COMPUTERS AND BIOMEDICAL RESEARCH 22, 297–313 (1989)

Performance Evaluation of Medical Expert Systems Using ROC Curves*

KLAUS-PETER ADLASSNIG† AND WERNER SCHEITHAUER‡

†*Department of Medical Computer Sciences and* ‡*2nd Department for Gastroenterology and Hepatology, University of Vienna, Garnisongasse 13, A–1090 Vienna, Austria*

Received January 28, 1988

This paper presents a performance evaluation of the diagnostic accuracy of the medical expert system CADIAG-2/PANCREAS. The study included 47 clinical cases from a university hospital with 51 diagnoses of pancreatic diseases (four patients had two pancreatic diseases). As gold standard, the histologically or clinically confirmed diagnoses were assumed. Performance was studied along three lines: (a) each case was evaluated twice, first, by restricting patient data to history, physical examination, and basic laboratory tests and, second, by utilizing the complete set of data including also special laboratory tests, US, X ray, CT-scan, ECG, and biopsy, if available; (b) considering CADIAG-2's hypotheses generation, each evaluation series was also carried out twice, first, by testing whether the gold standard was the first diagnosis in the ranked list of hypotheses and, second, whether the gold standard was among the hypotheses; (c) receiver operating characteristic (ROC) curves were determined by varying an internal threshold which determined the extent of CADIAG-2's diagnostic hypotheses generation. The evaluation showed that CADIAG-2's initial list of diagnostic hypotheses, based on patient history, physical examination, and basic laboratory tests, usually has already included the gold standard diagnosis and thus an application of CADIAG-2 at a very early stage of the diagnostic process seems achievable. Moreover, it turned out that given the complete set of patient's medical data the gold standard is usually ranked at the first place in the list of hypotheses, except for patients with chronic diseases where only unspecific findings are available. The last test series showed that ROC curves do not only allow optimal adjustment of the expert system's internal ad hoc decision criteria such as thresholds, weights, and scores but also provide a basis for better comparing the performance of different medical expert systems. © 1989 Academic Press, Inc.

1. INTRODUCTION

The evaluation of an expert system is a natural step that follows its design, development, and implementation. Basic research in medical expert systems aimed at finding new and better forms of knowledge representation, knowledge acquisition, and automated reasoning is still intensively pursued; however, some medical expert systems have already been applied as routine systems in clinical settings (HELP (*1*), PUFF (*2*), ONCOCIN (*3*), differential diagnosis of

* Presented in part at the MIE '87 Conference, Rome, Italy, September 22, 1987.

0010-4809/89 $3.00

acute abdominal pain (*4*)). There are other computer programs that are or will soon be ready for practical operation in the hospital or the physician's office (INTERNIST-I/CADUCEUS (*5*), QMR (*6*), RECONSIDER (*7*), CADIAG-2 (*8*, *9*)). Some of these systems have already been tested with several hundred clinical cases. The need for validation and verification of these expert systems prior and during practical operation grows and methods for evaluating performance, acceptability, and costs are searched for.

A number of methods evaluating certain aspects of medical expert systems, such as the systems' diagnostic accuracy, the consistency and completeness of their knowledge bases, their acceptance by the medical user, and their transportability to other locations have already been devised (*10*, *11*), yet further methods have to be developed.

The study on hand describes a performance evaluation of the diagnostic accuracy of CADIAG-2/PANCREAS[1] (*12*, *13*), a medical expert system for differential diagnosis of pancreatic diseases.

The performance of CADIAG-2/PANCREAS was studied with respect to three different aspects:

First, system's performance was determined (a) by generating diagnoses on the basis of an initial set of patient data containing only history items, signs from the general physical examination, and basic laboratory test results; and (b) by using the complete set of available data including the initial set as well as special laboratory test results and results from clinical investigations (US, X ray, CT-scan, ECG, and biopsy). This evaluation aimed at revealing differences in the hypotheses generation rooted in the availability of patient data from different examination areas. The comparison was done by calculating sensitivity and specificity rates as measures of the obtained accuracy.

Second, the system's performance was compared by applying two different definitions of what is considered to be a true positive result of the expert system. (Here, only the definition of what is a *diagnostic hypothesis* was varied; definitely confirmed diagnoses are of course true positive results.) Applied definitions were (a) true positive is if the gold standard is the first diagnosis in the ranked list of generated hypotheses; and (b) true positive is if the gold standard is among the hypotheses. This evaluation showed whether or under which circumstances the system indicates immediately the correct diagnosis or prompts the correct diagnosis only among others. Again, sensitivity and specificity rates were calculated to assess the accuracy of the expert system.

Third, in order to determine the overall degree of the system's performance, ROC curves were calculated by varying an internal threshold which influences the extent of CADIAG-2's diagnostic hypotheses generation. The system's designer and eventually the system's user is thus enabled to optimize ad hoc decision criteria such as thresholds, weights, and scores, which are often an inherent part of an expert system's reasoning mechanism.

[1] CADIAG stands for *C*omputer-*A*ssisted *DIAG*nosis.

2. The Medical Expert System CADIAG-2

The central goal of the CADIAG-2 project is the development of a medical consultation system for general internal medicine. Its underlying clinical issues are to assist in the differential diagnostic process (a) by indicating all possible diseases which might be the cause of patient's pathological findings, with special emphasis on rare diseases; (b) by offering further useful examinations to confirm or to exclude gained diagnostic hypotheses or to find stronger support for them; and (c) by indicating patients' pathological findings not yet accounted for by the expert system's proposed diagnoses.

After gaining experience with the medical expert system CADIAG-1 which was formally based on first-order predicate logic and pattern matching (*8*), a successor system CADIAG-2 was developed and implemented (*14, 15*). This system applies fuzzy set theory to model inherent vagueness of medical concepts and fuzzy logic to infer diagnostic conclusions. At present, CADIAG-2's knowledge base contains disease profiles and complex rules for about 295 diseases, among them 185 rheumatic diseases (69 joint diseases, 12 diseases of the spinal column, 38 diseases of soft tissue and connective tissue system, 45 diseases of cartilage and bone, 21 regional pain syndromes) (*16*) and 110 gastroenterological diseases (35 gall bladder and bile duct diseases (*17*), 10 pancreatic diseases (*12, 13*), 37 colon diseases, 28 diseases of the peritoneum).

The CADIAG-2 system is integrated into the medical information system WAMIS[2] of the Vienna General Hospital (*9*). This integration allows the collection of patient's findings for CADIAG-2 via the routine medical documentation and laboratory system of WAMIS. Through a data abstraction and aggregation process (*18*), patient data are made available to the CADIAG-2 system which tries to infer diagnoses from these abstracted findings in a data-driven manner. In addition, patient data not routinely collected in WAMIS can be added to CADIAG-2 through a man-machine interface which processes medical terms given in natural language. A word segmentation algorithm allows usage of medical synonyms and abbreviations; moreover, it accepts various orthographic variants and takes different medical suffixes into account (*19*).

The CADIAG-2 system was designed in such a way that three modes of application in our hospital are possible: (a) the screening and monitoring mode applied at a very early stage of the diagnostic process; (b) the consultation mode applied after complete data collection; and (c) the textbook mode without connection to the central patient data base.

CADIAG-2's diagnostic process is based on both stored disease profiles and rules (usually very complex ones such as the ARA criteria for rheumatic diseases (*20*)). Two relationships define the association between findings and diseases in these disease profiles, first, the necessity of occurrence of a certain finding with a disease (frequency of occurrence degree) and, second, its suffi-

[2] WAMIS is the German acronym for *W*iener *A*llgemeines *M*edizinisches *I*nformations-*S*ystem (Vienna General Medical Information System).

ciency to infer the disease (strength of confirmation degree). The same relationships are applicable to define the associations between the antecedents and consequents of rules.

The inference process of CADIAG-2 aims at generating one or more differential diagnoses and—at the same time—at excluding some or all remaining diagnoses. A diagnosis is either established as definitely confirmed or proposed as a diagnostic hypothesis to be confirmed or excluded after additional examinations are performed.

Diagnoses are indicated as definitely confirmed if pathognomonic findings were found in the patient or confirming rules were triggered by patient's findings. Because of the hierarchical relationships among diseases in CADIAG-2, diagnoses at a higher level in the disease hierarchy are confirmed as well if subdiagnoses are indicated as being confirmed.

Excluded diagnoses are established by either present excluding criteria or absent obligatory criteria. Excluding criteria may be single excluding findings, exluding rules or other, already established diagnoses which exclude other diagnoses. Findings and rule criteria defined to be obligatory present in the patient to establish a certain diagnosis but are definitely absent consequently exclude the respective diagnosis. Definitely excluded disease categories in the disease hierarchy cause also the exclusion of the entire set of the respective subdiagnoses, if any.

Diagnoses being confirmed and excluded at the same time—which might happen due to contradictory patient data and/or knowledge base errors—are termed diagnostic contradictions. They are displayed separately stating the reason of being established.

Diagnostic hypotheses are generated if a diagnosis is, first, neither confirmed, nor excluded, nor a contradictory result and, second, the strength of confirmation of at least one present finding, one triggered rule, or one already established subdiagnosis is equal or higher than a given threshold ε $(0 < \varepsilon < 1)$. Since the application of fuzzy set theory allows for mathematical modeling of borderline findings, the degree of presence of a finding (degree of membership in a fuzzy set) is combined with its strength of confirmation. If the resulting value, which is a measure of certainty of the concluded disease, lies between the threshold ε and unity (unity means full confirmation), the respective disease has to be taken into consideration as a diagnostic hypothesis. In addition, diagnostic hypotheses are ranked according to a score of support. This score is calculated on the basis of, first, the number of single findings present or present to a certain degree and having a relationship to the disease under consideration, second, the degree of presence of these findings, and third, the degrees for frequency of occurrence and strength of confirmation between these findings and the respective disease.

Diagnoses which are neither confirmed, nor excluded, nor diagnostic hypotheses, nor contradictory results are put into a category denoted by "not generated diagnoses." This allows the physician to obtain a complete survey of all diseases included into CADIAG-2's knowledge base.

In CADIAG-2, two forms of knowledge acquisition have been applied, first, acquisition of knowledge from medical experts and, second, semiautomatic acquisition of medical knowledge from a patient data base. Medical experts provide definitional and judgmental knowledge from textbooks and their own practical experience. The estimation of appropriate values for the frequency of occurrence and strength of confirmation degrees is assisted by an automatic procedure which calculates the respective values from stored patient records with known diagnoses (*21*).

Due to the large number of medical relationships contained in CADIAG-1 and CADIAG-2, intense efforts have been made to verify consistency and completeness of the respective knowledge bases. For CADIAG-1, a program was developed that verifies the internal consistency of the stored medical knowledge and—in case of inconsistencies—provides the line of reasoning for subsequent correction (*22, 23*). Because of the possible homomorphic mapping of CADIAG-2's finding-to-disease relationships into the finding-to-disease relationship categories of CADIAG-1, this program can partially be applied to CADIAG-2's knowledge base as well (*21*).

At present, extended clinical trials for testing the diagnostic accuracy in all differential diagnostic groups included into CADIAG-2 are in process. Results that have been obtained by applying the system to about 700 clinical cases are described in (*8, 12, 13, 16, 17*).

3. The CADIAG-2/PANCREAS System

The knowledge base of CADIAG-2/PANCREAS contains the profiles of 10 pancreatic diseases: *pancreatic cancer, acute pancreatitis, chronic pancreatitis, cystic pancreatic fibrosis, pancreatic pseudocyst, insulinoma, glucagonoma, Zollinger–Ellison syndrome, Verner–Morrison syndrome*, and *annular pancreas*. Complex rules were not defined for pancreatic diseases.

In order to establish the 10 disease profiles of pancreatic diseases, 327 findings (135 history items, 57 signs from the physical examination, 67 laboratory test results, 54 US, 6 X ray, 6 CT-scan, 1 ECG, and 1 biopsy finding) were applied. Altogether, 560 frequency of occurrence and 438 strength of confirmation degrees were entered. Some findings, such as patient's sex and his age category, were included into the respective disease profiles but with a frequency of occurrence degree only. In average, about 56 findings are contained in each profile (minimum 15 findings, maximum 120 findings). The complete disease profile of *pancreatic cancer* can be found in (*24*). In addition, 21 hierarchical relationships and 63 mutual exclusions among findings are contained in the knowledge base. There are no hierarchical relationships among diseases defined in the pancreas part of CADIAG-2.

In CADIAG-2/PANCREAS, the only two diagnoses that can be confirmed are *pancreatic cancer*, due to a *positive cytology by percutaneous aspiration biopsy* (*US/CT-guidance*) and *cystic pancreatic fibrosis*, due to an *abnormal finding by pilocarpine-iontophoresis*. Because of the lack of excluding criteria, excluded diagnoses as well as contradictory results cannot occur. Diagnostic hypotheses are established if one of the two pathognomonic findings are not fully present (i.e., the findings are uncertain) and/or the strength of confirmation between a present finding or a finding present to a certain degree and the respective diagnosis is between zero and unity, as is in the majority of cases, and—in addition—the resulting certainty values for the inferred diagnoses equal or surpass the threshold ε. A complete example of a diagnostic process in the area of pancreatic diseases can be found in (*24*).

4. Receiver Operating Characteristic (ROC) Curves

An appropriate way to determine the accuracy of an automated system for decision making is to compare its decisions with a gold standard. In terms of an expert system for medical diagnosis, the gold standard is the histopathological or surgical diagnosis—if available—or at least the opinion of recognized experts against which the program's performance should be assessed. The analysis of a series of decisions shows that four alternatives are possible.

- true positive decisions, that is, the system's decision is the "truly positive" standard;
- false positive decisions, that is, a decision offered by the system is not the standard; a "falsely positive" decision was made;
- true negative decisions, that is, the system did correctly avoid a decision that is not the standard; the system decided "truly negative";
- false negative decisions, that is, the system failed to make the correct decision; the result is "falsely negative."

These four alternatives can be entered into a 2 × 2 table and average accuracy ratios may then be computed (see Table 1).

Given a series of decisions, the accuracy of an automated system for decision

TABLE 1

2 × 2 Table for Calculating the Four Accuracy Ratios: Sensitivity, Specificity, False Positive Ratio, and False Negative Ratio (D, Decision or Diagnosis Present; $\bar{D}$, Decision or Diagnosis Absent)

	$D_{\text{gold standard}}$	$\bar{D}_{\text{gold standard}}$
$D_{\text{expert system}}$	a	b
$\bar{D}_{\text{expert system}}$	c	d
	$a + c$	$b + d$

Note. True positive ratio (sensitivity) = $a/(a + c)$; true negative ratio (specificity) = $d/(b + d)$; false positive ratio (1 − specificity) = $b/(b + d)$; false negative ratio (1 − sensitivity) = $c/(a + c)$.

making can now be determined by calculating the sensitivity and specificity rates or—which is equivalent—the respective false positive and false negative rates of the system.

However, a system's internal decision criterion such as a threshold usually influences the performance of the system. In order to uncover this influence, the accuracy rates are computed for several series of decisions, where the internal decision criterion is varied over its possible range. The graph of the true positive versus false positive ratios is the ROC curve in terms of the decision performance. The ROC curves quite effectively isolate the overall capacity for discrimination of a system from the specific degree of discrimination given a specific decision criterion (*25–27*). The ROC curves obtained in our study are given in Figs. 3–10. The figures are explained in detail in Section 5.2.

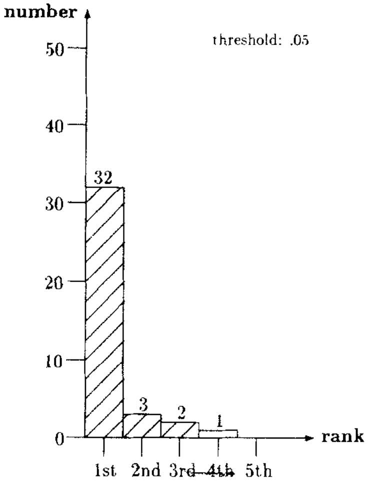
Moreover, it seems useful to explicitly provoke change of the performance of an expert system by varying these decision criteria, which are—in some form—usually an inherent part of an expert system and thus establish the systems' overall degree of discrimination. By analyzing the resulting ROC curves, it is possible to compare the systems' capacities for discrimination of different expert systems. The comparison between ROC curves can be carried out by calculating the areas under the ROC curves. Subsequently, these areas can be tested for statistically significant differences, as is reported in (*28–30*).

5. Evaluation of CADIAG-2/PANCREAS

5.1. Patient data. For this study, 47 patient records from the 2nd Department for Gastroenterology and Hepatology (Director: Professor Dr. G. Grabner) of the University of Vienna Medical School were available. Altogether, these patients showed 51 pancreatic diagnoses. The gold standard for most of them was the given histopathological or surgical diagnosis; in some cases, a reliable clinical diagnosis was available. There was a lack of cases with *cystic pancreatic fibrosis*, *glucagonoma*, *Verner–Morrison syndrome*, and *annular pancreas*, which occur very rarely in our department. In detail, the tested cases were as follows:

22 cases with pancreatic cancer;
11 cases with chronic pancreatitis;
6 cases with acute pancreatitis;
2 cases with pancreatic pseudocyst and with coincident chronic pancreatitis;
2 cases with pancreatic pseudocyst and with coincident acute pancreatitis;
3 cases with Zollinger–Ellison syndrome;
1 case with insulinoma;
a total of 47 cases with 51 pancreatic diagnoses.

The complete set of patient data contained about 200 findings of which about 30 findings were present and about 170 were marked by the physician as definitely absent. The reduced set of patient data usually comprised about 20 present and 150 definitely absent findings.



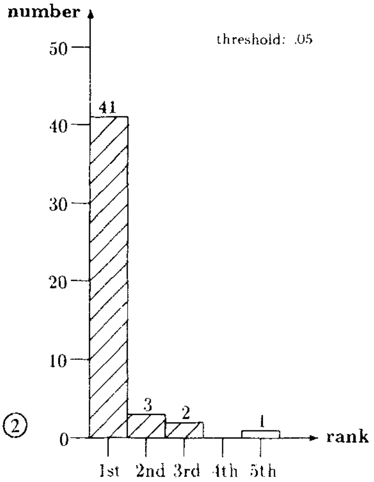


FIG. 1. Ranking of 51 clinical diagnoses being candidates for hypotheses generation in CADIAG-2. The hypotheses were established on the basis of patient history, physical examination, and basic lab tests only. The threshold ε was set to 0.05 allowing a liberal hypothesis generation. There were no confirmed diagnoses. In 13 cases, the system failed to establish the correct diagnosis as diagnostic hypothesis.

FIG. 2. Ranking of 48 clinical diagnoses being candidates for hypotheses generation in CADIAG-2. The hypotheses were established on the basis of complete patient data including history, physical examination, lab tests, and clinical investigations. The threshold ε was set to 0.05 allowing a liberal hypothesis generation. Three of the original 51 diagnoses were confirmed and not any longer candidates for hypothesis generation. In 1 case, the system failed to establish the correct diagnosis as diagnostic hypothesis.

5.2. Evaluation results. Figure 1 shows the ranking of the diagnoses proposed by CADIAG-2 in comparison with the 51 gold standard diagnoses where only the reduced set of patient data, which mostly consisted of more or less unspecific findings such as *fever, weakness, vomiting, increased blood sedimentation rate*, and so on, was applied. The threshold ε was set to 0.05, that is, a very low threshold allowing liberal hypotheses generation. There were no confirmed diagnoses. In 32 cases (62.8%), the gold standard was the first diagnosis in the ranked list of diagnostic hypotheses. In 6 cases (11.8%), the gold standard was at the second, third, or fourth place. In 13 cases (25.4%), the system failed to establish the gold standard as a diagnostic hypothesis at all. This can be explained by lack of specific findings in these cases, especially for patients with *chronic pancreatitis*.

In a next step, the same cases were tested but now with the complete set of patient data. Expectedly, the results improved because more data containing more specific information were available. In 3 cases, *pancreatic cancer* was confirmed by positive cytology and was no longer a candidate for diagnostic hypotheses generation. Figure 2 depicts the results of the remaining 48 cases in which the generation of the gold standard as diagnostic hypothesis was still possible. In 41 cases (85.4%), the gold standard was the first diagnosis in the ranked list of diagnostic hypotheses. In 6 cases (12.5%), the gold standard was at the second, third, fourth, or fifth place, and only in 1 case (2%), the gold standard was not among the generated hypotheses. Here, the reason was an incomplete medical record of the patient.

The results may be summarized in the following way: On the basis of patient's history, physical examination results, and results from basic lab tests, we obtained diagnostic accuracies of 62.8% if we demand that the correct diagnosis is at least the *first* hypothesis, and of 72.6% if we expect the correct diagnosis at least *among* the first three hypotheses. By adding results from further diagnostic investigations to the available finding list of the patient, the diagnostic results improved drastically, as can be expected. We obtained diagnostic accuracies of 86.3% (correct diagnosis either confirmed or the *first* hypothesis) and of 96.1% (correct diagnosis either confirmed or *among* the first three hypotheses).

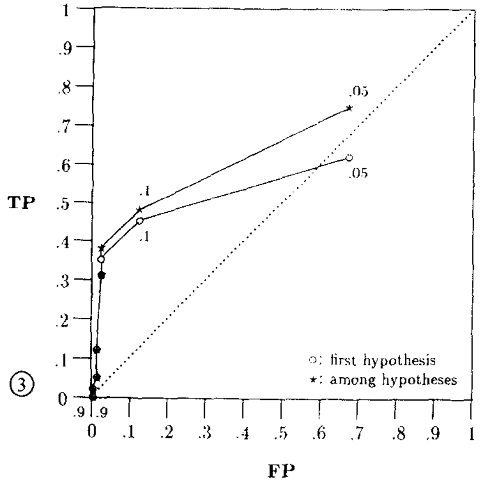
Figures 3–10 show the ROC curves obtained for various sets of gold standard diagnoses (Figs. 3 and 4 for all tested 51 diagnoses, Figs. 5 and 6 for 22 *pancreatic cancer* diagnoses, Figs. 7 and 8 for 13 *chronic pancreatitis* diagnoses, and Figs. 9 and 10 for 8 *acute pancreatitis* diagnoses). Each figure contains two curves. One of the curves, marked by stars, shows the ROC curves under the assumption that true positive means the gold standard is *among* the diagnostic hypotheses. The other curve, marked by circles, has as underlying assumption that true positive means the gold standard is the *first* hypothesis in the ranked list of diagnostic hypotheses, therefore a more rigid criterion.

As can be seen, the curve obtained by applying the more rigid decision criterion shows a worse performance than the other, but only in the section of the curves with a low threshold which allows many hypotheses to be generated. By making the threshold higher, the number of false positive results diminishes, so, if a hypothesis remains in the list it usually is the correct diagnosis.

Each pair of figures (Figs. 3 and 4, Figs. 5 and 6, Figs. 7 and 8, Figs. 9 and 10) were the result of testing the same cases but, first, with the reduced set of patient data (Figs. 3, 5, 7, and 9) and, second, with the complete set of patient data (Figs. 4, 6, 8, and 10).

The accuracy always increased by adding more specific data to the set of patient's findings, which concurs with the anticipated performance of such a system, especially in the case of *chronic pancreatitis* (compare Figs. 7 and 8). As can be seen from Fig. 7, *chronic pancreatitis* was systematically underrated (under the dotted line) if only unspecific medical data of the patient were available. Diseases such as *acute pancreatitis* or *pancreatic cancer* usually obtained a higher rank by CADIAG-2 in this early stage of the diagnostic process, but one might argue that this is even desirable because clarifying acute or prognostically serious illnesses should be given priority.

CADIAG-2/PANCREAS was designed in such a way that the threshold ε, which is the point of operation on the ROC curve, has as an initial value 0.10.



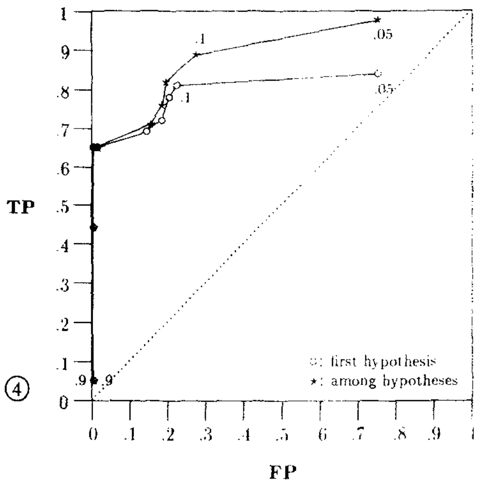


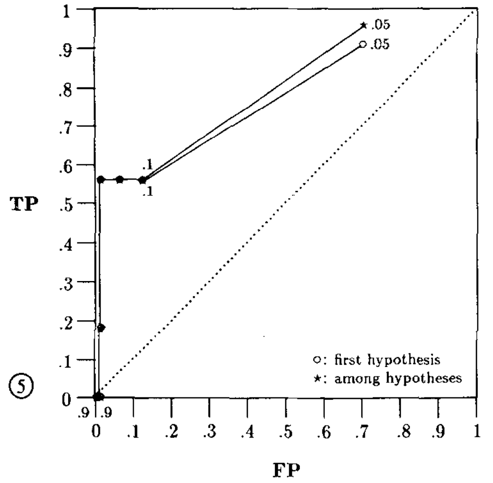
FIG. 3. ROC curves (TP, true positive ratio vs FP, false positive ratio) calculated from all 51 diagnoses being candidates for hypotheses generation. The hypotheses were established on the basis of patient history, physical examination, and basic lab tests only. Two decision criteria were applied: the gold standard diagnosis is the *first* hypothesis (line with circles), and the gold standard diagnosis is *among* the hypotheses (line with stars). The threshold ε assumed the values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. A result under the 45° dotted line is a systematic underrating of the correct diagnoses.

FIG. 4. ROC curves (TP, true positive ratio vs FP, false positive ratio) calculated from the remaining 48 diagnoses being candidates for hypotheses generation (in three cases, confirmed diagnoses were obtained). The hypotheses were established on the basis of complete patient data including history, physical examination, lab tests, and clinical investigations. Two decision criteria were applied: the gold standard diagnosis is the *first* hypothesis (line with circles), and the gold standard diagnosis is *among* the hypotheses (line with stars). The threshold ε assumed the values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. A result under the 45° dotted line would be a systematic underrating of the correct diagnosis.

However, this value can be actively varied by the physician during the actual consultation session and thus the sensitivity and specificity of the entire diagnostic system can be changed. The physician can adapt the threshold to the available patient data and has thus control over the extent of the hypotheses generation. This leads to an interactive optimization of the decision results in such a way that the physician may start with a low threshold to obtain a broad spectrum of possible diagnoses but increases the value in the subsequent iterations when more specific findings are available and the diagnostic possibilities narrow.

6. DISCUSSION

Computer-assisted diagnosis of pancreatic diseases has been an issue of clinical interest for more than 10 years. The following should give a brief overview on some of the attempts made so far.



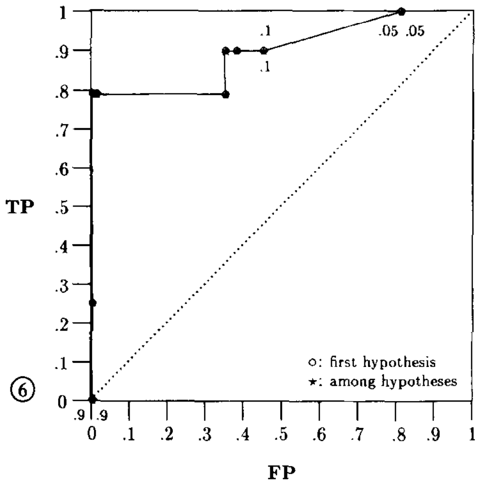
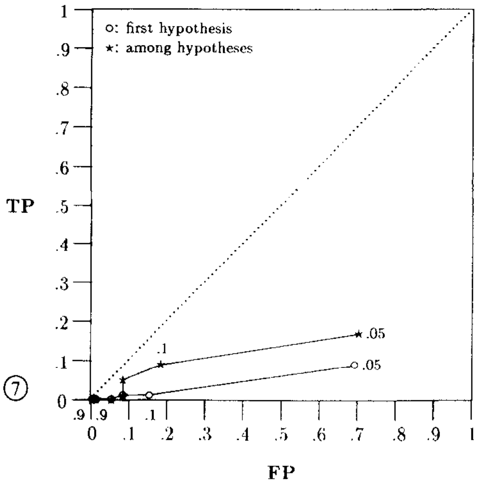


FIG. 5. ROC curves (TP, true positive ratio vs FP, false positive ratio) calculated from all 22 *pancreatic cancer* diagnoses being candidates for hypotheses generation. The hypotheses were established on the basis of patient history, physical examination, and basic lab tests only. Two decision criteria were applied: the gold standard diagnosis is the *first* hypothesis (line with circles), and the gold standard diagnosis is *among* the hypotheses (line with stars). The threshold ε assumed the values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. A result under the 45° dotted line would be a systematic underrating of *pancreatic cancer*.

FIG. 6. ROC curves (TP, true positive ratio vs FP, false positive ratio) calculated from the remaining 19 *pancreatic cancer* diagnoses being candidates for hypotheses generation (in three cases, *pancreatic cancer* was confirmed by *positive cytology*). The hypotheses were established on the basis of complete patient data including history, physical examination, lab tests, and clinical investigations. Two decision criteria were applied: the gold standard diagnosis is the *first* hypothesis (line with circles), and the gold standard diagnosis is *among* the hypotheses (line with stars). The threshold ε assumed the values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. A result under the 45° dotted line would be a systematic underrating of *pancreatic cancer*.

In studies carried out by Thurmayr *et al.* (*31–34*), several multivariate statistical approaches such as discriminante analysis, cluster analysis, and factor analysis were applied to aid in the diagnosis of pancreatic diseases. In these applications, disease entities to be discriminated were such as *pancreatic cancer*, *chronic, calcifying pancreatitis*, *chronic, recurring pancreatitis without radiological evidence of calcification*, and *acute pancreatitis*. Diagnoses were established on the basis of pancreatic function tests, where the number of test results taken into consideration ranges from 13 to 22. As is reported in (*33*), the correct classification rate obtained by applying a nonlinear discriminant analysis with 13 selected variables to different disease group arrangements was 46% for discriminating between six groups (32 cases with *normal pancreatic function test*, 21 cases with *pancreatic cancer*, 41 cases with *calcifying pancreatitis*, 64 cases with *chronic, recurring pancreatitis*, 36 cases with *acute gastric* or *duodenal ulcer*, and 48 cases with *chronic ulcer*), 77% for three groups (32



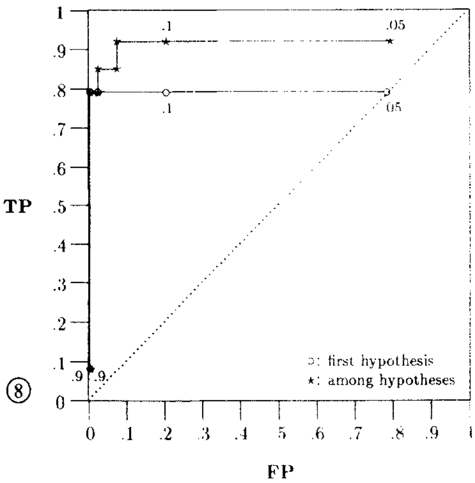


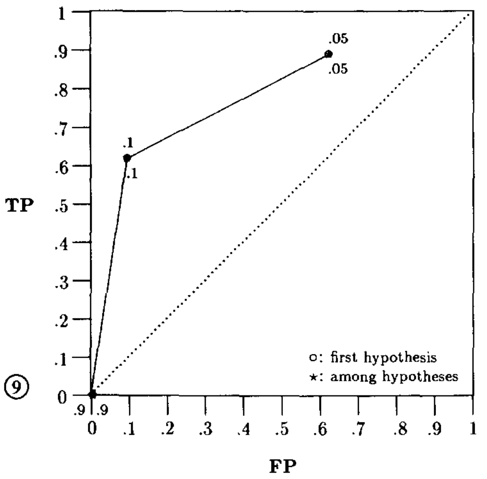
FIG. 7. ROC curves (TP, true positive ratio vs FP, false positive ratio) calculated from all 13 *chronic pancreatitis* diagnoses being candidates for hypotheses generation. The hypotheses were established on the basis of patient history, physical examination, and basic lab tests only. Two decision criteria were applied: the gold standard diagnosis is the *first* hypothesis (line with circles), and the gold standard diagnosis is *among* the hypotheses (line with stars). The threshold ε assumed the values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. A result under the 45° dotted line is a systematic underrating of *chronic pancreatitis*.

FIG. 8. ROC curves (TP, true positive ratio vs FP, false positive ratio) calculated from all 13 *chronic pancreatitis* diagnoses being candidates for hypotheses generation. The hypotheses were established on the basis of complete patient data including history, physical examination, lab tests, and clinical investigations. Two decision criteria were applied: the gold standard diagnosis is the *first* hypothesis (line with circles), and the gold standard diagnosis is *among* the hypotheses (line with stars). The threshold ε assumed the values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. A result under the 45° dotted line would be a systematic underrating of *chronic pancreatitis*.

healthy subjects, 126 cases with *pancreatic diseases*, 84 cases with *ulcer*), and 96% for a two-group discrimination (126 cases *with pancreatic diseases* and 126 cases *without pancreatic diseases*).

In (*35*), Durbec *et al.* report on an application of screening radiological signs to indicate pancreas pathologies. The screening was based on some selected binary variables found to be significant for the respective pathology. Four groups of pancreatic diseases were distinguished for this purpose: *pancreatic cancer, chronic calcifying pancreatitis, noncalcifying pancreatitis*, and *probable pancreatitis*.

Another application of computer-assisted diagnosis of pancreatic disorders is described in Boda and Pap (*36*). Multivariate statistics and pattern recognition methods were applied to separate patients with *pancreas insufficiency* from *healthy control subjects*. This two-group differentiation is carried out on the basis of 14 laboratory parameters. The application of a linear discriminant analysis to 71 cases yielded an accuracy of 92.2%, a nearest neighbor method



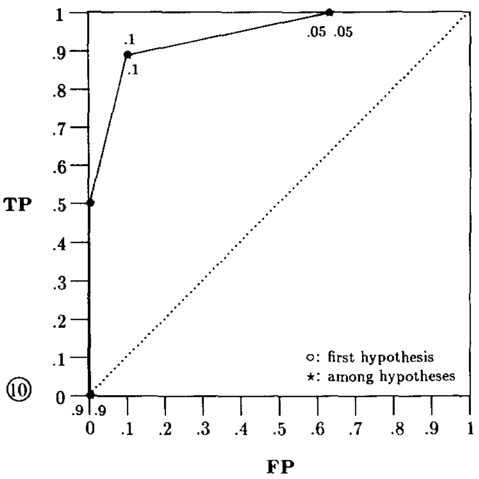


FIG. 9. ROC curves (TP, true positive ratio vs FP, false positive ratio) calculated from all 8 *acute pancreatitis* diagnoses being candidates for hypotheses generation. The hypotheses were established on the basis of patient history, physical examination, and basic lab tests only. Two decision criteria were applied: the gold standard diagnosis is the *first* hypothesis (line with circles), and the gold standard diagnosis is *among* the hypotheses (line with stars). The threshold ε assumed the values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. A result under the 45° dotted line would be a systematic underrating of *acute pancreatitis*.

FIG. 10. ROC curves (TP, true positive ratio vs FP, false positive ratio) calculated from all 8 *acute pancreatitis* diagnoses being candidates for hypotheses generation. The hypotheses were established on the basis of complete patient data including history, physical examination, lab tests, and clinical investigations. Two decision criteria were applied: the gold standard diagnosis is the *first* hypothesis (line with circles), and the gold standard diagnosis is *among* the hypotheses (line with stars). The threshold ε assumed the values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. A result under the 45° dotted line would be a systematic underrating of *acute pancreatitis*.

applied to 62 cases reached an 80.1% accuracy, and a centroid method with 66 cases showed an 85.7% accuracy. A reduction of the considered laboratory parameter to only 7 yielded a change of the accuracy rates to 88.3%, 80.5%, and 85.7%, respectively.

Apart from these applications dedicated to computer-assisted diagnosis of pancreatic diseases only, there are several other systems that include pancreatic diagnoses.

One of them is the system for diagnosing *acute abdominal pain* developed by de Dombal *et al.* (*4, 37, 38*). It contains *acute pancreatitis* as one of the major disease categories.

Systems that are aimed at medical diagnosis in the entire field of internal medicine necessarily contain also pancreatic diseases.

The INTERNIST-I/CADUCEUS system (*39–41*) (consequently also the QMR system (*6*) employing the same knowledge base as INTERNIST-I/CADUCEUS) contains at present 8 pancreatic diseases. In detail, these are *pan-*

creatitis acute, *pancreatitis chronic*, *pancreatic pseudocyst*, *carcinoma of body or tail of pancreas*, *carcinoma of head of pancreas*, *glucagonoma*, *insulinoma*, and *Zollinger–Ellison syndrome*.

The diagnostic prompting system RECONSIDER (*42–44*) comprises the largest number of pancreatic diseases. Twenty pancreatic diseases are included into RECONSIDER (from (*45*)): *Zollinger–Ellison syndrome*; *pancreas, aberrant*; *pancreas, abscess*; *pancreas, adenocarcinoma, body and tail*; *pancreas, annular*; *pancreas, atrophy*; *pancreas, adenocarcinoma, head*; *pancreas, cyst, congenital*; *pancreas, cystadenocarcinoma*; *pancreas, cyst, false*; *pancreas, cystadenoma*; *pancreas, cyst, true*; *pancreas, injury*; *pancreatitis, acute*; *pancreas, islet-cell tumor*; *pancreatitis, chronic*; *pancreatitis, hereditary*; *pancreas, postoperative*; *pancreas, hemorrhagic, acute*; and *Verner–Morrison islet-cell tumor syndrome*.

The approach taken in CADIAG-2/PANCREAS lead to the inclusion of 10 pancreatic diseases. After completion of the documentation and evaluation of the other differential diagnostic groups from the area of gastroenterology and hepatology (diseases of the esophagus, stomach, liver, gallbladder and biliary tract, intestine, colon, and peritoneum), we will merge these single groups to a larger system for gastroenterology and hepatology.

The heterogeneous number of pancreatic diagnoses included into the abovementioned computer systems is likely to result from the more or less controversial and continuously changing medical nomenclature that has been established according to etiology, anatomy, and histopathology (cf. ICD-9 (*46, 47*) and ICD-9/CM (*48, 49*)). For a computer-based diagnostic system, classification of diseases according to therapeutically and prognostically distinct entities seems most appropriate. Application of both the INTERNIST-I/CADUCEUS and the CADIAG-2 system will allow sufficient precise and clinically relevant differential diagnosis in the hospital and the physician's office.

Last but not least, an application of ROC curves to evaluate clinical performance is described in (*50*). The aim of those studies was to evaluate the effect of computer confidence of threshold levels and to assess clinical performance in the diagnosis of *acute appendicitis*.

The evaluation described in this paper was done to prepare the clinical application of CADIAG-2 in the field of gastroenterology and hepatology. CADIAG-2 is aimed at an application in two subsequent phases (as described in (*9*) in more detail):

(1) as an early, data-activated automatic screening and monitoring procedure for detecting pathological states in the patient, for generating diagnostic hypotheses, and for proposing further useful examinations; and

(2) as an on-line consultation system for the clinician which assists him in clarifying patient's disorders completely and in great detail.

The results obtained so far are considered to be satisfactory. We included 10 pancreatic disease entities into the expert system which—compared with other, similar systems—seem to be a good choice. It was possible to formally describe these diseases by disease profiles including findings from *all* areas of investigation. Two important medical relationships were selected to define the association between two medical entities: the frequency of occurrence (necessity) and the strength of confirmation (sufficiency) degrees. An inference mechanism combining a fuzzy logical and a heuristic approach to conclude diagnoses with degrees of certainty was applied. This algorithm also considered the possible vagueness of patient's medical findings. The diagnostic hypotheses ranking according to the heuristically calculated support score proved to be quite successful as was shown by the obtained accuracy rates. The overall performance of CADIAG-2 is reflected by ROC curves. This measure is independent of internal decision criteria, which—if changed—usually alter the outcome and thus the overall accuracy of a system. It is argued that it seems useful to provoke change of the performance of an expert system by varying internal decision criteria, which—in some form—are usually part of the system, and establish the system's general degree of discrimination. By analyzing these, one might say, systems' characteristic curves, it is possible to compare the systems' capacities for discrimination of different medical expert systems and to determine their optimal points of operation.

Acknowledgments

We thank J. Yen, G. Tusch, and W. Moser for many helpful discussions on the presented subject. We are also grateful to I. Gröger and G. Sedivy for their valuable assistance in the preparation of the patient data and the medical knowledge.

References

1. Pryor, T. A., Gardner, R. M., Clayton, P. D., and Warner, H. R. The HELP system. *J. Med. Syst.* **7,** 87 (1983).
2. Aikins, J. S., Kunz, J. C., Shortliffe, E. H., and Fallat, R. J. PUFF: An expert system for interpretation of pulmonary function data. *Comput. Biomed. Res.* **16,** 199 (1983).
3. Shortliffe, E. H., Scott, A. C., Bischoff, M. B., Campell, A. B., van Melle, W., and Jacobs, C. D. An expert system for oncology protocol management. *In* "Rule-Based Expert Systems—The MYCIN Experiments of the Stanford Heuristic Programming Project" (B. G. Buchanan and E. H. Shortliffe, Eds.), pp. 653–665. Addison–Wesley, Reading, MA, 1984.
4. de Dombal, F. T. Computers and decision-making: An overview for gastroenterologists. *In* "Computer Aid in Gastroenterology" (P. Rozen and F. T. de Dombal, Eds.), pp. 119–133. Karger, Basel, 1984.
5. Miller, R. A., Pople, H. E., and Myers, J. D. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. *N. Engl. J. Med.* **307,** 468 (1982).
6. First, M. B., Soffer, L. J., and Miller, R. A. QUICK (QUick Index to Caduceus Knowledge): Using the internist-1/caduceus knowledge base as an electronic textbook of medicine. *Comput. Biomed. Res.* **18,** 137 (1985).
7. Tuttle, M. S., Sherertz, D. D., Blois, M. S., and Nelson, N. "Expertness" from structured text? RECONSIDER: A diagnostic prompting program. *In* "Proceedings, Conference on Applied Natural Language Processing," pp. 124–131. Santa Monica, CA, 1983.
8. Adlassnig, K.-P., Kolarz, G., Scheithauer, W., Effenberger, H., and Grabner, G. CADIAG: Approaches to computer-assisted medical diagnosis. *Comput. Biol. Med.* **15,** 315 (1985).

9. Adlassnig, K.-P., Kolarz, G., Scheithauer, W., and Grabner, G. Approach to a hospital-based application of the medical expert system CADIAG-2. *Med. Inf.* **11,** 205 (1986).

10. Liebowitz, J. Useful approach for evaluating expert systems. *Exp. Syst.* **3,** 86 (1986).

11. Fieschi, M., and Joubert, M. Some reflections on the evaluation of expert systems in medicine. *Methods Inf. Med.* **25,** 15 (1986).

12. Adlassnig, K.-P., Scheithauer, W., and Grabner, G. Computerunterstützte medizinische Diagnostik und ihr Einsatz bei Pankreaserkrankungen. *Acta Med. Aust.* **11,** 125 (1984).

13. Adlassnig, K.-P., Scheithauer, W., and Grabner G. CADIAG-2/PANCREAS: An artificial intelligence system based on fuzzy set theory to diagnose pancreatic diseases. *In* "Proceedings, Third International Conference on System Science in Health Care," pp. 396–399, Springer-Verlag, Berlin, 1984.

14. Adlassnig, K.-P. A fuzzy logical model of computer-assisted medical diagnosis. *Methods Inf. Med.* **19,** 141 (1980).

15. Adlassnig, K.-P. Fuzzy set theory in medical diagnosis. *IEEE Trans. Syst. Man Cybern.* **SMC-16,** 260 (1986).

16. Kolarz, G., and Adlassnig, K.-P. Problems in establishing the medical expert systems CADIAG-1 and CADIAG-2 in rheumatology. *J. Med. Syst.* **10,** 395 (1986).

17. Akhavan-Heidari, M., and Adlassnig, K.-P. Preliminary results on CADIAG-2/GALL: A diagnostic consultation system for gallbladder and biliary tract diseases. *In* "Proceedings, Medical Informatics Europe '88," pp. 622–666, Springer-Verlag, Berlin, 1988.

18. Adlassnig, K.-P. Uniform representation of vagueness and imprecision in patient's medical findings using fuzzy sets. *In* "Proceedings, Cybernetics and Systems '88," pp. 685–692, Kluwer Academic Publishers, Dordrecht, 1988.

19. Adlassnig, K.-P., and Grabner, H. Verarbeitung natürlichsprachiger medizinischer Begriffe. *In* "WAMIS—Wiener Allgemeines Medizinisches Informations-Systems" (G. Grabner, Ed.), pp. 162–189, Springer-Verlag, Berlin, 1985.

20. Arnett, F. C., Edworthy, St. M., Bloch, D. A., McShane, D. J., Fries, J. F., Cooper, N. S., Healey, L. A., Kaplan, St. R., Liang, M. H., Luthra, H. S., Medsger, Th. A., Jr., Mitchell, D. A., Neustadt, D. A., Pinals, R. S., Schaller, J. G., Sharp, J. T., Wilder, R. L., and Hunder, G. G. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum.* **33,** 315 (1988).

21. Adlassnig, K.-P., and Kolarz, G. Representation and semiautomatic acquisition of medical knowledge in CADIAG-1 and CADIAG-2. *Comput. Biomed. Res.* **19,** 63 (1986).

22. Barachini, F. "Konsistenzprüfung von Wissensbasen medizinischer Expertensysteme." Ph.D. dissertation, Technical University Vienna, Vienna, 1984.

23. Barachini, F., and Adlassnig, K.-P. CONSDED: Medical knowledge base consistency checking. *In* "Proceedings, Medical Informatics Europe '87," pp. 951–956, EFMI, Rom, 1987.

24. Adlassnig, K.-P., Scheithauer, W., and Kolarz, G. Fuzzy medical diagnosis in a hospital. *In* "Fuzzy Logic in Knowledge Engineering" (H. Prade and C. V. Negoita, Eds.), pp. 275–294, Verlag TÜV, Rheinland, 1986.

25. Lusted, L. B. "Introduction to Medical Decision Making." Thomas, Springfield, IL, 1968.

26. Swets, J. A. The relative operating characteristic in psychology. *Science* **182,** 990 (1973).

27. Komaroff, A. L. The variability and inaccuracy of medical data. *Proc. IEEE* **67,** 1196 (1979).

28. Centor, R. M., and Schwartz, J. S. An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. *Med. Decision Making* **5,** 149 (1985).

29. Centor, R. M. A visicalc program for estimating the area under a receiver operating characteristic (ROC) curve. *Med. Decision Making* **5,** 139 (1985).

30. England, W. L. An exponential model for optimal threshold selection on ROC curves. *Med. Decision Making* **8,** 120 (1988).

31. Thurmayr, R., Otte, M., and Thurmayr, G. R. Computer aid for diagnosing pancreatic, hepatic and gastric diseases by pancreatic function test. *In* "Proceedings, MEDINFO 74," pp. 607–612, North-Holland, Amsterdam, 1974.

32. Thurmayr, R., Blomer, R. J., Forell, M. M., Jaffé, A., Otte, M., Raschewa, C., and Thurmayr, G. R. Computer aided diagnosis of pancreatic function tests in the routine situa-

tion. *In* "Decision Making and Medical Care" (F. T. de Dombal and F. Grémy, Eds.), pp. 175–182. North-Holland, Amsterdam, 1976.

33. Schäffer, J. B., Thurmayr, R., Thurmayr, G. R., and Otte, M. Auswertung von Daten aus dem Pankreasfunktionstest mit drei multivariaten Methoden: Clusteranalyse, nichtlineare Diskriminanzanalyse und Faktorenanalyse. *EDV Med. Biol.* **10,** 54 (1979).
34. Thurmayr, R., Thurmayr, G. R., Otte, M., and Forell, M. M. Computer aid for the screening test of pancreatic diseases. *In* "Computers and Mathematical Models in Medicine" (D. Cardus and C. Vallbona, Eds.), pp. 232–238. Springer-Verlag, Berlin, 1981.
35. Durbec, J.-.P., Cornée, J., and Berthezene, P. Data screening methods—Application to differential diagnosis in pancreatic pathology from radiological signs. *Methods Inf. Med.* **17,** 36 (1978).
36. Boda, K., and Pap, A. Diagnostics of pancreatic insufficiency using multivariate statistical and pattern recognition methods. *Comput. Biol. Med.* **14,** 91 (1984).
37. de Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P., and Horrocks, J. C. Computer-aided diagnosis of acute abdominal pain. *Brit. Med. J.* **2,** 9 (1972).
38. Leaper, D. J., Horrocks, J. C., Staniland, J. R., and de Dombal, F. T. Computer-assisted diagnosis of abdominal pain using "estimates" provided by clinicians. *Brit. Med. J.* **4,** 350 (1972).
39. Miller, R. A., Pople, H. E., and Myers, J. D. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. *N. Engl. J. Med.* **307,** 468 (1982).
40. Miller, R. A. INTERNIST-1/CADUCEUS: Problems facing expert consultant programs. *Methods Inf. Med.* **23,** 9 (1984).
41. Masarie, F. E., Miller, R. A., and Myers, J. D. INTERNIST-I Properties: Representing common sense and good medical practice in a computerized medical knowledge base. *Comput. Biomed. Res.* **18,** 458 (1985).
42. Tuttle, M. S., Sherertz, D. D., Blois, M. S., and Nelson, N. "Expertness" from structured text? RECONSIDER: A diagnostic prompting program. *In* "Proceedings, Conference on Applied Natural Language Processing," pp. 124–131. Santa Monica, CA, 1983.
43. Blois, M. S. "Information and Medicine." Univ. of California Press, Berkeley, 1984.
44. Nelson, S. J., Blois, M. S., Tuttle, M. S., Erlbaum, M., Harrison, P., Kim, H., Winkelmann, B., and Yamashita, D. Evaluation RECONSIDER—A computer program for diagnostic prompting. *J. Med. Syst.* **9,** 379 (1985).
45. Gordon, B. L. (Ed.). "Current Medical Information and Terminology." American Medical Association, Chicago, IL, 1971.
46. "International Classification of Diseases," 9th Rev., Vol. 1. World Health Organization, Geneva, 1975.
47. "International Classification of Diseases," 9th Rev., Vol. 2, Alphabetical Index. World Health Organization, Geneva, 1975.
48. "ICD-9-CM The International Classification of Diseases," 9th Rev., Clinical Modification, Vol. 1, Diseases-Tabular List. Edwards Brothers, Inc., Ann Arbor, MI, 1980.
49. "ICD-9-CM The International Classification of Diseases," 9th Rev., Clinical Modification, Vol. 2, Diseases-Alphabetic Index. Edwards Brothers, Inc., Ann Arbor, MI, 1980.
50. de Dombal, F. T., and Horrocks, J. C. Use of receiver operating characteristic (ROC) curves to evaluate computer confidence threshold and clinical performance in the diagnosis of appendicitis. *Methods of Inf. Med.* **17,** 157 (1978).