COMPUTER-AIDED MEDICAL DIAGNOSIS: LITERATURE REVIEW

W. ROGERS, B. RYACK and G. MOELLER

Naval Submarine Medical Research Laboratory, Naval Submarine Base, Groton, CT (U.S.A.)

(Received: 15 December, 1978)

SUMMARY

The difficulty of the medical diagnostic task and the advantages of the computer as an aid in this task are discussed. The general strategy and structure of any computer-aided system is presented, and the relationship of diagnostic accuracy to key variables involved in the development, test and use of a computer-aided diagnostic system is examined. These variables include: the computer algorithm, the source of the information used to develop the data base, the number and type of diseases under investigation, the number and type of indicants used, the source of the test sample, and the source of the validated diagnosis. A table of 58 empirically tested computer-aided medical diagnostic systems is presented; each system is summarised in relation to the variables mentioned above and diagnostic accuracy.

SOMMAIRE

On discute de la difficulté du travail de diagnostic médical et des avantages de l'ordinateur en tant qu'auxilliaire. On décrit la stratégie générale et la structure de tout système à ordinateur auxilliaire puis on examine la relation entre la précision du diagnostic, les variables fondamentales impliquées dans la conception, l'essai et l'utilisation d'un système à ordinateur auxilliaire. Ces variables comprennent: les algorithmes de calcul, la source d'information utilisée pour construire la base de données, le nombre et le type des maladies étudiées, le nombre et le type des indicateurs, la source de l'échantillon d'essai et celle des diagnostics confirmés. On présente un tableau de 58 systèmes de diagnostic médical à ordinateur auxilliaire; on décrit rapidement chaque système en tenant compte des variables signalées plus haut et de la précision du diagnostic.

Int. J. Bio-Medical Computing (**10**) (1979) 267–289

1. INTRODUCTION

Medical diagnosis is a difficult and complex task largely empirically based and poorly understood as an intellectual task. The gap between the information which exists for diagnosis and that accessible from memory is difficult to close even for a highly trained general practitioner with substantial daily exposure to many disorders. The computer-based systems described in this review were created to help the physician bridge this gap.

The impetus for this review comes from the still greater problem that arises when an individual trained in the allied health sciences must serve in a position normally assigned to a physician. The Navy hospital corpsman assigned to independent duty aboard a submarine is such an individual. As the only person aboard the submarine trained in medical science, he is totally responsible for solution of all medical problems that arise. Paradoxically, the youth and general good health of the crew aggravate the problem of diagnosis for the corpsman by severely limiting his practical experience with serious disease. The lower levels of training and experience of the corpsman relative to the physician imply that properly designed computer-based aids to diagnosis should be particularly useful in the corpsman's practice of submarine medicine.

The development of computer-aids to medical diagnosis has been enhanced by the fact that the computer has several inherent capabilities which seem ideally fitted to medical problem-solving. Paraphrasing Gorry and Barnett (1968), the principal advantages of the computer are its ability to: store large quantities of data without distortion over long periods of time; recall data, on receipt of the appropriate message, exactly as stored; perform complex logical and mathematical operations at very high speed; and display many diagnostic possibilities in an orderly fashion. A computer-aided diagnostic system could incorporate features to offset other limitations experienced by human diagnostic problem solvers. The limitations of man as an effective problem solver have been repeatedly demonstrated (Streufert, 1970; Newell and Simon, 1972; Janis and Mann, 1977). Newell and Simon (1972) found the limited capacity of short term memory to be a major deterrent to effective problem solving. It has been noted (Streufert, 1970) that in seeking and selecting data to evaluate an on-going situation men tend, on one hand, to gather information indiscriminately, resulting in an accumulation of more information than can be used effectively in problem solving, and on the other hand, to restrict search to only a limited subset of the alternatives relevant to the problem at hand. Janis and Mann (1977) have discussed the problems encountered by man as '... a reluctant decision maker—beset by conflict, doubts, and worry, struggling with incongruous longings, antipathies, and loyalties...'. In addition to compensating for the human limitations discussed above, computer-aided diagnosis promises needed insight into physicians' thought processes (Pauker *et al.*, 1976) and a resulting better understanding of the human diagnostic process.

While most computer-aided diagnostic systems have not been developed beyond the experimental stage, preliminary results indicate that the computer can be a useful diagnostic aid to the physician. The ability of many computer aided systems to empirically diagnose diseases as well as the average physician demonstrates that the computer does in fact possess those capabilities mentioned earlier which make it well suited as an aid in the diagnostic decision making process.

Although many different approaches and strategies have been employed to accomplish computer-aided diagnoses, the basic configuration of a typical system is well-defined. The components involved in developing, validating and using a computer-aided diagnostic system are shown in Fig. 1. The components include the computer data base, the computer algorithm, and an interactive program for communication between the machine and the user (i.e., physician, corpsman). The computer data base consists of disease-symptom relationships, disease probabilities, and depending on the particular system, other medical information pertinent to diagnoses and treatment of the particular diseases involved (i.e., drug interactions, further diagnostic tests). The computer algorithm is composed of the logical or statistical processes used to derive a solution to a diagnostic problem from the information included in the data base and the information obtained from the new patient through history, physical exam, laboratory tests, etc. An interactive program for man-machine communication allows the user who is unfamiliar with computer programming to interact with the computer in order to input the necessary patient information and to obtain the diagnostic output generated by the computer. In a well-planned system this interaction can include much more than simple data input and diagnostic output, as the user should be able to question the logic and data on which the computer bases a certain decision, or clarify a particular medical definition or laboratory procedure.

In the experimental stage of development, the accuracy of the computer generated diagnoses must be validated. Therefore, in addition to the major components of the computer-aided system already discussed, an external diagnosis and feedback loop is needed. By obtaining independent diagnoses from the most reliable

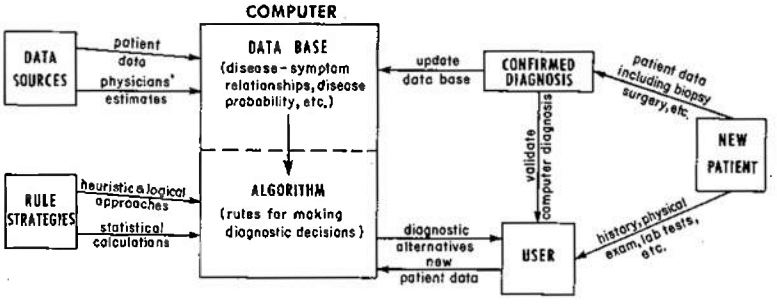

Fig. 1. Typical configuration of a computer-aided diagnostic system—development, test, and use.

source external to the computer system; one can assess the validity of the computer generated diagnoses. By using the validated disease-symptom information obtained from each new patient to update the data base, the system becomes dynamic and self-correcting.

While the basic external configuration of computer-aided diagnosis presented in Fig. 1 fits most present systems, the actual structure and content of the individual components vary greatly from system to system. There are enough computer diagnostic system applications differing in their internal structure and content, and in the results they produce, to make a summary and general comparison of the literature worthwhile. This may provide a basis for further development of computer-aided diagnostic techniques. This review concentrates specifically on reports which include empirical tests of a working computer diagnostic system. There is substantial literature which is primarily concerned with theoretical issues relevant to computer-aided diagnosis (Ledley and Lusted, 1959; Gorry and Barnett, 1968; Lusted, 1968; Croft, 1972; Gorry, 1973; Fisher *et al.*, 1975) which will not be covered here. The ability of a computer system to generate a state-of-the-art or better diagnostic accuracy is crucial to one concerned with the implementation of such a system. For this reason, the factors which seem significant for the implementation of an effective diagnostic system are summarised for each empirically tested system reviewed. The factors included in this summary analysis are: the computer algorithm; the source of the computer data base; diseases included in the data base; type and number of indicants (history, signs, symptoms, tests, etc.) included in the data base; the source of the test sample; and the method of validating the computer diagnosis. The diseases included in the data base are summarised in three ways: the general disease class to which they belong according to the International Classification of Diseases, Adapted for use in the United States—Eighth Revision (ICDA); the specific disease category or major symptom which best describes the disease problem explored; and the actual number of diseases included in the system. In addition, the first author, the year, and the diagnostic accuracy of each study are reported.

Fifty-eight studies have been reviewed in relation to the factors mentioned above. This review has been summarised in the Appendix. Each factor and its relation to diagnostic accuracy will be discussed individually.

2. THE COMPUTER ALGORITHM

The computer algorithm is considered by most researchers to be the heart of the computer diagnostic system. Many algorithms have been proposed, differing in their basic assumptions, method of attack, data requirements and adequacy of simulation of the diagnostic thought process. Although algorithm nomenclature is inconsistent from author to author, and it is difficult to accurately categorise

many algorithms, it is useful to dichotomise computer diagnostic algorithms for purposes of discussion. Algorithms will be referred to here as statistical or logical. Statistical refers to algorithms which calculate the most likely diagnosis from explicit statistical analysis of disease-symptom frequencies and disease probabilities. Logical refers to algorithms which usually proceed in a sequential branching fashion; a decision is made at each step based on a logical 'if A, then B' or similar type reasoning. This type of algorithm generally simulates the human diagnostic process more closely than the statistical models do. Although it has been argued that logical models are ultimately based on the statistical experience of clinicians, and in this sense are 'statistical' (Fisher *et al.*, 1975), we feel the absence of explicit statistical computations in these models makes them categorically different from those models generally referred to as 'statistical'.

2.1 *Statistical approaches*

Statistical approaches dominate the literature in computer-aided diagnosis. Three of the most common statistical models are: conditional probability based on Bayes' theorem; linear discriminant functions; and matching procedures. Conditional probability based on Bayes' theorem is the simplest and most widely used computer algorithm. A basic form of Bayes' theorem is:

$$P(D/S)=\frac{P(S/D)\times P(D)}{P(S)} \tag{1}$$

where P denotes the probability of occurrence, S represents all data about a patient in terms of symptoms, signs and diagnostic tests, D represents a disease, a set of diseases or a normal health state, and the notations $P(D/S)$ and $P(S/D)$ signify the probability of D given S and S given D respectively (Lusted, 1968). Strictly speaking, Bayes' theorem requires that $P(D)$, $P(S)$, and $P(S/D)$ be derived from subjective estimates. In practice, there is substantial disagreement between investigators, the values being derived subjectively for some systems and based on empirical data for others.

Linear discriminant functions, originally developed by Fisher (1936) for application to a taxonomic problem, basically distinguishes to which of 2 possible groups an individual belongs. This discrimination is based on a set of normally distributed measurements. In applications to medical diagnosis this discriminant function is calculated so as to give the lowest possible probability of misdiagnosis. Croft and Machol (1974) point out that there are a number of discriminant functions in use, including linear discriminant functions and Bayes' conditional probability, which are very similar both in their theoretical assumptions and empirical results.

Matching procedures basically compare a patient's symptom profile with every one in the data base or a calculated average symptom profile representative of each disease in the data base. The most common of the matching procedures involves the assignment of a weight to each symptom for each disease. The

symptoms of a new patient are then summed according to their weight for each disease. The disease which produces the largest ratio of the patient's weighted symptoms to the weighted sum of all characteristics for that disease is considered the correct diagnosis. This general procedure is referred to by a variety of names, the most common of which is weight summation (Birk *et al.*, 1974). The strategy in assigning weights to symptoms and the precise calculations used to determine the final diagnosis vary from study to study, but the general procedure is the same. Matching models also include a variety of other procedures such as template matching, and pattern recognition. The almost careless use of terms by some authors creates a major problem in analysis of the literature. Standardisation of algorithm nomenclature is sorely needed.

2.2 *Logical approaches*

Logical algorithms include flow charts, sequential questioning methods, and decision tree approaches. These methods all have the same basic structure and will all be generally referred to as decision tree models in this review. Basically, a decision tree model is patterned after the classical differential diagnostic procedure and consists of a sequence of questions or test nodes, decision nodes, and binary branches. Typically, 2 alternatives are possible at each question or test node, and the alternative chosen automatically leads to a specific branching logic and particular questions, tests, and decisions in the tree structure. Test results required, questions asked, and conclusions made are determined by the individual or collective knowledge of the designers. This knowledge, or data base, is gathered either from real life incidence data or from opinions and estimates gained from various sources. It need not be as extensive as a statistical algorithm data base.

The theoretical aspects of some of the proposed algorithms have been extensively reviewed (Croft, 1972; Croft and Machol, 1974; Fisher *et al.*, 1975). In addition, specific theoretical problems have been addressed by others. Rector and Ackerman (1975) discussed the advantages and disadvantages of a sequential vs. non-sequential decision model. Since both the human diagnostic process and disease manifestation are sequential, it is argued that computer-aided diagnosis should be sequential. Norusis and Jacquez (1975a, 1975b) discussed the problems met when assuming independence of symptoms with Bayes' conditional probability. They argued that models which assume independent symptoms necessarily produce a substantial increase over the minimum misclassification rate when even small symptom dependencies exist. They proposed practical alternate models which take into account the dependence of symptoms. Gorry *et al.* (1973) discussed the costs of misdiagnosis in their program for management of acute renal failure. They observed that most models assume symmetrical costs of misdiagnosis, even though some misdiagnoses are obviously more costly than others, both monetarily and from the standpoint of patient well-being.

In our review of computer diagnostic applications (see Appendix), it was found

that 60% of all the studies used an algorithm based on Bayes' theorem. Algorithms based on Bayes' theorem, linear discriminant functions, matching procedures, and decision trees accounted for nearly 90% of all systems reviewed.

Studies which compare different algorithms used in computer diagnosis generally fail to show significant differences in relation to diagnostic accuracy when all other factors are held constant. A comparison of a Bayes' model and discriminant functions analysis on patients with upper abdominal pain (Scheinok and Rinaldo, 1968) reported a 1% difference in diagnostic accuracy. Birk *et al.* (1974) compared Bayes' probability with a weight summation model and found a difference of 4% in diagnostic accuracy. These trends are supported by most of the studies testing the accuracy of more than 1 algorithm on the same set of data (Boyle *et al.*, 1966; Nordyke *et al.*, 1971; Fleiss *et al.*, 1972; Hirschfeld *et al.*, 1974). The general finding that several algorithms work equally well in relation to the accuracy of the system is best documented by Croft (1972). In comparing 10 statistical algorithms used in computer-aided diagnosis on the same set of data, Croft found a 13% difference in diagnostic accuracy. He considered this difference insignificant in relation to the diagnostic differences caused by other factors in computer-aided diagnosis.

The effect of other factors on diagnostic accuracy becomes apparent when one notes that similar algorithms, when applied in different studies to different sets of data, yield drastically different diagnostic accuracies. Reports using Bayes' theorem varied in accuracy from 57% obtained by Meerten *et al.* (1971) in the diagnosis of asthma, asthmatic bronchitis, chronic bronchitis and emphysema to 100% obtained by Wilson *et al.* (1965) in the diagnosis of gastric ulcers and by Spicer *et al.* (1973) in the diagnosis of Crohn's disease and proctocolitis. Studies using linear discriminant functions varied in accuracy from 49% obtained by Ross and Dutton (1972) in the diagnosis of upper gastrointestinal diseases to 100% obtained by Spicer *et al.* (1973) in the diagnosis of Crohn's disease and proctocolitis. Differences in the number and type of diseases diagnosed probably are the major cause of cross-study variability in diagnostic accuracy. The consistency of diagnostic accuracy when using different algorithms on the same data, combined with the variability found in using the same algorithm on different data, suggests that selection of the appropriate algorithm in itself does not guarantee development of an effective computer-based diagnostic system.

3. THE COMPUTER DATA BASE

Factors to consider in constructing the computer data base are: the source of information for the data base; the diseases included in the data base (ICDA class, disease category and number of diseases); and the number and type of indicants to be used.

3.1 *Source of information*

The source of information for the data base is of major significance, since its accuracy has a direct influence on the accuracy of the diagnostic system itself. Among the possible sources of the computer data base are: medical textbooks; physicians' and experts' opinions and estimates; and hospital and emergency room medical records. Using a Bayes' algorithm, Birk *et al.* (1974) and Leaper *et al.* (1972) reported that with all other factors held constant, data bases generated from medical records produced more accurate diagnoses than those generated from physicians' opinions and estimates. Leaper *et al.* (1972) reported a diagnostic accuracy of 91.1% with a data base generated from medical records, and 82.2% with a data base generated from physicians' estimates and opinions. Birk *et al.* (1974) reported accuracies of 84% and less than 70% under the same respective conditions. This gives strong support for a data base generated from medical records. Gustafson *et al.* (1971) give evidence, however, that a data base developed from subjective probabilities performs as well as a data base developed from actuarial probabilities, and requires less time and cost for development.

When using real-life data, it is preferable to use as large a sample as possible to assure that the disease-symptom frequencies computed for each disease are based on a sufficiently large number of patient records. However, gathering a sufficiently large data base is often an arduous and time consuming task. If the medical records are collected retrospectively, as is the case in most studies, they are often non-standardised, incomplete and difficult to interpret. If collected prospectively, the medical data can be recorded on standardised forms, eliminating the problems inherent in retrospective collection. Unfortunately prospective medical records can be collected only as fast as patients with the diseases under study are admitted to a particular medical facility.

3.2 *Diseases included in data base*

(a) *ICDA class:* Computer diagnostic systems have been applied to a wide range of disease categories. We have used the ICDA to categorise disease areas covered in the literature. The articles reviewed in this report span diseases included in 12 of the 17 major disease classes of the ICDA (Table 1). Although a wide-range of diseases is addressed in the aggregate by the reports reviewed, each individual computer diagnostic system typically includes a very narrow range of diseases, usually involving only one ICDA class. Only 2 articles reviewed attempt to diagnose a wide range of disease spanning several disease classes (Brodman and Van Woerkom, 1966; Birk *et al.*, 1974). Further, computer diagnostic systems developed to date have concentrated on a very small number of ICDA classes; 35 of the 54 articles reviewed involve 3 of the 17 ICDA classes. Thirteen studies deal with class III.—Endocrine, Nutritional and Metabolic Diseases, 10 studies involve class V.—Mental Disorders, and 12 studies explore class IX.—Diseases of the Digestive System.

TABLE 1
NUMBER OF ARTICLES IN COMPUTER-AIDED DIAGNOSIS RELATING TO EACH ICDA CLASSIFICATION. 58 STUDIES

ICDA classification		*Number of studies*
I	Infective and parasitic diseases	0
II	Neoplasms	0
III	Endocrine, nutritional and metabolic diseases	13
IV	Diseases of the blood and bloodforming organs	2
V	Mental disorders	10
VI	Diseases of the nervous system and sense organs	1
VII	Diseases of the circulatory system	5
VIII	Diseases of the respiratory system	2
IX	Diseases of the digestive system	12
X	Diseases of the genitourinary system	2
XI	Pregnancy, childbirth and the puerperium	1
XII	Diseases of the skin and subcutaneous tissue	3
XIII	Diseases of the musculoskeletal system and connective tissue	1
XIV	Congenital anomalies	0
XV	Perinatal morbidity and mortality conditions	0
XVI	Symptoms and ill-defined conditions	4
XVII	Accidents, poisonings and violence	0
Studies relating to a wide range of diseases		2
		58

It is of interest to note that within the 3 ICDA classes investigated by a significant number of studies, there is a marked correlation between the disease class and the kind of algorithm used to make the diagnoses. The major algorithms applied to disease classes III and IX are Bayes' probability and discriminant functions, while the decision tree is the most frequently used algorithm in studies of ICDA class V. Perhaps the choice of algorithm is, and should be, determined by the reliability of diagnosis for the disease studied. The statistical algorithms, which require extensive, detailed data bases, intuitively seem more appropriate for well-defined disease problems, while the more heuristic less formalised logic of the decision-tree type models seem better suited for disease categories which are not distinctly defined in terms of disease differentiation and disease-symptom profiles.

(b) *Disease category and number of diseases:* In the Appendix, disease category and number of diseases are treated separately, but for purposes of discussion it is convenient to combine them. The relationship of the disease category and number of diseases to diagnostic accuracy is not unique to computer-aided diagnoses. The fewer diseases and the more distinguishable they are from each other, the higher the resulting diagnostic accuracy, whether diagnosed by physician or computer. In computer diagnosis, the relationship of the type and number of diseases included in the data base to diagnostic accuracy becomes apparent in comparing studies which attempt to deal with a small number of diseases which are well-defined to studies which address a larger number of less well-defined diseases. Five studies which diagnose the metabolic status of thyroid dysfunction range in accuracy between 85% and 97%, while 6 studies involving diagnosis of abdominal pain

TABLE 2
COMPARISON OF COMPUTER-AIDED DIAGNOSIS STUDIES—ABDOMINAL PAIN VS. THYROID DYSFUNCTION

Study	*Algorithm*	*Number possible diagnoses*	*Diagnostic accuracy (%)*	
	Abdominal pain			
de Dombal, 1972	Bayes	8	91.8	
de Dombal, 1975	Bayes	6	77–85 (depending on information available)	
Horrocks, 1975	Bayes	4	85.4	
Rinaldo, 1963	Bayes	6	52	
Ross, 1972	Discriminant functions	8	49	
Scheinok, 1968	Bayes	6	57	
	Discriminant functions		56	
	Thyroid dysfunction			
Fitzgerald, 1966	Bayes	3	97.2	
Nordyke, 1971	Bayes		94.0	(Stage IV Scores)
	Discriminant functions	3	94.3	
	Pattern recognition		84.8	
Oddie, 1974	Bayes	3	96.8	
Overall, 1963	Bayes	3	93.3	
Winkler, 1967	Bayes	3	93	

vary in diagnostic accuracy between 49% and 92% (Table 2). The generally higher accuracy of the thyroid studies is assumed to be due to the smaller number of diseases and greater distinguishability among the diseases involved.

Although the disease category and number of diseases involved give some indication of the difficulty of the diagnostic problem, under closer scrutiny it becomes apparent that the way a particular category is divided into diagnostic alternatives is equal in importance. In computer diagnosis this division is sometimes arbitrary. As Oddie *et al.* (1974) point out, there is a large number of specific thyroid dysfunctions which could be differentiated, and in fact, the computer performs very poorly when attempting to diagnose the specific dysfunctions. However, most computer-aided systems deal with only the metabolic status of thyroid dysfunction and perform well in distinguishing among the diagnostic alternatives. This demonstrates that the diagnostic alternatives available are often more predictive of diagnostic accuracy than the disease category involved.

3.3 *Indicants included in data base*

Indicants are defined to include patient history, physical signs, symptoms, exam results, lab test results, or any other features of a patient's condition which could be considered manifestations of a particular disease. The number and type of indicants in the data base also affect diagnostic accuracy. It is obvious that the more powerful indicants (pathognomonic in the ideal case) one includes in the

system, the higher the diagnostic accuracy. This relationship is supported in a study (Nordyke *et al.*, 1971) in which diagnostic accuracy was assessed using different sets of indicants (Table 3).

However, practical considerations often prevent the inclusion of such powerful indicants as sophisticated lab tests and procedures. For example, in a program devised for corpsmen aboard submarines it would be useless to include results of lab tests that could not be administered aboard ship. Additionally, choice of indicants should be based on the balance between differentiating power and cost, where both monetary outlay and potential harmful effects of the diagnostic procedure are included in costs (Gorry *et al.*, 1973). Systems have been devised which use more costly indicants only when a definite diagnosis cannot be reached from less costly indicants (Gorry *et al.*, 1973; Pople *et al.*, 1975). There are also systems which produce high diagnostic accuracy using no complex, costly lab tests or procedures at all (de Dombal *et al.*, 1972; de Dombal *et al.*, 1975; Horrocks and de Dombal, 1975).

TABLE 3
NORDYKE *et al.* (1971)–METABOLIC STATUS OF THYROID DYSFUNCTION

Accuracies using Bayes' theorem	
Type of indicants used	*Diagnostic accuracy*
Stage 1	
History – 17 signs and symptoms + age + weight	79%
Stage 2	
History (Stage 1) + physical examination (tremor, skin feeling + pulse)	83%
Stage 3	
History + physical exam + thyroid palpation	89%
Stage 4	
History + physical exam + thyroid palpation + Achilles reflex time (ART)	94%
Stage 5	
All of above + 3 lab tests – T3RCU, 6 h ^{131}I uptake, and 24 h ^{131}I uptake	96%

Many researchers (Scheinok and Rinaldo, 1968; Burbank, 1969; Fleiss *et al.*, 1972; Birk *et al.*, 1974) have remarked that the number of indicants used in a system often can be reduced drastically without significantly affecting the accuracy of the system. Burbank (1969) found that reducing the number of indicants from 140 to 70 actually increased the diagnostic accuracy of the system when tested on a cross-validation sample. A study of chest pain (Pipberger *et al.*, 1968), showed that out of 498 information items under study, 55 was the maximum number required for effective differential diagnosis, and that nearly 90% of the total information available was either redundant or irrelevant both for the description of the disease entities and for their separation in a differential diagnoses. This suggests that the usefulness of many indicants for diagnosis has not been systematically analysed. Croft (1972) argues that no substantial improvement in computer

diagnosis is possible until clinical profiles of major diseases are more accurately defined. It does not seem likely that more accurate clinical profiles will be a major goal of medical research. Modern medical research emphasises identification of mechanisms and specific therapies rather than determination of clinical profiles based on patient history and physical examination.

4. TESTING THE SYSTEM

Once the algorithm and the factors involved in the construction of the data base have been optimally integrated into a functional and efficient computer-aided diagnostic system, the system must be tested. Test of the system requires an appropriate test sample and an independent criterion of the correct diagnosis for each patient.

4.1 *The test sample*

The test sample must consist of new patients whose medical records were not used to derive the information for the data base (the developmental sample). If the test sample and developmental sample are the same, the true diagnostic accuracy of the system will be unknown. Although use of a new test sample is fundamental to realistic assessment of a system's accuracy (Fisher *et al.*, 1975) many studies ignore this requirement and report diagnostic accuracies testing the system on the developmental sample. This procedure does, however, give an estimate of the best a particular system can be expected to do.

Two studies (Fleiss *et al.*, 1972; Hirschfeld *et al.*, 1974) clearly illustrate the effect of the test sample on the estimated accuracy of a computer-aided diagnostic system. Fleiss *et al.* (1972) reported that the statistical algorithms such as Bayes' or discriminant functions produce higher accuracy when tested on the developmental sample than when tested on a new sample from the same population. Statistical algorithms, by their curve fitting nature, minimise error for the particular sample they are developed from. Since any new sample will be somewhat different, the algorithms cannot be expected to perform as well on the new samples as they do on the developmental sample. These 2 studies also showed that a test sample from a different population will be less accurately diagnosed by these statistical models than the new sample from the same population. The inferiority of these statistical algorithms on new population data is explained by Fleiss *et al.* (1972) in relation to studies of mental disorders: 'The results ... illustrate the danger of applying numerical rules derived from a sample on one population to a sample of another population. Whenever the patterns of psychopathology change, as they well may between populations, the numerical constants of the statistical procedures appropriate to one are no longer appropriate to the other.' This would indicate that a diagnostic system based on actuarial data compiled from a par-

ticular population must be limited in its application to new cases from that same population. Logical algorithms, such as the decision tree, produce higher accuracy than statistical models when diagnosing patients from a new population (Fleiss *et al.*, 1972; Hirschfeld *et al.*, 1974). This is attributable to the fact that decision rules and disease-symptom relationships are not formulated from any one population or source for most decision tree systems.

4.2 *Method of validation*

When testing the system, the diagnoses of the new patients are usually confirmed by the most reliable source available for the particular ailment, such as histological exam, radiographic results, results found at biopsy, surgery or autopsy, or diagnosis based on retroactive assessment of all factors, including response to therapy. The meaningfulness of the reported accuracy of a system is greatly increased as the reliability of confirmed diagnoses increases. The diagnostic accuracy of the computer is usually stated as the percentage ratio of correct diagnoses to attempted diagnoses.

In most studies of mental disorders, accuracy of diagnoses is described by Kappa scores (K) or weighted Kappa scores (K_w) rather than by percentage correct. The Kappa statistic was developed by Cohen (1960, 1968) in recognition of the fact that professional consensus is largely the only source of validation for accuracy of psychiatric diagnosis. The Kappa measure is based on percentage agreement among authorities corrected for percentage agreement predicted from combinatorial theory and, often, for extent of disagreement among authorities. Kappa is a relative measure. A positive score represents some degree of agreement between computer and clinician, with a score of one equalling perfect agreement. Zero is equal to chance agreement, and negative scores represent less than chance agreement. The Kappa scores reported in the Appendix are more meaningful in the light of information supplied by Spitzer *et al.* (1974). They reported weighted Kappa scores produced by the amount of diagnostic agreement among well-trained clinicians when given precoded research protocols: the range of K_w scores was 0.25–0.80 with an average K_w of 0.45.

While percentage correct is important in medical diagnosis, it only becomes meaningful in real practice when it is compared to state-of-the-art diagnosis. If the computer diagnoses a particular set of cases with 90% accuracy and the average physician diagnoses the same cases with 80% accuracy, then the computer would be a valuable aid. If, however, those same cases are diagnosed by the average physician with 95% accuracy, then the computer offers no advantage. Unfortunately, for many disease categories, the precise state-of-the-art accuracy is not known. Therefore, when testing a computer-aided system it is useful to obtain physicians' diagnoses of the cases as well as the computer diagnoses and the validated diagnoses.

5. DISCUSSION

This review by no means covers the entirety of computer-aided diagnostic applications. Studies which did not directly specify several of the factors we investigated, studies which did not report systems tests and those dealing with systems whose indicants consisted entirely of sophisticated lab tests (e.g., the interpretation of electrocardiograms), were purposely excluded. Several studies which reflect stages of progression of one system by the same author or group are reported only once, usually as the particular report that contains the most information in relation to the factors investigated. In addition, there probably were relevant studies which were simply overlooked.

Nevertheless, the reports reviewed here are sufficient to gain a basic understanding of what is required for computer-aided diagnosis. In summary of the factors reviewed, the computer algorithm is certainly the most controversial area of computer-aided diagnosis. The superiority of a particular type of algorithm has not, to this point, been conclusively demonstrated. To discriminate the effective from the non-effective algorithms, and to progress towards the optimally performing algorithm(s), more comparative work is needed in testing different algorithms on the same data. In addition, to guarantee successful integration into the real-life medical sphere, the method by which an algorithm reaches a diagnostic decision must be visible to and understood by the physician. Shortliffe (1976) and Pople *et al.* (1975) emphasise the importance of the ability of the physician to question the logic and information on which the computer bases a particular decision. An entire segment of Shortliffe's (1976) MYCIN system is dedicated to answering questions presented by the user about its logic and medical information.

Croft (1972) and others feel the real improvement in the success of computer-aided diagnoses will come not with the slow sophistication of algorithms, but with the creation of more accurate disease-symptom profiles, obtained through the maintenance of large, standardised medical data bases. The computer, having no intuition or 'gut feelings', must make a diagnosis based on the measurable symptoms of the presenting patient and the known relationships of different symptoms and signs to different diseases. Therefore, accurate measurement and recording of the patients' symptoms along with precise knowledge of disease-symptom relationships will optimise the probability of the computer making a correct diagnosis.

It is obvious from this review that computer-aided diagnosis research should attack a wider variety of diseases and disease categories. Unfortunately, no computer-aided systems presently have the capability of diagnosing a large number of diverse diseases accurately. The number of diseases and symptoms involved in a wide-range system becomes overwhelming even for the computer. Patrick *et al.* (1974) have suggested dividing a wide range of diseases into subsystems so that subsequent to entering a small amount of critical information, the computer can

identify the most appropriate subset of diseases to evaluate. Thus the diagnostic problem becomes less complex than consideration of all the diseases simultaneously.

In reviewing the indicants used for computer-aided diagnosis as well as for contemporary clinical practice, it is apparent that more information is needed as to the actual utility of many signs and symptoms in differentiating diseases. As stated previously, this includes more accurate disease-symptom profiles, based on large numbers of documented cases.

Finally, the computer system must be tested in a real-time setting in order to: successfully demonstrate acceptibility to and compatibility with users; insure a state-of-the-art or better diagnostic accuracy on a sufficient number of new patients; and show an overall enhancement of the medical environment on a practical, technical and financial level. Positive results in all facets of such a field test will assure successful computer-aided diagnosis implementation on a real-time basis.

REFERENCES

ALPEROVITCH, A. and FRAGU, P., A suggestion for an effective use of a computer-aided diagnosis system in screening for hyperthyroidism, *Method. Inf. Med.*, **16** (1977) p. 93.

BIRK, R. E., ENDRES, L., MCDONALD, J. C., PROCTOR, L. D., RINALDO, J. A. and RUPE, C. E., Approach to a reliable program for computer-aided medical diagnosis, *Aerosp. Med.*, **45** (1974) p. 659.

BISHOP, C. R. and WARNER, H. R., A mathematical approach to medical diagnosis: application to polycythemic states utilizing clinical findings with values continuously distributed, *Comput. Biomed. Res.*, **2** (1969) p. 486.

BOUCKAERT, A., Computer-aided diagnosis of goitres in a cancer department, *Int. J. Bio-Med. Comput.*, **3** (1972) p. 3.

BOYLE, J. A., GREIG, W. R., FRANKLIN, D. A., HARDEN, R. M., BUCHANAN, W. W. and MCGIRR, E. M., Construction of a model for computer-assisted diagnosis: application to the problem of non-toxic goitre, *Q. J. Med.*, **35** (1966) p. 565.

Bricetti, A. B. and Bleich, H. L., A computer program that evaluates patients with hypercalcemia, *J. Clin. Endocrinol. Metab.*, **41** (1975) p. 365.

BRODMAN, K. and VAN WOERKOM, A. J., Computer-aided diagnostic screening for 100 common diseases, *J. Am. Med. Assoc.*, **197** (1966) p. 901.

BRUCE, R. A., PORTMAN, R. M., BLACKINN, J. R., LAMPPERT, R. V., HOFER, V. and MCGILL, R., Computer diagnosis of heart disease, *Proc. 5th IBM Med. Symp.*, 77 (1963).

BURBANK, F., A computer diagnostic system for the diagnosis of prolonged undifferentiating liver disease, *Am. J. Med.*, **46** (1969) p. 401.

COE, F. L., The performance of a computer system for metabolic assessment of patients with nephrolithiasis, *Comput. Bio-Med. Res.*, **5** (1974) p. 351.

COHEN, J., A coefficient of agreement for nominal scales, *Educ. Psychol. Meas.*, **20** (1960) p. 37.

COHEN, J., Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychol. Bull.*, **70** (1968) p. 213.

CROFT, D. J., Is computerized diagnosis possible?, *Comput. Bio-Med. Res.*, **5** (1972) p. 351.

CROFT, D. J. and MACHOL, R. E., Mathematical methods in medical diagnosis, *Ann. Biomed. Eng.*, **2** (1974) p. 69.

CROOKS, J., MURRAY, I. P. C. and WAYNE, E. J., Statistical methods applied to the clinical diagnosis of thyrotoxicosis, *Q. J. Med.*, **28** (1959) p. 211.

DE DOMBAL, F. T., CLAMP, S. E., LEAPER, D. J., STANILAND, J. R. and HORROCKS, J. C., Computer-aided diagnosis of lower gastrointestinal tract disorders, *Gastroenterology*, **68** (1975) p. 252.

DE DOMBAL, F. T., LEAPER, D. J., STANILAND, J. R. and HORROCKS, J. C., Computer-aided diagnosis of acute abdominal pain, *Br. Med. J.*, **2** (1972) p. 9.

ENGLE, R. L., FLEHINGER, B. J., ALLEN, S., FRIEDMAN, R., LIPKIN, M., DAVIS, B. I. and LEVERIDGE, L. L., HEME: A computer aid to diagnosis of hematologic disease, *Bull. N.Y. Acad. Med.*, **52** (1976) p. 584.

FELDMAN, S., KLEIN, D. F. and HONIGFELD, G., The reliability of a decision tree technique applied to psychiatric diagnosis, *Biometrics*, **28** (1972) p. 831.

FISCHER, M., Development and validity of a computerized method for diagnosis of functional psychoses (Diax), *Acta Psychiatr. Scand.*, **50** (1974) p. 243.

FISHER, R. A., The use of multiple measurements in taxonomic problems, *Ann. Eugen.*, **7** (1936) p. 179.

FISHER, M., FOX, R. I. and NEWMAN, A., Computer diagnosis of the acutely ill patient with fever and a rash, *Int. J. Dermatol.*, **12** (1973) p. 59.

FISHER, L., KRONMAL, R. and DIEHR, P., Mathematical aids to medical decision-making, in *Operations Research in health care*, L. J. Schuman, R. D. Speas, Jr. and J. P. Young, (Eds.), Johns Hopkins University Press, 1975, p. 365.

FITZGERALD, L. T., OVERALL, J. E. and WILLIAMS, C. M., A computer program for diagnosis of thyroid disease, *Am. J. Roentgenol.*, **97** (1966) p. 901.

FLEISS, J. L., SPITZER, R. L., COHEN, J. and ENDICOTT, J., Three computer diagnosis methods compared, *Arch. Gen. Psychiatry*, **27** (1972) p. 643.

FRASER, P., HEALY, M., ROSE, N. and WATSON, L., Discriminant functions in differential diagnosis of hypercalcemia, *Lancet*, **1** (1971) p. 1314.

FREEMON, F. R., Computer diagnosis of a headache, *Headache*, **8** (1968) p. 49.

FRIES, J. F., Experience counting in sequential computer diagnosis, *Arch. Int. Med.*, 126 (1970) p. 647

GLEDHILL, V. X., MATTHEWS, J. D. and MACKAY, I. R., Computer-aided diagnosis: a study of bronchitis, *Method. Inf. Med.*, **11** (1972) p. 228.

GLESER, M. A. and COLLEN, M. F., Towards automated medical decisions, *Comput. Biomed. Res.*, **5** (1972) p. 180.

Gorry, G. A., Computer-assisted clinical decision making, *Method. Inf. Med.*, **12** (1973) p. 45.

Gorry, G. A. and Barnett, G. O., Sequential diagnosis by computer, *J. Am. Med. Assoc.*, **205** (1968) p. 849.

GORRY, G. A., KASSIRER, J. P., ESSIG, A. and SCHWARTZ, W. B., Decision analysis as the basis for computer-aided management of acute renal failure, *Am. J. Med.*, **55** (1973) p. 473.

GUSTAFSON, D. H., KESTLY, J. J., GRIEST, J. H. and JENSEN, N. M., Initial evaluation of a subjective Bayesian diagnostic system, *Health Serv. Res.*, **6** (1971) p. 204.

HADLEY, T. P., GEER, D. E., BLEICH, H. L. and FREEDBERG, I. M., The use of digital computers in dermatologic diagnosis: computer-aided diagnosis of febrile illness with eruption, *J. Invest. Dermatol.*, **62** (1974) p. 467.

HIRSCHFELD, R., SPITZER, R. L. and MILLER, R. G., Computer diagnosis in psychiatry: a Bayes approach, *J. Nerv. Ment. Dis.*, **158** (1974) p. 399.

HORROCKS, J. C. and DE DOMBAL, F. T., Computer-aided diagnosis of 'dyspepsia', *Am. J. Dig. Dis.*, **20** (1975) p. 397.

Janis, I. L. and Mann, L., *Decision making*, The Free Press, 1977.

KNILL-JONES, R. P., STERN, R. B., GRIMES, D. H., MAXWELL, J. D., THOMPSON, R. P. H. and WILLIAMS, R., Use of a sequential Bayesian model in diagnosis of jaundice by computer. *Br. Med. J.*, **1** (1973) p. 530.

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, *Br. Med. J.*, 4 (1972) p. 350.

LEDLEY, R. S. and LUSTED, L. B., Reasoning foundations of medical diagnosis, *Science*, **130** (1959) p. 9.

LEONARD, M. S., KILPATRICK, K. E., FAST, T. B., MAHAN, P. E. and MACKENZIE, R. S., Automated diagnosis and treatment planning for craniofacial pain, *J. Dent. Res.*, **53** (1974) p. 1155.

LINCOLN, T. L. and PARKER, R. D., Medical diagnosis using Bayes' theorem, *Health Serv. Res.*, **2** (1967) p. 34.

LUSTED, L. B., *Introduction to Medical Decision Making*, Charles C. Thomas, 1968.

MEERTEN, R. J. VAN, DURINCK, J. R. and DEWIT, C., Computer guided diagnosis of asthma, asthmatic bronchitis, chronic bronchitis and emphysema, *Respiration*, **28** (1971) p. 399.

MELROSE, J. P., STROEBEL, C. F. and GLUECK, B. C., Diagnosis of psychopathology using stepwise multiple discriminant analysis, *Compr. Psychiatry*, **11** (1970) p. 43.

NEURATH, P. W., ENSLEIN, K. and MITCHELL, G. W., Design of a computer system to assist in differential preoperative diagnosis for pelvic surgery, *N. Engl. J. Med.*, **280** (1969) p. 745.

NEWELL, A. and SIMON, H. A., *Human Problem Solving*, Prentice-Hall, 1972.

NORDYKE, R. A., KULIKOWSKI, C. A. and KULIKOWSKI, C. W., A comparison of methods for the automated diagnosis of thyroid dysfunction, *Comput. Biomed. Res.*, 4 (1971) p. 374.

NORINS, A. L., Computers in dermatology, *Arch. Dermatol.*, **90** (1964) p. 506.

NORUSIS, M. J. and JACQUEZ, J. A., Diagnosis. I. Symptom nonindependence in mathematical models for diagnosis, *Comput. Biomed. Res.*, **8** (1975a) p. 156.

NORUSIS, M. J. and JACQUEZ, J. A., Diagnosis. II. Diagnostic models based on attribute clusters: a proposal and comparisons, *Comput. Biomed. Res.*, **8** (1975b) p. 173.
ODDIE, T. H., HALES, I. B., STIEL, J. N., REEVE, T. S., HOOPER, M., BOYD, C. M. and FISHER, D. A., Prospective trial of computer program for the diagnosis of thyroid disorders, *J. Clin. Endocrinol. Metab.*, **38** (1974) p. 876.
OVERALL, J. E. and HOLLISTER, L. E., Computer procedures for psychiatric classification, *J. Am. Med. Assoc.*, **187** (1964) p. 583.
OVERALL, J. E. and WILLIAMS, C. M., Conditional probability program for diagnosis of thyroid function, *J. Am. Med. Assoc.*, **183** (1963) p. 307.
PATRICK, E. A., MARGOLIN, G. and SANGHVI, V., Pattern recognition applied to early diagnosis of heart attacks, *Proc. Int. Med. Info. Processing Conf.*, Toronto, (1977).
PATRICK, E. A., STELMACK, F. P. and SHEN, L. Y., Review of pattern recognition in medical diagnosis and consulting relative to a new system model, *IEEE Trans. Syst. Man Cyber.*, (Jan, 1974) p. 1.
PAUKER, S. G., GORRY, G. A., KASSIRER, J. P. and SCHWARTZ, W. B., Towards the simulation of clinical cognition, *Am. J. Med.*, **60** (1976) p. 981.
PIPBERGER, H. V., KLINGEMAN, J. D. and COSMA, J., Computer evaluation of statistical properties of clinical information in the differential diagnosis of chest pain, *Method. Inf. Med.*, **7** (1968) p. 79.
POPLE, H., MYERS, J. and MILLER, R., Dialog: a model of diagnostic logic for internal medicine, *Proc. Int. Joint Conf. AI.*, Tbilisi, U.S.S.R., (1975) p. 848.
REALE, A., MACCACARO, G. A., ROCCA, E., D'INTINO, S., GIOFFRÉ, P. A., VESTRI, A. and MOTOLESE, M., Computer diagnosis of congenital heart disease, *Comput. Biomed. Res.*, **1** (1968) p. 533.
RECTOR, A. L. and ACKERMAN, E., Rules for sequential diagnosis, *Comput. Biomed. Res.*, **8** (1975) p. 143.
RINALDO, J. A., SCHEINOK, P. and RUPE, C. E., Symptom diagnosis. A mathematical analysis of epigastric pain, *Ann. Int. Med.*, **39** (1963) p. 145.
ROSS, P. and DUTTON, A. M., Computer analysis of symptom complexes in patients having upper gastrointestinal examinations, *Am. J. Dig. Dis.*, **17** (1972) p. 248.
SCHEINOK, P. A. and RINALDO, J. A., Symptom diagnosis. A comparison of mathematical models related to upper abdominal pain, *Comput. Biomed. Res.*, **1** (1968) p. 475.
SHORTLIFFE, E. H., *Computer-based Medical Consultations: MYCIN*, American Elsevier Publishing Co., Inc., 1976.
SLETTEN, I. W., ULETT, G., ALTMAN, H. and SUNDLAND, D., The Missouri standard system of psychiatry (SSOP), *Arch. Gen. Psychiatry*, **23** (1970) p. 73.
SMITH, W. G., A model for psychiatric diagnosis, *Arch. Gen. Psychiatry*, **14** (1966) p. 521.
SPICER, C. C., JONES, J. H. and JONES, J. E. L., Discriminant and Bayes analysis in the differential diagnosis of Crohn's disease and proctocolitis, *Method. Inf. Med.*, **12** (1973) p. 118.
SPITZER, R. L., ENDICOTT, J., COHEN, J. and FLEISS, J. L., Constraints on the validity of computer diagnosis, *Arch. Gen. Psychiatry*, **31** (1974) p. 197.
STREUFERT, S., Complex military decision making, *Nav. Res. Rev.*, **23/9** (1970) p. 12.
TAKAYAMA, J., Automatic diagnosis of congenital heart diseases by electronic computer, *Med. J. Osaka Univ.*, **20** (1969) p. 179.
TAYLOR, T. R., SHIELDS, S. and BLACK, R., Study of cost-conscious computer-assisted diagnosis in thyroid disease, *Lancet*, **2** (1972) p. 79.
TEMPLETON, A. W., LEHR, J. L. and SIMMONS, C., The computer evaluation and diagnosis of congenital heart disease, using roentgenographic findings, *Radiology*, **87** (1966) p. 658.
WARNER, H. R., TORONTO, A. F., VEASEY, L. G. and STEPHENSON, R., A mathematical approach to medical diagnosis, *J. Am. Med. Assoc.*, **177** (1961) p. 177.
WILLIAMSON, J., WHALEY, K., DICK, W. C. and ANDERSON, J. A., Computer-assisted diagnosis in keratoconjunctivitis sicca, *Trans. Ophthalmol. Soc. U.K.*, **91** (1971) p. 147.
WILSON, W. J., TEMPLETON, A. W., TURNER, A. H. and LODWICH, G. S., The computer analysis and diagnosis of gastric ulcers, *Radiology*, **85** (1965) p. 1064.
WINKLER, C., REICHERTZ, P. and KLOSS, G., Computer diagnosis of thyroid diseases, *Am. J. Med. Sci.*, **253** (1967) p. 27.
WORTMAN, P. M., Medical diagnosis: an information processing approach, *Comput. Biomed. Res.*, **5** (1972) p. 315.

APPENDIX

Summary of 58 computer-aided diagnosis studies

ICDA CLASS	STUDY	ALGORITHM	SOURCE OF DB	#-INDICANTS	#-DX'S	TS[a]	DISEASE/SYM.	SOURCE OF CORRECT DX	ACCURACY
	ALBEROVITCH (1977)	BAYES	359 MR	9 CL. + FREE THYROXIN INDEX	2	117 MR	THYROID	CL. JUDGEMENT + LAB TESTS	91%
III.	CROOKS (1959)	W. SUMMATION	182 MR AND EARLIER CASES	23	2	289 MR	THYROTOXICOSIS	RADIOIODINE RESULTS, LAB TESTS, RTT, PO	86%
E N D O C R I N O L O G I C	FITZGERALD (1966)	BAYES	1379 MR	27 POSSIBLE	3	500 MR	THYROID	CL. JUDGMENT, LAB TESTS + FOLLOW-UP	97.2%
	NORDYKE (1971)	BAYES	2405 MR	24	3	1688 MR FROM DB	THYROID	CL. JUDGMENT, LAB TESTS, RTT	94.0% 94.3%
		L.D.F.							84.8%
		PATTERN RECOGNITION							(STAGE IV)
	ODDIE (1974)	BAYES	CL. DATA + EARLIER STUDY	19 (METABOLIC)	3	1166 MR	THYROID	RTT, PATHOLOGICAL + HISTOLOGICAL EXAMS	96.8%
				10 (SPECIFIC)	10	29 MR			34.5%
	OVERALL (1963)	BAYES	879 MR	21 POSSIBLE	3	450 MR FROM DB	THYROID	CL. JUDGMENT + RTT OVER 6 MOS. DURATION	93.3%
	WINKLER (1967)	BAYES	CASES FROM EARLIER STUDY	17 CL.+ 5 TESTS	3	100 MR	THYROID	UNSPECIFIED	95%
				17 CL.		105 MR			91%
ANOMALIES	BOUCKAERT (1972)	BAYES	ROUTINE CL. INVESTIGATIONS	30- CL., LAB TESTS, RADIOISOTOPE RESULTS	5	86 MR	GOITRES	HISTOPATHOLOGICAL EXAM, RTT	83%-CL. DATA 94%-ALL DATA
METABOLIC	BOYLE (1966)	BAYES	155 MR	30	3	88 MR	NON-TOXIC GOITRE	HISTOLOGICAL EXAM, RTT, COMPREHENSIVE CL. ASSESSMENT	83.0% 85.2%
		LIKELIHOOD FUNCTION							

A	TAYLOR	BAYES	155 MR FROM	30 POSSIBLE	3	67 MR	NON-TOXIC	HISTOLOGICAL EXAM	89.6%
B	(1972)	(COST-	BOYLE (1966)				GOITRE		(NO COMPLEX
O		CONSCIOUS)							TESTS-67%)
L	BRICCETTI	DEC. TREE	CL.EST.,	DEPENDS ON AN-	6	51 MR	HYPER-	PATHOLOGICAL EXAM OR	92.1%
I	(1975)		M. LITERATURE	SWERS TO QUES.			CALCEMIA	ENDOCRINOLOGISTS' DX	
C	FRASER	W.L.D.F	128 MR	5-7 BIOCHEMICAL	4-5	218 MR	HYPER-	PO, OPERATION,	90.4%
	(1971)					103-DB	CALCEMIA	NEOCROPSY	
	GLESER	DEC. TREE	19010 MR	57 POSSIBLE	2	9505 MR	DIABETES	CL. JUDGMENT	---
	(1972)			16 USED			MELLITUS		
IV. BLOOD	BISHOP	BAYES	250 MR	14	2	103 MR	POLYCYTHEMIC	HEMATOLOGISTS' DX	95%
AND BLOOD	(1969)						STATE	FROM FOLLOW-UP EXAM	
FORMING	ENGLE	BAYES	PREVIOUS STUD-	SELECTED FROM	40	44 MR	HEMATOLOGIC	PHYSICIANS' DX WITH	86.8% (1ST
ORGANS	(1976)		IES, CL. EST.	585 FINDINGS				ALL DATA AVAILABLE	2 DX'S)
	FELDMAN[b]	TEST SAMPLE	EXPERIENCE +	11 PATIENT	4	153 MR	MENTAL	CLINICIANS' DX AFTER	.508
	(1972)	DEC. TREE	153 MR (TS)	DESCRIPTORS		(TS)		EVALUATION OF ALL	-----
		CVS DEC.	EXPERIENCE +	DERIVED FROM				FACTORS + RTT	.254
V.		TREE	152 MR (CVS)	40 ITEMS					
	FISCHER	DEC. TREE	EXPERT EST. +	39 SYM. GROUPS	13	1202 MR	FUNCTIONAL	CLINICIANS' DX AFTER	73%
	(1974)		OPINIONS				PSYCHOSIS	2 YEAR FOLLOW-UP	
	FLEISS[c]			25 COMBINATIONS		1) DB			1) 2) 3)
M	(1972)	BAYES	454 MR	OF ITEMS FROM	11	2) 286 MR	MENTAL	CLINICIANS' DX	.56 .43 .20
E		L.D.F.		94 TOTAL ITEMS		3) 435 MR-			.56 .47 .28
N		DEC. TREE				MATERNITY			.42 .48 .36
T	HIRSCHFELD[b]	BAYES	417 MR	26 COMBINATIONS		1-212 MR	MENTAL	CLINICIANS' DX	1).59 2).28
A	(1974)			OF ITEMS FROM	8	2-107 MR			3).25 4).30
L		DEC. TREE		150 TOTAL ITEMS		3-277 MR			1).40 2).25
						4-121 MR[d]			3).39 4).26

ICDA CLASS	STUDY	ALGORITHM	SOURCE OF DB	#-INDICANTS	#-DX'S	TS[a]	DISEASE/SYM.	SOURCE OF CORRECT DX	ACCURACY
	MELROBE[b] (1970)	STEPWISE MULTIPLE D.F.	413 MR	70	14	255 MR	MENTAL	CLINICIANS' DX	.388 (1ST 3 DX'S)
0		DEC. TREE							.305
I	OVERALL (1964)	O.F. BAYES PROFILE ANAL-YSIS-2 TYPES	EXPERT EST. + OPINIONS	16 SYM. AREAS	13	489 SIMU-LATED MR	PSYCHOTIC	CLINICIANS' DX	HIGH AGREE-MENT WITH CL. DX
S									
O									
R									
D	SLETTEN (1970)	STEPWISE MULTIPLE O.F.	857 MR	32 USED FROM POSSIBLE 56	12	DB 858 MR	MENTAL	CLINICIANS' DX	54% 48%
E									
R	SMITH (1966)	BAYES	RATING OF 14 DIAGNOSTICIANS	41	38	30 MR	MENTAL	CLINICIANS' DX	87%
S	SPITZER[c] (1974)	DEC. TREE	300 MR, EXPERT OPINIONS	UNSPECIFIED # FROM 2 CL.FORMS	42	100 MR	MENTAL	EXPERTS' DX	.45
	WORTMAN (1972)	DEC. TREE	VERBAL REPORTS OF NEUROLOGIST	147	16	20 SIMU-LATED MR	CEREBELLAR SYNDROME	NEUROLOGISTS' DX	95%
VI. NERVOUS SYSTEM + SENSE ORGANS	WILLIAMSON (1971)	L.D.F.	77 MR	17	2	420 MR	KERATOCON-JUCTIVITIS SICCA	FULL OPHTHALMOLOGICAL EXAM	98.6%
	BRUCE (1963)	BAYES	294 MR + M. LITERATURE	259	22	119 MR	HEART	ANGIOCARDIOGRAPHS, SURGERY OR NECROPSY	35%-VALVULAR
			120 MR + M. LITERATURE	202	9	76 MR			60%-CONGENITAL
VII.	REALE (1968)	BAYES	1184 MR	46 SYM. GROUPED INTO 25 SETS	94	DB 125 MR	HEART	CATHER. AND/OR OPERATION, AUTOPSY	81.6% 60%
CIRCULATORY	TAKAYAMA (1969)	W. MATCHING	137 MR	75	6	DB[e]	HEART	SURGERY OR AUTOPSY	96%
SYSTEM	TEMPLETON (1966)	BAYES	231 MR	20 ROENTGENO-GRAPHIC SIONS	9	OB	HEART	AUTOPSY, SURGERY, CATHER., ANGIOGRAPHY	78%

	WARNER (1961)	CONDITIONAL PROBABILITY	1035 MR (EST. WHEN MR NOT AVAILABLE)	50	33	36 MR	HEART	CATHER. AND/OR FINDINGS AT SURGERY	EQUAL TO DX OF 3 CARDIOLOGISTS
VIII.	DLEDHILL (1972)	W. MATCHING	161 MR	3778 QUESTIONS POSSIBLE	2	DB[e]	BRONCHITIS	CLINICIANS' DX	71.4%
RESPIRATORY SYSTEM	MEERTEN (1971)	BAYES	703 MR	15	4	DB	BRONCHITIS	PHYSICIANS' DX BASED ON LAB TESTS, ETC.	57.5%
		MULTIPLE D.F.		12			EMPHYSEMA		60.0%
		L. REGRESSION		12			ASTHMA		61.5%
	DEDOMBAL (1972)	BAYES	600 MR FROM EARLIER STUDIES	ABOUT 33	8	304 MR	ACUTE ABDOMINAL PAIN	HISTOPATHOLOGICAL EXAM	91.8%
	DEDOMBAL (1975)	BAYES	642 MR	ABOUT 33	6	301 MR	LOWER GI	HISTOPATHOLOGICAL EXAM	85%
IX.	HORROCKS (1975)	BAYES	278 MR	26	4	122 MR	DYSPEPSIA	HISTOPATHOLOGICAL EXAM	87.7%
						76 MR			81.6%
D	RINALDO (1963)	BAYES	204 MR	8	6	96 MR	EPIGASTRIC PAIN	RADIOGRAPHIC DX, OCCASIONAL BIOPSY	52%
I									
D	ROSS (1972)	L.D.F.	PAST MR, TEXTS	48 SYM. CATEGORIES	8	1046 MR	UPPER GI	RADIOLOGIC DX	49%
E									
S	SCHEINOK (1968)	BAYES	UNSPECIFIED	11	6	300 MR	UPPER ABDOMINAL PAIN	RADIOGRAPHIC DX	57%
T		D.F. ANALYSIS							56%
1	BURBANK (1969)	BAYES	52 MR	70	6	DB[e]	LIVER	SURGICAL, SERIAL LIVER BIOPSIES, PO NEDCROPSY	77% (98.1%-SURGICAL\M.
V									
E	CROFT (1972)	10 MOST USED D.F.	1991 MR	50	20	437 MR	LIVER	BIOPSY, SURGERY OR AUTOPSY	51%-64% FOR 10 D.F.
	KNILL-JONES (1973)	BAYES	309 MR	102 POSSIBLE	11	65 MR	JAUNDICE	CL.+ LAB DATA, BIOPSY, LAPAROTOMY, NECROPSY, FOLLOW-UP DATA	69% 89%-SURGICAL\M.
S									
Y									

ICDA CLASS	STUDY	ALGORITHM	SOURCE OF DB	#-INDICANTS	#-DX'S	TS[a]	DISEASE/SYM.	SOURCE OF CORRECT DX	ACCURACY
S	LINCOLN	BAYES	UNSPECIFIED #	16	10	40 MR	LIVER	BIOPSY OR AUTOPSY	82.5% (1ST
T	(1967)		OF MR						3 DX'S)
E	SPICER	BAYES		7	2	DB	CHROHNS	CLINICIANS' DX	
M	(1973)	L.D F.	97 MR	6			DISEASES		100%-ALL
		MAXIMUM LIKE-		9			PROCTOCOLITIS		ALGORITHMS
		LIHOOD D.F.							
	WILSON	BAYES	93 MR	17 STATISTIC-	2	14 MR	GASTRIC	HISTOLOGICAL EXAM	100%(BENIGM
	(1965)			ALLY VALID			ULCERS		\MALIGNEMT)
	COE	DEC. TREE	UNSPECIFIED	CL. DATA + 8	6	122 MR	NEPHROLITHI-	NEPHROLOGISTS' DX-	94.6%
X.	(1974)			TEST RESULTS			ASIS	ALL DATA AVAILABLE	
GENITO-	GORRY	BAYES (COST-	EXPERT EST. +	31 POSSIBLE	14	33 SIMU-	ACUTE RENAL	NEPHROLOGISTS' DX	94% WITH .9
URINARY	(1973)	CONSCIOUS)	OPIMIONS	7.7-AVERAGE		LATED MR	FAILURE		CERTAINTY
SYSTEM				8.7-AVERAGE					100% -.95
									CERTAINTY
XI. PC+P	NEURATH	STEPWISE	>500 MR	26 MOST DISCRI-	9	425 FROM	GYNECOLOGICAL	PATHOLOGISTS' REPORT	66%
	(1969)	D.F.		MINATING SYM.		DB			
	FISHER	W. SUMMATION	CL. OPIMIONS	18	16	34 MR	FEVER + RASH	CLINICIANS' DX	85%
XIII.	(1973)		+ IMPRESSIONS						
SKIN AND	HADLEY	TEMPLATE	TEXTBOOKS +	34	18	62 MR	FEBRILE ILL-	CLINICIANS' DX	90% LISTED
SUBCUTANEOUS	(1974)	BAYES	EXPERT OPINIONS				NESS WITH		81%->.99
TISSUE			M. LITERATURE				ERUPTIONS		CERTAINTY
	HORIMS	TEMPLATE	TEXTBOOKS, MR,	ABOUT 200	ABOUT	25 MR	DERMATOLOGIC	DERMATOLOISTS' DX	72%
	(1964)		EXPERIENCE	QUESTIONS	300			USING LAB TESTS, ETC.	
XIII. MUSCU-	FRIES	SEQUENTIAL	CL. EXPERIENCE	35 POSSIBLE	35	190 MR	ARTHRITIS	CL. DX OF	96%
LOSKELETAL	(1970)	QUESTIONING		QUESTIONS				RHEUMATOLOGIST	
SYSTEM + CON-		FLOW CHART							
NECTIVE TISSUE									

	FREENON	DEC. TREE	H.G. WOLFF	14 QUESTIONS	3	20 MR	HEADACHE	CLINICIANS' DX	80% (100%-
XVI.	(1968)		TEXT						W. SYM.)
SYMPTONS	LEONARD	L. PATTERN	250 MR(PART-	295 POSSIBLE	17	PART OF	CRANIOFACIAL	CLINICIANS' DX	89.7%
AND ILL-	(1974)	RECOGNITION	OB, PART- TS)			250 MR	PAIN		
DEFINED	PATRICK	BAYES	247 MR	17	3	DB	CHEST PAIN	3-DAY COURSE OF	80.0%
CONDITIONS	(1977)							ENZYMES + 12 LEAD ECG	
	PIPBERGER	D.F.	1238 MR	6-47	3	DB	CHEST PAIN	CL. DX- ONLY CLEAR	74.6%
	(1968)	ANALYSIS						CUT CASES WERE USED	
STUDIES	BIRK	BAYES	12000 MR	533 POSSIBLE	53	1996 MR	WIDE RANGE	SENIOR CLINICIANS' DX	79.8%
INVOLVING	(1974)	W. SUMNATION	COVERING 99	25-AVERAGE #					84.4%
WIDE RANGE			DISEASES	USED					
OF DISEASES	BRODMAN	W. SUMNATION	PREVIOUS	150 QUESTIONS	100	252 MR	COMMON	PHYSICIANS' DX	70% (COMMON
	(1966)		STUDIES, ETC.	POSSIBLE			DISEASES		DISEASES)

ABBREVIATIONS:

BAYES- ANY ALGORITHM BASED ON BAYES' THEOREM
CATHER.- CATHERIZATION
CL.- CLINICAL
CVS- CROSS-VALIDATION SAMPLE
DB- INFORMATION USED TO DEVELOP THE DATA BASE
DEC.- DECISION
D.F.- DISCRININANT FUNCTIONS
DX- DIAGNOSIS
ECG- ELECTROCARDIOGRAPH
EST.- ESTIMATES
GI- GASTROINTESTINAL
L.- LINEAR
M.- MEDICAL
MR- MEDICAL RECORDS
PC+P- PREGNANCY, CHILDBIRTH & PUERPERIUM
PO- PROLONGED OBSERVATION
RTT- RESPONSE TO THERAPY
SYM.- SYMPTOMS
TS- TEST SAMPLE
W.- WEIGHT(ED)

[a] MEDICAL RECORDS USED FOR TEST SAMPLE ARE NOT THOSE USED TO DEVELOP THE DATA BASE UNLESS OTHERWISE SPECIFIED.

[b] REPORTS DIAGNOSTIC ACCURACY USING KAPPA SCORES.

[c] REPORTS DIAGNOSTIC ACCURACY USING WEIGHTED KAPPA SCORES.

[d] 212 MR- FROM DEVELOPMENTAL SAMPLE. 107 MR- FROM SAME POPULATION AS DEVELOPMENTAL SAMPLE.
277 MR- FROM NEW POPULATION (MATERNITY). 121 MR- FROM NEW POPULATION (PATIENTS OF ITALIAN DESCENT).

[e] EACH RECORD TESTED INDIVIDUALLY, WITH REMAINING RECORDS USED AS DATA BASE, UNTIL ALL RECORDS HAVE BEEN TESTED.