FUZZY SET THEORY IN MEDICINE
Part II

Klaus-Peter Adlassnig

Department of Medical Computer Sciences
1090 Vienna , Austria

4. MEDICAL EXPERT SYSTEM CADIAG-2

CADIAG-2 is intended to be the physician's active partner in diagnostic situations. It makes possible a man-machine partnership in order to unite the experience, creativeness and intuition of a physician with the knowledge-based computational power of a computer. The general structure of CADIAG-2 is shown in Figure 1.

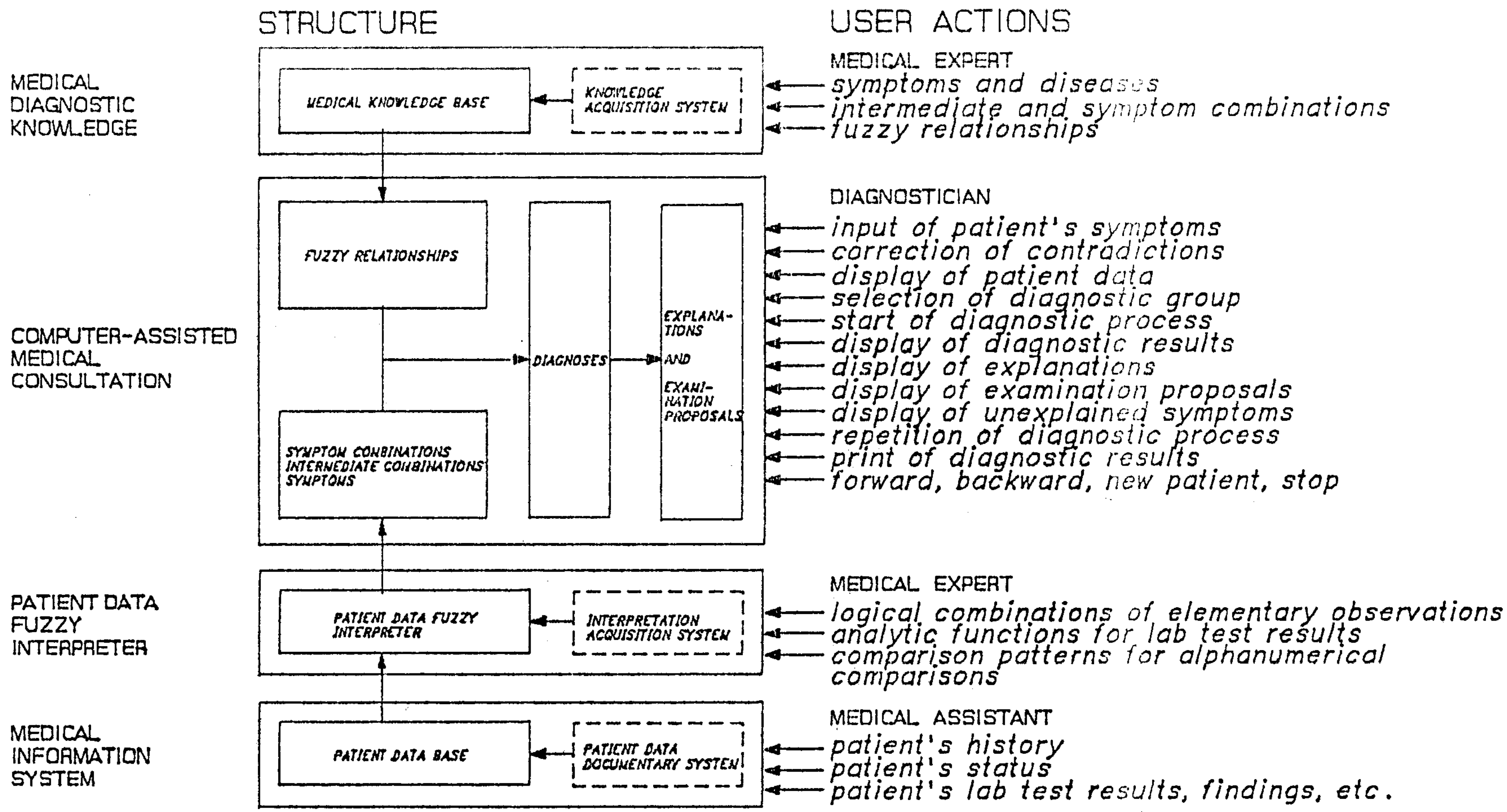


FIGURE 1: Structure of CADIAG-2 with connection to a medical information system
(dashed lines mark components effective before starting the consultation)

4.1. Representation of medical knowledge

CADIAG-2 considers four medical entities:

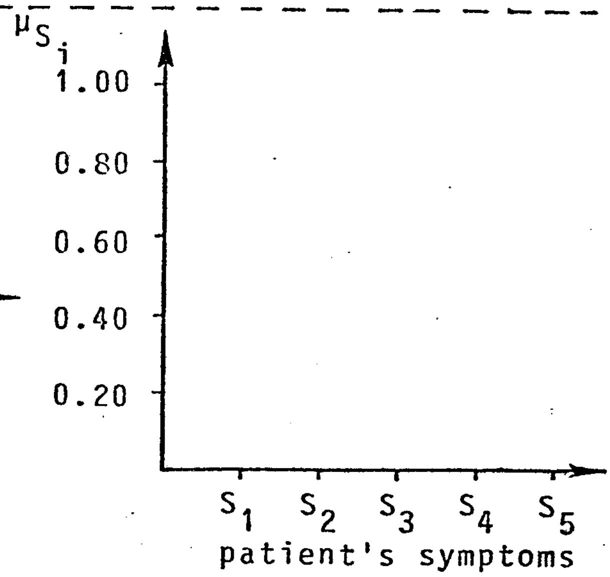
- symptoms, signs, test results, findings (S_i)
- diseases, diagnoses (D_j)
- intermediate combinations (IC_k)
- symptom combinations (SC_l).

Symptoms S_i take their values μ_{S_i} in $[0,1]\cup\emptyset$. The value μ_{S_i} indicate the degree to which S_i applies to the patient ($\emptyset$: S_i not yet examined). In the sense of fuzzy set theory, μ_{S_i} expresses the grade of membership of the patient's symptom manifestation S_i.

Example:

quantitative value		fuzzy values	
measured potassium level of 5.3 mmol/liter	fuzzy interpreter →	potassium, highly decreased	$S_1=0.00$
		potassium, decreased	$S_2=0.00$
		potassium, normal	$S_3=0.40$
		potassium, increased	$S_4=0.60$
		potassium, highly increased	$S_5=0.00$

5.3 $\xrightarrow[\text{interpreter}]{\text{fuzzy}}$



A binary fuzzy relationship $R_{PS} \subset \Pi \times \Sigma$ is established, defined by $\mu_{R_{PS}}(P_q, S_i) = \mu_{S_i}$ for patient P_q where $P_q \varepsilon \Pi = \{P_1, \ldots, P_r\}$ and $S_i \varepsilon \Sigma = \{S_1, \ldots, S_m\}$.

Diseases or diagnoses respectively also take their values in $[0,1] \cup \emptyset$. Fuzzy values $0.00 < \mu_{D_j} < 1.00$ qualify diagnoses as hypotheses. The values $\mu_{D_j} = 1.00$ or $\mu_{D_j} = 0.00$ mark confirmed or excluded diagnoses respectively. Diagnoses not yet considered as a diagnostic result of any kind have values $\mu_{D_j} = \emptyset$. Formally, $R_{PD} \subset \Pi \times \Delta$ defined by $\mu_{R_{PD}}(P_q, D_j) = \mu_{D_j}$ for patient P_q where $D_j \varepsilon \Delta = \{D_1 \ldots D_n\}$ is established.

Intermediate combinations were introduced to model pathophysiological states of patients. They are fuzzy logical combinations of symptoms and diseases.

Symptom combinations are combinations of symptoms, diseases and intermediate combinations. Both entities take their values μ_{IC_k} or μ_{SC_l} respectively in $[0,1] \cup \emptyset$ ($\emptyset$:not yet determinable). The relationship $R_{PSC} \subset \Pi \times K$ is defined by $\mu_{R_{PSC}}(P_q, SC_l) = \mu_{SC_l}$ for patient P_q where $SC_l \varepsilon K = \{SC_1 \ldots SC_t\}$ formally describes present, partly present and absent symptom combinations at the patient.

The fuzzy logical connectives are defined by:

- conjunction

$$x_1 \wedge x_2 = \begin{cases} \mathrm{MIN}(x_1, x_2) & \text{if } x_1 \varepsilon [0,1] \text{ and } x_2 \varepsilon [0,1] \\ \emptyset & \text{if } x_1 = \emptyset \text{ and/or } x_2 = \emptyset \end{cases}$$

- disjunction

$$x_1 \vee x_2 = \begin{cases} \mathrm{MAX}(x_1, x_2) & \text{if } x_1 \varepsilon [0,1] \text{ and } x_2 \varepsilon [0,1] \\ x_1 & \text{if } x_1 \varepsilon [0,1] \text{ and } x_2 = \emptyset \\ x_2 & \text{if } x_1 = \emptyset \text{ and } x_2 \varepsilon [0,1] \\ \emptyset & \text{if } x_1 = \emptyset \text{ and } x_2 = \emptyset \end{cases}$$

- negation

$$\bar{x}_1 = \begin{cases} 1 - x_1 & \text{if } x_1 \varepsilon [0,1] \\ \emptyset & \text{if } x_1 = \emptyset \end{cases}$$

The following relationships between medical entities are considered in CADIAG-2:

- symptom-disease relationships (S_iD_j)
- symptom combination-disease relationships (SC_lD_j)
- symptom-symptom relationships (S_iS_j)
- disease-disease relationships (D_iD_j).

The relationships are characterized by two aspects:

- frequency of occurrence (O)
- strength of confirmation (C).

The relationships between medical entities are presented in form of relationship rules with associated relationship tupels. The general form is

IF (premise) THEN (conclusion) WITH (O,C).

The relationship tupels (O,C) contain numerical and/or linguistic fuzzy values (see ZADEH [3]) μ_o and/or λ_o or μ_c and/or λ_c respectively.

The linguistic values λ_o and λ_c cover fuzzy intervals. Reasonable numerical representatives for λ_o and λ_c were selected in order to make fuzzy inferences (see paragraph 4.2.) easily possible. How linguistic fuzzy values, numerical intervals and representatives were chosen is described in more detail in ADLASSNIG and KOLARZ [8] and ADLASSNIG [9].

Table 1: Linguistic fuzzy values, numerical intervals and numerical representatives for frequency of occurrence and strength of confirmation.

frequency of occurrence			strength of confirmation		
value λ_o	interval	representative μ_o	value λ_c	interval	representative μ_c
always	[1.00,1.00]	1.00	always	[1.00,1.00]	1.00
almost always	[0.99,0.98]	0.99	almost always	[0.99,0.98]	0.99
very often	[0.97,0.83]	0.90	very strong	[0.97,0.83]	0.90
often	[0.82,0.68]	0.75	strong	[0.82,0.68]	0.75
medium	[0.67,0.33]	0.50	medium	[0.67,0.33]	0.50
seldom	[0.32,0.18]	0.25	weak	[0.32,0.18]	0.25
very seldom	[0.17,0.03]	0.10	very weak	[0.17,0.03]	0.10
almost never	[0.02,0.01]	0.01	almost never	[0.02,0.01]	0.01
never	[0.00,0.00]	0.00	never	[0.00,0.00]	0.00
unknown	∅	∅	unknown	∅	∅

Examples:

- IF (ultrasonic of pancreas is pathological)
  THEN (pancreatic carcinoma)
  WITH (0.75=often, 0.25=weak).
- IF (tophi)
  THEN (gout)
  WITH (0.25=seldom, 1.00=always).
- IF (lower back pain $\wedge$ limitation of motion of the lumbar spine $\wedge$ diminished chest expansion $\wedge$ male patient $\wedge$ age between 20 and 40 years)
  THEN (ankylosing spondylitis)
  WITH (-,0.90=very strong).

The relationship values μ_o and μ_c are interpreted as values of fuzzy relationships between premises and conclusions:

- S_iD_j -occurrence relationship $R^o_{SD} \subset \Sigma \times \Delta$
- S_iD_j -confirmation relationship $R^c_{SD} \subset \Sigma \times \Delta$
- SC_lD_j-occurrence relationship $R^o_{SCD} \subset K \times \Delta$
- SC_lD_j-confirmation relationship $R^c_{SCD} \subset K \times \Delta$
- S_iS_j -occurrence relationship $R^o_{SS} \subset \Sigma \times \Sigma$
- S_iS_j -confirmation relationship $R^c_{SS} \subset \Sigma \times \Sigma$
- D_iD_j -occurrence relationship $R^o_{DD} \subset \Delta \times \Delta$
- D_iD_j -confirmation relationship $R^c_{DD} \subset \Delta \times \Delta$

4.2. Concept of fuzzy logical inference

The compositional rule of inference proposed by ZADEH [4] and introduced in medical diagnosis by SANCHEZ [12,13] serves as inference mechanism. It accepts fuzzy descriptions of patient's symptoms and infers fuzzy descriptions of patient's diseases by means of fuzzy relationships between the medical entities described in the previous section.

For the deduction from symptoms S_i present, partly present or absent at the patient P_q to diseases D_j present, possibly present or absent, three different compositional rules of inference are computed:

- S_iD_j-confirmation composition (composition 1)

$$R^1_{PD} = R_{PS} \circ R^c_{SD} \tag{1}$$

defined by

$$\mu_{R^1_{PD}}(P_q,D_j) = \max_{S_i} \text{MIN}[\mu_{R_{PS}}(P_q,S_i);\mu_{R^c_{SD}}(S_i,D_j)]$$

- S_iD_j-non-confirmation composition (composition 2)

$$R^2_{PD} = R_{PS} \circ (1-R^c_{SD}) \tag{2}$$

defined by

$$\mu_{R^2_{PD}}(P_q,D_j) = \max_{S_i} \text{MIN}[\mu_{R_{PS}}(P_q,S_i);1-\mu_{R^c_{SD}}(S_i,D_j)]$$

- S_iD_j-non-symptom composition (composition 3)

  $$R^3_{PD}=(1-R_{PS})oR^o_{SD} \quad (3)$$

  defined by

  $$\mu_{R^3_{PD}}(P_q,D_j)=\underset{S_i}{MAX}\ MIN[1-\mu_{R_{PS}}(P_q,S_i);\mu_{R^o_{SD}}(S_i,D_j)]$$

- The following diagnostic results are obtained:
- confirmed diagnoses

  $$\text{if } \mu_{R^1_{PD}}(P_q,D_j)=1.00 \quad (4)$$

- diagnostic hypotheses

  if

  $$0.10\leq\mu_{R^1_{PD}}(P_q,D_j)\leq 0.99 \quad (5)$$

  The boundary of 0.10 is a heuristic value.
  It precludes diagnoses with very low evidence.
- excluded diagnoses (6)

  $$\text{if } \mu_{R^2_{PD}}(P_q,D_j)=1.00$$

  or

  $$\text{if } \mu_{R^3_{PD}}(P_q,D_j)=1.00 \quad (7)$$

Symptom combination-disease inferences (compositions 4, 5 and 6) are performed and interpreted in an analogous way. Symptom-symptom inferences (compositions 7, 8 and 9) are computed in order to complete patient's symptom patterns. Disease-disease inferences (compositions 10, 11 and 12) are carried out in order to confirm super-terms of diseases from present sub-terms or to exclude entire sub-areas if the super-term is excluded.

4.3. Acquisition of medical knowledge

The knowledge acquisition system is capable of aquiring medical entities and relationships between them. In CADIAG-2, relationships are stored as numerical fuzzy values from [0,1]. Two ways can be taken to acquire medical knowledge:
- linguistic documentation by medical experts
- statistical evaluation of a patient data base.

Relationships can be gathered linguistically by using predefined linguistic values for determining the linguistic variables frequency of occurrence O and strength of confirmation C (cf. Table 1). Thus empirical, judgmental and definitorial knowledge is acquired.

CADIAG-2 relationships have the important property to be interpreted statistically. It is

- frequency of occurrence

$$\mu_o = \frac{F(S_i \cap D_j)}{F(D_j)} = F(S_i/D_j) \tag{8}$$

- strength of confirmation

$$\mu_c = \frac{F(S_i \cap D_j)}{F(S_i)} = F(D_j/S_i) \tag{9}$$

where

- $F(S_i \cap D_j)$: absolute frequency of occurrence of S_i and D_j
- $F(D_j)$: absolute frequency of occurrence of D_j
- $F(S_i)$: absolute frequency of occurrence of S_i
- $F(S_i/D_j)$: conditional frequency of S_i with D_j
- $F(D_j/S_i)$: conditional frequency of D_j with S_i.

By using definitions (8) and (9), extended statistical evaluations of medical relationships already acquired or not yet acquired can be carried out by means of patient data with known diagnoses.

4.4. Diagnostic process

4.4.1. Patient's symptoms

Patient's symptoms can be entered into CADIAG-2 in three ways (described in detail in ADLASSNIG [9]):

- natural language input of symptoms S_i

- natural language input of keywords that trigger whole segments of symptoms S_i
- patient data base access and transfer of patient's symptoms already stored via a fuzzy interpreter.

Natural language input of symptoms S_i such as "High fever", "Increased GOT", "Blood stool positive" is realized by a symptom search algorithm with a word segmentation algorithm embedded that allows the use of synonyms and abbreviations, orthographic variants and different flexions of words.

Input of keywords such as "Present complaints", "Previous complaints", "Blood count", "Ultrasonic" causes the whole sections of the symptom thesaurus to be displayed. Subsequently, fuzzy values can be attached to these symptoms by the physician.

The existence of a patient data base in a medical information system with patient data already stored suggests the automatic transfer from the data base to CADIAG-2. This transfer includes the fuzzy interpretation of patient's data. The fuzzy interpreter contains definitions about the assignment of fuzzy values to observations, lab test results and even simple alphanumerical texts.

After patient's symptoms are gathered, symptom-symptom inferences are performed. The established symptom list contains every necessary information such as fuzzy value, origin (original; inferred), predefined symptom class (routine; by special order; invasive or expensive), numerical value, units and date of observation. Afterwards, the symptom list is checked for contradictions.

4.4.2. Patient's symptom combinations

Intermediate combinations are evaluated in the next step. After passing the check for contradictions, fuzzy values for symptom combinations are computed. The associated lists are now as complete as possible and without contradictions.

4.4.3. Confirmed diagnoses

Patient fuzzy values $\mu_{D_j}=1.00$, e.g. confirmed diagnoses D_j for P_q, are determined by (9)

$$\mu_{D_j}=1.00 \text{ if} \begin{cases} \mu_R 1_{PD}(P_q,D_j)=1.00 \\ \text{or} \\ \mu_R 4_{PD}(P_q,D_j)=1.00 \end{cases} \qquad (9)$$

4.4.4. Excluded diagnoses

Patient fuzzy values $\mu_{D_j}=0.00$, e.g. excluded diagnoses D_j for P_q, are calculated by (10)

$$\mu_{D_j}=0.00 \text{ if} \begin{cases} \mu_R 2_{PD}(P_q,D_j)=1.00 \\ \text{or} \\ \mu_R 3_{PD}(P_q,D_j)=1.00 \\ \text{or} \\ \mu_R 5_{PD}(P_q,D_j)=1.00 \\ \text{or} \\ \mu_R 6_{PD}(P_q,D_j)=1.00 \end{cases} \qquad (10)$$

Now disease-disease relationships allow the inference of further diagnoses (confirmed or excluded):

$$\mu_{D_j}=\begin{cases} 1.00 \text{ if } \mu_R 10_{PD}(P_q,D_j)=1.00 \\ 0.00 \text{ if} \begin{cases} \mu_R 11_{PD}(P_q,D_j)=1.00 \\ \text{or} \\ \mu_R 12_{PD}(P_q,D_j)=1.00 \end{cases} \end{cases} \qquad (11)$$

4.4.5. Diagnostic hypotheses

Method 1:

Patient fuzzy values μ_{D_j} with $0.10 \leq \mu_{D_j} \leq 0.99$ indicate diagnostic hypotheses. They are determined by (12).

$$\mu_{D_j}=\mathrm{MAX}\left[\mu_R 1_{PD}(P_q,D_j);\mu_R 4_{PD}(P_q,D_j);\mu_R 10_{PD}(P_q,D_j)\right] \quad (12)$$

$$\text{if}\ \begin{cases} 0.10 \leqq \mu_R 1_{PD}(P_q,D_j) \leqq 0.99 \\ \text{and/or} \\ 0.10 \leqq \mu_R 4_{PD}(P_q,D_j) \leqq 0.99 \\ \text{and/or} \\ 0.10 \leqq \mu_R 10_{PD}(P_q,D_j) \leqq 0.99 \end{cases}$$

Method 2:

Because the value μ_{D_j} calculated by (12) is independent of the number of rules that can be applied for D_j, a powerful heuristic function was introduced which considers the number of present but not confirming criteria for D_j and calculates a number of points PN_{D_j}. These numbers of points are helpful for the immediate judgment among the diagnostic hypotheses although the ultimate aim should be the confirmation of diseases.

$$PN_{D_j} = \sum_{i=1}^{m^*} \left[\alpha \cdot \mu_R o_{SD}(S_i,D_j) + \beta \cdot \mu_R c_{SD}(S_i,D_j)\right] \quad (13)$$

where m^* is the number of present symptoms that occur in the definition of D_j. It is $\alpha+\beta=1.00$. At present the values are fixed with $\alpha=0.09$ and $\beta=0.91$, e.g. the strength of confirmation determines PN_{D_j} ten times stronger than the frequency of occurrence.

4.4.6. Explanation of diagnostic results

The acceptance of diagnostic results by the physician is highly dependent on the ability of CADIAG-2 to explain its diagnostic output. On request, reasonings for confirmed diagnoses, excluded diagnoses and diagnostic hypotheses are presented. The reasonings consist of the name of the medical entities, their definitions,their measured and fuzzy values, and the relationships to the diagnostic output.

4.4.7. Plan for patient's further examination

One of the main objectives of CADIAG-2 is to provide an iterative consultation starting with simple, easy-to-examine and cheap patient data. Usually some diagnostic hypotheses can be inferred and further examinations are necessary to confirm or deny these hypotheses. On the basis of the medical knowledge contained in the knowledge base of CADIAG-2, proposals for patient's further examination can be made. The selected symptoms to be examined are those with high relationships to diagnostic hypotheses.

4.4.8. Patient's unexplained symptoms

Clearly confirmed diagnoses and remaining diagnostic hypotheses together should explain any pathological symptom, sign or lab test result of the patient. Unexplained data (usually) indicate further diseases that have to be taken into consideration in further diagnostic steps.

5. RESULTS

5.1. Rheumatological diseases

CADIAG-2/RHEUMA has been partly tested with real patient data from a rheumatological hospital. 169 cases of rheumatoid arthritis, SJÖGREN's disease, systemic lupus erythematodes, REITER's disease, and sclerodermia yielded a 77.16% correctness. In order to calculate this correctness, the clinical diagnoses established by the responsible clinician of the rheumatological hospital and the confirmed diagnoses offered by CADIAG-2 were compared. The clinical diagnoses were assumed to be correct. Well-defined confirming symptom combinations were available to calculate these results. Most of the cases in which the clinical diagnosis could not be confirmed were cases that

- represented not the first hospitalization of the patient but only a control stay to check the effectiveness of drugs already administered
- represented cases with early stages of the disease under consideration; in almost every one of these cases a diagnostic hypothesis was generated.

5.2. Pancreatic diseases

CADIAG-2/PANCREAS was tested with 31 real patient records. The final clinical diagnoses of these patients were not confirmed by histological findings. Nevertheless, they were assumed to be correct and an objective criterion to evaluate CADIAG-2.

Pancreatic carcinoma was confirmed twice. The pathognomonic finding "Specific abnormal pancreatic biopsy" was available which has a strength of confirmation μ_c=1.00 for pancreatic carcinoma.

Diagnostic hypotheses were generated, and the heuristic number of points yielded the basis for evaluation. Table 2 shows the results.

Table 2: Comparison of diagnostic hypotheses with final diagnoses.

Final diagnosis is in ...%	the diagnostic hypotheses with the ... number of points
50.0% of cases	highest number of points
71.4% of cases	highest or second highest number of points
82.2% of cases	highest, second highest and third highest number of points
89.2% of cases	highest, second highest, third highest and fourth highest number of points
10.8% of cases	no proposal of a diagnostic hypothesis

ACKNOWLEDGMENTS

The author gratefully acknowledges the contributions of G. KOLARZ, M.D., and W. SCHEITHAUER, M.D., for the medical documentation in rheumatology and pancreatic diseases.

REFERENCES

[1] ZADEH, L.A.: Fuzzy Sets. Information and Control 8 (1965) 338-353.

[2] ZADEH, L.A.: A Fuzzy-Algorithmic Approach to the Definition of Complex or Imprecise Concepts. In BOSSEL, H., S. KLACZKO, N. MÜLLER (Eds.): System Theory in the Social Sciences. Birkhäuser Verlag Basel-Stuttgart, 1976, 202-282.

[3] ZADEH, L.A.: Linguistic Variables, Approximate Reasoning and Dispositions. Med. Inform. 8 (1983) 173-186.

[4] ZADEH, L.A.: Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. IEEE Transactions on Systems, Man, and Cybernetics, Vol.SMC-3, No.1, January 1973, 28-44.

[5] BELLMAN, R.E., L.A. ZADEH: Local and Fuzzy Logics. Memorandum NO.ERL-M584, Electronics Research Laboratory, College of Engineering, University of California, Berkeley 94720, 11 May 1976.

[6] ADLASSNIG, K.-P.: A Survey on Medical Diagnosis and Fuzzy Subsets. In GUPTA, M.M., E. SANCHEZ (Eds.): Approximate Reasoning in Decision Analysis. North-Holland Publishing Company Amsterdam-New York-Oxford, 1982, 203-217.

[7] ADLASSNIG, K.-P.: A Fuzzy Logical Model of Computer-Assisted Medical Diagnosis. Meth.Inform.Med. 19 (1980), 141-148.

[8] ADLASSNIG, K.-P., G. KOLARZ: CADIAG-2: Computer-Assisted Medical Diagnosis Using Fuzzy Subsets. In GUPTA, M.M., E. SANCHEZ (Eds.): Approximate Reasoning in Decision Analysis. North-Holland Publishing Company Amsterdam-New York-Oxford, 1982, 219-247.

[9] ADLASSNIG, K.-P.: Ein Computerunterstütztes Medizinisches Diagnosesystem unter Verwendung von Fuzzy Teilmengen. PhD.diss.Technical University of Vienna, Vienna, 1983.

[10] Encyclopaedia Britannica. Encyclopaedia Britannica Inc. Chicago, 1975.

[11] Lexikon der Kybernetik. Akademie-Verlag Berlin, 1981.

[12] SANCHEZ, E.: Compositions of Fuzzy Relations. In GUPTA, M.M., R.K. RAGADE, R.R. YAGER (Eds.): Advances in Fuzzy Set Theory and Applications. North-Holland Publishing Company Amsterdam-New York-Oxford, 1979, 421-433.

[13] SANCHEZ, E.: Medical Diagnosis and Composite Fuzzy Relations. In GUPTA, M.M., R.K. RAGADE, R.R. YAGER (Eds.): Advances in Fuzzy Set Theory and Applications. North-Holland Publishing Company Amsterdam-New York-Oxford, 1979, 437-444.