2}$2>...]) ...] ;	adds or updates records
UPDATE	<file specification> <obj$_i$ name><dp$_j$ name> <v$_{ij}$> ;	sets a new diagnostic parameter value in a record

Data selecting

SELECT	<file specification> {<number> / <list>} <logical condition> ;	selects a record according to a logical condition

Content review

DIRECTORY ;		shows the file names, the number of diagnostic parameters, the number of records in the files
TYPE	<file specification> [<dp$_{j_1}$ name> [,dp$_{j_2}$ name> ...]] [<obj$_{i_1}$ name> [,<obj$_{i_2}$ name> ...]];	explicites the whole record or diagnostic parameter values

Knowledge acquisition

STATISTIC	<analysis type> ;	performs a statistic procedure
KNOWLEDGE	<file specification> <obj$_i$ name> <file specification> <logical condition>;	accumulates data from all records satisfying a logical condition and inserts the result into the file with medical knowledge
KNOWLEDGE	<file specification> obj$_i$ name> <file specification> <dp$_j$ name> <<logical conditions>;	accumulates data from all records satisfying a logical condition weighted by confidence degree and inserts the result into the medical knowledge file

Here are some clarifications to the given notation:

1)

<file specification>	-	the file name;
<number>	-	a number is to be typed;
<list>		a list is to be typed;
<max number obj>	-	maximal numbers of patients or of diseases;
<logical condition>	-	is (<arg> <log oper>) [,(<arg><log oper>)...], where:
<arg>	-	is {<dp$_j$> / <obj$_i$>} <comparison oper> {numerical constant / alphanumeric constant};
comparison oper>	-	.LT., .GT., .LE., .GE.,.EQ., .NE., .BE., NB., where: BE. - belongs to the interval. NB. - does not belong to the interval;
<log oper>	-	.AND., .OR., .NOT., .END., where; .END. - ends a logical expression;
<analysis type>	-	is {<statistics> / <crosstabulation> / <binary correlation> / <partial correlation>; / <multiple regression> / <factor>;}
<statistics>	-	gives mean, standard error, standard deviation, variance kurtosis, skewness, range, minimum, maximum;
<crosstabulation>	-	displays table of relationships between two or more diagnostic parameters;
<binary correlation>	-	gives Pearson and rank-order correlation;
<partial correlation>	-	calculates partial correlation;
<multiple regression>	-	performs multiple regression analysis;
<factor>	-	performs factor analysis;

2) When the data are replaced from one file to another, their type is converted to the type of the corresponding data in the second file, if it is possible;
3) Some commands dealing with linguistic and graphic data gather additional information with the help of questions not described here.

As far as the command KNOWLEDGE (WKNOWLEDGE) is the key point of the approach, it will be discussed in more detail. This command can be used for obtaining knowledge from the accumulated patients' data presented numerically, linguistically or graphically. The following technique is applied: all rows in the patient matrix corresponding to the same disease are processed together and a new row is obtained and added to the knowledge matrix. This row is easily converted to a new production rule of the knowledge base.

For numeric data the mean obtained by (psi) function for Hampel estimate [3] is applied.

For graphic data, pattern recognition methods are used. For every graphic a number of parameters characterizing the graphics are entered by the user or automatically calculated through user-given rules (e.g. maximum, minimum, number of peaks, frequency, etc.) According to the values of these parameters the graphics obtained as values of a graphic diagnostic parameter are attached to one of the user defined classes (e.g. classes of cardiograms corresponding to the different heart pathologies). The mean is obtained as the class with the maximum graphics included.

For linguistic data a fuzzy-theory approach is applied. Linguistic data are converted to fuzzy values [1], and after some operations the result is converted back to the linguistic form. The fuzzy values are represented as 4-tuples $n=(a,b,\alpha,\beta)$, where a,b is the closed interval in which the membership function is equal to 1.0, α is the left bound and β the right bound. Figure 2 illustrates this characterization.

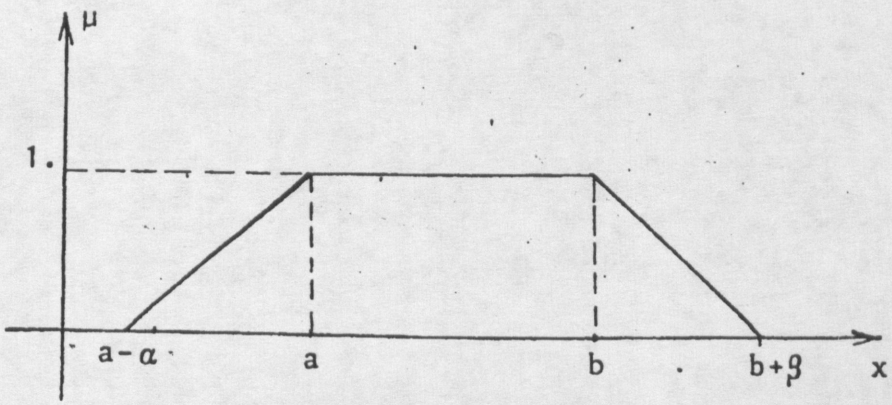

Figure 2. Characterization of fuzzy value $n=(a,b,\alpha,\beta)$

A basic set of fuzzy values is used to represent the range of the linguistic values. All linguistic values are approximated with fuzzy values from the basic set. The mean of the linguistic values is formed as the mean of the correspondent fuzzy values (the weighted mean - as the weighted mean of the correspondent fuzzy values). The operations: addition, multiplication, division of two fuzzy values are defined as follows:

$m=(a,b,\alpha,\beta)$ and $n=(c,d,\gamma,\delta)$

$m+n=(a+c,b+d,\alpha+\gamma,\beta+\delta$

$m*n=(ac,bd,a\gamma+c\alpha-\alpha<ga>,b\partial+d\beta+b\delta)$ if $a,b,\alpha,\beta,c,d,\gamma,\delta>0$.

After obtaining the resulting fuzzy values, the nearest fuzzy value from the basic set is to be found and interpreted, having in mind the following order among fuzzy values:

$$\text{if} \quad m=(a,b,\alpha,\beta) \quad \text{and } n=(c,d,\gamma,\delta)\text{, then}$$

$$m<n \text{ if } \begin{cases} b<d & \text{or} \\ b=d \text{ and } \beta<\delta & \text{or} \\ b=d \text{ and } \beta=\delta & \text{and } a<c \text{ or} \\ b=d \text{ and } \beta=\delta & \text{and } a=c \text{ and } \alpha<\delta \end{cases}$$

In all the described cases knowledge is obtained by considering all the cases that concern the disease with the new one, not the last obtained mean with the new case.

Implementation of KAS for IBM PC XT

The system is developed in Pascal. The first version of the system was developed just to explore the autonomous functioning of such an information system [4], [5]. At present the system functioning simultaneously with d-Base III is being experimented. The graphical data processing part has not really been applied because the hardware system is not equipped with visual input devices.

Example

In the Appendix an example is given in which:

1) A file named 'patients' is created which includes patients' data for the following diagnostic parameters : age, temperature, leucocytosis, exaggeration of the liver, diagnosis, confirmity.
2) Five records with patients' data are entered in the file 'patients'
3) A file named 'knowledge' is created which includes medical knowledge data for the following diagnostic parameters: age, temperature, leucocytosis, exaggeration of the liver.
4) The cases (records) which include "mononucleosis" as diagnosis in the file 'patients' are accumulated (in accordance with 'confirmity' parameter) and are added to the file 'knowledge'. The linguistic value handling is illustrated.

5) Selecting and visualising all the records from the file 'patients' for which the temperature is higher than 38°C, the age is in the interval <7,20>, and the degree of confirmity is greater than or equal to 'more or less confirmed'.
6) Visualizing the contents of the file 'knowledge'.

Conclusion

The standardization of the captured information structure makes possible exchanges of knowledge bases or simultaneous creation of several knowledge bases for expert consultation systems.

The standardization of the HIS subsystems makes them transportable from one HIS to another, as far as the hard/software system allows it. By the approach described in the paper such problems as: the way to combine different expertises, the way to compromise between experts' opinions, the verification of the knowledge (internal conflicts, incompleteness, inaccuracy, level of details), etc., are partially eliminated or at least made easier.

References

[1] Bonissone P.Y.A., Fuzzy Set Based Linguistic Approach: Theory and Applications, In: *Approximate Reasoning in Decision Analysis*, Gupta M.M., Sanchez F., (Eds.), North-Holland, Amsterdam, 1982, pp.329-339

[2] Welkbank M., *A Review of Knowledge Acquisition Techniques for Expert Systems*, British Telecom Research Laboratories technical reports, Martlesham Heath, Ipswich, 1983.

[3] *BMP D Statistical Software*, Dept. of Biomechanics, University of California, Los Angeles, 1981.

[4] Stanchev P., One Aspect of Building Medical Expert Consultation Systems, In: *Medical Decision Making*, Bemmel J., Gremy F., Zvarova J. (Eds), North-Holland, Amsterdam, 1985, pp. 74-80

[5] Stanchev P., Stancheva E., Kaloianov K., Decision Information System, In: *Proceedings of the Thirteenth Spring Conference of The Union of Bulgarian Mathematicians*, G. Gerov (Ed.), Publishing House of the Bulgarian Academy of Sciences, Sofia, 1984, pp.431-439

Appendix

```
*create
PLEASE ENTER:
FILE SPECIFICATION ? patients
NUMBER OF DIAGNOSTIC PARAMETERS ? 6
NAMES OF DIGNOSTIC PARAMETERS AND DATA TYPE, PARTED BY COMMA,
USING DIGITS:
1(FIXED INTEGER) , 2 (FIXED REAL) , 3 (FIXED TEXTUAL) , 4 (FIXED GRAPHIC) ,
5 (INTERVAL INTEGER) , 6 (INTERVAL REAL) , 7 (INTERVAL TEXTUAL) ,
8 (VECTOR INTEGER) , 9. (VETOR REAL)
DIAGRAM PARAM 1 ? age, 1
DIAGRAM PARAM 2 ? temperature, 2
DIAGRAM PARAM 3 ? Leucocytosis, 2
DIAGRAM PARAM 4 ? exagg_liver, 3
DIAGRAM PARAM 5 ? diagnosis, 3
DIAGRAM PARAM 6 ? confirmity, 3
MAXIMAL NUMBER OF OBJECTS ? 500

*insert
PLEASE ENTER:
FILE SPECIFICATION ?  patients
DO YOU WANT TO ENTER A NEW OBJECT, Y OR N ? y
NAME OBJECT 1 ? Peter
VALUES OF THE DIAGNOSTIC PARAMETERS FOR THIS OBJECT :
AGE                 ? 5
TEMPERATURE         ? 38.5
LEUCOCYTOSIS        ? 11
EXAGG_LIVER         ? too much exaggerated
DIAGNOSIS           ? mononucleosis
CONFIRMITY          ? nearly confirmed
DO YOU WANT TO ENTER A NEW OBJECT, Y OR N ? y
NAME OBJECT 2 ? Jim
VALUES OF THE DIAGNOSTIC PARAMETERS FOR THIS OBJECT :
AGE                 ? 7
TEMPERATURE         ? 39.4
LEUCOCYTOSIS        ? 11.2
EXAGG_LIVER         ? considerably exaggerated
DIAGNOSIS           ? mononucleosis
CONFIRMITY          ? nearly confirmed
DO YOU WANT TO ENTER A NEW OBJECT, Y or N ? y
NAME OBJECT 3 ? Mery
VALUES OF THE DIAGNOSTIC PARAMETERS FOR THIS OBJECT :
AGE                 ? 14
TEMPERATURE         ? 38
LEUCOCYTOSIS        ? 11.9
EXAGG_LIVER         ? extremely exaggerated
DIAGNOSIS           ? mononucleosis
CONFIRMITY          ? more or less confirmed
DO YOU WANT TO ENTER A NEW OBJECT, Y OR N ? y
NAME OBJECT 4 ? Ann
VALUES OF THE DIAGNOSTIC PARAMETERS FOR THIS OBJECT :
AGE                 ? 14
TEMPERATURE         ? 38.1
```

```
LEUCOCYTOSIS        ? 9.5
EXAGG_LIVER         ? suspected not normal
DIAGNOSIS           ? chickenpox
CONFIRMITY          ? nearly confirmed
DO YOU WANT TO ENTER A NEW OBJECT, Y OR N ? y
NAME OBJECT 5 ? Helen
VALUES OF THE DIAGNOSTIC PARAMETERS FOR THIS OBJECT :
AGE                 ? 10
temperature         ? 39.1
LEUCOCYTOSIS        ? 12
EXAGG_LIVER         ? more or less exaggerated
DIAGNOSIS           ? mononucleosis
CONFIRMITY          ? not confirmed
DO YOU WANT TO ENTER A NEW OBJECT, Y OR N ? n
DO YOU WANT TO ENTER A NEW VALUE FOR AN OLD OBJECT, Y OR N ? n

*create
PLEASE ENTER:
FILE SPECIFICATION  ? knowledge
NUMBER OF DIAGNOSTIC PARAMETERS ? 4
NAMES OF DIAGNOSTIC PARAMETERS AND DATA TYPE, PARTED BY COMMA,
USING DIGITS :
1 (FIXED INTEGER), 2 (FIXED REAL), 3 (FIXED TEXTUAL), 4 (FIXED GRAPHIC),
5 (INTERVAL INTEGER) , 6 (INTERVAL REAL) , 7 (INTERVAL TEXTUAL) , .
8 (VECTOR INTEGER) , 9. (VECTOR REAL)
DIAGRAM PARAM 1 ? age, 2
DIAGRAM PARAM 2 ? temperature,6
DIAGRAM PARAM 3 ? leucocytosis,6
DIAGRAM PARAM 4 ? exagg_liver,3
MAXIMAL NUMBER OF OBJECTS ? 100

*wknowledge
PLEASE ENTER:
MEDICAL KNOWLEDGE FILE SPECIFICATION ? knowledge
NAME OF OBJECT ? mononucleosis
PATIENTS' DATA FILE SPECIFICATION ? patients
NAME OF THE DIAGNOSTIC PARAMETER CONFIDENCE DEGREE ? confirmity
LOGICAL EXPRESSION
ARG 1 ? diagnosis .EQ. mononucleosis
LOGICAL OPERATOR ? .END.
DIAGNOSTIC PARAMETER DIAGNOSIS
STANDARD BASIC SET FOR LINGUISTIC VARIABLES, Y OR N ? y
1. (0,0,0,0)          2. (0,0,0,.2)          3. (0,.1,0,.2)
4. (.2,.2,.2,.2)      5. (.5,.5,.2,.2)       6. (.8,.8,.2,.2)
7. (.9,1,.2,0)        8. (1,1,.2,0)          9. (1,1,0,0)
PLEASE, LIST THE LINGUISTIC VARIABLES AND POINT TO THE
CORRESPONDENT NUMBER FROM THE BASIC SET ?
absolutely normal - 1
comparatively normal - 2
more or less normal - 3
suspected not normal - 4
more or less exaggerated - 5
obviously exaggerated - 6
too much exaggerated - 7
considerably exaggerated - 8
```



```
extremely exaggerated - 9
.END.
DIAGNOSTIC PARAMETER CONFIRMITY
STANDARD BASIC SET FOR LINGUISTIC VARIABLES, Y OR N ? y
1. (0,0,0,0)          2. (0,0,0,.2)          3. (0,.1,0,.2)
4. (.2,.2,.2,.2)      5. (.5,.5,.2,.2)       6. (.8,.8,.2,.2)
7. (.9,1,.1,0)        8. (1,1,.2,0)          9. (1,1,0,0)
PLEASE, LIST THE LINGUISTIC VARIABLES AND POINT TO THE
CORRESPONDENT NUMBER FROM THE BASIC SET ?
absolutely rejected - 1
nearly rejected - 2
more or less rejected - 3
nor confirmed - 4
more or less confirmed - 7.
nearly confirmed - 8
absolutely confirmed - 9
.END.

*select
PLEASE ENTER:
FILE SPECIFICATION ? patients
TO PRINT THE NUMBER OF CASES, Y OR N ? n
TO PRINT THE LIST OF THE CASES, Y OR N ? y
ARG 1 ? temperature .GT. 38
LOG OPER ? .AND.
ARG 2 ? age .BE. 7,20
LOG OPER ? .AND.
ARG3 ? confirmity .GE. more or less confirmed
LOG OPER ? .END.
DIAG PARAM:
            1     2     3                          4              5            6
OBJ:
Jim         7   39.4  11.2  considerably exaggerated  mononucleosis  nearly confirmed
Ann        14   38.1   9.5      suspected not normal     chickenpox  nearly confirmed
DIAG PARAM :
1.AGE
2.TEMPERATURE
3.LEUCOCYTOSIS
4.EXAGG_LIVER
5.DIAGNOSIS
6.CONFIRMITY

*type
PLEASE ENTER:
FILE SPECIFICATION ? knowledge
DIAGNOSTIC PARAMETERS ? all
OBJECTS ? all
DIAG PARAM:
                      1          2           3                     4
OBJ:
mononucleosis       9.0  38.0/39.4  11.0/12.0  too much exaggerated
DIAG PARAM:
1.AGE
2.TEMPERATURE
3.LEUCOCYTOSIS
```

(72.) * Stanchev P., “Medical Knowledge Acquisition for Expert Consultation System", book chapter in "***Medical Expert Systems***", Chytel M., Engelbrecht R. (eds.), Sigma press, London 1988, 153-163.