1991:006 749

DECISION SUPPORT FOR THE DIFFERENTIAL DIAGNOSIS OF JAUNDICE

Robert Segaar

DECISION SUPPORT FOR THE DIFFERENTIAL DIAGNOSIS OF JAUNDICE

Beslissingsondersteuning bij
de diagnostiek van geelzucht

Proefschrift

ter verkrijging van de graad van Doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de Rector Magnificus
Prof. Dr. C.J. Rijnvos
en volgens besluit van het College van Dekanen.
De openbare verdediging zal plaatsvinden op
woensdag 23 januari 1991 om 15.45 uur

door

Robert Willem Segaar

geboren te Leiden

Promotiecommissie

Promotoren :	Prof. dr. ir. J.D.F. Habbema
	Prof. J.H.P. Wilson
Overige leden :	Prof. dr. ir. J.H. van Bemmel
	Dr. J. Hilden

Voor Inge, Martijn en Marlous

Voor mijn moeder

Ter nagedachtenis aan mijn vader

Table of Contents

Chapter 1

Preface

Diagnostic models

Making a diagnosis is one of the most important activities in medicine. Once established, a diagnosis guides rational therapeutic action and provides a basis for prognosis. In most areas of medicine, extensive knowledge is available about the diagnostic outcomes. Compared to this knowledge, little is known about the intellectual actions that guide the diagnostic activity of clinicians. Diagnostic *models* capture abstractions of the diagnostic process. The fact that (part of) the diagnostic process can be abstracted in models does not imply that the intellectual activities of clinicians will all be identical (Sim81). This is not required, and perhaps even undesirable. Capturing diagnostic processes into models is a way to document what is known about the diagnostic problem. It contributes to our understanding of the structure and the contributing elements of the diagnostic problem (Cha79). This offers a framework that promotes constructive communication. Finally, diagnostic models can be applied to real-world problems. In that way the diagnostic model becomes a diagnostic aid. This can be useful too when diagnostic aids support or complement human performance on diagnostic tasks.

Diagnostic aids

Many examples of aids supporting the diagnostic process can be found in the literature, but all diagnostic aids converge in a few basic methods. The nature of a diagnostic aid can be found by looking at its *structure* and at its *task*. Figure 1.1 shows a simplified structure of a diagnostic aid. The figure shows that a diagnostic aid implements a synthesis between new information (medical data obtained from a patient) and background information, relevant

to the problem. So, diagnostic aids can differ in the input (patient information) required, the nature and origin of the background information, and the mechanism used to accomplish a synthesis.

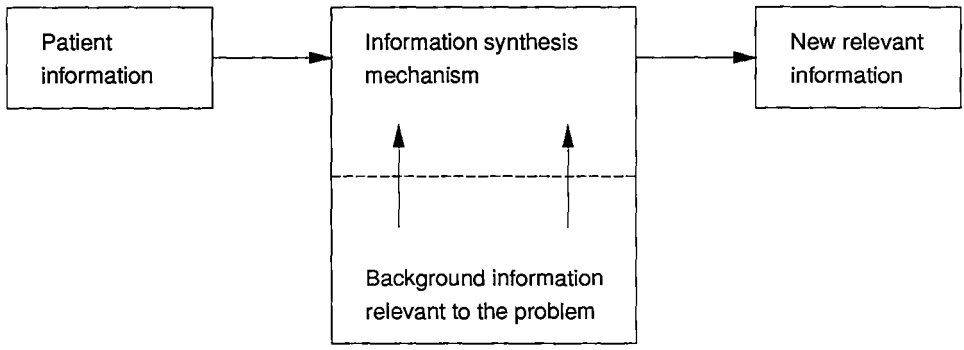


Figure 1.1. Basic structure of a diagnostic aid

For its synthesis task, the diagnostic aid needs **input information**. Such information comes from patient history, physical examination, laboratory testing, imaging techniques, pathology, and so on. The information required can only be a subset of all information that could be made available. Therefore constraints are to be made. The simplest way is to impose limitations on the domain for which the aid operates. We took that approach, in choosing the diagnostic outcomes of jaundice as domain for our study. In diagnostic aids dealing with broad domains such as Internal Medicine (Mil82), the constraints must result from the 'synthesis' method employed.

The **background information** of diagnostic aids always builds on previous experience. If it results from an analysis of data collections, it may embody disease probabilities (both unconditional and conditional on the presence of certain symptoms), or coefficients used in statistical discriminant models (Bro81, Mor84). Background information can also emerge from a formalization of recommendations of experienced clinicians. Common methods to elicitate such expert knowledge are interviews, consensus meetings and Delphi methods. A well-known approach in the context of computerized diagnostic aids is referred to as 'knowledge engineering'

(Fei84). The key difference of these methods used to solicit expert experience is *the way feedback is provided* during the process of knowledge elicitation. The use of experts is another way to obtain estimates on various probabilities, similar to those arising from data analysis. In knowledge engineering, experts can assist in building causal models of diseases (Kim83), that can also be used as background information for diagnostic aids. Background information always reflects features that are local to the source of the experience, usually a patient population studied. This is of no concern if the aid is used for that source population, but on transfer to other locations, problems may arise (Pos86).

All diagnostic aids generate **output information** complying with the diagnostic classification scheme used. Within such a scheme, diagnostic aids can assign probabilities to the diagnoses under consideration. A diagnosis assigned the highest probability can then be interpreted as 'the' diagnosis of the aid. In other instances the aid can just simply put out one or more diagnoses that match the input data. In addition, some aids provide information explaining how the output information could result from the synthesis between input information and background information.

Several methods are used to implement the **synthesis** function of diagnostic aids. Uncertainty plays a dominant role in most, if not all classification problems in medicine, and it is therefore natural to rely on methods that address uncertainty explicitly. Typical examples of such methods are Bayes' rule (Bro81), and discriminant models (Mor84). Other approaches use causal reasoning as strategy (Kim83). Most diagnostic aids require mathematical computations. Sometimes these computations can be presented to a clinician in such a way that the clinician can do without computer assistance. As computations may be error prone in practice, it is no surprise that most diagnostic aids are implemented as computer programs, either by default, or afterwards. In that case, the use of diagnostic aids is also referred to as 'computer-assisted diagnosis'.

Jaundice

All diagnostic aids referred to in this study have the diagnostic process starting from the symptom jaundice as their subject. This choice is justified by several arguments. Diseases underlying the symptom jaundice often have serious consequences for a patient. Jaundice also is a relatively frequent

symptom. Different diagnoses in jaundice have different therapeutic options and, therefore, a reliable diagnosis is important in determining further therapy. There is a reasonable consensus concerning the structure of the diagnostic problem, and the nomenclature of diagnostic categories into which a jaundiced patient can be classified. At the end of the diagnostic process, most jaundiced patients can be classified into the diagnostic categories with confidence. The presence of such a 'gold standard' is a first and obligatory condition for relevant research. Finally, it is simple to specify the characteristics of a population for whom results of the study may be generalized.

The diagnosis of jaundice itself can be implemented as a stepwise process of refinement that ends in establishing a final detailed diagnosis. Along these steps, more global diagnostic categories are used. During the diagnostic process, several points are passed where decisions follow interpretation of information. Usually, these decisions deal with selection of subsequent diagnostic tests. The diagnostic aids studied in this thesis comply to this approach. Consequently, they address only a part of the diagnostic process, in particular the *early stages*. Furthermore, their goal is not focussed on *establishing a final diagnosis* but, instead, they are directed towards assistance in the selection of subsequent diagnostic tests. The work in this thesis is best understood within that context.

Goals: Performance and Transfer

A major concern in the use of diagnostic aids is their performance for diagnostic tasks (Sha77, Buc84, Hil90). As no diagnostic aid claims perfection - neither do clinicians - we take an interest in the amount of errors made. Although it is worthwhile to examine the assumptions and methods employed within the aid critically as a start, it will be insufficient to provide answers. Relevant performance measurement can only be obtained from controlled application of the diagnostic aid. Such an evaluation requires a test population of patients with the characteristics of the anticipated setting. We explained earlier that all diagnostic aids build on previous experience, usually a specific patient population. In the selection of a test population, this is important. If a test population is drawn from the same population on which that experience builds (e.g., the same hospital), there will be no problem. But if it is decided to evaluate a diagnostic aid in a distant setting, problems may develop. It then becomes an additional issue whether the

diagnostic features of the source population also apply to that distant population. This makes the question of *transfer* relevant: how do diagnostic aids developed at one setting, perform at some other setting (Zol77, Dom81). Together with the question of performance, this will be another main theme of this thesis. As spin-off of these topics, the reader will also find information regarding issues such as:

- information synthesis methods used in diagnostic aids,
- sources of the background information included in diagnostic aids,
- measures used to assess the performance of diagnostic aids,
- the extent to which observed 'behavior' of clinicians corresponds to normative outcomes of diagnostic models.

Structure of this thesis

We will now project these themes on their comprehending chapters. Evaluation of diagnostic models is the recurrent theme. Figure 1.2 provides a schematic overview of the evaluations in this thesis.

Chapter 2 describes the implementation of two data collections on jaundiced patients in Rotterdam. Most subsequent chapters (4-8) in this thesis build on these data collections.

Chapter 3 presents a selection of descriptive data from these two data collections. They describe the jaundiced patients seen at the Internal Medicine II department of the University Hospital Dijkzigt. This should provide sufficient information to readers to decide whether our conclusions can be generalized elsewhere.

Chapter 4 describes results of an evaluation of the HEPAR expert system with Rotterdam data. The HEPAR expert system is a rule-based expert system. It allows diagnostic classification of patients into detailed categories.

Chapter 5 presents the evaluation of the COMIK algorithm with our data. The COMIK algorithm is a diagnostic aid based on a logistic discriminant model. To maintain the performance of this aid upon use in Rotterdam, special adaptations were necessary. The adaptations were not exceptional, and relevant for transfer of other aids to other locations as well.

Chapter 6 shows the results of the application of a diagnostic flowchart. The flowchart operates on the same diagnostic classification as the COMIK algorithm.

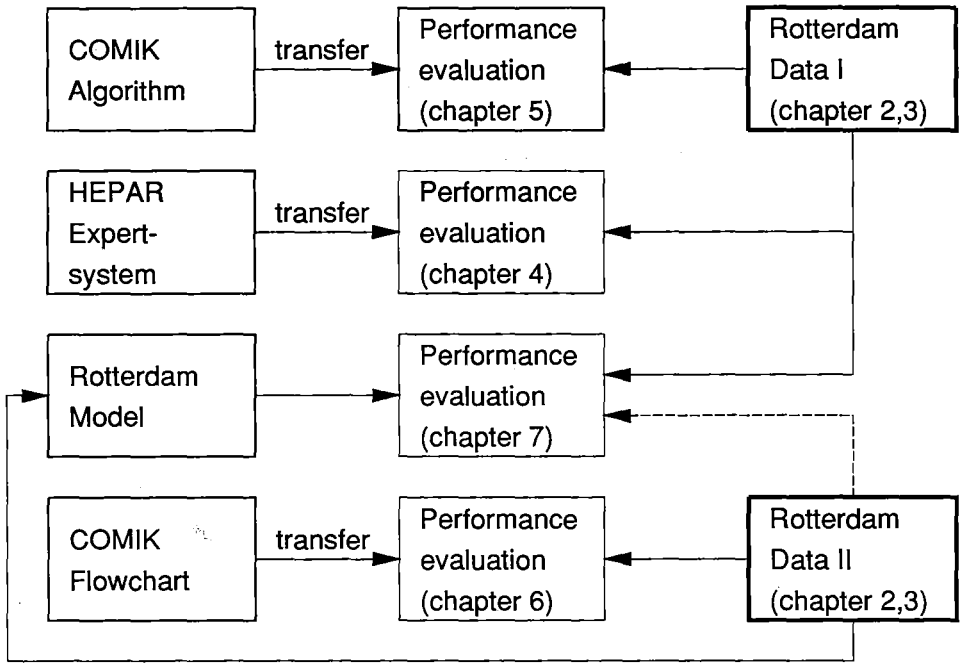


Figure 1.2. Schematic overview of this thesis

Chapter 7 reports the development and evaluation of a diagnostic aid based on our collected data. The exercise enabled us to compare results of our diagnostic aid with those of other diagnostic aids. In addition, the parameters included of the aid could be compared to those of other aids.

Chapter 8 investigates the way test selection takes place in clinical practice. The chapter deals with those diagnostic tests *not* used on a routine basis. The COMIK algorithm was used to predict the final diagnosis, given the (early) information from the patient only. Together with a proposal for tests usage, it is possible to recommend particular tests as worthwhile in the context of the information available. A comparison between the recommended and the actual test-behavior will be made and the implications discussed.

Chapter 9 describes a computer implementation of the COMIK algorithm. Although the COMIK algorithm was intended for paper-and-pencil use, computations can be error-prone; this stimulated the development of a computerized version of the COMIK algorithm. The structure and operation of this Computer Icterus Program (COMIP) will be shown.

Chapter 10 provides the general discussion to this thesis. From issues unresolved, suggestions for future research will be given.

References

Buc84 Buchanan BG, Shortliffe EH. Evaluating performance. The problem of evaluation. In: Buchanan BG, Shortliffe EH, eds. Rule based expert systems: The MYCIN experiments of the STANFORD heuristic programming project. Reading, Massachusets: Addison Wesley Publishing Company 1984: 571-88.

Bro81 Brown GW. Bayes' formula. Conditional probability in clinical medicine. Am J Dis Child 1981; 135: 1125-9.

Cha79 Chandrasekaran B, Gomez F, Mittal S, Smith J. An approach to medical diagnosis based on conceptual structures. In: Proceedings of the 7th IJCAI. Volume 1, 1979; 134-42.

Dom81 De Dombal FT, Staniland JR, Clamp SE. Geografical variation in disease presentation. Does it constitute a problem and can information science help? Med Decis Making 1981; 1: 59-69.

Fei84 Feigenbaum EA. Knowledge engineering. The applied side of artificial intelligence. In: Pagels, ed. Computer culture. The scientific, intellectual and social impact of the computer. New York: The New York Academy of Sciences, 91-107. (Annals of the New York Academy of Sciences Volume 426).

Hil90 Hilden J, Habbema JDF. Evaluation of clinical decision aids. More to think about. Med Inform 1990; 15: 275-84.

Kim83 Kim JH, Pearl J. A computational model for causal and diagnostic reasoning in inference systems. In: Proceedings of the 8th IJCAI. Volume 1, 1983; 190-3.

Mil82 Miller-RA, Pople HE, Myers JD. INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982; 307: 468-76.

Mor84 Morton BA, Teather D, Du Boulay GH. Statistical modeling and diagnostic aids. Med Decis Making 1984; 4: 339-48.

Pos86 Poses RM, Cebul RD, Collins M, Fager SS. The importance of disease prevalence in transporting clinical prediction rules. Ann Intern Med 1986; 105: 586-91.

Sha77 Shapiro AR. The evaluation of clinical predictions: A method and initial application. N Engl J Med 1977; 296: 1509-14.

Sim81 Simon HA. Studying human intelligence by creating artificial intelligence programs. Am Scientist 1981; 69: 300-9.

Zol77 Zoltie N, Horrocks J, De Dombal F. Computer assisted diagnosis of dyspepsia. Report on transferability of a system with emphasis on early diagnosis of gastric cancer. Meth Inf Med 1977; 16: 89-92.

Chapter 2

Data collection

Introduction

In this thesis, two data collections of jaundiced patients have been used. Both come from the department of Internal Medicine of the University Hospital Dijkzigt in Rotterdam. Since most chapters only provide brief reports on these data collections, a more comprehensive description is given in this chapter.

Rotterdam I and II data

Our first data collection ('Rotterdam-I'), was used to evaluate two diagnostic aids: the HEPAR expert system (described in Chapter 4), and the COMIK algorithm (described in Chapter 5). The number of patients aimed at was arbitrarily set at 100. After completion of the first data collection, and the subsequent evaluation of the diagnostic aids, we decided to collect data on a larger series of jaundiced patients for evaluation of other diagnostic aids (Chapter 6) and also to develop a diagnostic model, based on local patient data (Chapter 7). This data collection ('Rotterdam-II') had no preset aim at a particular number of patients. When data collection was stopped due to time restrictions, 214 patients had been entered.

Patient selection criteria

The choice of patient selection criteria is important in application of diagnostic aids. In principle, when using diagnostic aids in other centers, patient selection criteria applied should be identical to those used by the authors of the aid. Thus, for our evaluation study, the choice of criteria was largely dependent on the criteria used by the authors of the aids. Three diagnostic aids were considered for evaluation: a flowchart, an algorithm and an expert system. Two of these, the flowchart and the algorithm, used well-defined patient selection criteria, whereas patient selection criteria for

the expert system were more global. We decided to comply with the well-defined criteria, and to restrict evaluation of the expert system to the same population. The final patient protocol in the Rotterdam data collections was thus similar to the protocol used by the COMIK group (Mat84, Mal88) used for the flowchart and the algorithm. Due to this decision, the diagnostic aid that we developed from local data (Chapter 7), also used the same patient selection criteria. The criteria used were:

- patients, 15 years of age or older, with
- visible signs of jaundice, and
- a serum-bilirubin exceeding 17 μmol/liter

As the diagnostic aids were developed to assist in the initial diagnosis of jaundice, the criteria should ideally be obtained on admission of the patient to hospital. This was however not always possible for several reasons:

- some patients were referred by other hospitals,
- some patients were admitted previously,
- part of our information came from patients files in which the timing of observations in relation to admission was not always precisely stated.

The best definition of our selection criteria is, therefore, that the patient satisfied the criteria during the early days of his current episode of illness. Readmissions of the same patient within one data collection were not allowed. One-third of the patients in the first data collection had recurrent episodes of jaundice for which they were readmitted, and entered both data collections.

Variables

Variables required for the evaluation of the diagnostic aids were the core of both data collections. In addition we recorded variables considered potentially useful for further analysis. For their selection we relied on expert opinions. The major differences between the two data collections result from our desire to document the diagnostic process in more detail during the second data collection. Especially on previous patient history, hematology, ultrasound, liver biopsy and final diagnosis there was a considerable increase in the number of variables used in the second data collection. Some minor adjustments reflect experiences from the first data collection.

We present a brief description of each global category and the number of variables collected within each global category (table 2.1). An exhaustive listing of all variables on which we collected data would occupy too much space.

Personal administration variables refer to patient identification, birthday, sex and age. **Admission** variables refer to type of admission (readmission, referral), reason for admission, the referring authority and so on.

In the context of patient history, **previous history and risk** variables will refer to drug and alcohol usage, hepatitis risk factors, such as profession, drug addiction and transfusions. It also codes important previous diagnoses and relevant items regarding family history. **General history** variables refer to general well-being, anorexia, weightloss, fever, malaise and fatigue. **Jaundice-specific history** variables refer to the jaundice, the presence of nausea, pruritus, vomiting, dark-stained urine, discolored feces and so on. **Pain** variables refer to pain, in the abdomen or in the back. **Systemic symptom** variables refer to myalgia, burning eyes, dry mouth, arthralgia, Raynauds' symptom, etcetera. **Liver failure** variables refer to the presence of dullness, melena, bruises, and increased abdominal girth. **Remaining aspect** variables finally refer to symptoms like dizziness, headache, dyspnea and so on. In addition there are **timing of major symptoms** variables that describe the timing of the major symptoms over time.

In the context of the physical examination, **impression** variables refer to a judgment of the illness, nutritional status and so on. **Jaundice-specific** variables refer to the depth of jaundice and scratch marks. **Abdomen** variables refer to findings of liver, gallbladder, spleen, and the presence of ascites or tumours. **Liver-failure** variables refer to flapping tremor, hepatic fetor, edema, bruises, visible collaterals and disturbed consciousness. **Chronic liver disease** variables refer to xanthelasmata, spider naevi, testicular atrophy, palmar erythema, gynecomastia and so on. **Specific etiologic clue** variables refer to the presence of Kayser Fleisher rings, alcoholic fetor, bronzed skin, butterfly erythema and so on.

In the context of laboratory procedures, **hematology** variables refer to hemoglobin, erythrocytes, leukocytes, white-cell counts and erythrocyte sedimentation rate (ESR). **Hemostasis** variables refer to platelet counts, Cefaloplastin time (APTT), prothrombin time (PT), Thrombotest (TT), Normotest (NT), fibrinogen and so on. **Standard biochemistry** variables refer to standard biochemistry like Ureum, Creatinine, Uric acid, Sodium and so on. **Markers and special biochemistry** variables refer to Copper excretion, Alfafetoprotein (AFP), Carcinoembryogenic antigen (CEA) and so on.

Immunology variables refer to antibodies against mitochondria, smooth muscle, DNA, lysosomal membrane, and to IgA, IgM and IgG. **Viral serology** variables refer to hepatitis B surface antigen (HBsAg), antibodies (HBsAb) and other hepatitis B markers, hepatitis A immunoglobulin (IgM HA) and so on.

In the context of radiological procedures, **ultrasound** variables refer to findings at ultrasound investigation. We made a distinction between *descriptive findings* and *interpretative findings*. **X-ray esophagus** variables refer to the detection of esophageal varices. **Angiography** variables refer to the detection of hepatoma, and the detection of tumours, or determination of their extension. **Computer tomography (CT) scan** variables refer to anomalies of the liver and the pancreas. **Percutaneous transhepatic cholangiography (PTC)** variables refer to bile duct obstruction and its site and possible etiology.

In the context of interventionist procedures, **endoscopy** variables refer to the presence of varices and gastric ulceration. **Endoscopic retrograde cholangiography and pancreaticography (ERCP)** variables refer to the visualisation of the bile ducts, the presence of bile-duct obstruction, its site and possible etiology.

From the remaining procedures, **Electro-encephalography (EEG)** variables refer to the grade of encephalopathy. **Liver biopsy** variables refer to a global description of findings from liver biopsy. **Laparotomy** variables refer to findings at laparotomy, and **autopsy** variables refer to findings at autopsy.

Final diagnosis variables refer to the final diagnoses in several classification systems, the presence of uncertainty regarding the final diagnosis, and reasons for remaining uncertainty. In addition final outcomes like death during admission were recorded.

Final diagnosis

To facilitate the interpretation of the concept 'final diagnosis' throughout this thesis, it will be defined as **the disorder held most responsible for the current episode of jaundice of a patient**. We have chosen a four level classification scheme to encode final diagnoses. For the first three levels we adopted the three level classification system of the diagnostic aids originated by the COMIK group (Mat84, Mal88). Our motivations to do so, were similar to those in choosing patient selection criteria. On the first level patients are classified into surgical- (also called 'obstructive') and medical (also called

Table 2.1. Variables (questionnaire items) in the two databases on jaundice. The table shows numbers of variables per category.

Category	Rotterdam I N=100 database	Rotterdam II N=214 database
Personal administration	5	5
Admission	3	7
Previous history/risks/drug usage	63	135
Patient history: general	4	7
Patient history: jaundice specific	7	8
Patient history: pain	5	5
Patient history: systemic symptoms	7	6
Patient history: liver failure	5	5
Patient history: remaining aspects	4	7
Timing of major symptoms	5	11
Examination: impression	3	3
Examination: jaundice specific	3	2
Examination: abdomen	7	7
Examination: liver failure	8	8
Examination: chronic liver disease	6	6
Examination: specific etiologic clues	8	8
Hematology	7	23
Hemostasis	2	9
Standard biochemistry	21	23
Markers/special biochemistry	6	13
Immune serology	6	11
Viral serology	8	11
Liver scintigraphy	6	0
Ultrasound	15	27
X-ray esophagus	2	2
Endoscopy	0	7
Angiography	2	7
CT scan	2	4
ERCP	2	4
PTC	2	4
EEG	0	2
Liver biopsy	2	35
Laparotomy	2	2
Autopsy	2	2
Final diagnosis	22	48
Total for the previous categories	252	464

'non-obstructive') causes of jaundice. Obstructive will be synonymous to extrahepatic obstruction in cases of ambiguity. The first level classification can be subdivided further into acute- and chronic disorders for the non-obstructive/medical causes, and into benign and malignant disorders for the obstructive/surgical causes. Second level diagnoses are subdivided further into 23 detailed third level diagnostic categories. Finally we used free text to encode the fourth level diagnoses. A fourth level diagnosis provided the greatest level of detail allowed for by the information available. Examples of fourth level diagnoses are hepatitis A, hepatitis B, and Cytomegalovirus infections, which are classified as acute viral hepatitis on the third level, as acute non-obstruction or acute medical on the second level, and as non-obstruction or medical on the first level. Table 2.2 shows the three level COMIK classification.

The final diagnoses in our study were obtained from an experienced hepatologist. A final diagnosis for each patient was based on a revision of all information known of that patient. For this assignment he relied on original information in the patient files, and not on the information in our database, since details may be lost in coding variables, or simply not present. This information included discharge letters, which were very helpful in the reconstruction of ongoing chronic disease.

The three level COMIK classification of table 2.2 was used in the evaluation of an algorithm (Chapter 5) and a flowchart (Chapter 6). Free text diagnoses (fourth level) were used to evaluate the HEPAR expert system (Chapter 4). In addition, we recorded information that documented the diagnostic process itself. Some typical examples of these 'diagnostic' variables are:

- whether a diagnosis was suspected at admission
- the diagnostic tests that were used to confirm diagnoses
- the reasons for difficulties experienced in the diagnostic process
- the presence of multiple diagnoses related to jaundice
- the presence of diagnoses not related to jaundice

Some of these variables will be discussed in more detail in Chapter 3.

Patient selection criteria

Patients for the first data collection were selected from discharge letters of the Internal Medicine II department of Dijkzigt hospital. The aim was to collect data on 100 patients. Data collection was started with the patients admitted in 1985, working backwards in the archives of 1984.

Table 2.2. Three level diagnostic classification scheme for jaundice

1st level	2nd level	3rd level
Medical	Acute medical	Acute viral hepatitis
		Drug hepatitis
		Alcoholic hepatitis
		Chronic persistent hepatitis
		Septicemia
		Postoperative jaundice
		Heart failure
		Congenital hyperbilirubinemia
		Hemolysis
	Chronic medical	Alcoholic cirrhosis
		Posthepatitic cirrhosis
		Cryptogenic cirrhosis
		Primary biliary cirrhosis
		Chronic active hepatitis
		Hepatocellular carcinoma
Surgical	Benign surgical	Choledocholithiasis
		Acute cholangitis
		Pancreatitis
		Iatrogenic bile duct lesion
		Secondary biliary cirrhosis
	Malignant surgical	Bile duct carcinoma
		Pancreatic carcinoma
		Liver metastasis

Patients for the second data collection were obtained from two sources. A first group of patients was selected from discharge letters in the archives of the Internal Medicine II Department of Dijkzigt hospital, similar to the way in which the first data collection was implemented. We worked our way back in these archives starting with the most recent admissions, going back until July 1985. Meanwhile, we collected data from newly admitted patients to the Internal Medicine department II of Dijkzigt. Patients data were collected until December 1987.

Once a patient was considered suitable for participation in the data collection, the patient record was requested from the central administration of Dijkzigt hospital. The procedure was time-consuming, due to the dynamics of normal patient management which have first priority. Especially for the newly admitted patients, the resulting delay between patient discharge and availability of the patient record was long. On the other hand, it ensured that all diagnostic information gathered during admission was available from the patient record. For both data collections, a questionnaire was used. On these questionnaires, data from each patient were entered. Patient information was collected from patient records, and from data available through the hospital information system. The patient record was used to obtain data on patient history, physical examination, morphology, and some specialized laboratory data. Most laboratory data came from the hospital information system. Patient information was anonymous, using the patient identification number (PID nr) to detect possible readmissions. After verification, the questionnaire data were entered into a computer-database, thus allowing further processing.

Discussion

Due to the large number of variables on which we collected our data, and the fact that both data collections were essentially retrospective, data on many items were missing. We marked these items as missing data in our database. Most missing data belonged to 'invasive' diagnostic procedures (liver biopsy, ERCP), but ultrasound also was not always performed. It appears that clinicians use the information gathered so far to decide whether or not to request an additional test. It is therefore inadequate to assume that missing data arise by chance. The subject of missing data as a consequence of test selection in clinical practice is studied in Chapter 8.

Although hospital information systems (HIS) provide good services for patient registration, we soon experienced major disadvantages in their use on retrospective data collections. As a first drawback we noticed that old laboratory data were not accessible through the HIS. This forced us to go back to the 'paper' patient files for these patients once more. Another disadvantage originated from the fact that the HIS was still under development. As a consequence, the availability and presentation of some data through the hospital information system changed over time. We also had to experience that, with data collections spanning several years and including many parameters, laboratory procedures may change. This was discovered only by chance. Such changes in procedures remain 'hidden' within the data of the HIS, and may confuse researchers.

The state of information in medical records has been questioned (Bur89). Whether patient records suffice for a data collection, depends on the type of research. In our case, early information regarding patient history and physical examination at admission had highest priority. Such data are obtained through patient records reasonable well. Documentation of the diagnostic process following admission, or data on symptoms, signs and physical examination of a patient during admission were less sufficient. Also, the intellectual steps of the diagnostic process could not be recovered from the data in patient records. We agree with Burnum (Bur89) that the medical record is going through a critical period. Whether medical informatics will save the medical record from further decay, or perhaps induce even further decline, remains an issue.

References

Mat84 Matzen P, Malchow-Møller A, Hilden J, Thomsen C, Svendsen LB, Gammelgaard J, Juhl E. Differential diagnosis of jaundice: a pocket diagnostic chart. Liver 1984; 4: 360-71.

Mal88 Malchow-Møller A, Thomsen C, Hilden J, Matzen P, Mindeholm L, Juhl E. A decision tree for early differentiation between obstructive and non-obstructive jaundice. Scand J Gastroenterol 1988; 23: 391-401.

Bur89 Burnum JF. The misinformation era: the fall of the medical record. Ann Intern Med 1989; 110: 482-4.

Chapter 3

Data description

Introduction

In this chapter we provide information on some relevant properties of the patients admitted to Dijkzigt hospital. Together with the procedural description of the data collections in Chapter 2, the reader can use this information in deciding whether our conclusions are likely to be valid elsewhere. The history of the two data collections has been discussed in Chapter 2. We start with a comparison of the two data collections. The second data collection includes more patients (N=214), and provides more detail when compared to the first data collection. Therefore, the second (Rotterdam-II) data collection is used to describe the patient population examined in this thesis.

A comparison of the two databases

We present a comparison of the two databases on general criteria like sex (table 3.1), age distribution (table 3.2), and the final diagnoses assigned to the patients (table 3.3). Since we know that both data collections resulted from the same population, we refrain from a statistical assessment of possible differences (which would reflect secular trends in the spectrum of patients admitted to Dijkzigt hospital).

For the Rotterdam-I data collection the mean age was 59 years (range 18 - 82 years) for the men and 58 years (range 21 - 90 years) for the women. For the Rotterdam-II data collection the mean age was 57 years (range 15 - 91 years) for the men and 58 years (range 21 - 90 years) for the women.

Table 3.1. Sex distribution in the Rotterdam data collections I & II

	Rotterdam-I data		Rotterdam-II data	
Sex	Patients	%	Patients	%
Males	54	54%	106	50%
Females	46	46%	108	50%
Total	100	100%	214	100%

Table 3.2. Age distribution in the Rotterdam data collections I & II

	Rotterdam-I data		Rotterdam-II data	
Age (years)	Patients	%	Patients	%
15 - 30	7	7%	22	10%
31 - 50	13	13%	41	19%
51 - 64	51	51%	74	35%
65 and above	29	29%	77	36%
Total	100	100%	214	100%

Table 3.3. Final diagnoses in the Rotterdam data collections I & II

	Rotterdam I data		Rotterdam II data	
Final diagnoses	Patients	%	Patients	%
Acute Medical	9	9%	37	17%
Chronic Medical	51	51%	82	38%
Benign Surgical	14	14%	34	16%
Malignant Surgical	26	26%	61	29%
Total	100	100%	214	100%

Patient flow

In addition to diagnostic variables, the data collections contained many variables that documented patient flow. From these variables we reconstructed the referral path of the patients in the Rotterdam II study. This is an important characteristic of the patients in the data collections, as Dijkzigt hospital often operates as a tertiary referral hospital. Three types of referrals were distinguished: patients referred by the general practitioner, patients referred by an internist, and patients referred by an other medical specialist ('other' in table 3.4). If there was no information in the patient file on referral, and neither on previous medical encounters, it was assumed that the patient was referred by the general practitioner.

Table 3.4. Referral in the Rotterdam-II data collection

Referring authority	Patients	%
General practitioner	73	34%
Internist	125	58%
Other specialists	16	8%
Total	214	100%

We also coded the type of admission for the Rotterdam-II data collection. We distinguished between:

- New admissions: this category included both patients not previously admitted to the Dijkzigt hospital and patients previously admitted to Dijkzigt but for **other** reasons than their current disease episode
- Readmissions: this category included only patients which had been admitted to Dijkzigt hospital before, for the same disease.
- Secondary referrals: this category included only patients which had been admitted to an other hospital for their current disease but were referred to Dijkzigt hospital for further diagnostics or therapy.

Results are shown in table 3.5, which shows that the majority of patients had been admitted to the University hospital before or were referred from other hospitals. Only one third of the patients are new admissions. This finding is compatible with the role Dijkzigt hospital plays as a secondary or tertiary referral hospital.

Table 3.5. Type of admission in the Rotterdam-II data collection

Type of admission	Patients	%
New admission	71	33%
Readmission	53	25%
Secondary referral	90	42%
Total	214	100%

The reason for admission (shown in table 3.6) showed a pattern similar to that of the admission type. We distinguished the following reasons for admission:

- Initial diagnostics: this category was coded if no diagnostic investigation had been initiated for a patient.
- Continued diagnostics: this category was coded if some diagnostic workup had been performed, but a final diagnosis had not been established.
- Therapy: this category was coded if a patient was admitted for therapy with a final diagnosis known.
- Social indications: We used this code for patients who were admitted for social reasons only.

The length of stay varied markedly. Benign-obstructive disorders usually have the shortest length of stay. Patients with malignant obstruction stay longer, although their length of stay does not exceed six weeks. Chronic non-obstructive disorders are responsible for the admissions of long duration.

Table 3.6. Reason for admission in the Rotterdam-II data collection

Reason for admission	Patients	%
Initial diagnostics	74	35%
Continued diagnostics	65	30%
Therapy	74	35%
Social indication	1	.5%
Total	214	100%

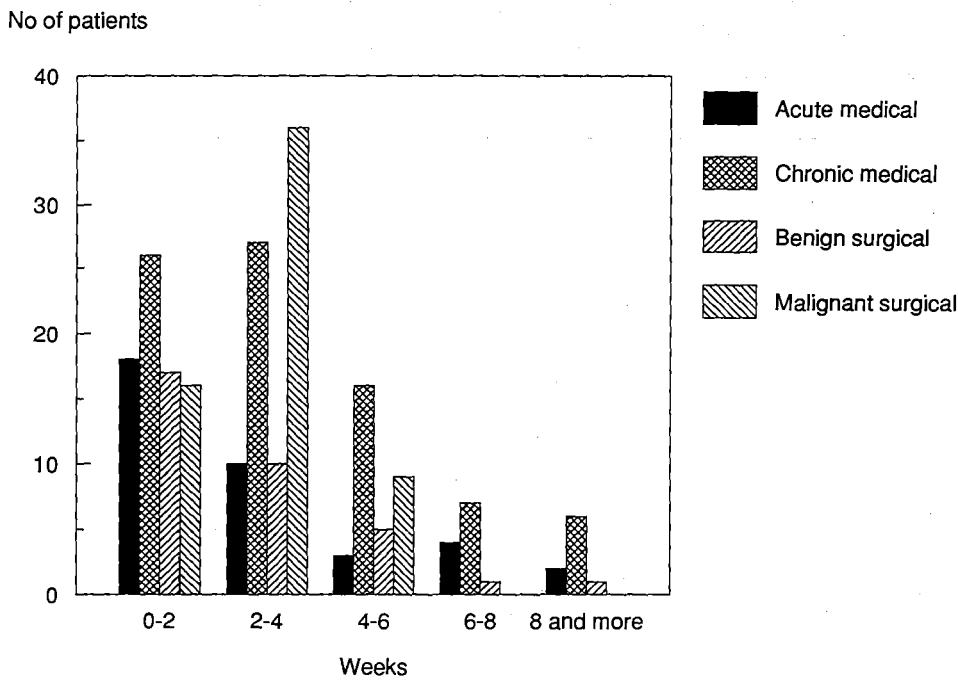


Figure 3.1. Length of stay for the main diagnostic categories

Diagnostic testing

For each patient, many diagnostic tests were performed. Some biochemical tests, like bilirubin and alkaline phosphatase, were done for every patient. Other tests like ultrasound were used in most, but not all patients. Tests like ERCP (endoscopic retrograde cholangiography and pancreaticography), PTC (percutaneous transhepatic cholangiography) and others were performed only in a minority of patients. Such differences in test usage mainly result from a selection by clinicians based on data available. We explore these differences in more detail in Chapter 8.

Table 3.7. Ultrasound and diagnostic tests subject to selection in the Rotterdam-II data collection. They are tabulated against the final diagnosis: AM = acute medical, CM = chronic medical, BS = benign surgical, MS = malignant surgical. The table shows the number of patients (% = percentage) for which the test was performed.

	Final diagnosis							
Diagnostic procedure	AM		CM		BS		MS	
Ultrasound	27	(73%)	64	(78%)	23	(68%)	53	(87%)
X-esophagus	4	(11%)	20	(24%)	1	(3%)	1	(2%)
CT scan	4	(11%)	4	(5%)	7	(21%)	27	(44%)
Angiography	0	(0%)	2	(2%)	1	(3%)	17	(28%)
PTC	0	(0%)	0	(0%)	1	(3%)	1	(2%)
Endoscopy	0	(0%)	26	(32%)	1	(3%)	3	(5%)
ERCP	1	(3%)	2	(2%)	27	(79%)	42	(69%)
Liver biopsy	9	(24%)	47	(57%)	5	(15%)	4	(7%)
Total	37	(100%)	82	(100%)	34	(100%)	61	(100%)

In table 3.7 we present data on the use of ultrasound and of procedures that are evidently subject to selection. This table shows data on the non-invasive radiological procedures (abdominal ultrasound investigations, esophagus radiology and computer tomography (CT) scanning), invasive radiological procedures (abdominal angiography and PTC), interventionist

procedures (endoscopy and ERCP) and finally liver biopsy. To provide insight in the test selection mechanisms, we tabulated them against the final diagnoses of the patients.

Diagnostic uncertainty

In Chapter 2 we defined the 'final diagnosis' as **the disorder held most responsible for the current episode of jaundice of a patient**, thereby allowing multiple diagnoses. In the Rotterdam-II data collection we recorded information concerning additional diagnoses, and reasons for diagnostic uncertainty (see Chapter 2). Collection of such information may be unconventional, presentation is even more. It is nevertheless our believe that it is worthwhile to document such properties of our data.

The number of diagnoses relevant for jaundice in each patient are shown in table 3.8. Of the 182 patients with one diagnosis relevant for jaundice in that table, 21 had a second diagnosis unrelated to the jaundice.

Table 3.8. Number of diagnoses related to jaundice in each patient.

Number of diagnoses	Patients	%
One	182	85
Two	31	14
Three	1	.5
Total	214	100

All patients of the Rotterdam-II data collection could be classified into the second level classification. In some patients, there remained some uncertainty regarding the third level diagnosis. We also gathered information on the reasons why precise statements were impossible in those patients. In table 3.9 we present a summary of the problems encountered.

Table 3.9. Reasons for uncertain final diagnoses

Reason for uncertainty	Patients	%
No uncertainty	162	76.7
Patient refused investigation	1	.5
Technical problems	26	12.1
Patient died and autopsy refused	3	1.4
Contradictory findings	16	7.5
Atypical course of disease	2	0.9
Other	4	1.9
Total	214	100

Discussion

The Rotterdam data collections reflect the population presenting at the Internal Medicine II department of Dijkzigt hospital. As data in this chapter show, this population is unlikely to be typical of jaundiced patients presenting in other hospitals in the Netherlands. Patients are admitted to Dijkzigt from various sources - directly from the general practitioner, from the outpatient department and from other hospitals. The latter two sources result in:

- a selection towards patients with serious complications, usually due to chronic disease
- a selection towards patients with a complicated diagnostic process

The main concern remains whether results in this thesis, that build on the Dijkzigt patient population, can be generalized to other hospitals. The selection towards chronic patients implies a change of relative incidences encountered. We anticipate the subsequent chapters in stating that such differences may indeed be relevant. More quantitative information on the relevance of these differences, and possible ways to deal with them will be presented in the subsequent chapters. The selection towards complicated cases in Dijkzigt however, is likely to result in a decline in performance of

diagnostic aids if studied in Dijkzigt. From that point of view, results obtained from diagnostic aids studied in Dijkzigt are likely to represent an estimate of the minimal performance that one can anticipate.

In general, the consequences of using a population that differs from a 'standard' population are felt in particular in the development of a diagnostic aid, and will be less prominent in the evaluation of diagnostic aids on non-standard populations. This results from the fact that in modeling features of non-standard populations, there is a risk that non-standard features enter the model. Upon evaluation of such an aid on a standard population, these features will not be present and therefore performance will be unfavorable: in fact the model is bad. In the opposite case the model will reflect features from the standard population. During evaluation on a non-standard population, these features will also be present, although perhaps less prominent. Results will however remain acceptable. Therefore, the conclusions in this thesis that relate to evaluation of diagnostic aids in our view can be generalized. Whether this also applies to a diagnostic model based on our data (developed in Chapter 7) is open to discussion.

Chapter 4

Evaluation of an expert system for hepatic disease

Abstract

Expert systems are a recent approach to diagnostic classification problems. A breakthrough came with the development of the MYCIN system at Stanford. Since then, many MYCIN-like expert systems have been constructed, covering a broad range of - usually diagnostic - problems. One such a MYCIN-like system is the HEPAR system for the diagnosis of diseases of the liver and the biliary tract. Of the many systems constructed, only a minority have actually been evaluated. We therefore collected data from 101 consecutive patients in the Rotterdam University Hospital Dijkzigt that allowed evaluation of the HEPAR system. We discuss the results obtained from HEPAR on the 101 Rotterdam patients and the pitfalls in the development of expert systems that affect their future application.

Patients and methods

Expert systems

In our view, artificial intelligence (A.I.) is a branch of psychology that studies (human) intelligence through simulation experiments on the computer (Sim81). Initially, one aimed at the development of methods for general purpose problem solving. Later on, attention focussed on problem solving within narrow domains. From the latter experiments, a special architecture of computer programs emerged. These programs are referred to as *production systems.* Probably the best known program was MYCIN (Sho76). Using these production systems, it seemed possible to achieve expert performance on various complicated tasks. As a result the systems were also being referred to as expert systems. These developments are recent. For example, the National Library of Medicine's MEDLARS system for searches of current medical literature, did not have a separate listing on artificial intelligence until 1984 (Kin87).

The success for A.I. research started when researchers realized that problem solving in narrow domains was an area where breakthroughs could be anticipated. Since then, most A.I. research follows this direction. Focussing on narrow domains implies that systems incorporate large amounts of high quality information. This is analogous to real-world experts, who are also specialists on restricted domains. As knowledge within the domain develops, subspecialists appear, thus stabilizing the amount of knowledge managed by the single expert. One difference between the expert system as domain specialist and a human domain expert is the performance on problems within other domains. Human experts tend to show a type of behavior known as *graceful degradation.* It implies a gradual decline in performance on problems, as the distance between their domain of expertise and the problem domain increases. Human experts are also aware of this process. Expert systems lack such behavior. The question whether expert systems actually mimic expert behavior is difficult to answer. It assumes that we know how human experts solve their problems. Although quite a lot is known about the mental processes of experts, most of this knowledge is hard to implement in a useful way. One may even question whether it is desirable to mimic expert behavior. In a practical setting, the major goal is to achieve optimal *performance* on particular tasks, not expert *behavior.*

The process of creating computer programs is usually referred to as *software engineering.* Expert systems are computer programs too. In a sense it is correct to state that building an expert system is another type of software engineering. However, the steps involved in the construction of expert systems differ from conventional software engineering in many respects. In the construction of conventional computer programs, most time is spend coding algorithms that manipulate data. In the construction of expert systems, most time is spend coding *knowledge.* Actual computer-programming in the usual sense may even be unnecessary. Special computer programs called *expert system shells, empty shells,* or *shells* are commercially available, and enable the user to build expert systems without programming effort. An expert system shell provides several components summarized in table 4.1. These components are general to all expert systems.

Table 4.1. Components provided by an expert system shell

Component	Function
User interface	Provides interaction with the user. It translates production rules in intelligible syntax. It accepts and checks input from the user.
Inference engine	The "reasoning" part of the expert system. It tries to confirm a number of goal hypothesis (e.g. the diagnosis of the patient) through application of the production rules on external information (e.g. patient information) supplied by the user (e.g. doctor).
Explanation facility	Provides the user with feedback about the reasoning in the inference engine.

The body of knowledge coded by the user is usually referred to as a *knowledge base.* The process that results in a knowledge base is often referred to as *knowledge engineering,* and the persons involved are referred to as *knowledge engineers.* Examples of knowledge representation formalisms are *production rules, semantic networks* and *frames.* Such knowledge representation formalisms must be interpretable by the expert system (shell) and intelligible to the knowledge engineer. Only production rules will be considered here. Production rules, also called *if-then* rules, are

an intuitively appealing way to formalize knowledge. Informally they state that if all conditions specified hold (i.e. that the statements are "true"), one or more actions take place. An example of a production rule of the MYCIN system is given in figure 4.1. When several production are applied in a sequence, a deductive reasoning process is emulated.

Figure 4.1. A production rule from the MYCIN knowledge base (Buc84)

IF: 1) The infection is primary-bacteremia, and
2) The site of the culture is one of the sterile sites, and
3) The suspected portal of entry of the organism is the gastrointestinal tract,

THEN: There is suggestive evidence (.7) that the identity of the organism is Bacteroides.

The formulation of that same production rule for MYCIN was:

```
PREMISE: ($AND (SAME CNTXT INFECT PRIMARY-BACTEREMIA)
               (MEMBF CNTXT SITE STERILESITES)
               (SAME CNTXT PORTAL GI))
ACTION: (CONCLUDE CNTXT IDENT BACTEROIDES TALLY .7)
```

The MYCIN expert system (Sho76), developed by Shortliffe in Stanford, was a prototype for many of the expert systems build later. MYCIN was intended as a consultative tool to provide advice on the diagnosis and treatment of infectious diseases, in particular in the blood (Buc84). The HEPAR system (Luc89) is a good representative of a MYCIN-like expert system. HEPAR is designed as a diagnostic program in liver disease. The HEPAR system emulates the software developed for the MYCIN system, and uses a similar knowledge representation. Even within the small domain covered by HEPAR, several other expert systems exist (Tor85, Cha83, Les84).

Patients

We collected data from 101 consecutive patients admitted to the Rotterdam University Hospital Dijkzigt between 1984 and 1985. In the data collection, a future evaluation of the expert system was anticipated by inclusion of all parameters employed by the expert system. Data collection was retrospective. All patients satisfied three criteria at admission : aged over 15 , visible signs of jaundice and a serum bilirubin exceeding 17 μmol/l. The data collection is referred to as Rotterdam-I, because it was followed by a second data collection (Rotterdam-II) that included 214 patients.

A final diagnosis was coded in several levels of detail. At a global level patients were classified into obstructive and non-obstructive causes of jaundice. Non-obstruction was then divided further into acute and chronic non-obstruction, and obstruction was divided further into benign and malignant obstruction. Finally a detailed diagnosis (like Mirizzi syndrome, or primary biliary cirrhosis) was coded. To establish a final diagnosis we used expert opinion (J.H.P. Wilson). The expert preferably used clinical chemical tests and/or pathoanatomical findings, usually biopsy, ERCP, laparotomy or autopsy. In a few cases, some doubt remained about the detailed final diagnosis, but all patients could be classified into the global diagnostic categories.

Results

The constructors of the HEPAR system conducted an evaluation of the system with Rotterdam-I data. Of the 101 patients available for an assessment of the HEPAR system, seven had to be withdrawn because they suffered from diseases not covered by the knowledge base of the system, or because only diagnoses at a global level (like malignant obstructive disease) could be established. There is little doubt that these cases belonged to a category of patients that is hard to classify for human experts and for expert systems.

Based on the patient data entered, most diagnostic expert systems - including HEPAR - generate a list of possible diagnoses. Each diagnosis on the list is associated with a certainty factor. By definition, we take the diagnosis that was assigned the highest certainty factor as the final diagnosis of the expert system. From a comparison of this diagnosis with the gold standard diagnosis, an error rate results. A less critical way to assess the output produced by expert systems, is to check whether the gold standard

diagnosis is on the list generated by the system. HEPAR provides diagnoses at several levels of detail: a global level that differentiates obstructive from non-obstructive causes, a global level that differentiates benign from malignant causes, and finally a detailed diagnosis. At each level performance was evaluated.

Results of HEPAR on the Rotterdam-I data are shown in table 4.2. For 95% of the patients a differentiation into obstructive and non-obstructive causes of jaundice was obtained. Of these patients, 85% were classified correctly. Also, 65% of the patients could be classified as either benign or malignant. Of these patients, 92% were correctly classified. The system established a detailed diagnosis in 80% of the patients. Of these patients, 80% had a correct final diagnosis. For most classified patients, HEPAR produced several diagnoses (average number of diagnoses = 2.9). In 87% of the classified patients, the gold standard diagnosis was on the list generated by the HEPAR system.

Table 4.2. Results of the HEPAR system for the Rotterdam-I test population. Seven patients were excluded.

Conclusion	Correct n (%)	Incorrect n (%)	Unclassified n (%)	Total n (%)
Type of hepatobiliary derangement	76 (81)	13 (14)	5 (5)	94 (100)
Benign/malignant nature of disorder	56 (60)	5 (5)	33 (35)	94 (100)
Final diagnosis	60 (64)	15 (16)	19 (20)	94 (100)

The impact of missing data on the performance of the system was also explored. For this experiment, two subsets of the original data were used. A first subset contained only data on symptoms, signs, hematology and bloodchemistry, and excluded serology and ultrasound data. A second subset contained only data from the medical interview and physical

examination. Results from this experiment are shown in table 4.3, where table 4.2 serves as a reference. The general conclusion is that the system maintains the performance at the cost of many more unclassified patients.

Table 4.3. The effect of incomplete data on the diagnostic performance of the HEPAR system for the patients of the Rotterdam-I data collection. Seven patients were excluded.

A: Using only data concerning symptoms, signs, hematology and blood-chemistry (no data from ultrasound or serology presented).
B: Using only data from patient history and physical examination

Conclusion	Correct (%)		Incorrect (%)		Unclassified (%)	
	A	B	A	B	A	B
Obstruction versus non-obstruction	81	60	14	17	5	23
Benign versus malignant causes	60	56	5	3	35	40
Final diagnosis	36	35	12	10	52	55

The Rotterdam-I data were also used to evaluate other diagnostic aids, which allows a comparison to be made. One such aid was the COMIK algorithm (Mat84, Seg88). The global classification scheme used by the HEPAR system is somewhat different from that of the COMIK algorithm. HEPAR uses two binary classifications: non-obstruction versus obstruction, and benign versus malignant causes of illness. The COMIK algorithm uses a fourway classification into acute non-obstruction, chronic non-obstruction, benign obstruction, and malignant obstruction. With the data of our evaluation at hand (Seg88) it is possible to compare results of HEPAR and the COMIK algorithm of those two tasks, using the HEPAR classification scheme into benign and malignant. These result are shown in table 4.4.

Table 4.4. A comparison between HEPAR and the COMIK algorithm on the classification of the Rotterdam-I patients. All patients (100) were classified by the COMIK algorithm, 7 patients out of 101 were excluded in the evaluation of HEPAR.

Classification	HEPAR	COMIK
Obstruction versus non-obstruction	76 (81%)	83 (83%)
Benign versus malignant causes	56 (60%)	86 (86%)

Discussion

As can be seen from the above results, HEPAR manages to classify the Rotterdam patients reasonable well, despite the fact that the system is still under development. As the Rotterdam patient data have also be used to evaluate other diagnostic aids (Seg88), it is possible to compare results between systems. HEPAR had a performance comparable to the COMIK chart (Mat84, Seg88) in classifying the global type of jaundice. An interesting finding of the evaluation of HEPAR is that differentiation between benign and malignant causes of jaundice is particularly difficult. This finding is compatible with our evaluation of the COMIK algorithm (Chapter 5) and a local diagnostic model (Chapter 7). These evaluations too showed a less favorable performance on the classification into benign and malignant categories. This classification task probably is too difficult to accomplish given early information only, and may require ERCP, and more advanced imaging techniques like CT scanning, or angiography.

A striking feature of many publications on expert systems is the description of the target population. Systems may be presented as diagnostic aids for patients with dementia, or diseases of the liver and the biliary tract. This presupposes that the patient has already been preclassified, thus requiring knowledge that is only available once the diagnostic process is (nearly) completed: an example of circular reasoning. This probably is a reflection of the fact that most authors pay too little attention to an operational application of the expert systems. A preferred way to describe a target population for diagnostic problems will use simple symptoms and signs only. Therefore, patients with dyspepsia (Spi87), jaundice (Mat84), abdominal

pain (Dom85), and also asymptomatic patients(!) are acceptable descriptions of possible target populations. In the evaluation of HEPAR described above, the problem of the description of the target population was avoided because the Rotterdam data collection used jaundiced patients only. Whether results of HEPAR also apply to the domain anticipated by the constructors of the system (diseases of the liver and the biliary tract), cannot be inferred from the evaluation with the Rotterdam patients. There is an other problem associated with the use of global domains instead of symptoms and signs. It is very difficult to separate domains in medicine (like for example diseases of the liver and the biliary tract) from other domains (like gastroenterology, hematology and even cardiology). Consequently, the knowledge base should in fact also include knowledge of these related domains. This contradicts the approach of restricted domains. It also imposes restrictions on the communication of the constructors of expert systems towards end-users. Information on limitations of the domain can only be communicated through examples.

Due to the incremental development of their knowledge base and the fact that perfection is never achieved, authors of most expert systems state that the system is still under development. The (usually implicit) assumption that future changes in the system will also result in improvements in performance can be erroneous. In the early stages of the development of expert systems, global knowledge describing the domain enters the knowledge base. Such knowledge is usually less susceptible to features of a local patient population. As the knowledge base continues to be developed, experience from patients incorrectly classified by the expert system contributes to further adaptations of the knowledge base. This knowledge is dependent on features of the local patient population and, as refinement proceeds, also becomes susceptible to random variation. Of course, the procedure resembles in a way the stepwise development of statistical discriminant models. The main problem, however, is that there is no formal stopping rule for further changes in the knowledge base. In that way, systems result with too much emphasis on particularities of the patients at hand. This problem is aggravated by the fact that most expert systems are granted a local evaluation only, whereas part of the problems of such "over-fitted" diagnostic aids only become apparent with transfer to other sites.

The knowledge representation scheme of MYCIN-like expert systems (the production rule formalism) is highly appealing at first sight. Nevertheless, all knowledge representation formalisms exhibit restrictions of some kind in the long run. The production rule formalism is awkward to represent problems

that are handled quite elegant by multivariate statistical techniques. Interpretation of the results of multiple clinical chemical laboratory tests is one field where the production rule formalism results in an ungraceful representation of the available knowledge. A far more important limitation of the production rule formalism results from its inability to handle uncertainty. The certainty factor theory developed by Buchanan and Shortliffe (Buc84) is at its best an ad hoc extension to solve the problem. Other approaches (Lau88, Spi89) offer better perspectives, but do not fit into the present type of knowledge engineering. The question whether MYCIN-like systems can be reliably used, depends on the contents and structure of the knowledge base. It is therefore conceivable that systems that primarily use logic reasoning are less susceptible to the peculiarities of the certainty factor calculus than systems that actually employ the certainty factor calculus to model the domain. The difference of these approaches will be obvious as soon as some patient data are missing. Logic systems will maintain performance at the cost of many unclassified patients, whereas systems that rely on the certainty factor calculus, will show a decline in performance. The experiment on missing data in the HEPAR system (table 4.1) suggests that HEPAR uses logic reasoning as its basic strategy because it is less susceptible to missing data.

Since the development of the MYCIN system, many other expert systems have been build, emulating the software techniques used in MYCIN and using the same knowledge representation schemes. With that technology at hand, a broad range of other diagnostic problems was tackled. The HEPAR system also belongs to this category. Despite all activity focussed on the development of new systems, only few systems passed a careful evaluation (Buc84, Yu79a, Yu79b, Aik80). One reason for this discrepancy is the fact that data collection for evaluation is cumbersome, and intellectually less appealing than construction of new systems. There is little doubt that, with many unevaluated expert systems available, and many more being under construction, time now has come to shift attention from development to evaluation.

References

Aik80 Aikins JS. Prototypes and production rules: A knowledge representation for computer consultations. Stanford: Department of Computer Science Stanford University, 1980: report no STAN-CS-80-814.

Buc84 Buchanan BG, Shortliffe EH, eds. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Reading: Addison-Wesley, 1984.

Cha83 Chandrasekaran B, Mittal S. Conceptual representation of medical knowledge for diagnosis by computer: MDX and related problems. Advances in Computers, volume 22. New York: Academic Press, 1983; 217-93.

Dom85 de Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer aided diagnosis of acute abdominal pain. In: Computer assisted medical decision making. Volume 1. New York: Springer Verlag, 1985: 159-69.

Kin87 Kinney EL. Medical expert systems. Who needs them? Chest 1987; 91: 3-4.

Lau88 Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. J Royal Stat Soc B 1988; 50: 157-224.

Les84 Lesmo L, Marzuoli M, Molino G, Torasso P. An expert system for the evaluation of liver functional assessment. J Med Syst 1984; 8: 87-101.

Luc89 Lucas PJF, Segaar RW, Janssens AR. HEPAR, an expert system for the diagnosis of disorders of the liver and the biliary tract. Liver 1989; 9: 266-75.

Mat84 Matzen P, Malchow-Møller A, Hilden J, Thomsen C, Svendsen LB, Gammelgaard J, Juhl E. Differential diagnosis of jaundice: a pocket diagnostic chart. Liver 1984; 4: 360-71.

Seg88 Segaar RW, Wilson JHP, Habbema JDF, Malchow-Møller, Hilden J, van der Maas PJ. Transferring a diagnostic decision aid for jaundice. Neth J Med 1988; 33, 5-15.

Sho76 Shortliffe EH. Computer-based medical consultation: MYCIN. New York: Elsevier, 1976.

Sim81 Simon HA. Studying human intelligence by creating artificial intelligence programs. Am Scientist 1981; 69: 300-9.

Spi89 Spiegelhalter DJ. A unified approach to imprecision and sensitivity of beliefs in expert systems. In: Kanel LN, Levitt TS, Lemmer JF, eds. Uncertainty in Artificial Intelligence 3. Amsterdam: Elsevier Science Publishers B.V., 1989.

Spi87 Spiegelhalter DJ, Crean GP, Holden R, Knill Jones R. Taking a calculated risk: Predictive scoring systems in dyspepsia. Scand J Gastroenterol 1987; 128(supp): 152-60.

Tor85 Torasso P. Knowledge based expert systems for medical diagnosis. Stat Med 1985; 4: 317-25.

Yu79a Yu VL, Fagan LM, Wraith SM, Clancey WJ, Scott AC, Hannigan J, Blum RL, Buchanan BG, Cohen SN. Antimicrobial selection by computer: a blinded evaluation by infectious disease experts. JAMA 1979; 242: 1279-82.

Yu79b Yu VL, Buchanan B, Shortliffe EH, Wraith SM, Davis R, Scott AC. Evaluating the performance of a computer-based consultant. Comp Prog Biomed 1979; 9: 95-102.

Chapter 5

Transferring a diagnostic decision aid for jaundice

Segaar RW, Wilson JHP, Habbema JDF,
Malchow-Møller A, Hilden J, van der Maas PJ.
Neth J Med, 1988; 33: 5-15.

Abstract

To facilitate the interpretation of diagnostic data in jaundice, formal algorithms may be of use. The algorithm developed by the Danish COMIK group has proved useful in various centers. It assigns diagnostic probabilities to the four major categories of jaundice: acute non-obstructive, chronic non-obstructive, benign obstructive and malignant obstructive.

This study reports the results of applying the algorithm to data from 100 consecutive jaundice patients admitted to the Rotterdam Dijkzigt Hospital in 1984 and 1985. Initial results were somewhat inferior to those obtained in Copenhagen. There are two reasons for this. Firstly, large differences in disease incidences, especially of chronic medical cases, were found. Secondly, the normal biochemical ranges show large variation between hospitals, even if the units used are the same. Correction methods for these differences were developed. Afterwards, the performance of the algorithm was comparable to that obtained elsewhere. We conclude that decision support systems can be transferred and discuss how local adaptations can be made to retain optimal performance.

The aim of the algorithm is to assign diagnostic probabilities to each of the four classes, based on data from a specific patient. To facilitate application of the algorithm without advanced computational facilities a chart was designed on which the scores of the symptoms, signs, and laboratory results could be marked and the resulting diagnostic probabilities computed (Mat84a). The chart is shown in table 5.2.

Table 5.1. COMIK classification of final diagnoses.

Acute medical (non obstructive)	Chronic medical (non-obstructive)
Acute viral hepatitis	Alcoholic cirrhosis
Drug hepatitis	Posthepatitic cirrhosis
Alcoholic hepatitis	Cryptogenic cirrhosis
Chronic persistent hepatitis	Primary biliary cirrhosis
Septicemia	Chronic active hepatitis
Postoperative jaundice	Hepatocellular carcinoma
Heart failure	
Congenital hyperbilirubinemia	
Hemolysis	
Benign surgical (obstructive)	**Malignant surgical (obstructive)**
Choledocholithiasis	Bile duct carcinoma
Acute cholangitis	Pancreatic carcinoma
Pancreatitis	Liver metastasis
Iatrogenic bile duct lesion	
Secondary biliary cirrhosis	

Table 5.2. Part A.	Med vs Sur	Acu vs Chr	Ben vs Mal		Med vs Sur	Acu vs Chr	Ben vs Mal
Age							
31 - 64 years	+7	+5		Spiders	-6	+11	
>= 65 years	+12	+5					
Jaundice due to previous cirrhosis	-7	+8		Ascites	-3	+6	
Cancer in GI tract, pancreas or bile system or breast	+10		+7	Liver surface nodular		+5	
Leukemia or malignant lymphoma	-13			Gallbladder: Courvoisier	+16		+11
				Firm or tender	+5		
Previous biliary colics or proven gallstones	+3	+7	-7	S-bilirubin > 200 µmicromol/l	+5	-5	+5
In treatment for congestive heart failure		-5		S-alk. posph 400 - 1000 U/l	+6		
				> 1000 U/l	+11		+6
Present history > 2 weeks			+7	S-ASAT 40 - 319 U/l		+5	
				> 320 U/l	-10	+1	+6
Upper abd. pain: severe	+9		-6	Clotting factors (PP) < 0.55		+8	+5
slight or moderate	+4			0.56-0.70		+5	+5
Fever:				S-LDH			
without chills		-3	-5	> 1300 U/l		-5	+7
with chills		-6	-10				
Intermittent jaundice	+5		-5				
Weight loss (> 2 kg)			+4	SUM B	...	...	...
Alcohol				SUM A	...	...	...
1-4 drinks per day	-4			Constant	-17	-19	-9
> 5 drinks per day	-4	+4		Copenhagen priors	-2	-2	1
SUM A	...	...	...	Total score	...	...	...

Table 5.2. COMIK chart for early differential diagnosis of jaundice. Copenhagen version. **Part B.**

Algorithm for differential diagnosis of jaundice

The algorithm overleaf is composed of variables selected from a database comprising 1000 jaundiced patients.

Instructions

1. Circle scores corresponding to symptoms and findings present

2. Sums a and b are calculated from the encircled scores

3. A total score is calculated for each column by addition of sum a, sum b, the constant and the 'prior' terms

4. If in the first column (medical vs surgical) the total score is *negative*, *medical* jaundice is more likely. The probability is read from the adjacent scale. The corresponding probability of surgical jaundice is 1 - the probability of medical jaundice. A *positive* score signifies *surgical* jaundice as more likely. The probability is read from the scale and the complementary probability of medical jaundice is 1 - the probability of surgical jaundice. Thus, a total score of + 6 signifies an 80% probability of surgical and a 20% probability of medical jaundice.

5. The probabilities are entered below:

Medical	= ..... (A),	Surgical	= ..... (B)
Acute	= ..... (C),	Chronic	= ..... (D)
Benign	= ..... (E),	Malignant	= ..... (F)

6. The probabilities of the patient belonging to the four main categories are:

Acute medical :	(A) * (C) =
Chronic medical :	(A) * (D) =
Benign surgical :	(B) * (E) =
Malignant surgical :	(B) * (F) =

Table 5.2, Part C. Translation of scores into probabilities

Scores	Probability	Scores	Probability
0	0.50	10	0.91
1	0.56	11	0.93
2	0.61	12	0.94
3	0.67	13	0.95
4	0.72	14	0.96
5	0.76	15	0.97
6	0.80	16-18	0.98
7	0.83	19-22	0.99
8	0.86	>= 23	1.00
9	0.89		

Patients

The patient population used for this retrospective study consisted of 100 consecutive clinical admissions to the Department of Internal Medicine II at Dijkzigt Hospital, Rotterdam between 1984 and 1985. To be included in the study a patient had to satisfy three criteria at admission : aged over 15 yr, visible signs of jaundice and a serum bilirubin exceeding 17 µmol/l. Approximately 50 patients satisfying these criteria are admitted each year.

A final diagnosis was coded according to an internationally accepted classification and based on clinical chemical tests and/or pathoanatomical findings, usually biopsy, ERCP, laparotomy or autopsy. In some cases, some doubt remained about the final diagnosis, but all patients could be classified into the four major categories.

The patient material, summarized in table 5.3, includes 100 consecutive cases of jaundice satisfying the criteria mentioned earlier. There were 54 males (mean age 59 yr, range 18 - 82) and 46 females (mean age 58 yr, range 21 - 90). The chronic medical and malignant surgical disease categories were the most common ones in this population.

Table 5.3. Characteristics for the 100 Rotterdam jaundiced patients and a comparison with the Copenhagen patients.

Value	Rotterdam (%)	Copenhagen(%)
Sex		
Females	54	60
Males	46	40
Duration of complaints		
< 2 weeks	14	45
> 2 weeks	86	55
Age(yr)		
15 - 30	7	16
31 - 50	13	18
51 - 64	51	26
> 65	29	40
Diagnostic categories		
Acute non-obstructive	9	37
Chronic non-obstructive	51	22
Benign obstructive	14	17
Malignant obstructive	26	22
Other	0[1]	2

Clinical Chemical Tests

Considerable differences exist between the normal ranges of the clinical chemical tests in Dijkzigt Hospital, Rotterdam and at Hvidovre Hospital, Copenhagen. Although some studies indicate the impact of systematic errors is not very important (Mat84b, Hil80), a preliminary analysis revealed that

1 No patients were left out due to diagnostic uncertainty

correction was necessary. A simple adjustment method was chosen which aligns the medians of the normal-range interval only. All computations were done using the corrected values for the biochemical parameters. The normal ranges from either hospital and the correction factors adopted are summarized in table 5.4.

Table 5.4. Normal ranges for biochemical parameters for Dijkzigt and Hvidovre hospitals, and the resulting transfer factor.

Parameter	Normal range Dijkzigt (Rotterdam)	Normal range Hvidovre (Copenhagen)	Transfer factor
Bilirubin (µmol/l)	2 - 12	5 - 17	1.6
Alkaline phosp. (U/l)	25 - 75	51 - 275	3.3
ASAT (U/l)	5 - 30	10 - 40	1.4
LDH (U/l)	160 - 320	200 - 450	1.4

Results

Initial Application

The results of application are summarized in table 5.5. Two measures of performance were used. The percentage of cases classified correctly by the algorithm's best bid (the category assigned highest probability) was used as a simple measure of performance. We also evaluated the performances for those cases in which a confident diagnosis, i.e. a probability > 0.80, was achieved. In fact, several other statistics, discussed in (Hab78), could be used. The COMIK chart contains 3 algorithms for binary classification. In the evaluation of performance we looked at the binary classifications and four-way classifications separately.

Missing Data

The chart published by the COMIK group requires that data are complete. Unpublished factors to handle missing data were available to us but

the effect on the performance was negligible. This is a reflection of the relative 'completeness' of our database. We draw any conclusions about the beneficial effects that this extension might have for less complete databases.

Table 5.5. Performance of the COMIK algorithm in the Rotterdam patients for different classification subtasks.

	Copenhagen incidences		Rotterdam incidences	
	best bid (%)	confident (%)	best bid (%)	confident (%)
Obstruction vs non-obstruction	82 %	89 %	87 %	93 %
Acute vs chronic	80 %	87 %	95 %	97 %
Benign vs malignant	75 %	78 %	72 %	77 %
Classification in 4 categories	70 %	88 %	77 %	90 %

Adjustments for local disease incidence

The COMIK chart uses the relative incidences from the Copenhagen database. The results of initial application using Copenhagen incidences are presented in table 5.5. Table 5.4 already showed that marked differences exist between the relative occurrence of the two non-obstructive types of jaundice in Rotterdam and Copenhagen. Correction for the local incidences is, therefore, indicated. This can be done by replacing the score terms derived from the Danish incidences by their Rotterdam counterparts. In the chart given in table 5.2 they are shown as separate constants marked 'Copenhagen priors'; in the original chart (Mat84a) the incidences were combined with the preceding constants. One must replace the Copenhagen incidence terms (shown as the triplet -2,-2,1 in table 5.2) by the Rotterdam terms (-2,8,3) reflecting the disease incidences from table 5.3. The computation of the triplet employs the following conditional frequencies:

acute cases :	9 / (51 + 9)	= 15 %	(table 5.3)
chronic cases :	51 / (51 + 9)	= 85 %	
benign cases :	14 / (14 + 26)	= 35 %	
malignant cases :	26 / (14 + 26)	= 65 %	

The score system used in the chart employs decimal logarithms of probability odds, multiplied by 10 for convenience. The corrected incidence triplet is found by taking the appropriate logarithms:

medical 60 % vs surgical 40 % -> 10 * $Log_{10}(40/60)$ = -2,
acute 15% vs chronic 85% -> 10 * $Log_{10}(85/15)$ = 8,
benign 35% vs malignant 65% -> 10 * $Log_{10}(65/35)$ = 3,

The results of applying the COMIK rule with these Rotterdam relative incidences are also presented in table 5.5. A classification matrix, showing the final classification using the Rotterdam incidences is given in table 5.6.

Tables 5.6a & 5.6b. Results of the COMIK chart using Rotterdam relative incidences. Columns: final clinical diagnosis; Rows: the diagnostic class assigned the highest probability by the algorithm. AM = acute medical, CM = chronic medical, BS = benign surgical, MS = malignant surgical

Table 5.6a. All patients

	Clinical diagnosis				
Algorithmic diagnosis	AM	CM	BS	MS	Total
Acute Medical	7	2	2	2	13
Chronic Medical	1	44	4	3	52
Benign Surgical	0	0	6	1	7
Malignant Surgical	1	5	2	20	28
Total	9	51	14	26	100

Table 5.6b. Only patients in whom some diagnostic category was assigned a probability of 80 % or more.

	Clinical diagnosis				
Algorithmic diagnosis	AM	CM	BS	MS	Total
Acute Medical	2	1	0	0	3
Chronic Medical	0	38	0	0	38
Benign Surgical	0	0	1	0	1
Malignant Surgical	0	3	1	9	13
Total	2	42	2	9	55

Table 5.7. Performance of COMIK four-group algorithm in different countries (percentages correct classification, percentage confident diagnoses in parenthesis).

Classification	Rotterdam (Dijkzigt)	Copenhagen (Hvidovre & Frederiksberg)	Stockholm	Mexico (ISSSTE)
Number of patients	N = 100	N = 108	N = 144	N = 1000
Best bid i.e. highest probability	77%	77%	76%	89%
confident, i.e. P > 0.80	90% (55%)	92% (56%)	93% (41%)	?? (??%)

A Comparison

The COMIK algorithm has been tested in several countries. A comparison of our results with those of Mexico, Sweden and Denmark is shown in table 5.7. The Swedish and Mexican studies employed easy corrections,

similar to those discussed in the previous sections, in order to correct for differences in incidence and laboratory practice. We conclude that the performance is comparable to that in Sweden and Denmark.

Discussion

Most clinical knowledge is based on research projects carried out at places other than the location of the medical practitioner. This makes the problem of transferral a universal one. Also, a clinician should address the problem of what adaptations have to made for a successful transfer to his own clinic. It is precisely this problem that has been studied in the present paper.

Almost 10 years have passed since data collection for the COMIK pocket diagnostic chart started. Published in 1984, it has now been tested in various centers. All studies show that the algorithm has preserved most of its accuracy, even when considerable differences in population composition are present. Some authors report that transfer of a diagnostic algorithm from a patient population on which it is based, to another setting, may induce a serious decline in performance (Bje76). This may be due to several causes, some of them related to the composition of the population under study, like hospital referral patterns and disease incidences, or the distribution of symptoms and signs (DeD81). We found no evidence of such a decline in performance with the COMIK algorithm once the obvious local adaptations were made.

Some people may be dissatisfied with these results. Their view may, however, be biased by the diagnostic accuracy that is obtained once more invasive diagnostic procedures have been used. Better performance, based on such early data, is hard to achieve. The COMIK algorithm is no alternative for invasive diagnostic tests. The purpose is to have an early assessment of the patient and to guide the selection of subsequent tests.

Large-scale data collection takes a long time and some diagnostic procedures may change in the mean time. This also applies to the COMIK algorithm, which does not include information available from ultrasound and hepatitis serology, both among the commonest investigations nowadays. For this reason our future research question will be whether incorporation of these techniques in the COMIK algorithm will improve its classification tasks.

The Rotterdam Dijkzigt Internal Medicine population differs considerably from the original Copenhagen population, probably as a consequence of difference in referral patterns. The impact of these referral patterns and resulting differences in disease distribution on diagnostic management has hardly been recognized. Diagnostic techniques that perform well on one location, may turn out to be less effective on an other location. The concept that diagnostic performance may depend on disease distributions, as shown for the COMIK algorithm, is generally applicable. A diagnostic strategy should only be preferred with these concepts in mind.

The statistical score-based approach to decision making as employed by the COMIK chart is one out of many conceivable methods. Other approaches like decision trees and expert systems (Luc85) may be also appropriate for these purposes. Looking into the future it cannot be expected that they will play a major role in routine medical decision making. Nevertheless, they all converge in their ability to impose a structure upon a problem, and focus attention to those elements with a maximum contribution to the solution of the problem. This is rewarding, not only for clinical use, but also for educational purposes.

References

Bje76 Bjerregaard B, Brynitz S, Holst-Christensen J, Kalaja E, Lund-Kristensen J, Hilden J, De Dombal FT, Horrocks JC. Computer-aided diagnosis of the acute abdomen: a system from Leeds used on Copenhagen patients. In: De Dombal FT, Grémy F, eds. Decision Making and Medical Care. Amsterdam: North Holland Publishing Company, 1976: 165-71.

Boo81 Boom RA, Gil D, Maass R, Manrique G. The differential diagnosis of obstructive jaundice based on a logarithmic index of alkaline phosphatase and total cholesterol values. Med Decis Making 1981; 1: 227-37.

Boo86 Boom R, Gonzalez C, Fridman L, Alaya JF, Realpe JL, Morales P, Quintero R. Looking for 'Indicants' in the Differential Diagnosis of Jaundice. Med Decis Making 1986; 6: 36-41.

Bro81 Brown G. Bayes' Formula. Conditional probability and clinical medicine. Am J Dis Child 1981; 135: 1125-9.

DeD81 De Dombal FT, Staniland JR, Clamp SE. Geographical variation in disease presentation. Does it constitute a problem and can information science help? Med Decis Making 1981; 1: 59-69.

Hab78 Habbema JDF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis. Parts 1 - 3: Meth Inf Med 1978; 17: 217-46. Parts 4 - 5: Meth Inf Med 1981; 20: 80-100.

Hil80 Hilden J, Matzen P, Malchow-Møller A, Bryant S. Precision requirements in a study of computer-aided diagnosis of jaundice (The COMIK study).
Scand J Lab Invest 1980; 40(supp 155): 125-8.

Kni73 Knill-Jones RP, Stern RB, Girmes DH, Maxwell JD, Thompson RPH.
Williams R. Use of sequential Bayesian model in diagnosis of jaundice by computer. Br Med J 1973; 1: 530-3.

Lin83 Lindberg G, Nilsson L, Thulin L. Decision theory as an aid in the diagnosis of cholestatic jaundice. Acta Chir Scand 1983; 149: 521-9.

Lin87 Lindberg G, Thomsen C, Malchow-Møller A, Matzen P, Hilden J. Differential diagnosis of jaundice: applicability of the Copenhagen pocket chart proved in Stockhom patients. Liver 1987; 7: 43-9.

Luc85 Lucas PJF, Janssens AR. Medische expert systemen: hulpmiddel bij diagnose en therapie. Ned Tijdschr Geneeskd 1985; 129: 160-5.

Mat84a Matzen P, Malchow-Møller A, Hilden J, Thomsen C, Svendsen LB, Gammelgaard J, Juhl E. Differential diagnosis of jaundice: a pocket diagnostic chart. Liver 1984; 4: 360-71.

Mat84b Matzen P, Hilden J, Thomsen C, Malchow-Møller A, Juhl E. Does test quality influence the outcome of algorithmic classification of jaundiced patients? Scand J Lab Invest 1984; 44(supp 171): 35-40.

Sai85 Saint-Marc Girardin MF, le Minor M, Alperovitch A, Roudot-Thoraval F, Metreau J-M, Dhumeaux D. Computer-aided selection of diagnostic tests in jaundiced patients. Gut 1985; 26: 961-7.

Ste73 Stern RB, Maxwell JD, Knill-Jones RP, Thompson RPH, Williams R. Use of computer-assisted model in diagnosis of drug hypersensitivity jaundice. Br Med J 1973; 2: 767-9.

Ste75 Stern RB, Knill-Jones RP, Williams R. Use of computer program for diagnosing jaundice in district hospitals and specialized liver unit. Br Med J 1975; 2: 659-62.

Whe79 Wheeler PG, Theodossi A, Pickford R, Laws J, Knill-Jones RP, Williams R. Non-invasive techniques in the diagnosis of jaundice. Ultrasound and computer. Gut 1979; 20: 196-9.

Chapter 6

Evaluation of a flowchart for early diagnosis of jaundice

Segaar RW, Wilson JHP, Habbema JDF.
Submitted for publication, 1990.

Abstract

The Copenhagen flowchart for the differential diagnosis of jaundice differentiates obstructive from non-obstructive causes of jaundice, based on information of 13 variables. With additional information, patients with non-obstructive jaundice can be further classified into acute and chronic causes, and patients with obstructive jaundice can be classified into benign and malignant causes of jaundice. The information required by the flowchart is usually available within 24 to 48 hours.

There was no information regarding evaluations of the flowchart outside Copenhagen, and therefore we conducted an evaluation of the flowchart with data from 214 consecutive patients admitted to Dijkzigt hospital in Rotterdam. At initial application, it turned out that adjustment procedures for biochemistry parameters were necessary. Before correction, diagnostic performance was 73 % correct for classification into obstruction and non-obstruction. After corrections for differences in reference ranges, it improved to 83% correct diagnoses. For the classification into acute and chronic non-obstruction we found 67% correct classifications, and for classification into benign and malignant obstruction 79%, both after corrections. In spite of these corrections made, the results are inferior to those found in Copenhagen. We conclude that the transfer of the Copenhagen flowchart to Rotterdam was no success. The results contrast our experience with another diagnostic aid from Copenhagen (the COMIK algorithm) which yielded favorable results after transfer. Possible explanations for the results will be presented.

Introduction

Many new diagnostic techniques have become available for diagnostic assessment of patients presenting with jaundice. What technique to choose remains a matter of experience, and crucially depends on an interpretation of early patient information, like history and physical examination. Lindberg (Lin83) and Safran (Saf88) have integrated this issue into a decision analytic framework emphasizing the importance of an early probabilistic assessment of patients presenting with jaundice. To facilitate the interpretation of early diagnostic information, flowcharts can be used.

Flowcharts (also referred to as decision trees) are designed for the management of circumscribed problems, for example the symptom "jaundice". They provide a graphical tree-like summary of relevant information for the problem. At each branchpoint in these trees, relevant questions are formulated. Starting from the root of the flowchart, a user follows the paths through the flowchart, according to the answers on the diagnostic questions in the chart. End-points usually coincide with a diagnostic classification, or a suggestion for action. In 1984, the British Medical Journal published an extensive series of flowcharts, covering a wide range of problems (see references Sco84, She84, Orm84, Jam84a, Jam84b, McH84a, McH84b, Tow84 for a random sample). Looking at publications found in other journals (for example Phi87, Por87, Gif88 and Mac88) and in books (for example Ber85, Bre87 and Wei87) it shows that plenty examples of flowcharts can be found in the literature.

Flowcharts often originate from expert experience. In that case they reflect the opinion of the author, which may not always coincide with that of the user. An other way to derive flowcharts is through consensus meetings. In that case, the resulting flowchart may anticipate acceptance from the major part of the professional group. Statistical analysis of databases is a third alternative. From an analysis of data collected on symptoms, signs, test results and final diagnoses, we can derive flowcharts for diagnostic purposes. This removes subjective elements from the compilation of a flowchart. Such an approach was used by the COMIK group (Mal88). They published a flowchart for classification of adult jaundice, based on analysis of a database including 1002 patients presenting with jaundice. Earlier flowcharts for jaundice (for example Tha77, Ost75, Fis81), covered all age categories from ranging newborns, through older children, to adults. These flowcharts were however not derived in a formal manner.

The COMIK flowchart has been evaluated in Copenhagen, where it displayed a good performance, but we found no studies that evaluated the flowchart outside Copenhagen. In an earlier study (Seg88) we evaluated an other diagnostic aid for jaundice from Copenhagen, the COMIK chart. The COMIK chart uses logistic regression for classification of jaundice. That study indicated that transfer of diagnostic aids to other centers can be associated with a decline in performance, unless local adjustment procedures are implemented. This motivated us to conduct an evaluation study of the flowchart in Rotterdam. With the present study we want to answer to following questions:

- what is the performance of the COMIK flowchart in Rotterdam?
- how does the performance compare to that of the COMIK algorithm in Rotterdam?
- how does the performance compare to the performance in Copenhagen?
- what is the impact of referral patterns on the diagnostic classification from the flowchart?

Patients and Methods

The COMIK flowchart for classification of jaundice

In the diagnostic classification used for the flowchart, adult jaundice is subdivided into four groups. First, there is a two way subdivision into non-obstructive (medical) and obstructive (surgical) causes of jaundice. A next level of classification subdivides medical into acute and chronic causes, and surgical into benign and malignant causes. This flowchart provides a classification of jaundice along these four major groups.

The flowchart for classification of jaundice is subdivided into three parts. The first part is used for the differentiation between obstruction and non-obstruction and uses 13 variables. The flowchart for differentiation between benign and malignant obstruction uses 7 variables, and the flowchart for differentiation between acute and chronic non-obstruction uses 5 variables. Some variables, for example serum-bilirubin, are used several charts. The variables, their usage and the cut-off points for continuous variables are shown in table 6.1. The flowcharts that correspond to the three classifications tasks are shown in figures 6.1 - 6.3. In these flowcharts, results of classification of the 214 Rotterdam patients are shown too.

Application of a flowchart for an individual patient is straightforward. Starting from the left of the flowchart a user follows branchpoints until an endnode is reached. At each branchpoint the decision in which direction to proceed depends on the answer to the diagnostic question presented at that branchpoint. Endnodes of the flowchart are associated with a diagnostic category. Ending in one of the endnodes corresponds to classification of the patient into the diagnostic category itemized there. Although there are many endnodes in one flowchart, they refer to only two diagnostic categories. If a user encounters a branchpoint for which diagnostic information is missing, the patient is left unclassified.

First, the flowchart for differentiation between obstructive and non-obstructive causes (figure 6.1) is applied. As a result, the patient is classified into one of these two categories. If necessary, patients can be classified further using the flowchart for obstruction (figure 6.2), or the flowchart for non-obstruction (figure 6.3), depending on the result of application of the first flowchart.

The COMIK flowchart also has a refinement that is not seen in other flowcharts. It provides two types of diagnostic endnodes: endnodes that are called *certain* and endnodes that are called *doubtful*. This allows a rough probabilistic assessment of the outcome of application of the flowchart. The rationale behind this partition is that doubtful endnodes require further diagnostic testing before classification. In fact, the development of the flowchart has historically been closely connected to the development of a probabilistic algorithm for the differential diagnosis of jaundice, which is based on a logistic discrimination analysis of a large data collection (Mat84).

Table 6.1. Variables used in the flowcharts for differentiation of jaundice. ME = Medical, AM = Acute Medical, CM = Chronic Medical, SU = Surgical, BS = Benign Surgical, MS = Malignant Surgical. + indicates that a test is used for a specific differentiation.

		Differentiation		
Variable	Cut-off value (if relevant)	ME vs SU	AM vs CM	BS vs MS
Age	above 65 years	+		
Previous gallstone manifestations	present/absent	+	+	
Previous malignant disease	present/absent	+		
Exposure to hepatitis	present/absent	+		
Severe abdominal pain	present/absent	+	+	
Upper abdominal pain with radiation to the back	present/absent	+		
Palpable gallbladder or local tenderness	present/absent	+		
Ascites and/or increased abdominal girth	present/absent	+		+
One or more typical spider naevi	present/absent	+		
Serum Bilirubin	> 320 µmol/l	+	+	
Serum Alkaline Phosphatase	> 411 U/l	+		
Serum Alkaline Phosphatase	> 965 U/l	+		
Serum aspartate aminotransferase	> 320 U/l	+		+
HBsAg serology	positive/negative	+		
First symptom	> 2 weeks		+	+
Fever with chills	present/absent		+	
Weightloss	> 2 kg		+	+
Liver surface nodular	present/absent		+	
Alcohol consumption > 50 gram daily for a period > 5 years	present/absent			+

Figure 6.1. The flowchart for differentiation between obstructive and non-obstructive causes of jaundice and results for 214 Rotterdam patients. obs: the flowchart diagnosis is grey zone obstruction; nobs: the flowchart diagnosis is grey zone non-obstruction; OBS: the flowchart diagnosis is obstruction; NOBS: the flowchart diagnosis is non-obstruction; nn C: number(nn) of patients correctly classified; nn F: Number(nn) of patients falsely classified

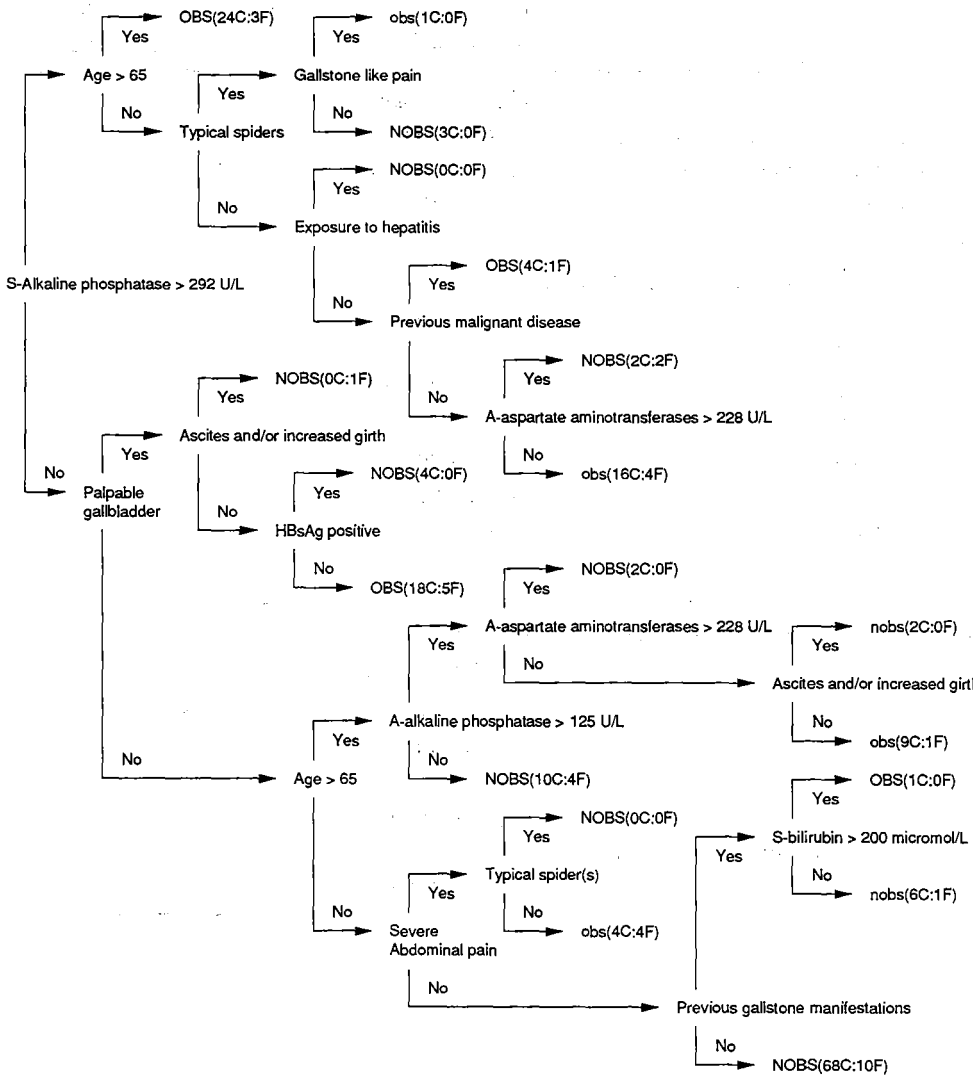


Figure 6.2. The flowchart for differentiation between acute and chronic causes of non-obstructive jaundice and results for 214 Rotterdam patients. acu: the flowchart diagnosis is grey zone acute non obstruction; chr: the flowchart diagnosis is grey zone chronic non obstruction; ACU: the flowchart diagnosis is acute non obstruction; CHR: the flowchart diagnosis is chronic non obstruction; nn C = number(nn) of patients correctly classified; nn F = number(nn) of patients falsely classified

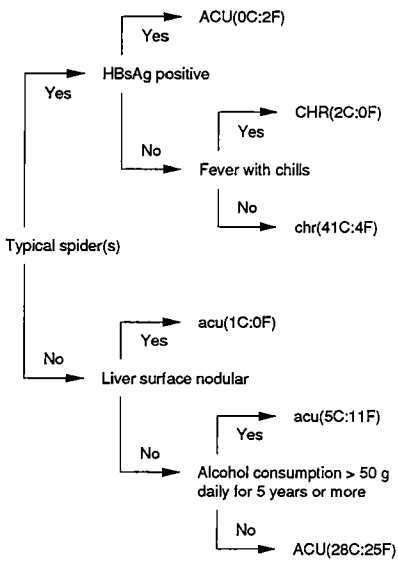
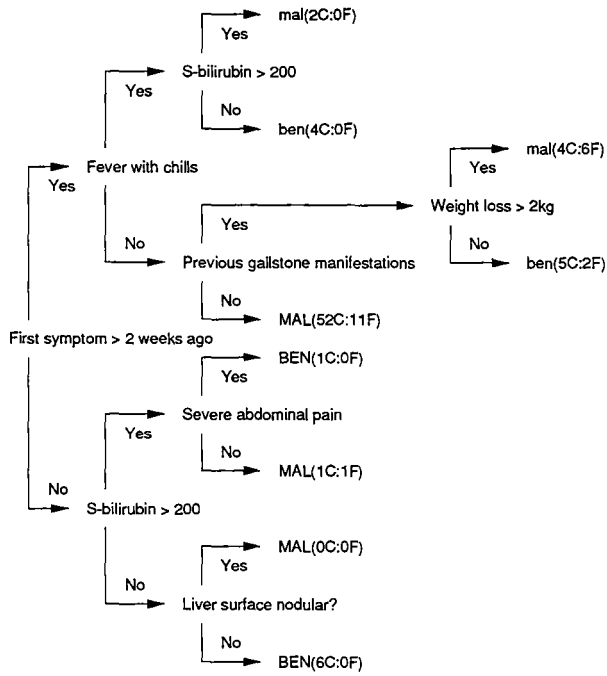


Figure 6.3. The flowchart for differentiation between malignant and benign causes of obstructive jaundice and results for 214 Rotterdam patients. ben: the flowchart diagnosis is grey zone benign obstruction; mal: the flowchart diagnosis is grey zone possible malignant obstruction; BEN: the flowchart diagnosis is benign obstruction; MAL: the flowchart diagnosis is malignant obstruction; nn C: number(nn) of patients correctly classified; nn F: number(nn) of patients falsely classified.



- age above 15 years,
- visible jaundice, and
- serum-bilirubin above 17 μmol/liter

Data obtained from patient history, physical examination and diagnostic tests were recorded. A clinical diagnosis was coded later, using the disease classification of the COMIK algorithm (Mat84), that is also employed for the flowchart. Although we could not establish a confident detailed final diagnosis for all patients, patients could be classified into the four main diagnostic categories (table 6.2).

All patient data were used in the evaluation of the flowchart that classifies into obstruction and non-obstruction. In addition, patient data were subdivided into two subsets, based on their final diagnoses. A first subset of patients with non-obstructive diagnoses provided a test population for the flowchart that classifies non-obstructive cases. In the same way, a second subset with obstructive diagnoses provided a test population for the flowchart that classifies obstructive cases.

Results

Before correction for biochemistry, the flowchart classified 73% of the cases correctly as obstruction and non-obstruction. Based on our experience with the evaluation of the COMIK algorithm (Seg88), it was obvious that corrections for differences in reference-ranges of biochemistry were also necessary for the flowchart. We aligned the medians of the reference ranges of the Copenhagen laboratory and our laboratory. Reference-ranges and the transfer factors that were derived are shown in table 6.3. After correction for biochemistry, the performance was improved to 83% correct diagnoses. The improvement is consistent with our previous experience on the transfer of a diagnostic algorithm for jaundice (Seg88).

To evaluate the two other flowcharts, we the other two subsets described in the patient section. The flowchart for classification into benign and malignant obstruction was used after correction for biochemistry. Results of the evaluation are summarized in table 6.4.

Table 6.2. Final diagnoses in 214 jaundiced patients

Category	Final diagnoses	no of patients
Acute medical	Acute viral hepatitis	14
N=37 (17 %)	Drug hepatitis	11
	Alcoholic hepatitis	0
	Chronic persistent hepatitis	0
	Septicemia	6
	Post-operative jaundice	0
	Heart failure	3
	Congenital hyperbilirubinemia	0
	Hemolysis	3
	Other	0
Chronic Medical	Alcoholic cirrhosis	31
N=82 (38 %)	Posthepatitic cirrhosis	6
	Cryptogenic cirrhosis	8
	Primary biliary cirrhosis	12
	Chronic active hepatitis	15
	Hepatocellular carcinoma	8
	Other	2
Benign surgical	Choledocholithiasis	22
N=34 (16 %)	Acute cholangitis	1
	Pancreatitis	2
	Iatrogenic bile duct lesion	1
	Secondary biliary cirrhosis	5
	Other	3
Malignant Surgical	Bile duct carcinoma	21
N=61 (29 %)	Pancreatic carcinoma	23
	Liver metastasis	13
	Other	4
N=214 (100%)	Total	214

Table 6.3. Reference ranges in Copenhagen and Rotterdam, and derived transfer factors.

Parameter	Reference range in Rotterdam	Reference range in Copenhagen	Transfer factor
Bilirubin (μmol/l)	2 - 12	5 - 17	1.6
Alkaline Phosp. (U/l)	25 - 75	51 - 275	3.3
ASAT (U/l)	5 - 30	10 - 40	1.4
LDH (U/l)	160 - 320	200 - 450	1.4

Table 6.4. Results of application of the flowcharts on for differentiation of jaundice.

Classification	Correct (%)	Total(%)	Unclassified
Obstruction vs non-obstruction	174 (83 %)	210 (100 %)	4
Acute vs chronic	77 (67 %)	115 (100 %)	4
Benign vs malignant	75 (79 %)	95 (100 %)	0

We also evaluated the performance on the four-way classification. It implies that for each patient, two flowcharts are used. First the main flowchart for differentiation into obstruction or non-obstruction is applied. Depending on the outcome of this application, one of the two other flowcharts is applied. Results of this four-group classification are compared to the actual diagnosis in table 6.5. From this table it becomes apparent that failures on the classification of chronic non-obstructive and benign obstructive cases are responsible for the majority of misclassifications.

Table 6.5. Results of the four-way classification of the flowchart for classification of jaundice. Columns: final clinical diagnosis. Rows: diagnostic class selected by the flowchart. AM = Acute medical, CM = Chronic medical, BS = Benign surgical, MS = Malignant surgical. Results: 131/210 = 63.8 % correctly classified

	Clinical diagnosis				
Flowchart diagnosis	AM	CM	BS	MS	Total
Acute Medical	27	26	3	14	70
Chronic Medical	2	42	1	0	45
Benign Surgical	2	4	16	1	23
Malignant Surgical	3	9	14	46	72
Unclassified	3	1	0	0	4
Total	37	82	34	61	214

A comparison

In table 6.6 we compare the performance of the flowchart in Rotterdam with the performance found in Copenhagen. The table shows that for all classifications studied, results of the flowchart turn out less favorable in Rotterdam than in Copenhagen. The main reasons for the misclassifications were shown in table 6.5. It seems that the flowchart for differentiation between obstruction and non-obstruction maintains reasonable results, whereas the flowchart for differentiation of acute and chronic non-obstruction performs quite unfavorable.

Table 6.6 also shows a comparison of the performance of the flowchart and the COMIK algorithm on these same patients. Evaluation of the COMIK algorithm on the N=214 database is an extension of an earlier study (Seg88). The performance of the COMIK algorithm is superior to the performance of the flowchart on both the two- and four-group classifications. We applied the COMIK algorithm after the same corrections for the biochemistry and using relative disease incidences in Rotterdam, estimated from an earlier data-collection.

Table 6.6. Results of the COMIK algorithm and the flowchart compared. AM = Acute medical, CM = Chronic medical, BS = Benign surgical, MS = Malignant surgical

	Percentage correct classified			
Diagnostic aid	ME - SU	AM - CM	BS - MS	Fourway[1]
Flowchart with ROTTERDAM data (N=214)	83 %	67 %	79 %	62 %
Flowchart with Copenhagen database (N=982)	87 %	89 %	86 %	77 %
Flowchart with Copenhagen testdata (N=108)	91 %	no data	no data	72 %
COMIK algorithm with ROTTERDAM data (N=214)	90 %	87 %	78 %	77 %

The impact of referral on classification performance

The Dijkzigt patient population involves many secondary referrals. This may imply a selection towards complex cases. We therefore also evaluated differences in performance of the flowchart that classifies into obstruction and non-obstruction on three distinct groups:

- new admissions (N=71)
- re-admitted patients (N=53)
- secondary referrals from other hospitals (N=90)

1 Four-way refers to the performance on classification into four groups

Small differences in performance were found, see table 6.7. Given to the small numbers of patients in each group, these differences are too small to conclude that significant differences exist.

Table 6.7. Results of application of the flowchart for differentiation between obstruction and non-obstruction for three referral groups after corrections for biochemical parameters.

Referral group	Correct (%)	Total	Unclassified
New admissions	60 (86%)	70 (100%)	1
Re-admitted patients	41 (82%)	50 (100%)	3
Secondary referrals	73 (81%)	90 (100%)	0

A comparison of confident and grey-zone endnodes

Endnodes of the flowcharts may be of a "confident", and a "grey-zone" type. The interpretation is that confident diagnoses are correct in a greater percentage of the cases than grey-zone diagnoses are. In this way, the flowchart allows for some probabilistic interpretation. The proposal for this labeling came from the authors of the flowchart, and was based on their analysis of data from the Copenhagen patients.

We examined whether results on "confident" endnodes were indeed better than results on "grey zone" endnodes for Rotterdam patients. On "grey zone" endnodes we found 80 % correct classification, whereas "confident" endnodes showed 84% correct classification for the flowchart for obstruction versus non-obstruction. We doubt whether it is advisable to use confident endnodes only, if only such a small difference exists, since a large proportion of patients was left unclassified. Similar results and conclusions were obtained for the flowchart for classification into benign and malignant obstruction. For the flowchart on acute versus chronic non-obstruction, we found only one node with good discrimination. We conclude that the distinction between confident and grey-zone endnodes is a feature that is hard to preserve after transfer.

Discussion

The diagnostic performance of the flowchart is inferior to that of the COMIK algorithm. This seems especially due to a poorer performance on differentiation between acute and chronic types of non-obstructive jaundice (see tables 6.5 and 6.6). Several explanations come to mind. First of all, performance of the COMIK algorithm (Seg88) improved from corrections in relative disease incidences. No such corrections are readily available for the flowchart. We believe that this accounts for a large part of the differences found.

Looking at the flowchart for differentiation between acute and chronic non-obstruction (figure 6.1), a more clinical explanation can be obtained. The figure shows that only in one case a patient could be classified correctly because a nodular surface was found on examination. We suspect that on careful examination of the chronic non-obstructive patients probably more livers with a nodular surface might be detected. Such discrepancies illustrate the problem of transfer of decision aids which (partially) rely on features with different local traditions in use and interpretation. Theodossi (The81) and Lindberg (Lin81) studied this issue in more detail.

Missing parameters put forth problems in the use of flowcharts. Although provisions to deal with missing values were discussed by the authors of the flowchart (Mal88), they require complex computations, preventing use on a routine basis. This is a disadvantage if many patients have information on one or more parameters missing. In the present evaluation study, only 4 patients had one or more required parameters missing. This will not have influenced the results of our evaluation. Summarizing on the issue of performance, we conclude that from a performance point of view, the COMIK algorithm should be preferred to the flowchart, both in Copenhagen and in Rotterdam. But when ease of use of flowcharts is appreciated, and application limited to the major classification into obstruction and non-obstruction, the flowchart for the differential diagnosis of jaundice is a useful addition to techniques that assist the diagnostic work-up of individual jaundiced patients. Its applicability will further increase once techniques like ultrasound and viral serology are incorporated.

Referral patterns can play an important role. They determine the patient mix presenting at a specific hospital. Secondary referrals for example, may be a selection of complex cases which may be especially difficult for the decision aid under study. But such patients usually also have a longer history with a complete set of diagnostic data which makes them easier to deal with. As a consequence, it is difficult to predict what the performance for a specific

referral subgroup will be. Although we found no strong evidence for differences between referral subgroups in our study, selection patterns towards complex cases may indeed decrease performance of decision aids developed elsewhere.

Future research on formalized diagnostic tools, like the COMIK flowchart and the COMIK algorithm, will disclose to which extend their use can be improved and fields of application extended. Routine clinical use of the diagnostic aids discussed in a clinical setting has until now been limited. This is partially due to the fact that jaundice is only one out of many diagnostic problems presenting at an Internal Medicine department. When limited to jaundice it may therefore not be expected that much experience will develop. Emotional objections against formal diagnostic aids are an other reason. Diagnostic aids often use data that - although acknowledged as relevant within the domain - differ from the description of prototype patients. This makes it hard for clinicians to integrate them in their diagnostic decision making process.

The major attraction of the flowchart approach to medical decision making is that it provides insight in the "internals" of its strategy without additional help. Although statistical approaches can be extended to provide similar information (Seg89), a flowchart is superior in that respect. The flowchart approach can be applied to a wide range of problems. Although often used for classification purposes, like the jaundice problem, the use of flowcharts also extends to "what-to-do-if" applications (Kom78, Woo80). Flowcharts focus more on the structure of the clinical decisions, than on the contents of the problem. In that way they are a skeleton upon which additional knowledge can be build. The introspection required from experts to construct flowcharts, is one of the best ways to elicitate their knowledge (Fei74). It also makes them suitable candidates to train medical and para-medical personnel (Kom74, Sox73, Gre74). Application should however not restricted to these groups. Even experienced workers may at moments benefit from the use of flowcharts. The mere fact that they can serve as reminders (McD76), thus preventing things from being overlooked, especially when operating under time constraints (Gol82). In this way we may improve the quality of the care that we provide our patients.

References

Ber85 Berman S, ed. Pediatric decision making. Philadelphia: B.C. Decker, 1985.

Bre87 Bready LL, Smith RB, eds. Decision making in anaesthesiology. Philadelphia: B.C. Decker, 1987.

Fei74 Feinstein AR. An analysis of diagnostic reasoning. III. The construction of clinical algorithms. Yale J Biol Med 1974; 47: 5-32.

Fis81 Fisher MG, Gelb AM, Weingarten LA. Cholestatic jaundice in adults. Algorithms for diagnosis. JAMA 1981; 245: 1945-8.

Gif88 Gifford RW Jr. An algorithm for the management of resistant hypertension. Hypertension 1988; 11(supp 3 pt 2): II101-5.

Gol82 Goldman L, Weinberg M, Weisberg M, Olshen R, Cook EF, Sargent RK. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. N Engl J Med 1982; 307: 588-96.

Gre74 Greenfield S, Bragg FE, McCraith DL, Blackburn J. Upper-respiratory tract complaint protocol for physician extenders. Arch Intern Med 1974; 133: 294-9.

Jam84a Jamieson M. Clinical algorithms. Headache. Br Med J 1984; 288: 1281-3.

Jam84b Jamieson M. Clinical algorithms. Loss of vision. Br Med J 1984; 288: 1523-6.

Kom74 Komaroff AL, Black WL, Flatley M, Knopp RH, Reiffen B, Sherman H. Protocols for physician assistants: management of diabetes and hypertension. N Engl J Med 1974; 290: 307-12.

Kom78 Komaroff AL, Pass TM, McCue JD, Cohen AB, Hendricks TM, Friedland G. Management strategies for urinary and vaginal infections. Arch Intern Med 1978; 138: 1069-73.

Lin81 Lindberg G. Effects of observer variation on performance in probabilistic diagnosis of jaundice. Meth Inf Med 1981; 20: 163-8.

Lin83 Lindberg G, Nilsson L, Thulin L. Decision theory as an aid in the diagnosis of cholestatic jaundice. Acta Chir Scand 1983; 149: 521-9.

Mac88 MacKinnon SE, Dellon AL. A surgical algorithm for the management of facial palsy. Microsurgery 1988; 9: 30-5.

Mal88 Malchow-Møller A, Thomsen C, Hilden J, Matzen P, Mindeholm L, Juhl E. A decision tree for early differentiation between obstructive and non-obstructive jaundice. Scand J Gastroenterol 1988; 23: 391-401.

Mat84 Matzen P, Malchow-Møller A, Hilden J, Thomsen C, Svendsen LB, Gammelgaard J, Juhl E. Differential diagnosis of jaundice: a pocket diagnostic chart. Liver 1984; 4: 360-71.

McD76 McDonald CJ. Protocol-based computer reminders, the quality of care and the non-perfectability of man. N Engl J Med 1976; 295: 1351-5.

McH84a McHardy KC. Clinical algorithms. Weakness. Br Med J 1984; 288: 1591-4.

McH84b McHardy KC. Clinical algorithms. Incoordination. Br Med J 1984; 288: 1668-9.

Orm84 Ormerod AD. Clinical algorithms. Syncope. Br Med J 1984; 288: 1219-22.

Ost75 Ostrow JD. Jaundice in older children and adults. Algorithms for diagnosis. JAMA 1975; 234: 522-6.

Phi87 Phillip M, Bashan N, Smith CP, Moses SW. An algorithmic approach to the diagnosis of hypoglycemia. J Pediatr 1987; 110: 387-90.

Por87 Portenoy RK, Lipton RB, Foley KM. Back pain in the cancer patient: an algorithm for evaluation and management. Neurology 1987; 37: 134-8.

Saf88 Safran C, Greenes RA, Bynum TE, Kierstead ML. Diagnosis of cholestasis: an analytic view. Med Decis Making 1988; 8: 102-9.

Sco84 Scott AK. Clinical algorithms. Management of epilepsy. Br Med J 1984; 288: 986-7.

Seg88 Segaar RW, Wilson JHP, Habbema JDF, Malchow-Møller A, Hilden J, van der Maas P. Transfer of a diagnostic decision aid for jaundice. The application of the COMIK pocket diagnostic chart for the differential diagnosis of jaundice. Neth J Med 1988; 33: 5-15.

Seg89 Segaar RW, Wilson JHP, Habbema JDF, Hilden J. A computer aid for early diagnostic classification of jaundice (the COMIP program). Comp Meth Prog Biomed 1989; 28: 131-6.

She84 Shepherd DI. Clinical algorithms. Sensory disturbances. Br Med J 1984; 288: 1147-9.

Sox73 Sox HC Jr, Sox CH, Tompkins RK. The training of physician's assistants: the use of a clinical algorithm system for patient care, audit of performance and education. N Engl J Med 1973; 288: 818-24.

Tha77 Thaler MM. Jaundice in the newborn. Algorithmic diagnosis of conjugated and unconjugated hyperbilirubinemia. JAMA 1977; 237: 58-62.

The81 Theodossi A, Knill-Jones RP, Skene A, Lindberg G, Bjerregaard B, Holst-Christensen J, Williams R. Inter-observer variation of symptoms and signs in jaundice. Liver 1981; 1: 21-32.

Tow84 Towler HMA. Clinical algorithms. Dizziness and vertigo. Br Med J 1984; 288: 1739-43.

Wei87 Weisberg LA, Strub RL, Gargia CA, eds. Decision making in adult neurology. Philadelphia: B.C. Decker 1987.

Woo80 Wood RW, Tompkins RK, Wolcott BW. An efficient strategy for managing acute respiratory illness in adults. Ann Intern Med 1980; 93: 757-63.

Chapter 7

A diagnostic model for the classification of jaundice

Segaar RW, Wilson JHP, Habbema JDF.
Submitted for publication, 1990.

Abstract

There are several ways in which the differential diagnosis of jaundice can be supported. One is to rely on diagnostic models that result from analysis of large datasets collected elsewhere. Such models may be sensitive to transfer, requiring careful adjustment procedures. Another approach is to develop local models. But large datasets may not be at hand locally. We therefore investigate what is to be preferred: local application of a carefully transferred remote diagnostic model, or development of a local model based on limited data.

In data from 214 patients admitted to a Rotterdam hospital with jaundice, we explore the value of early diagnostic information (history, physical examination and simple biochemistry) for the differential diagnosis of jaundice. We will consider four global categories into which patients can be classified: acute- and chronic non-obstruction, and benign- and malignant obstruction. In several steps we fit a logistic discriminant model to the data. The resulting local diagnostic model is evaluated in a group of 100 new patients. Results are compared with an other diagnostic model (the COMIK algorithm) which is transferred from Copenhagen to Rotterdam. The performance of the local model is shown to be comparable to that of the COMIK model. It also appears that the local model shares many variables with the COMIK model. We conclude that there is no real difference between the two alternatives from a performance point of view. Consequently, the choice depends on the willingness to collect data locally, and the ability the make local adjustments to transferred models.

Introduction

The differential diagnosis of jaundice has often been studied using probabilistic or other formal methods (Kni73, Ste73, Ste75, Sai85, Boo81, Boo86, Mat84). Several arguments motivate this choice of subject. Jaundice is a common symptom, and most underlying causes have serious consequences. Finally, it is possible to establish a final diagnosis with reasonable certainty.

Published models for probabilistic classification of jaundice, like the COMIK algorithm (Mat84), show good performance, even after transfer to other sites (Seg88). The COMIK model benefits from the fact that it builds on a large number (1002) of patients, collected in a prospective way. Our evaluation (Seg88) also showed that good performance following transfer can only be maintained after careful adjustments.

Development of a model, based on patient data from a local hospital, is another alternative. In such a model, characteristic features of the local patient population will be fully utilized. The diagnostic tests incorporated in the model are also compatible with local management of patients. The resulting model will thus be 'optimal' for a local hospital. But such models may have to rely on small sized datasets, collected retrospectively, as it involves considerable efforts to implement large prospective data collections like COMIK.

Since both alternatives are defensible, information on performance will have to settle the matter. This motivated us to develop a local model for classification of jaundice. With information on the performance of the COMIK model at hand, this exercise allows for a comparison between the performance of our local model and that of the COMIK model. We therefore choose a comparable procedure to build the model. This paper describes the development of the probabilistic model, from data analysis, to an evaluation of the resulting logistic model. With the results at hand we can provide an answer to the question: will a decision aid based on limited local data outperform a carefully transferred decision aid that builds on a large prospective data collection?

Patients and methods

A diagnostic classification of jaundice

Throughout this chapter we will refer to a classification into four diagnostic categories (Mat84). It is based on the pathophysiological division of jaundice into non-obstructive (also called 'medical')- and obstructive causes (also called 'surgical'). The classification of jaundice into medical and surgical is a reflection of the historical situation, where the surgeon usually took care of obstruction and the internist took care of non-obstruction. Non-obstructive jaundice can be subdivided into acute and chronic. For this division we adopt the interpretation of the COMIK category, that chronic non-obstruction implies impaired liver function, either present or to be anticipated in the future. In a similar way obstruction can be subdivided into benign- and malignant categories. According to this classification scheme a patient can be classified into one of the following categories:

- acute non-obstructive ('acute medical' or A.M.)
- chronic non-obstructive ('chronic medical' or C.M.)
- benign obstructive ('benign surgical' or B.S.)
- malignant non-obstructive ('malignant surgical' or M.S.)

From these global diagnostic categories, a further differentiation into detailed diagnostic categories can follow. In this paper we will focus on the classification into four categories only.

Patients

We collected retrospective data from 100 consecutive patients admitted to the department of Internal Medicine II of Dijkzigt Hospital Rotterdam, between 1984 and 1985. Criteria to enter the study were a serum-bilirubin above 17 μmol/liter, visible jaundice, and an age of 15 years or more. From patients meeting these criteria, a large number of variables were recorded and stored in a database. A final diagnosis was coded long after patients were discharged from the hospital. It was based on a revision of all information, and preferably used morphological evidence (biopsy, autopsy). Although it was not possible to establish a detailed final diagnosis in all patients, all patients could be satisfactory classified into one of the four global categories. The data collection was called the Rotterdam-I, because it was

followed by a similar data collection (Rotterdam-II) that included 214 patients admitted to the department of Internal Medicine II between 1985 and 1987. Patient criteria for this data collection were identical to the Rotterdam-I data collection. A schematic overview of the study is given in figure 7.1.

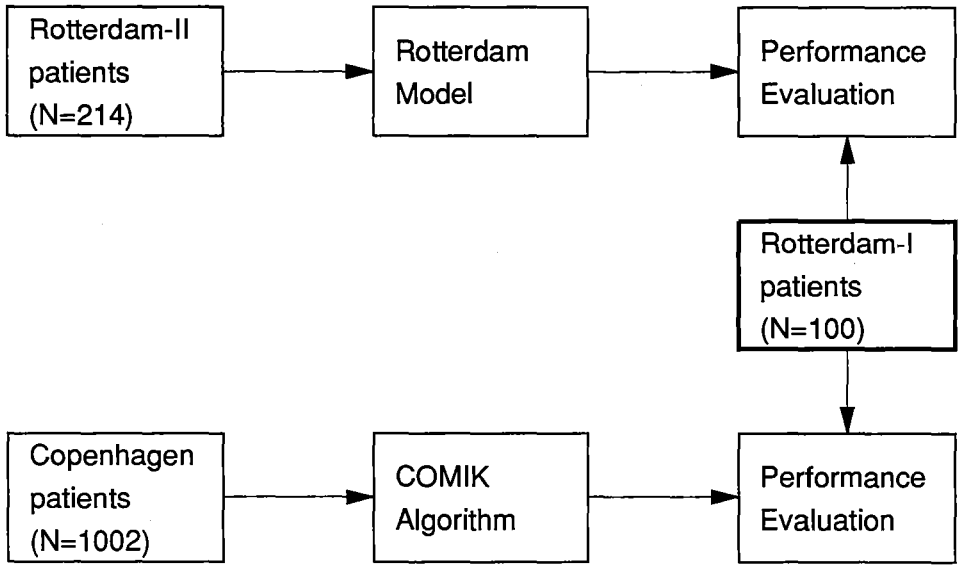


Figure 7.1. Structure of the study

The design of the model

Models for classification into four categories can be build in various ways. We will follow the pathophysiological classification, separating obstruction from non-obstruction first, and then acute from chronic, and benign from malignant causes. This implies the necessity of three submodels for binary classification:

- a classifier for obstruction versus non-obstruction; This classifier computes the two probabilities P(obstruction) and P(non-obstruction)

- a classifier for acute- versus chronic non-obstruction; Given non-obstruction, this classifier computes P(acute) and P(chronic)
- a classifier for benign- versus malignant obstruction; Given obstruction, this classifier computes P(benign) and P(malignant)

To establish a classifier into four categories requires multiplication of the probabilities generated by the three classifiers. So the probability function for acute non-obstructive disease becomes:

P(acute medical) = P(acute) * P(non-obstructive)

Similar constructs apply to the other three diagnostic categories. The classifiers obey basic probabilistic formulae. Since the diagnostic categories are mutually exclusive, the computed probabilities add up to 1.0.

Analysis

For the development of the model we use data from the Rotterdam-II data collection as 'training sets'. All (N=214) patient data are used for the construction of the classifier into obstruction versus non-obstruction. Patient data are also split into subsets to provide training sets for the other two classifiers. One subset contains only patients with non-obstructive diseases (N=119) and is used as a training set for the construction of the classifier for acute versus chronic non-obstruction. The other subset contains only patients with obstructive diseases (N=95). This subset is used as a training set for the construction of the classifier for benign versus malignant obstruction. The Rotterdam-II data collection includes several hundred variables. All variables are candidate for each of the three classifiers. Only the variables which yield independent diagnostic information are relevant to build a classifier, and therefore selection is required. Before analysis, all variables are dichotomized. The choice of cut-off points is left over to the selection process, through creation of derived binary variables.

A first selection step is based on the amount of missing values for single variables. Missing values result from a diagnostic selection process, and the relative importance of those variables cannot be assessed properly. At the risk of ignoring relevant variables, it is worthwhile to put a limit to the number of missing values allowed. We have chosen an arbitrary 10 % threshold for each variable in a training set. Exclusion in one training set does not imply

exclusion in other sets by default. Variables in excess of 10 % missing values usually belong to more advanced diagnostic techniques like special biochemistry, ultrasound, ERCP, and biopsy.

The second step involved a further selection of candidate variables. Stepwise logistic procedures, like the ones implemented in the BMDP computer programs are less applicable for a selection out of large numbers of variables. A first selection was therefore done with INDEP (Gel81), a computer program intended for stepwise forward selection of prognostic or diagnostic categorical variables. Selections are made through construction of discriminant models based on Bayes rule under the assumption of conditional independence. The addition of a new variable to the model depends on an improvement of a quadratic performance criterion. For a full discussion on performance criteria we refer to Hab78 and Gel81. Prior probabilities for the classifier were estimated from the dataset. To evaluate the performance of the models, the leaving one out (jackknife) method was used. For each classifier we arbitrarily limited the number of variables to be selected to 20, that were passed to the next selection step.

In the final selection step we used the Logistic Regression (LR) module of the BMDP statistical package. The LR module selects or excludes variables in a stepwise manner through forward selection with backward deletion. Variables are included only if their inclusion improved the classification as measured by the p-value of the log likelihood below a chosen significance level ($p < 0.10$). Once included, removal of variables was considered only if the associated decrease in performance was small ($p > 0.15$). When selection stops a logistic model results. Logistic regression does not assume independence between variables, and dependencies are incorporated into the model, whereas this was not the case in the selection procedure of INDEP.

Results

Six variables were selected for the classifier for obstruction versus non-obstruction. For the classifier for acute versus chronic non-obstruction four variables were selected, and for the classifier for benign versus malignant seven variables were selected. In the tables 7.1 - 7.3, we show for each classification the variables, their associated regression coefficients and the constant term.

Table 7.1. Constant term, variables and coefficients for the classification into non-obstructive and obstructive causes of jaundice. The model computes the probability for obstruction. Therefore, positive numbers indicate evidence for obstruction, and negative for non-obstruction.
‡ : the variable is also used in the COMIK model (Mat84).

Symbol	Variable	Coefficient
β_0	'Constant term'	-2.5
β_1	Previous cirrhosis present ‡	-2.0
β_2	Aged above 56 ‡	+1.4
β_3	LDH above 344 ‡	-1.1
β_4	Alkaline Phosphatase above 211 ‡	+1.0
β_5	Abdominal pain as first symptom ‡	+0.9
β_6	Ascites present ‡	-1.2

Table 7.2. Constant term, variables and coefficients for the classification into acute and chronic non-obstructive causes of jaundice. The model computes the probability for chronic non-obstruction. Therefore, positive numbers indicate evidence for chronic non-obstruction, and negative for acute non-obstruction.
‡ : the variable is also used in the COMIK model (Mat84).

Symbol	Variable	Coefficient
β_0	'Constant term'	-1.4
β_1	Experienced previous liver disease	+1.4
β_2	Fever present ‡	-1.4
β_3	Aged below 71 ‡	+1.6
β_4	ASAT above 177 ‡	-1.7

To illustrate the use of the model, we describe how a fictitious patient will be classified. The patient is a 50 year old men who develops jaundice that was not previously known. He has complaints for some weeks now.

Table 7.3. Constant term, variables and coefficients for the classification into benign and malignant obstructive causes of jaundice. The model computes the probability for malignant obstruction. Therefore, positive numbers indicate evidence for malignant obstruction, and negative for benign obstruction. ‡ : the variable is also used in the COMIK model (Mat84).

Symbol	Variable	Coefficient
β_0	'Constant term'	-2.7
β_1	Severe abdominal pain present ‡	-1.3
β_2	Weightloss present ‡	+1.4
β_3	Nausea as first symptom	-2.0
β_4	Bilirubin above 120 ‡	+1.2
β_5	Vomiting as first symptom	-1.4
β_6	Duration of complaints > 14 days ‡	+1.3
β_7	Haemoglobin less than 9.0	+1.0

Beside his jaundice, the only other relevant symptom on physical examination is ascites. Laboratory tests show that serum-bilirubin is 81 μmol/l, ASAT is 123 U/l, LDH is 216 U/l, alkaline phosphatase is 89 U/l and hemoglobin is 9.3 mmol/l.

The probability for the diagnostic category as computed by a logistic regression model results from the formula

$$P = \frac{\exp(t)}{1+\exp(t)} \qquad \text{where} \qquad t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p$$

P is the probability computed by the model, β_0 is the constant term and β_1 through β_p are the coefficients. The reader can verify that exp(t) equals P/1-P, i.e. the odds of the diagnostic category, and therefore t is equal to the logarithm of the odds. The variable X_p takes a value depending on the answer to the question formulated at variable p. It takes a value of +1 if the answer to the question is 'yes' or 'present', and a value of -1 if the answer is 'no' or 'not present'. For example in the model of table 7.1, X_1 takes a value of +1 if there is previous cirrhosis, and -1 if there is no previous jaundice.

Looking at the model that computes the probability for obstruction (table 7.1), we see that only for the variable 'Ascites present' (β_6) the answer to the question listed is 'yes', whereas for the other variables the answer is 'no' or 'not present'. Consequently, only X_6 takes a value of +1, whereas X_1 thru X_5 take a value of -1. For the classification into obstruction and non-obstruction we then compute:

$$t = -2.5 + (-2.0 \times -1) + (+1.4 \times -1) + (-1.1 \times -1) + (+1.0 \times -1) + (+0.9 \times -1) + (-1.2 \times +1)$$

and therefore

$$P(obstruction) = \frac{\exp(-3.9)}{1 + \exp(-3.9)} = 0.02$$

and consequently,

$$P(non\text{-}obstruction) = 1.0 - P(obstruction) = 0.98$$

We leave it for the reader to figure out that

$P(chronic) = 0.87$ and $P(acute) = 0.13$ (since t = 1.9)

and that

$P(benign) = 0.43$ and $P(malignant) = 0.57$ (since t = -0.3)

From multiplication of these probabilities, it follows that

$P(acute\ non\text{-}obstruction) = 0.13$
$P(chronic\ non\text{-}obstruction) = 0.85$
$P(benign\ obstruction) = 0.01$
$P(malignant\ obstruction) = 0.01$

The category assigned by far the highest probability here is chronic non-obstruction (P = 0.85). By convention we take that category as 'model diagnosis', which was correct in this case, because the patient turned out to have alcoholic cirrhosis.

Evaluation

First we evaluated the performance of the classifiers on the training sets derived from the N=214 Rotterdam-II data collection. In this way, we could evaluate each classifier separately. By convention we decide that the category assigned highest probability is the diagnosis of the model. In addition, we examined the performance on classification into four categories, resulting from concurrent application of all three classifiers. As we indicated earlier, probabilities for each of the four diagnostic categories result from multiplication of the three binary classifiers with each other. Results are shown in table 7.4.

These results are too optimistic, because the same patients that were used for the construction of the classifiers were also used for testing them. We therefore evaluated the classifiers with the 'test population' from the N=100 Rotterdam-I data. Here too we evaluated the performance of the three separate models for classification, and on classification into four groups (Table 7.5). In that table we also present results of the COMIK algorithm on the same patients, which are almost identical. Note that the benign versus malignant classification is especially problematic for both models.

Results for the classification into four categories can be displayed as a four by four classification matrix (Table 7.6). Such a presentation provides more information than the (non) error rate, because it shows where most classification errors occur: the benign surgical category appears to be the real problematic one, with 8 errors out of 14 patients.

Table 7.4. Results of classification into two- and four categories for 214 patients of the Rotterdam-II 'training population'.

Classification	Correct (%)	False (%)	Total (%)
Medical vs surgical	194 (91 %)	20 (9 %)	214 (100 %)
Acute vs chronic	107 (90 %)	12 (10 %)	119 (100 %)
Benign vs malignant	86 (91 %)	9 (9 %)	95 (100 %)
Into four categories	174 (81 %)	40 (19 %)	214 (100 %)

Table 7.5. Results of classification into two- and four categories for 100 patients of the Rotterdam-I 'test population'. For a comparison we also show the performance of the COMIK algorithm on the same patients.

Classification	Correct (%)	False (%)	Total (%)	COMIK
Medical vs surgical	87 (87 %)	13 (13 %)	100 (100 %)	87%
Acute vs chronic	54 (90 %)	6 (10 %)	60 (100 %)	95%
Benign vs malignant	31 (77 %)	9 (23 %)	40 (100 %)	72%
Into four categories	76 (76 %)	24 (24 %)	100 (100 %)	77%

Table 7.6. Results of the model for classification of jaundice into four categories. No threshold was applied. Columns: Final clinical diagnosis. Rows: The diagnostic class selected by the model. AM=Acute medical, CM=Chronic medical, BS=Benign surgical, MS=Malignant surgical. Results: 76/100 = 76 % correctly classified

	Clinical diagnosis				
Model diagnosis	AM	CM	BS	MS	Total
Acute Medical	7	4	0	2	13
Chronic Medical	0	43	3	1	47
Benign Surgical	0	1	6	3	10
Malignant Surgical	2	3	5	20	30
Total	9	51	14	26	100

It is possible to define a threshold level. If the probability of a diagnostic category exceeds for example 0.80, the diagnosis of the model for such a patient can be considered as 'confident'. Such a threshold approach will increase the diagnostic performance of the model, at the cost of less patients being classified. In an operational setting, special actions can be taken for

'grey zone' patients not classified by the model. Results of an evaluation with a 0.80 threshold are given in table 7.7. Observe the now dramatic behavior of the benign surgical group.

The overall result (88 % correct classifications) is almost identical to the result we obtained in the evaluation of the COMIK algorithm (Seg88) on the same test population. There we found 90 % correct classifications, with 55 % of the patient classified, using the same p = 0.80 threshold.

Table 7.7. Results of the model for classification into four categories. Only confident diagnoses are accepted, i.e. the model diagnosis should exceed a 0.80 probability. Columns: Final clinical diagnosis. Rows: The diagnostic class selected by the model. AM=Acute medical, CM=Chronic medical, BS=Benign surgical, MS=Malignant surgical. Results: 52 % of the patients classified, of which 46/52 = 88 % correctly classified.

	Clinical diagnosis				
Model diagnosis	AM	CM	BS	MS	Total
Acute Medical	1	0	0	0	1
Chronic Medical	0	31	0	1	32
Benign Surgical	0	0	1	1	2
Malignant Surgical	0	1	3	13	17
Not classified	8	19	10	11	48
Total	9	51	14	26	100

Discussion

In our study, we aimed at a comparison of a transferred diagnostic aid developed elsewhere, with a diagnostic aid based on local data. We therefore developed a model, based on a retrospective local dataset. We evaluated the model with an additional dataset. The performance of the local model turns out to be similar to the performance of the COMIK model (Seg88), using the same test population. The analogy between variables included in

our model and variables used for the COMIK algorithm (Mat84) is remarkable and also reassuring (table 7.1 - 7.3). It seems that both the Rotterdam and the Copenhagen analyses have selected variables that are intrinsically connected to the diagnosis of jaundice. Some Copenhagen variables did not enter our models. For the Copenhagen variables relating to the presence congestive heart failure and leukemia that may be a reflection of referral patterns. For other variables we must speculate that it results from differences in the sizes of the databases (1002 versus 214).

Many 'prototypical' variables associated with chronic liver impairment like spider naevi, palmar erythema, and disturbed consciousness were not selected in our model. To understand this, one has to appreciate the fact that such variables are associated with prototype patients, which may not necessarily be the best starting point for optimal classification of a group of jaundiced patients. Ultrasound and viral serology were also not included in the model. The reason for their exclusion was the amount of missing data that exceeded our maximum criteria on all binary classifications. Clinicians use available information to decide whether such tests may be worthwhile. Consequently, the diagnostic relevance of variables with many missing values may be overestimated.

The classification of benign surgical patients is quite disappointing. In our evaluation of the COMIK algorithm with the same test-population (Seg88), we had a similar experience. It therefore seems likely that classification of benign surgical patients is a difficult task, probably because these patients have less 'typical' characteristics. The finding that we were not able to improve the COMIK results, may be due the fact that the Rotterdam model was based on a smaller data collection and thus more susceptible to random variation. The COMIK algorithm was based on a larger number of patients and therefore less susceptible to such variations. Apparently, this was not compensated for by the fact that the Rotterdam model takes into account the characteristics of the local population. It can be anticipated that the transfer of the Rotterdam model to other locations will be associated with some decline in performance (DeD81). This may especially be true for the classification into acute and chronic non-obstruction, and for the classification into benign and malignant obstruction, since these classifiers were based on the least numbers of patients. From tables 7.5 and 7.6 it follows that classification of obstructive jaundice into benign and malignant causes is too difficult given early diagnostic information only. For correct classification of these patients, one probably has to rely on additional techniques like ultrasound.

Based on our comparison, there is no preference for development of local models or transfer of existing models. A decision will therefore depend on additional arguments. If one is able to conduct large size prospective data collections, then development of a local model is to be preferred. If one is not in a position to conduct such large data collections, transfer of a diagnostic aid developed elsewhere will be a realistic alternative. But such a transfer must at the very least include a careful assessment of the local relative incidences and biochemistry, and provide corrections if necessary (Seg88). The ambition of the European community to collect data on jaundice in a 'concerted action' (Lav88) may also be a worthwhile experiment. Such data collections minimize random variation, at the cost of geographical differences. Whether such an approach really is to be preferred remains to be proven. The data of the 214 patients of the Rotterdam-II data collection were among the first ones that were entered in database of the EEC project.

References

Boo81 Boom RA, Gil D, Maass R, Manrique G. The differential diagnosis of obstructive jaundice based on a logarithmic index of alkaline phosphatase and total cholesterol values. Med Decis Making 1981; 1: 227-37.

Boo86 Boom R, Gonzalez C, Fridman L, Alaya J, Realpe JL, Morales P, Quintero R. Looking for the 'indicants' in the differential diagnosis of jaundice. Med Decis Making 1986; 6: 36-41.

Ded81 De Dombal FT, Staniland JR, Clamp SE. Geographical variation in disease presentation. Does it constitute a problem and can information science help? Med Decis Making 1981; 1: 59-69.

Gel81 Gelpke GJ, Habbema JDF. User's manual for the INDEP SELECT discriminant analysis program. Leyden: Department of Medical Statistics, State University of Leyden, 1981.

Hab78 Habbema JDF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis.
Parts 1 - 3 Meth Inf Med 1978; 17: 217-46.
Parts 4 - 5 Meth Inf Med 1981; 20: 80-100.

Kni73 Knill-Jones RP, Stern RB, Girmes DH, Maxwell JD, Thompson RPH, Williams R. Use of sequential Bayesian model in diagnosis of jaundice by computer. Br Med J 1973; 1: 530-3.

Lav88 Lavelle SM, COMAC BME, European Community. Concerted action on objective medical decision making. Computer aided clinical diagnosis of jaundice. Technical announcement from the EC.

Mat84 Matzen P, Malchow-Møller, Hilden J, Thomsen C, Svendsen LB, Gammelgaard J, Juhl E. Differential diagnosis of jaundice: a pocket diagnostic chart. Liver 1984; 4: 360-71.

Sai85 Saint-Marc Girardin MF, le Minor M, Alperovitch A, Roudot Thoraval F, Metreau J-M, Dhumeaux D. Computer-aided selection of diagnostic tests in jaundiced patients. Gut 1985; 26: 961-7.

Seg88 Segaar RW, Wilson JHP, Habbema JDF, Malchow-Møller, Hilden J, van der Maas PJ. Transferring a diagnostic decision aid for jaundice. Neth J Med 1988; 33, 5-15.

Ste73 Stern RB, Maxwell JD, Knill-Jones RP, Thompson RPH, Williams R. Use of computer-assisted model in diagnosis of drug hypersensitivity jaundice. Br Med J 1973; 2: 767-9.

Ste75 Stern RB, Knill-Jones RP, Williams R. Use of computer program for diagnosing jaundice in district hospitals and specialized liver unit. Br Med J 1975; 2: 659-62.

Chapter 8

Test selection in jaundice: A comparison between physician behavior and a diagnostic model

Segaar RW, Wilson JHP, Habbema JDF.
Submitted for publication, 1990.

Abstract

The results of an observational study in clinical test-selection behavior are presented. In the study, tests selected by clinicians for jaundiced patients are tabulated against a probabilistic estimate of the diagnosis, based on the COMIK algorithm. For most tests, the behavior of clinicians can be predicted from the probabilistic estimate. We formulate a proposal for diagnostic test usage in jaundice. From this proposal, and the algorithmic estimate we can predict the 'test selection behavior'. We show that for most tests, the predicted 'test selection behavior' is consistent with the observed test selection behavior at a statistical level. Discussions of discrepancies between prediction and observation, and reasons for deviations from general guidelines, provide insight, and add new dimensions towards teaching a rational approach to the jaundiced patient.

Introduction

Diagnostic strategies in jaundice differ among clinicians, in particular for specialized diagnostic tests like ERCP (endoscopic retrograde cholangiography and pancreaticography) and liver biopsy. This is caused by differences in perception of the likelihood of possible causes, and of the anticipated value of diagnostic information from those tests. Not all available tests will be done on a routine basis for all patients. In this paper, the choice of tests, not performed on a routine basis, is referred to as 'test selection behavior'.

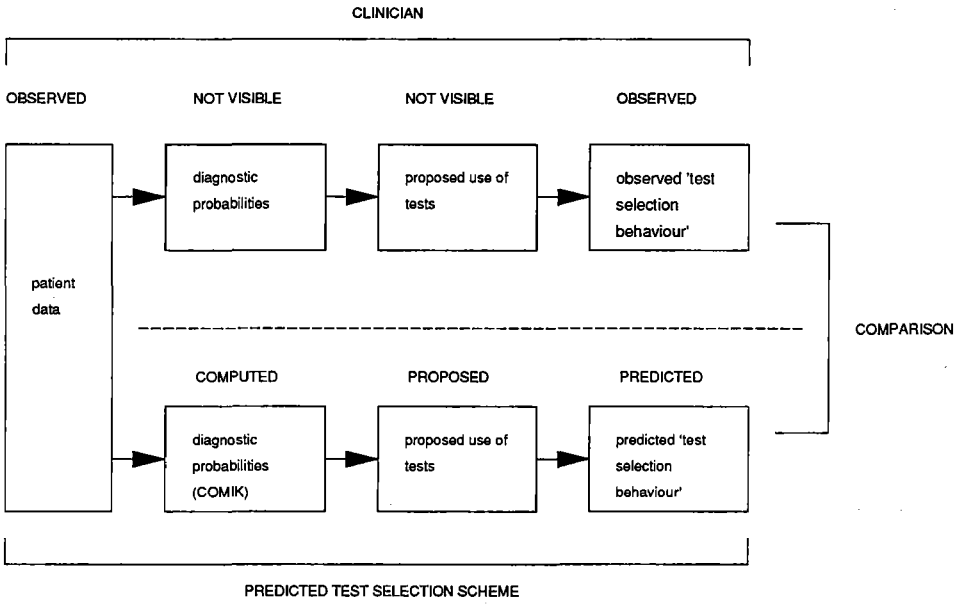
In the observational part of our study, we record diagnostic data available from the patient, including the selection of specialized tests by the clinicians. The mental steps between are invisible to us. We assume that the clinician uses accessible patient information, to estimate relative likelihoods for the diagnostic categories under consideration. Under these circumstances, test ordering is a synthesis of the estimated likelihoods and judgments about what tests are especially suited for confirming or rejecting diagnostic categories.

In addition, we investigated the test selection that would result when the 'unobservable steps' would be substituted by a formalized approach. To simplify matters, we assume that diagnosis in jaundice is a two step process. In the first step, data are collected on a routine basis in all patients. Using these data, patients are classified into a few global diagnostic categories. In the second step a selection of specialized diagnostic tests is made, conditional on the category into which a patient was classified in the first step. To implement the formal substitution, we describe a proposal for rational test usage, based on the possible diagnostic outcomes. Using the COMIK algorithm for probabilistic differential diagnosis of jaundice (Mat84), we can classify patients into the major diagnostic categories, based on early diagnostic data from individual patients. From a combination of the probabilistic diagnosis and the proposal for rational test usage, we can then propose diagnostic tests, or alternatively predict clinical test selection behavior for individual patients. Finally, the prediction of test selection behavior can be compared to the data on behavior from the observational part of the study. The hypothesis to be tested was: do the proposals from the formalized approach correspond to the observed behavior of clinicians.

The lay-out of the evaluation is shown in figure 8.1. Agreement between the tests proposed by the formalized approach, and the test behavior

observed from clinicians is assessed statistically. Results of the project are presented and commented on. Implications of our results for medical practice, in particular medical education are discussed

Figure 8.1. The strategy for comparing test selection behavior of physicians with a predicted test selection scheme.



Patients & Methods

Patient material

In the observational part of the study we collected data from consecutive patients admitted to the Department of Internal Medicine II of Dijkzigt Hospital Rotterdam between July 1985 and December 1987, satisfying three criteria: aged 15 year or older, visible jaundice, and a serum-bilirubin above 17 µmol/liter. Data obtained from patient history, physical examination and results and usage of diagnostic tests were recorded. This resulted in a large

database containing many variables for each patient. Test ordering of clinicians was not influenced by the study. A clinical diagnosis was coded later according to internationally acceptable criteria, using the disease classification proposed by the COMIK group (Mat84).

The COMIK algorithm

Each patient was classified probabilistically into four diagnostic categories using the COMIK algorithm. The algorithm assigns probabilities to the four major diagnostic categories of jaundice: acute medical, chronic medical, benign surgical and malignant surgical. The COMIK algorithm uses information from patient history, physical examination, and routine biochemical data, available within 24-48 hours following admission, to assign probabilities to each of the diagnostic categories mentioned above. The diagnostic performance of the COMIK algorithm in Rotterdam is good (Seg88) and similar to the performance obtained in Denmark (Mat84), Mexico (Boo86), and Sweden (Lin87). For a full discussion of the algorithm, and the classification of the diseases into the four categories we refer to the original paper by Matzen (Mat84).

A proposal for diagnostic use of tests

We investigated only specialized tests not performed on a routine basis. For jaundice, such tests are usually specialized biochemistry, viral serology, autoimmune serology, ultrasound, ERCP, endoscopy, angiography, EEG, liver biopsy and so on. Based on a review of current textbooks on hepatology and the advice of an experienced hepatologist, we formulated indications for diagnostic usage of each test. These indications describe the use of such a specialized diagnostic test, given the fact that a patient belongs to one of the four major diagnostic categories. We will not consider use of specialized tests like ultrasound for the classification into the four major diagnostic groups. For that classification task, only the COMIK algorithm will be used. Table 8.1 summarizes our proposal for test usage. For each major diagnostic category, the table shows whether or not we consider a test useful in further diagnostic activities when that diagnostic category is assumed certain for a patient. The motivation for testing will be discussed in the result section. Table 8.1 also shows the expected percentage of patients for whom the test would be requested, if the model-strategy had been implemented, and the percentage of patients in which the test was actually done.

Table 8.1. Proposed use for the tests included in the study. Columns 1 - 4 : a + sign indicates that the test is to be done if early diagnostic information indicates that probability for a diagnostic category is high. Column 5 : patients for whom the COMIK algorithm indicates that the test might be relevant as a percentage of all patients. The application of the COMIK algorithm for our population shows that 14% of the patients have highest probability for acute medical, 36% for chronic medical, 12% for benign surgical and 38% for malignant surgical. Column 6 : the patients for whom the test was actually performed as a percentage of all patients. The number of patients is 214.

	Proposed use of tests					
	acute medical	chronic medical	benign surgical	malign. surgical	% of tests expected	% of tests done
Trombotest	-	+	-	-	36%	80%
Normotest	-	+	-	-	36%	76%
Fibrinogen	-	+	-	-	36%	56%
Ammonia	-	+	-	-	36%	34%
Iron	-	+	-	-	36%	32%
Iron binding capacity	-	+	-	-	36%	29%
Ferritin	-	+	-	-	36%	25%
Amylase serum	-	-	+	-	12%	24%
Amylase urine	-	-	+	-	12%	2%
Ceruloplasmin	-	+	-	-	36%	20%
AFP	-	+	-	-	36%	23%
CEA	-	-	-	+	38%	8%
Alpha 1 antitrypsin	-	+	-	-	36%	20%
HBsAg	+	+	-	-	50%	56%
HBsAb	+	-	-	-	14%	42%
IgM HA	+	-	-	-	14%	34%
CMV Ab	+	-	-	-	14%	39%
AB mitochondria	-	+	-	-	36%	39%
AB smooth muscle	-	+	-	-	36%	35%
AB parietal cells	-	+	-	-	36%	22%
AB DNA	-	+	-	-	36%	13%
AB liver membrane	-	+	-	-	36%	20%
Anti nuclear factor	-	+	-	-	36%	30%
LE cell phenomenon	-	+	-	-	36%	7%

Table 8.1. Continued

	Proposed use of tests					
	acute medical	chronic medical	benign surgical	malign. surgical	% of tests expected	% of tests done
IgG	-	+	-	-	36%	33%
IgA	-	+	-	-	36%	32%
IgM	-	+	-	-	36%	33%
Ultrasound	-	+	+	+	86%	78%
X-esophagus	-	+	-	-	36%	12%
Endoscopy	-	+	-	-	36%	14%
Angiography	-	-	-	+	38%	9%
CT scan	-	-	-	+	38%	20%
ERCP	-	-	+	+	38%	34%
PTC	-	-	+	+	50%	1%
EEG	+	-	-	-	14%	14%
Liver biopsy	+	+	-	-	50%	30%

Assumptions

In the evaluation, several assumptions were made. We will briefly itemize them. As a whole, the assumptions provide a background why the use of our formal model for test proposal was appropriate under these circumstances.

Assumption 1: For patients presenting with jaundice, we assume that testing is primarily diagnostic, to reveal, or rule out the etiology of the disease of the patient. This is a simplification, and one has to bear in mind that the motivation to order a test may not always be primarily diagnostic. Other reasons for testing can sometimes be more prominent, such as the assessment of severity (Sim85) and prognosis (Mil85, Chr85) of a disease.

Assumption 2: We assume that probabilistic techniques are an appropriate technique to generate a diagnostic summary of patient information gathered within 48 hours.

Assumption 3: We assume that the COMIK algorithm is an adequate tool to generate such a probabilistic estimate, based on early diagnostic information.

Assumption 4: We assume that the diagnostic testing done is redundant, that is, if more than one test will provide the same diagnostic information, no selection is made and all tests are done.

Assumption 5: We assume that 'rational' selection of a test is subject to the following conditions:
(1) the test is regarded as relevant for a diagnostic category and
(2) that the diagnostic category is assigned highest probability by a probabilistic diagnostic classifier such as the COMIK algorithm.

Results

For each diagnostic test in the study we used a logistic regression model with a binary variable indicating whether or not this test had been done as dependent variable, and the COMIK probabilities as explaining variables. We tested both for a quadratic and a linear relationship. A quadratic relationship was usually inferior to the linear relationship, and hardly ever changed conclusions. Therefore, only linear relationships will be considered. The results of the statistical analyses are shown in table 8.2. This table shows the relation between the observed probability of a test being requested, and the algorithmic estimate of diagnostic probabilities. The signs indicate whether a high algorithmic probability is associated with an increased (+), or decreased (-) probability of testing. To simplify matters, table 8.2 shows significance levels only. From table 8.2, it can be observed:

- whether selection behavior for a test can be predicted (following from the significance level)
- in which diagnostic context tests are used (following from the diagnostic categories showing significant relations).

Table 8.2. Relation between the probability of a test being requested and the algorithmic estimates of diagnostic probabilities. The signs + and - are based on logistic regressions with the test being requested as dependent variable and the algorithmic probability as explaining variable. Critical values

for the (difference in) scaled deviance are 3.84 (for p=0.05 : *) and 6.64 (for p=0.01 : **). Signs : + indicates increased testing with increased algorithmic probability, - indicates the opposite.

	Explaining COMIK probabilities							
	p(AM)		p(CM)		p(BS)		p(MS)	
Test	sign	scd	sign	scd.	sign	scd	sign	scd
Trombotest	-	*	+	ns	-	**	+	**
Normotest	-	**	+	*	-	**	+	**
Fibrinogen	-	**	-	ns	-	ns	+	**
Ammonia	-	*	+	**	-	**	-	**
Iron	-	ns	+	**	-	**	-	*
Iron binding capacity	-	ns	+	**	-	**	-	ns
Ferritin	-	ns	+	**	-	*	-	**
Amylase serum	+	ns	-	**	+	*	+	ns
Amylase urine	-	ns	-	ns	+	ns	+	ns
Ceruloplasmin	+	ns	+	**	-	**	-	**
AFP	-	*	+	**	-	*	-	ns
CEA	-	**	-	ns	-	*	+	**
Alpha 1 antitrypsin	-	ns	+	**	-	*	-	**
HBsAg	+	**	+	**	-	**	-	**
HBsAb	+	*	+	**	-	ns	-	**
IgM HA	+	*	+	ns	-	ns	-	ns
CMV Ab	+	*	+	ns	-	ns	-	*
AB mitochondria	+	ns	+	**	-	**	-	**
AB smooth muscle	+	ns	+	**	-	**	-	**
AB parietal cells	-	ns	+	**	-	**	-	**
AB DNA	+	ns	+	*	-	ns	-	ns
AB liver membrane	+	ns	+	**	-	*	-	**
Anti nuclear factor	+	ns	+	**	-	*	-	**
LE cell phenomenon	-	ns	+	*	-	ns	-	ns

Table 8.2. Continued

	Explaining COMIK probabilities							
	p(AM)		p(CM)		p(BS)		p(MS)	
Test	sign	scd	sign	scd.	sign	scd	sign	scd
IgG	+	ns	+	**	-	**	-	**
IgA	+	ns	+	**	-	**	-	**
IgM	+	ns	+	**	-	**	-	**
Ultrasound	-	ns	-	ns	-	*	+	**
X-esophagus	+	ns	+	**	-	**	-	**
Endoscopy	-	*	+	**	-	ns	-	**
Angiography	-	ns	-	**	-	ns	+	**
CT scan	-	ns	-	**	-	ns	+	**
ERCP	-	**	-	**	+	**	+	**
PTC	-	ns	-	ns	-	ns	+	*
EEG	-	**	+	**	-	ns	-	**
Liver biopsy	-	ns	+	**	-	**	-	**

In the sections below, results for all diagnostic tests used in the study will be discussed. First we present our proposal for diagnostic testing on each diagnostic test or group of diagnostic tests. This *proposed* test selection behavior of the model (summarized in table 8.1), is compared to the *observed* test selection behavior (summarized in table 8.2). The proposed testing matches the observed testing, when proposed usage from table 8.1 is confirmed through a significant positive (+) relation in table 8.2. Matches and discrepancies will be briefly discussed.

Biochemistry

Trombotest® (Nygard), Normotest® (Nygard) and Fibrinogen can be used for multiple purposes. They can be used to assess severity of an impaired liver function. In addition, they are used as a routine workup before surgical intervention to evaluate bleeding tendency. It was thus not possible to make a specific proposal for their use. The pattern we found clearly shows that testing is performed in the absence of acute

non-obstructive disease, and in the presence of malignant obstructive disease. This implies that their use is primarily to assess liver function in chronic liver disease and before surgery in malignant obstruction.

Ammonia

is primarily used to assess severity of liver failure. The hypothesis is that testing will be done when probability of chronic medical disease is high. The pattern found is in agreement with this hypothesis.

Iron, Iron binding capacity and Ferritin

are used as diagnostic tests in the case of chronic liver disease. They are also used in the analysis of anemia which may be a confounding situation. The pattern we found confirms the use as a diagnostic test in chronic liver disease.

Amylase (urine and serum)

are diagnostic tests in suspected pancreatitis. Since this diagnosis is classified as benign surgical in COMIK terminology, testing should be done whenever the probability for benign surgical is high. The pattern we found is not so specific.

Ceruloplasmin

is used as a diagnostic test in chronic liver disease. The hypothesis is that testing will be done when the probability for chronic medical disease is high. The pattern found is in agreement with this hypothesis.

Alphafetoprotein (AFP) is a diagnostic test for hepatoma, which often arises after prolonged (i.e. chronic) medical liver disease. The hypothesis is that testing will be done when the probability for chronic medical disease is high. The pattern found is in agreement with this hypothesis.

Carcino-embryonic antigen (CEA)

is marker for (relapsing) malignant disease. The hypothesis is that testing will be done when the probability for malignant surgical disease is high. The pattern found is in agreement with this hypothesis.

Alpha-1-antitrypsin

is used to detect alpha-1-antitrypsin deficiency which can be a cause of chronic liver disease. The hypothesis is that testing will be done when the probability for chronic medical disease is high. The pattern found is in agreement with this hypothesis.

Viral serology

Hepatitis B serology (HBsAg and HBsAb) are diagnostic for both acute- and chronic medical categories. The hypothesis is that testing will be done when the probability for acute or chronic medical disease is high. The pattern found is in agreement with this hypothesis.

Hepatitis A serology (IgM HA)
is used as a diagnostic test in acute liver disease. The hypothesis is that testing will be done when the probability for acute medical disease is high. The pattern found is less specific, showing a tendency to use it primarily when the probability for medical disease is high, but not predominantly in the acute medical cases, suggesting that the test is either being used to discriminate between acute and chronic disease, or is being used incorrectly.

Cytomegalo virus serology (CMV Ab)
is used as a diagnostic test in acute liver disease. The hypothesis is that testing will be done when the probability for acute medical disease is high. The pattern found is less specific, showing a tendency to use it primarily when the probability for medical cases is high, similar to the IgM HA test.

Autoimmune serology/biochemistry

The hypothesis for autoimmune serology (several antibodies (AB) and the LE-cell test) and biochemistry (IgG IgM and IgA) was that testing will be done if the probability for chronic medical disease is high. For all tests within this category this hypothesis is confirmed. For most tests (with an exception for antibodies against parietal cells and the LE cell test) there is a tendency towards use in acute cases too.

Remaining tests

Ultrasound
is primarily diagnostic for the obstructive disease categories, since it visualizes bile duct obstruction. It is also used to assess gravity of disease in chronic medical disease, through the demonstration of portal hypertension and ascites. Finally it might be used on a protocol basis in all cases of jaundice. Hypothesis with regard to the use of ultrasound are hard to formulate, except that it may be less useful in acute medical

jaundice. The pattern that emerges shows that ultrasound is used if the probability for malignant disease is high, and - unexpectedly - if the probability for benign surgical disease is low. For the medical categories no clear pattern was found.

X esophagus

can reveal the presence of the esophagus varices that can accompany prolonged (chronic) liver disease. It is primarily a technique to assess disease severity. The hypothesis is that testing will be done when the probability for chronic medical disease is high. This hypothesis is confirmed. There is also a tendency towards use in acute cases.

Endoscopy

can reveal the presence of the esophagus varices that can accompany prolonged (chronic) liver disease. It also offers the opportunity for therapeutic intervention. As a consequence the hypothesis is that testing it will be done when the probability for chronic medical disease is high. This hypothesis is confirmed.

Angiography

can be used to determine the extend of malignant disease to determine operability. Its use as a diagnostic test is limited. As a consequence the hypothesis is that testing it will be done when the probability for malignant obstructive disease is high. This hypothesis is confirmed.

CT (Computer Tomography) scanning

is used as a diagnostic test, and as a test to assess the extend of a (malignant) disease. The hypothesis is that testing it will be done when the probability for malignant obstructive disease is high. This hypothesis is confirmed.

ERCP (Endoscopic retrograde cholangiography and pancreatography)

is a diagnostic and therapeutic procedure for obstructive (surgical) disease. The hypothesis is that testing it will be done when the probability for both obstructive disease categories is high. This hypothesis is confirmed.

PTC (Percutaneous Transhepatic Cholangiography)

is a diagnostic procedure for obstructive disease. The hypothesis is that testing it will be done when the probability for obstructive disease is high. In the few cases where it was used, the hypothesis was not confirmed. PTC seems to be limited to malignant disease.

EEG (Electroencephalogram)

is a procedure to assess the severity of chronic medical disease. The hypothesis is that testing it will be done when the probability for chronic medical disease is high. This hypothesis is confirmed.

Liver biopsy

is both a diagnostic procedure and a procedure to measure disease progression. It is predominantly done in chronic medical cases, but it may be relevant in some acute cases too. The hypothesis is that testing it will be done when the probability for chronic- and - to a lesser extent - acute medical disease is high. The hypothesis is confirmed for the chronic medical cases but not for the acute cases.

Discussion

The idea to use linear models to describe clinicians' judgments in medicine is not new (Wig88). Other authors (Sim85, Mil85, Chr85) have focussed primarily on more definite end-points like a final diagnosis, and treatment decisions. To our knowledge, these models have never been used to predict test selection behavior, using early diagnostic information. We think that the early diagnostic period, in which the number of possible diagnoses is at its largest, and information at its minimum, is the most challenging part of the diagnostic process.

Our proposal for test usage can be subject to local opinions. Local opinions can very well reflect exclusive characteristics of the local patient population, such as relative incidences of disease, and the distribution of symptoms and signs over the population. One of the major advantages of this type of study is that the differences are made explicit. Even with the limited data presented, the reader can formulate other proposals for test usage and verify them against our results.

In the evaluation of test behavior, we studied each diagnostic test as a separate entity. The finding that a test is generally applied in the proposed category of patients, does not necessarily mean that test selection by the doctor was as strict or optimal as it could be. Since many tests overlap in their diagnostic characteristics, we assumed that redundant testing would be present. The finding that test usage could indeed be predicted for individual tests, indicates that on many occasions tests which supply similar information were applied simultaneously. This finding can be explained in various ways. First of all, physicians prefer simultaneous tests in an attempt to reduce the time of diagnosis. It may also be that physicians, although capable of selecting applicable tests based on the early information available, feel that they are not really in a position to select the best tests at that stage. In such cases, dedicated test-selection tools as suggested by Saint Marc Girardin et al (Sai85) may be useful.

Chapter 9

A computer aid for early diagnostic classification of jaundice (The COMIP program)

Segaar RW, Wilson JHP, Habbema JDF, Hilden J.
Comp Meth Prog Biomed 1989; 28: 131-6.

Abstract

We describe the Computer Icterus Program (COMIP), a computer assisted diagnosis (CAD) program which is designed to facilitate the early choice of a diagnostic strategy in cases of jaundice. To anticipate transfer to other centres, the COMIP program offers facilities to make local adjustments for relative disease incidences and for laboratory data. This is a useful extension for CAD systems.

Introduction

Jaundice is a symptom that requires further investigation until a cause is found. The first diagnostic steps include history, physical examination and simple laboratory data. Thereafter more risky and expensive investigations like ERCP (endoscopic retrograde cholangiopancreaticography) or biopsy can be chosen, based on the interpretation of the information available. The diagnostic categories in the early stage are summarized in figure 9.1.

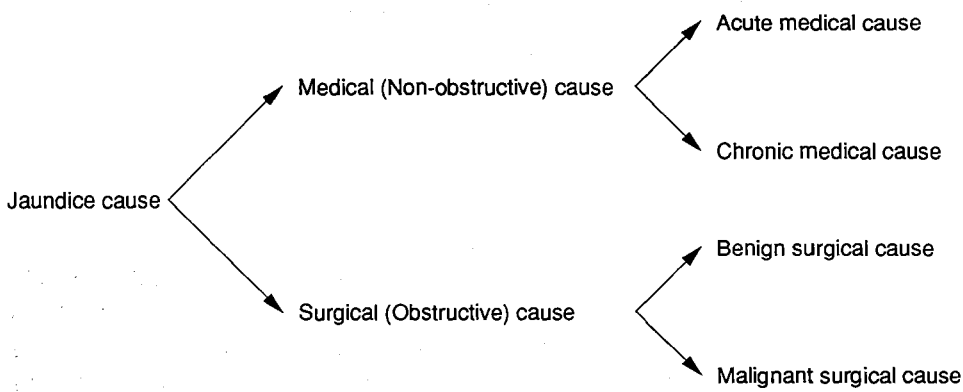


Figure 9.1. Conceptual model of the diagnostic categories for jaundice classification. The COMIK algorithm calculates the probabilities for each of the six categories.

Lindberg (Lin83) has shown that to optimize between patient risk and diagnostic efficacy requires a good probabilistic assessment of the diagnostic categories (figure 9.1) to which the patient might belong. Carefully collected and analyzed databases can indicate which symptoms and signs are most suited for this task. Whenever multiple symptoms and signs are to be used together it becomes necessary to attach quantitative weights to each symptom.

Using these concepts the Danish COMIK group developed a decision-making algorithm based on a large data collection (1002 patients) of jaundiced patients (Mat84). The result of their analysis was transcribed into a compact format and called the COMIK pocket diagnostic chart. It is intended for paper and pencil use. The chart is applicable to jaundiced patients above 15 years of age, with a bilirubin above 17 μmol/liter. For a full methodological discussion we refer to Matzen et al. (Mat84). Subsequent

evaluation of the COMIK algorithm in Denmark (Mat84), Mexico (Boo86), Sweden (Lin87) and the Netherlands (Seg88) revealed that the performance of the algorithm remained stable. Nevertheless some local adjustments for optimum performance were indicated both for disease incidences, which show a large geographical variation (DeD81), and for biochemistry results, which may have different reference ranges. In applying the COMIK chart in Rotterdam we also experienced that, using paper and pencil, the calculations were error prone. This stimulated the development of a computer program which deals with these problems in an explicit way. The resulting program can thus be regarded not only as a computerized version of the COMIK chart, but also as a model of how quantitatively defined diagnostic and prognostic rules can be implemented for transfer to other circumstances. Other authors (Boo86) took a similar approach, but focussed on educational topics.

Computational methods and theory

The COMIK algorithm

The discussion of the COMIK algorithm will be brief. For more detailed information the reader can consult the sample run (Appendix). The COMIK algorithm requires information about 21 parameters describing a patient. The parameters cover history, physical examination and simple laboratory results. The parameters are used in three classification rules, based on a logistic statistical model. The classification rules calculate the probability that the cause of jaundice is medical, acute and benign. Figure 9.2 shows how the probabilities that match the conceptual classification scheme of figure 9.1 are derived.

Figure 9.2. Computations involved in the derivation of the diagnostic probabilities in the COMIK algorithm for jaundice classification.

P(medical)	=	classification rule 1 (logistic model)		
P(acute)	=	classification rule 2 (logistic model)		
P(chronic)	=	classification rule 3 (logistic model)		
P(surgical)	=	1.0 - P(medical)		
P(chronic)	=	1.0 - P(acute)		
P(malignant)	=	1.0 - P(benign)		
P(acute medical)	=	P(acute)	*	P(medical)
P(chronic medical)	=	P(chronic)	*	P(medical)
P(benign surgical)	=	P(benign)	*	P(surgical)
P(malignant surgical)	=	P(malignant)	*	P(surgical)

Adjustments

Within the COMIK classification rules the relative disease incidences of the Copenhagen database are present as constants. Adjustment to local incidences is achieved by changing the Copenhagen constants into the local constants. The COMIP program computes the new constants from the local incidence percentages as supplied by the user. The quantitative results of biochemical parameters are reduced to a binary or at most ternary scale before use in the COMIK algorithm. The cutoff points are closely related to the 95 % reference ranges of the Copenhagen laboratory. If reference ranges of the local hospital differ, the cutoff points should be adapted to preserve discrimination (Seg90). We took a simple approach where the COMIP program computes a conversion factor which adjusts the mid-reference ranges only. Cutoff points are adjusted using this conversion factor. Although the reduction into binary data could be automated too, we decided to leave this feature out, to remain compatible with the paper and pencil version of the original COMIK chart.

Program description

Using the COMIP program is a two step process: installation and application. Installation is started only on initial execution of the program. Later on the installation procedure can be invoked on request.

Installation

First the COMIP program prompts the user to supply estimates of the local relative disease incidences and the 95 % normal range intervals for the biochemistry, see section **Adjustments**. For both sets of data an option to settle for the Copenhagen values is available.

Application

The application part consists of three major subroutines:

- data entry
- calculation
- printing of results

Data entry is simple and fast. For each of the 21 parameters the program allows the choice of one item out of a menu of two or three only. On completion of data entry a semi-natural language summary is generated for the purpose of error checking. The user is asked for his approval and in case of disagreement corrections can be made. The actual computation is straightforward. We decided not to compute the exact probability formula, but instead we used the conversion table accompanying the published COMIK algorithm. We think the minor inaccuracy is justified by a full compatibility with the chart, which in turn facilitates interpretation and error checking. The results of the computation are available to the user in various ways. A simple table showing the computed probabilities is the default. A full table showing all computations involved is also available. Additional graphs showing the development of the probabilities in relation to the information supplied are available on request. For both the four-way and the 'triple two-way' classifications these graphs illustrate in a simple manner how each finding contributed to the final probability. If necessary, graphs can also be generated using a simple 25 x 80 character terminal screen.

Structurogram

Figure 9.3 shows the COMIP program in a structurogram

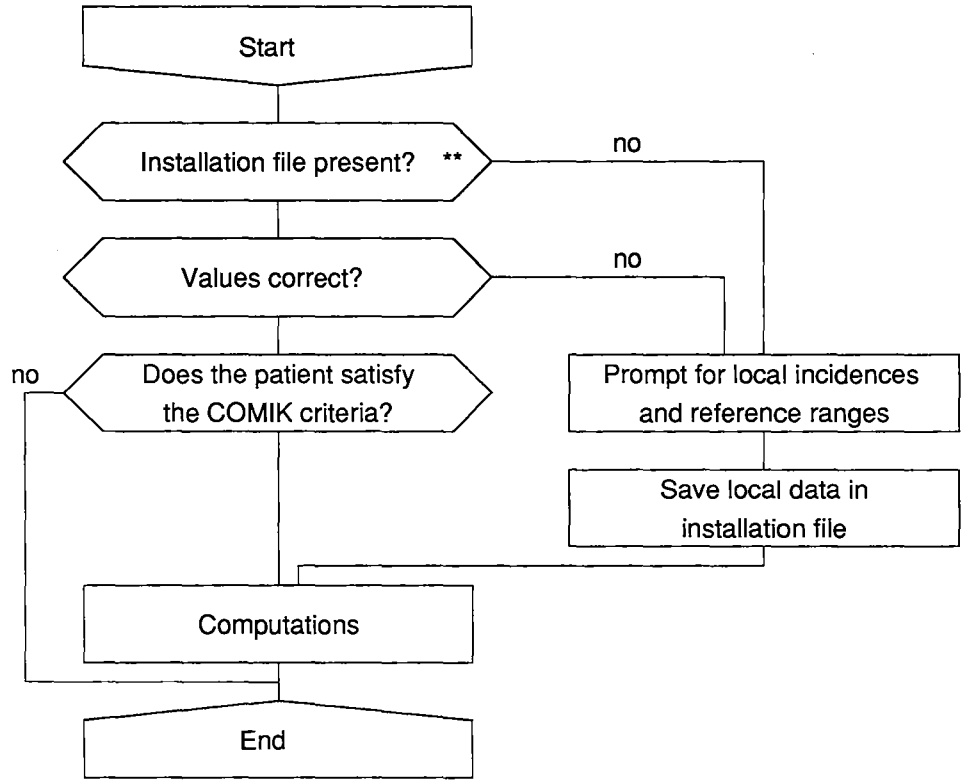


Figure 9.3. COMIP structurogram. Part a. ** All tests except this one to be answered by the user.

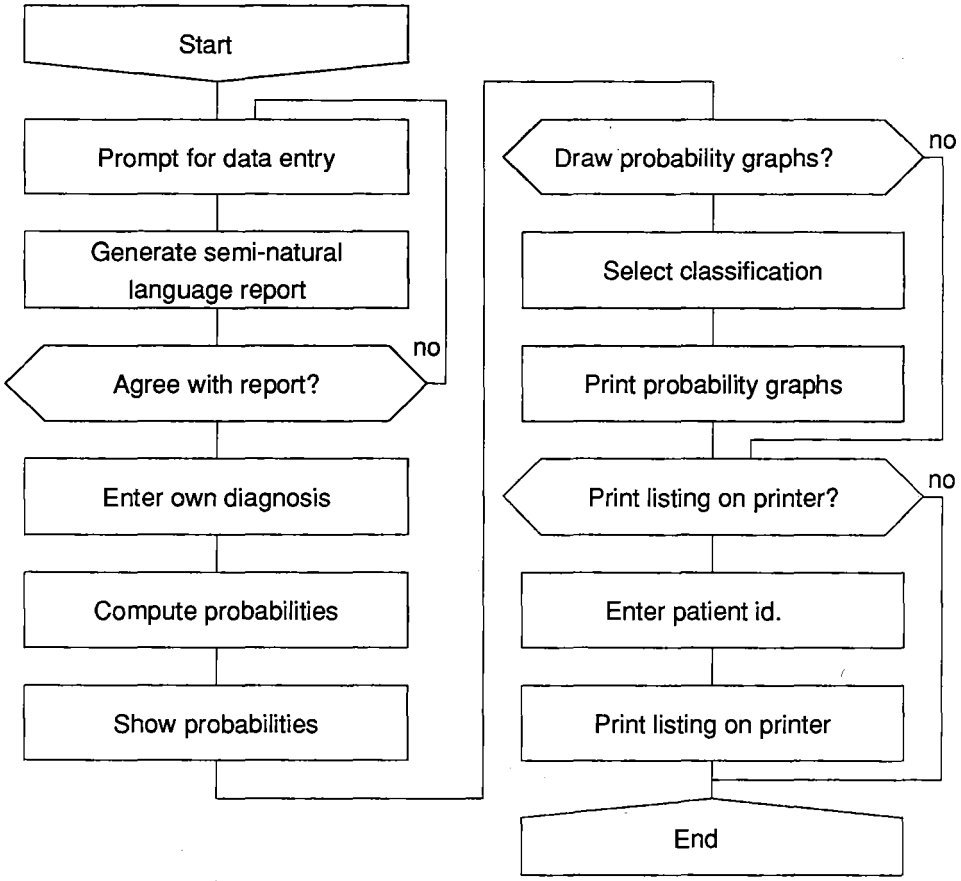


Figure 9.3. COMIP structurogram. Part b: Computations.

Sample run

A sample run of the COMIP program with comment is shown in the Appendix. In the example of the installation run we show the incidences that match the jaundiced population of the Internal Medicine II department of Dijkzigt hospital, Rotterdam. The reference ranges are those used by the clinical laboratory. In the application example we use data from a patient

that was proved later to suffer from acute viral hepatitis, to be classified as 'acute medical' in COMIP terminology. All output applies to this specific patient.

Conclusion

Despite the attention which computer assisted diagnosis programs have received in recent years, they have had little impact on routine clinical decision making. Ease of application and improved sensitivity to the distribution of disease and symptoms may help facilitate the use of such a program.

The COMIP program is a prototype of such a decision support system for early decision making in jaundice. The algorithm underlying COMIP has been based on solid data gathered in the COMIK study and evaluated on multiple occasions. The program presented here is a by-product of this study. The COMIP program forces the user to pay attention to concepts underlying transportability of computer-assisted diagnosis systems.

Hardware and software specifications

COMIP is intended for use on a small microcomputer. Memory requirements are limited and it will operate on any IBM/XT compatible computer using MSDOS version 2.0 or upwards with 5 1/4 inch 360 Kb floppy disks. An optional printer can be connected to obtain a hardcopy of the results. If graphs are also to be printed the printer has to be an industry standard compatible (Epson MS/FX 80 compatible) one.

The COMIP program is written in the C language. Transfer to hardware equipment lacking graphical facilities has been anticipated. The program has also been translated into PASCAL. The C source code is compatible with both Microsoft C Version 4.0 and Turbo C Version 1.0. The PASCAL source code is compatible with Turbo Pascal Version 3.0.

Availability

A floppy disk containing the COMIP program is available from the author on request at administration costs only.

Information

Information can be obtained either by direct mail or by computer-network:

SEGAAR@HROEUR51.BITNET

References

Boo86 Boom R, Gonzalez C, Fridman L, Alaya J, Realpe JL, Morales P, Quintero R. Looking for the 'indicants' in the differential diagnosis of jaundice. Med Decis Making 1986; 6: 36-41.

DeD81 De Dombal FT, Staniland JR, Clamp SE. Geographical variation in disease presentation. Does it constitute a problem and can information science help? Med Decis Making 1981; 1: 59-69.

Lin83 Lindberg G, Nilsson L, Thulin L. Decision theory as an aid in the diagnosis of cholestatic jaundice. Acta Chir Scand 1983; 149: 521-9.

Lin87 Lindberg G, Thomsen C, Malchow-Møller A, Matzen P, Hilden J. Differential diagnosis of jaundice. Applicability of the Copenhagen pocket chart proved in Stockholm patients. Liver 1987; 7: 43-9.

Mat84 Matzen P, Malchow-Møller A, Hilden J. Differential diagnosis of jaundice: a pocket diagnostic chart. Liver 1984; 4: 360-71.

Seg87 Segaar RW. The COMIP reference manual Version 1.1. Technical report. August 1987. Rotterdam: Department of Public Health and Social Medicine, Erasmus University, 1987.

Seg88 Segaar RW, Wilson JHP, Habbema JDF, Malchow-Møller A, Hilden J, van der Maas PJ. Transferring a diagnostic decision aid for jaundice. Neth J Med 1988; 33: 5-15.

Seg90 Segaar RW, Wilson JHP, Habbema JDF, Malchow-Møller A, Hilden J, van der Maas PJ. The impact of reference-range differences on algorithmic classification of jaundiced patients. Submitted for publication.

Appendix

Sample run of the COMIP program.

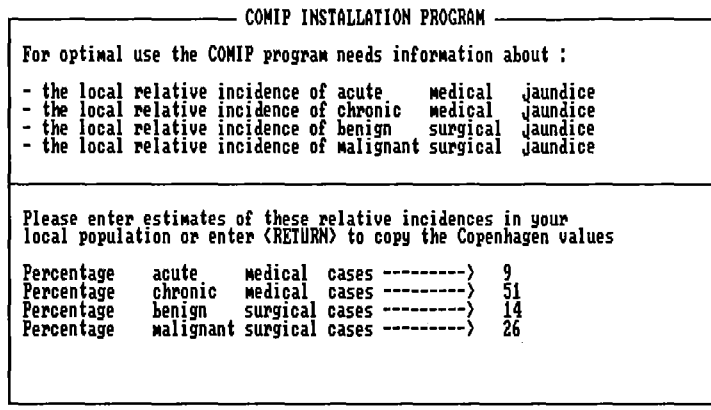


Figure 9.4a. The COMIP installation program sets the values for the relative incidences.

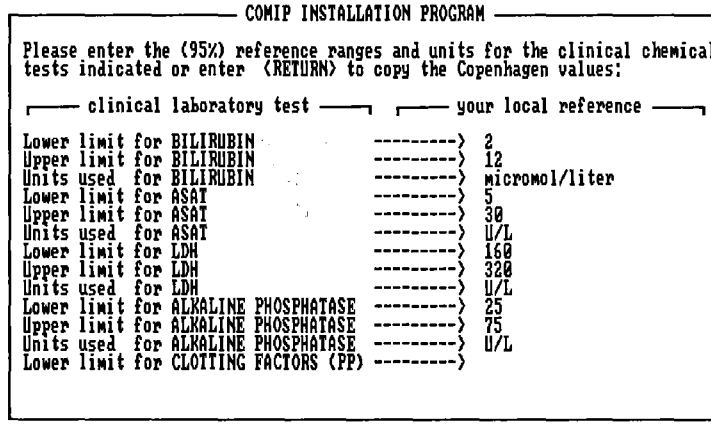


Figure 9.4b. The COMIP installation program sets the laboratory reference ranges.

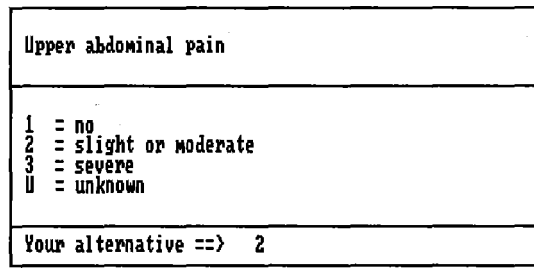


Figure 9.4c. Data entry in the COMIP program (only two out of 21 diagnostic questions are presented).

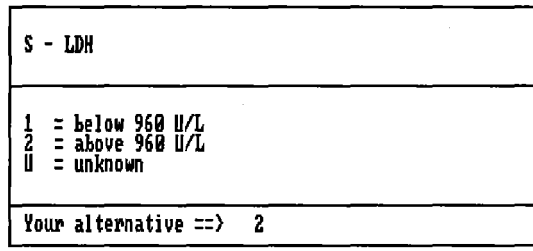


Figure 9.4d. Data entry in the COMIP program showing reduction of biochemistry data to a binary scale.

```
Age : between 15 and 30 years

Previous history is negative for:
- cirrhosis
- cancer
- leukemia
- biliary colics / proven gallstones
- congestive heart failure

Duration : shorter than 2 weeks

Present history is negative for:
- intermittent jaundice

Present history is positive for:
- abdominal pain (slight/moderate)
- fever (without chills)
- weightloss above 2 kg
- alcohol (1 - 4 drinks per day)

Physical examination is negative for:
- spiders
- ascites
- nodular liver surface
- palpable gallbladder

Clinical chemistry shows
- bilirubin above 127 micromol/l
- alk phos 123 - 307 U/L
- asat above 224 U/L
- clot fact above 0.72 PP
- ldh above 960 U/L

                                Do you agree (Y/N) ?
```

Figure 9.4e. Data entry is completed with a semi-natural language summary which should be verified before proceeding.

```
If a clinical diagnosis is already suspected it can be
entered below for a comparison with the probability calculations

 acute medical causes
A  Acute viral hepatitis
B  Drug hepatitis
C  Alcoholic hepatitis
D  Chronic persistent hepatitis
E  Septicemia
F  Postoperative jaundice
G  Heart failure
H  Congenital hyperbilirubinaemia
I  Hemolysis

 chronic medical causes
J  Alcoholic cirrhosis
K  Posthepatic cirrhosis
L  Cryptogenic cirrhosis
M  Primary biliary cirrhosis
N  Chronic active hepatitis
O  Hepatocellular carcinoma

 benign surgical
P  Choledocholithiasis
Q  Acute cholangitis
R  Pancreatitis
S  Iatrogenic bile duct lesion
T  Secondary biliary cirrhosis

 malignant surgical causes
U  Bile duct carcinoma
V  Pancreatic carcinoma
W  Liver metastasis

Enter RETURN if unknown

Enter your choice => ___
```

Figure 9.4f. A preliminary clinical diagnosis can be entered using the COMIK classification. It plays no role plays no role in the actual computation, but serves as an unbiased comparison.

```
Sum-scores                          Computed probabilities
Medical(-) vs Surgical (+) -18      P(Medical)  =  0.98    P(Surgical) =  0.02
Acute  (-) vs Chronic  (+) -23      P(Acute)    =  1.00    P(Chronic)  =  0.00
Benign (-) vs Malignant(+)  11      P(Benign)   =  0.07    P(Malignant)=  0.93

Computed COMIK probabilities of the 4 main categories are
        Category                                  Computed probability
Acute      Medical  causes of jaundice      Probability (P) =   0.98
Malignant  Surgical causes of jaundice      Probability (P) =   0.02
Benign     Surgical causes of jaundice      Probability (P) =   0.00
Chronic    Medical  causes of jaundice      Probability (P) =   0.00

Most likely category according to COMIK rule is Acute     Medical  (p=0.98)
This category is   52.7 times as likely as    Malignant Surgical (p=0.02)

Clinical diagnosis so far is                 : Unknown

Press any key to continue....
```

Figure 9.4g. Default summary of the results produced by the COMIP program showing both the COMIP diagnosis (diagnosis assigned highest probability) and the clinical diagnosis (if entered).

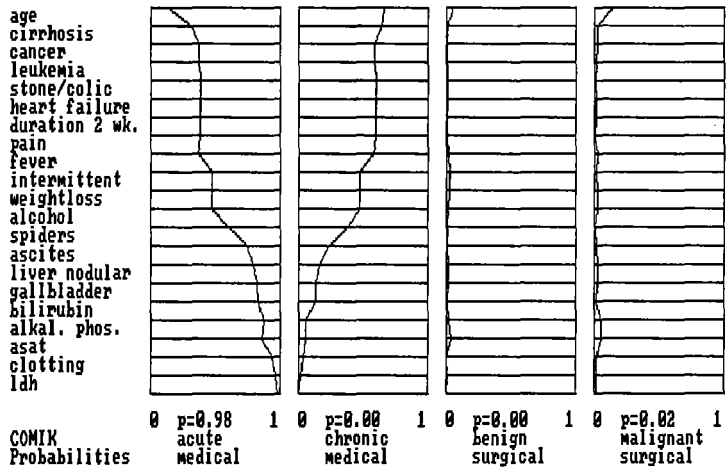


Figure 9.4h. Results can also be visualized in graphs. An example of a four-way probability graph is shown.

sequential nature of the diagnostic workup that generally takes place in the clinic. In Chapter 8 we also conclude that the sequential approach taken by most clinicians is quite similar to the approach that would have resulted from application of formal diagnostic aids.

Gold standard

A diagnostic gold standard is crucial in the evaluation of diagnostic aids. Diagnostic gold standards serve as a comparison diagnosis for the diagnostic outcomes of the diagnostic aids evaluated. Ideally, a gold standard in the context of diagnostic models has the following characteristics:

- it has an operational definition
- it uses a classification that is generally accepted
- the classification can be applied in a consistent way
- it makes a true statement about the diagnosis
- its truth can be verified through several methods, independent of the way in which the gold standard is derived, and independent of the information utilized within the diagnostic aid.

Most gold standards are imperfect. This applies to medicine, but also to other fields. From research in physics on particle masses (Par76) it is known that gold standards on particle masses have shown a variation over the years. Under all circumstances, doubt concerning gold standards makes results obtained difficult to interpret. We will therefore verify whether the ideal situation regarding gold standards, also applies to our gold standard.

Operationality of the gold standard: In Chapter 2 we discussed the gold standard as it was used in this thesis. We defined it as *the disorder held most responsible for the current episode of jaundice of a patient.* We found this a useful definition since it allowed for a coexistence of several diagnoses capable of causing jaundice, of which only one was considered responsible for the current episode of jaundice. Furthermore, it is plausible that subsequent therapeutical actions will be directed towards the disorder most responsible the present episode of jaundice. Focussing on the disorder most responsible for the current episode of jaundice also ensures that we focus on the most relevant part of the diagnostic problem.

Acceptance: The diagnostic classification scheme used by the COMIK group (shown in table 5.1), that we used throughout this thesis, is certainly not the only diagnostic classification of jaundice conceivable. Except for one diagnosis (sclerosing cholangitis) we had no problems mapping the final diagnoses of our patients onto these classifications. Also, the coordinating committee of the concerted action on jaundice (the EC Euricterus project), has used in principle the same classification, although they refined the classification with an additional level.

Objectivity: Classification for the gold standard was based on a review of all information gathered for a patient, interpreted by an experienced hepatologist. In the final review, information obtained from techniques like ERCP, biopsy, findings at laparotomy and autopsy, played an important role.

Consistency: With the COMIK classification we could encode our patients with relative ease. From this we cannot infer whether the classification will also be used in a consistent manner when applied clinically. Differences in utilization of a diagnostic classification, derive from inter-observer variation (differences in the way clinicians perceive information), and from inadequacies in the classification used.

Precision: We were able to establish an exact classification of all patients on the global four group level. At the more detailed level (21 diagnoses for the COMIK chart) we encountered some patients where it was impossible to make a precise statement about the diagnosis at that level.

Independence: While establishing a final diagnosis, the expert had to use information that was also utilized by some of the models. This of course bears a risk of circular reasoning. But the final diagnosis of the expert primarily relied on invasive techniques like ERCP and biopsy. These techniques played no role in the diagnostic aids, and we therefore consider the risk of circular reasoning of minor importance.

Evaluation methods

In its basic form, *evaluation of a diagnostic aid should assess the agreement between a diagnostic gold standard, and diagnostic results*

obtained from the aid. Preferably, the observed analogy should be expressed using quantitative methods because in that way, aids can be compared. All diagnostic aids studied by us deal with probability in some way. It would therefore be conceivable to use other methods to evaluate performance instead of error rates. In the literature, some relevant overviews on evaluation of probabilistic diagnostic aids can be found (Hab78, Hil78a, Hil78b, Hab79, Hab81a, Hab81b, Hil90). Aiming at a comparison of several diagnostic aids however, it was decided to rely only on simple quantitative measures applicable to all aids. From this decision, the choice of the (non) error rate as our principal measure followed.

Results of the evaluations

We expressed results of our evaluations as simple error rates. With these results at hand the crucial question remains what meaningful interpretation such results have in clinical practice. To answer the question, we are in need of references describing the diagnostic classification performance achieved by clinicians on the basis of the data used by the diagnostic aids. But data obtained through studies like that of Malchow-Møller (Mal87) on such clinical performance are rare. In that respect, the diagnostic aids too often serve as their own reference. One way to proceed is to gather more data on quantitative assessments of the diagnostic performance of clinicians like those of Malchow-Møller (Mal87). An other approach may come from the more advanced evaluation methods.

Other techniques for evaluation

The evaluations in this thesis do not address the fundamental question: will clinicians equipped with diagnostic aids achieve better diagnostic results compared to clinicians deprived of diagnostic aids. From an operational point of view, such comparisons are to be preferred. Such comparisons are practically achievable and produce intriguing answers, at the cost of great efforts. But even after such evaluations, some unresolved questions remain. One such question addresses the gains in utility experienced by patients, from different diagnostic strategies. These gains in diagnostic utility are the 'true' endpoints to measure the impact of diagnostic activity. Assessing diagnostic activity in that way goes far beyond the diagnostic performance assessed in this thesis. Attempts to assess diagnostic activity using true

endpoints may also be difficult to implement, for example because differences in outcome due to different diagnostic procedures are bound to be diluted by the many activities not strictly under control, like the choice of therapeutic actions possible.

Acceptance of the fact that diagnostic accuracy is a 'surrogate' endpoint substituted for improvement in patient utility (the 'true' endpoint) helps to evaluate the role of the diagnostic process in medicine. For some of the non-directed diagnostic activities like pre-operative screening, the concept may indeed elicit a first realization of its limitations, for other diagnostic activities the concept may be helpful in promoting changes from which patients benefit.

Users, transition to usage, and usage

In the previous chapters we made some references to the potential usage and the potential users of the diagnostic aids discussed there. We recognize that our statements have been implicit or incomplete at times, which justifies a more detailed discourse in this section.

Potential users

All our evaluations have been hospital-based, and from the work in this thesis we conclude that the use of diagnostic aids may be useful in specialist-based patient care. This conclusion also extends to tertiary referral hospitals, like Dijkzigt hospital. A conclusion might be that there is a place for diagnostic aids in hospitals only, and not in primary care. We will address the opportunities for diagnostic aids in primary care from two perspectives. First we look at the methods employed for the diagnostic aids. Next we look at the subject jaundice.

Our diagnostic approach itself, especially its focussing on summarization of information early in the diagnostic process, would also fit into the needs of general practitioners. Of course, the issue of relative incidences has to be addressed again. In Chapter 6 we showed that the flowchart showed little difference in performance on patients referred by a general practitioner, as compared to those referred by an internist. These findings justify optimism about the perspectives of the aids in primary care.

But the symptom jaundice is associated with diagnoses that often require specialized treatment. For that reason, diagnostic activities for jaundice mainly take place outside the primary care sector, usually in a department of internal medicine or a surgical department. As most therapeutic alternatives for jaundice are likely to remain hospital-based, it seems reasonable to expect that most of the diagnostic activity for jaundice will also continue to take place there. The diagnostic aids for jaundice can then guide the referral of the general practitioner to either the internist, or to the surgeon.

In conclusion we believe that the diagnostic approaches described in this thesis can be extended to diagnostic problems in primary care, provided that all procedures for careful implementation of diagnostic aids are fulfilled. Especially relative incidences of diseases have to be considered carefully. Jaundice may not be the best diagnostic problem to start with in primary care. Better problems to work on are not hard to find (Hod78). Especially those diagnostic dilemmas where diagnostic and therapeutic options remain within the scope of the general practitioner are suitable candidates.

Transition to usage

The transition of a diagnostic aid from a research tool, into a tool applied in clinical practice is a goal aimed at in most research on diagnostic aids. Despite prolific literature on diagnostic aids developed and applied in protected research environments, we find that literature describing routine application of diagnostic aids is rare. This also applies to the aids considered here: none of them is applied on a routine basis. Two explanations for this discrepancy come to mind. One explanation may be that transfer of diagnostic aids to non-experimental clinical settings is frequently unsuccessful. An other explanation may be that only few attempts for such transfers are actually made. Both hypotheses deserve further exploration.

Application of a diagnostic aid may turn out to be unsuccessful in practice, due to a poor performance on its diagnostic task, or to difficulties in its application. A decline in performance is always conceivable, in the same way drug treatments turn out less successful in practice as compared to their clinical trials. Such declines in performance are explained by the fact that original patient criteria are applied less strict in practice. But even when performance is not an issue, application may face difficulties. Translation of research within the scope of the abstraction 'diagnostic classification', to the clinical environment is anything but trivial. The attempts made to solve diagnostic classification problems in a proper way impose restrictions on the

resulting diagnostic aids. As a result, these aids may sometimes have a 'toy-like' appearance. In the clinical environment the products of such research have to compete with other, more pragmatic considerations, unless the research was planned with a pragmatic scope in mind. Pragmatic incentives were certainly part of our research, as we focussed on fast and simple diagnostic information, used early in the diagnostic process, guiding the selection of subsequent diagnostic tests.

In other cases, no attempt for clinical application is even made. This does not imply a failure by default. The introduction of diagnostic aids always requires adaptations of some sort in the clinical environment. Once these adaptations have been implemented, significant improvements in clinical management are possible, even *without* explicit use of the aid. The final improvements obtained from application of the diagnostic aid may then be small compared to the improvements already achieved.

Potential usage

'Diagnostic performance' has been a central issue in the foregoing chapters. In spite of the fact that the diagnostic aids discussed there are used in a process that leads to establishing a final diagnosis, *establishing a final diagnosis was not the intended purpose of the diagnostic aids discussed in this thesis.* All aids are used early in the diagnostic process, and their goal is *to guide further diagnostic activities.* In that context, the gold standard diagnosis is necessary only as a surrogate measure for an assessment of the quality of the aid in *guiding further diagnostic activity*, and not as an *assessment of the capability to establish a final diagnosis.* Looking upon our evaluations as a method to assess the performance of the aids as tools to achieve a final diagnosis must be regarded inadequate and will result in an disappointment. Nevertheless, such an error is often made. We take an opportunity once more to prevent such inadequate inferences to be drawn from our work. The previous statements also imply that the diagnostic aids discussed in this thesis are no substitute for further diagnostic activities. Additional diagnostic actions will remain necessary.

It it also worthwhile to look again at the sequencing of events when diagnostic aids are applied. We must distinguish between the operation of the aid itself, and the environment in which it serves its goal. Our diagnostic aids are intended for use within *sequential* diagnostic strategies. Looking at the data in our database, we can see that on many parameters data are missing, particularly in invasive, or costly procedures. This points at the fact

that clinicians actually use sequential strategies, through which some diagnostic procedures are performed, and others not. The information that comes forth from application of the aids should therefore influence the sequence of diagnostic actions that follows. The confident diagnoses ($P > 0.80$) are a good basis to plan focussed diagnostic activities on. In addition, the concept of confident and doubtful diagnoses in the COMIK chart, and of confident endnodes and doubt endnodes in the flowchart, can help to achieve better performance, by indicating situations where one is in need of supplemental information.

The operation of the diagnostic aids themselves is subject to variation. The COMIK chart is operated in a non-sequential manner. The HEPAR system on the other hand operates in a semi-sequential manner. But as the HEPAR system too relies on simple non-invasive, no-risk diagnostic procedures, the typical problems that are manifest in the implementation of a sequential strategy, do not occur here. It would make hardly any difference to operate the HEPAR system in a non-sequential way.

Truly sequential systems can be designed, and such usually involves a decision-analytic approach in which diagnostic probabilities and patient utilities for the various outcomes, are used to optimize the expected utility. Theoretically, such approaches are to be preferred whenever applicable. The greatest gains to be made however, can be found at the end of the diagnostic process, when invasive, inconvenient or costly procedures come into sight. The diagnostic aids that we evaluated in the foregoing chapters leave these techniques out. From this we may expect that, although limited benefits can be anticipated, sequential strategies at this (early) part of the diagnostic process will not substantially improve the performance of the systems *in their current state*. Once our systems are enhanced further, incorporating techniques like ERCP, and biopsy, the opposite becomes true. Such systems indeed require decision analytic approaches in order to claim optimality. Here some interesting future outlooks appear. The simplest approach to develop such systems will be through decision analysis of some global strategies, that find their implementation in the system. A more advanced approach would be to incorporate the decision analytic process into the reasoning strategies of the diagnostic aids. In that way, strategies can be optimized using the utility structure of the patient. In the diagnostic process such utilities are often expressed in extremes only, like unwillingness of the patient to undergo a procedure. The incorporation of decision analytic techniques into diagnostic aids is a complicated matter. Such is however compatible with the finding that most developments in this field, finally all converge in being complex.

Use in medical education

There are several parallels between the activities that take place in medical education and in research on diagnostic aids. Both activities rely on abstraction. In medical education abstraction is necessary to verbalize the messages, whereas in diagnostic aids it is necessary to end up with functional systems. For the same reasons, medical education and diagnostic aids benefit from restricted domains.

In contrast to diagnostic aids however, medical education pays little attention to the choice of options in the context of some optimality criterion. This is an area in need of more attention in medical education, but perhaps also in clinical practice. In our opinion, diagnostic aids can successfully be employed in medical education. From their use, many concepts related to diagnostic decision making under uncertainty, can be introduced. Examples of such concepts are prior- and posterior probabilities, test characteristics and thresholds to be applied. Although abstract, these concepts are worthwhile teaching to students.

Transfer of diagnostic aids

On transfer of diagnostic aids to other (geographical) locations, one has to keep in mind that several factors may disturb the performance of these aids when applied elsewhere. Differences in reference ranges are in fact only a symptom, indicating that some features of the laboratory procedures are different. As a consequence, our approach to these differences was at its best *symptomatic*. We conclude that the transfer of a diagnostic aid can be improved through simple symptomatic adjustments for the differences in reference values for clinical chemical tests. This conclusion seems to contradict the earlier studies of the COMIK group (Hil80, Mat84) in which systematic errors were shown to be of no importance. But the *systematic* perturbations tested for by Matzen (Mat84) were +40 % and -40 %, so the factor 3.3 for alkaline phosphatase (230%, Copenhagen vs Rotterdam) is far beyond this range.

Symptoms and signs of the patient also contribute to diagnostic information. Patient history and physical examination are very important for an optimal diagnostic assessment of the patient, early in the diagnostic process. Differences between hospitals in the performance on these items may also be responsible for a decline in the performance of diagnostic aids when used elsewhere. Such differences in the performance on history-taking and

physical examination have been described before for the Dyspepsia System (Lin87). In our evaluation of the flowchart (Chapter 6) we found that the discriminative performance of information on the nodularity of the liver surface was not optimal. This may be a result of differences in the traditions on physical examination, similar to those in Lin87.

The problem of differences in local relative incidences is general to all clinical activities that make assumptions about them. To be optimal in the diagnostic context, every hospital might have to implement a somewhat different diagnostic procedure, based on the relative incidences encountered. It might therefore be recommended that clinicians in each hospital try to assess the relative incidences encountered. From that information, they should infer whether the diagnostic strategies implemented are in correspondence to these population characteristics.

The future

From the previous chapters, a reader may conclude that future perspectives for diagnostic aids are uncertain. If one is prepared to perceive the development of diagnostic aids as a high-tech method to cultivate intellectual capacities regarding a particular diagnostic area, then research on diagnostic aids remains worthwhile. Studies like ours help focus on the essential components of diagnostic problems. They also help to promote a better communication and interpretation of information. This is mandatory also in other activities like medical training programs. The question remains whether the clinical introduction of diagnostic aids is always an endpoint to aim at. From what is known now, the introduction of diagnostic aids into clinical settings not only depends on further technical improvement of the aids themselves, but also - and perhaps even more - on changes in the clinical practice that inhibit introduction now. There is a fair chance that, once those changes and improvements are realized, clinical practice will adopt some of the reasoning underlying the decision aids. The actual introduction of the diagnostic aids will then be a confirmation of changes that have actually been taking place. The changes will have been stimulated by research on decision aids and will provoke definite and measurable improvements in the management of our patients. Since this is an ultimate endpoint worth aiming at, we may expect benefits from research on diagnostic aids now, but also in the future.

References

Hab78 Habbema JDF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis. Part 1. The problem, descriptive tools, and measures based on classification matrices. Meth Inf Med 1978; 17: 217-26.

Hab81a Habbema JDF, Hilden J. The measurement of performance in probabilistic diagnosis. Part 4. Utility considerations in therapeutics and prognostics. Meth Inf Med 1981; 20: 80-96.

Hab81b Habbema JDF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis. Part 5. General recommendations. Meth Inf med 1981; 20: 97-100.

Hil78a Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis. Part 2. Trustworthiness of the exact values of the diagnostic probabilities. Meth Inf Med 1978; 17: 227-37.

Hil78b Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis. Part 3. Methods based on continuous functions of the diagnostic probabilities. Meth Inf Med 1978; 17: 238-46.

Hil80 Hilden J, Matzen P, Malchow-Møller A, Bryant S. Precision requirements in a study of computer aided diagnosis of jaundice (The COMIK study). Scand J Clin Lab Invest 1980; 40(supp 155): 125-8.

Hil90 Hilden J, Habbema JDF. Evaluation of clinical decision aids. More to think about. Med Inform 1990; 15: 275-84.

Hod78 Hodgkin K. Towards earlier diagnosis in primary care. 4th ed. Edinburgh: Churchill Livingstone, 1978.

Lin87 Lindberg G, Seensalu R, Nilsson LH, Forsell P, Kager L, Knill-Jones RP. Transferability of a computer system for medical history taking and decision support in dyspepsia. A comparison of indicants for peptic ulcer disease. Scand J Gastroenterol 1987; 128(supp): 190-6.

Mal87 Malchow-Møller A, Mindeholm L, Rasmussen HS, Rasmussen B, Wilhelmsen F, Petersen-JS, Jørgensen S, Hilden J, Thomsen C, Matzen P, et al. Differential diagnosis of jaundice: junior staff experience with the Copenhagen pocket chart. Liver 1987; 7: 333-8.

Mat84 Matzen P, Hilden J, Thomsen C, Malchow-Møller A, Juhl E and the COMIC study group. Does test quality influence the outcome of algoritmic classification of jaundiced patients. Scand J Clin Lab Invest 1984; 44(supp 171): 35-40.

Par76 Particle data group. Reviews of particle properties. Rev Modern Physics 1976; 48(supp): S1-S20.

Tea81 Teach RL, Shortliffe EH. An analysis of physician attitudes regarding computer-based clinical consultation systems. Comp Biomed Res 1981; 14: 542-58.

Summary

Making a diagnosis is always a challenge for the clinician who faces a jaundiced patient. A diagnosis provides a basis for prognosis, and guides therapeutic action. Modeling of the diagnostic process provides an understanding of the contributing elements. In addition, diagnostic models can be applied to real world situations. This can be useful when the model supports or complements the clinician on particular tasks. In that way, the diagnostic model becomes a diagnostic aid. Computers can often be successfully used to implement diagnostic models, hence the term 'computer assisted diagnosis'.

The performance of diagnostic aids is always an important issue. Most diagnostic aids are developed at one place, whereas they are used at an other place. In the development of diagnostic models, one often models features that are characteristic to the local patient population. These features may be different for patients elsewhere. There is a chance that there will be a decline in performance of the diagnostic aids due to these differences, and that adjustment procedures are required. These are the crucial questions addressed in this thesis.

The symptom jaundice is a good starting point to model diagnostic processes. There is a reasonable consensus regarding the diagnostic categories. Diseases underlying jaundice can have serious consequences for the patient. Jaundice also is a relatively frequent symptom. These properties, together with practical, historical and organizational motives explain our choice of jaundice.

Two retrospective data collections on jaundiced patients admitted to the department of Internal Medicine of Dijkzigt hospital are used throughout this thesis. A first series included 100 patients, whereas a second contained 214 patients. With these data collections at hand, we could answer the questions formulated. Especially the second data collection provided detailed information on the diagnostic parameters of the patients. There we also recorded the referral patterns. These data confirm that Dijkzigt hospital operates as a hospital with many secondary and tertiary referrals. The implementation of the data collections will be outlined in chapter 2, whereas chapter 3 present some background data on the patient populations.

Chapter 4 describes the evaluation of the HEPAR system. The HEPAR system is an expert system for diagnosis of the liver and the biliary tract, developed in the Netherlands (Leyden, Amsterdam). Expert systems are computer programs which come forth from research on artificial intelligence (A.I.). They have an unconventional architecture, in which there is a complete separation of the knowledge, and the procedures that apply the knowledge to accomplish particular goals. Expert systems are often used for classification problems, like diagnostic problems in medicine. The performance of the HEPAR system was evaluated with data from the Rotterdam patients. The impact of missing data on the results was also studied. The conclusion is that HEPAR is capable of generating diagnostic statements that closely match the actual state of the patient.

Chapter 5 describes the evaluation of the COMIK algorithm for jaundice diagnosis. The COMIK algorithm was developed in Copenhagen to provide a diagnosis of jaundice, based on information available from patient history, physical examination, and simple biochemistry. Using this information, the algorithm assigns probabilities to the four major groups of jaundice: acute non-obstruction (like viral hepatitis), chronic non-obstruction (like liver cirrhosis), benign obstruction (like gallstones), and malignant obstruction (like pancreatic cancer).

The most striking feature of the COMIK algorithm is its presentation to the user. The Copenhagen group designed a paper and pencil version of the model, that allows clinicians to make some simple computations that result in an assignment of probabilities to the diagnostic categories under consideration. The algorithm has been evaluated in several countries now, and all evaluations showed that the algorithm maintained its performance. We describe the evaluation of the algorithm in Rotterdam and conclude that the algorithm remains valuable there too. To maintain adequate performance however, we had to make local adjustments. These adjustment procedures were necessary, because of differences between the Copenhagen and in Rotterdam hospitals. Both for the reference ranges for biochemistry, and for the relative incidences adjustments were necessary. The differences encountered are no exception, but in fact relevant to any transfer of a diagnostic aid between hospitals.

In chapter 6 we evaluate the performance of a flowchart for jaundice diagnosis. Flowcharts are often used to structure problems. Such problems may be diagnostic, but flowcharts are also used for other problems. The development of the flowchart was related to that of the COMIK algorithm. Also its diagnostic classification scheme, and our method of evaluation were similar. The performance of the flowchart was inferior to that of the COMIK

algorithm. A possible explanation may be that we were not able to make the adjustments for the relative incidences of the diagnostic categories, in the same way we made adjustments to the algorithm.

In chapter 7 we use our data collections to develop and evaluate a local diagnostic model. To develop the model, we used data from the largest (N=214) dataset. In the analysis we selected variables based on their contribution to diagnostic classification. We compared the variables of the final model with those of the COMIK algorithm. It turns out that many variables of our model were also present in the COMIK model. Especially if one takes into account the large number of candidate variables, the differences between the patient characteristics at both locations, and the modest size of our dataset (214 versus 1002 in Copenhagen) we may conclude that the variables that appear in both models contain the core of diagnostic information for jaundice. There seems to be little difference between a carefully transferred diagnostic aid developed elsewhere, and the development a model based on local data.

In chapter 8 we compare the tests ordered by clinicians (diagnostic test behavior) with those recommended by formal models. For this comparison we use the COMIK algorithm to provide probabilistic assessments of the patients. Together with a protocol for test usage, these probabilities enable us to propose diagnostic tests. We conclude that it is indeed possible to predict test selection behavior of clinicians. This implies that clinicians are capable of selecting relevant tests. It is however also likely that the test selection was redundant to some extent. We discuss the potential use of this information in an educational context.

The the COMIK algorithm was originally designed for 'paper and pencil' use, but the computations turned out error prone in practice. Especially on repeated application on series of patients these errors became obvious. It was therefore rational to computerize these activities. In chapter 9 we describe a computerized version of the COMIK algorithm, COMIP: a **COM**puter **I**cterus **P**rogram. The program is compatible with our experience that local adjustments are often necessary. The program enables the user to make adjustments for biochemistry and relative incidences before actual application. In addition, we provided a graphical display of the contribution of the relevant parameters to the final probability. Due to these extensions, the program can be considered a prototype for probabilistic diagnostic aids.

In the final chapter of this thesis, we pick up those issues that are common to all foregoing chapters. Some problems related to the concepts 'diagnosis' and 'golden standard' will be discussed. We also elaborate on

the techniques for evaluation of diagnostic aids. We provide our views on anticipated users, and the anticipated use. Finally we take a look into the future of diagnostic aids.

Samenvatting

Het stellen van de juiste diagnose is steeds weer een uitdaging voor de arts bij wie de patiënt met geelzucht zich presenteert. Het stellen van een diagnose kan verstrekkende gevolgen hebben voor de kiezen therapie en de prognose. Door het diagnostisch proces te modelleren neemt het inzicht in de relevante aspecten toe. Daarnaast bestaat de mogelijkheid om modellen toe te passen. Daarbij wordt de in modellen geformaliseerde kennis toegepast op concrete situaties. Dit laatste kan nuttig zijn doordat de geoperationaliseerde modellen de arts ondersteunen of aanvullen. Op deze wijze groeit het diagnostisch model uit tot een diagnostisch hulpmiddel. Bij dit laatste wordt vaak en met succes gebruik gemaakt van de computer. Er wordt dan ook van computer-ondersteunde diagnostiek gesproken.

Een belangrijke vraag die men zich bij de toepassing van diagnostische hulpmiddelen steeds moet stellen betreft de te verwachten kwaliteit. Deze vraag krijgt een extra dimensie wanneer men bedenkt dat diagnostische hulpmiddelen vaak op een bepaalde lokatie ontwikkeld worden, terwijl ze op een andere lokatie toegepast worden. Tijdens de ontwikkeling van een diagnostisch hulpmiddel wordt meestal uitgegaan van de eigenschappen van een aanwezige patiëntenpopulatie, die niet noodzakelijkerwijs overeenkomen met de eigenschappen van patiënten elders. Het is dus goed mogelijk dat de kwaliteit van het diagnostisch hulpmiddel hierdoor verslechtert en dat er aanpassingen nodig zijn. Het zijn vooral deze onderwerpen die in deze dissertatie nader uitgewerkt worden.

Het symptoom geelzucht is om diverse redenen een goed uitgangspunt voor het modelleren van een diagnostisch proces. Er bestaat consensus over de aard van de te onderscheiden ziekte-entiteiten. Het geïsoleerde symptoom kan grote consequenties voor de patiënt hebben. Daarnaast is geelzucht zeker geen zeldzaam symptoom. Het zijn deze overwegingen, welke samen met overwegingen van meer practische, organisatorische en historische aard, bepalend zijn geweest voor de keuze van het onderwerp geelzucht.

In deze dissertatie is sprake van twee retrospectieve gegevensverzamelingen bij patiënten met geelzucht. Het betrof hier patiënten welke op de afdeling Interne Geneeskunde van het Dijkzigt ziekenhuis in Rotterdam werden opgenomen. Een eerste serie omvatte 100 patiënten, terwijl een latere 214 patiënten telde. Met behulp van deze gegevensverzamelingen konden de geformuleerde vraagstellingen onderzocht worden. Met name in de tweede gegevensverzameling werd de diagnostiek van deze patiënten uitgebreid in kaart gebracht. Ook werd het verwijzingspatroon geregistreerd. Deze gegevens bevestigen het bestaande vermoeden dat het Dijkzigt ziekenhuis veel secundaire en tertiaire verwijzingen kent. De opzet van de gegevensverzamelingen wordt uitgebreid besproken in hoofdstuk 2, terwijl in hoofdstuk 3 een aantal karakteristieken van de patiëntenpopulaties gepresenteerd worden.

In hoofdstuk 4 wordt de evaluatie van het HEPAR-systeem besproken. Het HEPAR-systeem is een in Nederland (Leiden/Amsterdam) ontwikkeld expertsysteem voor de diagnostiek van aandoeningen van lever en galwegen. Expertsystemen zijn computerprogramma's welke voortkomen uit het onderzoek naar kunstmatige intelligentie (artificial intelligence of A.I.). Expertsystemen kenmerken zich door een - ten opzichte van de gangbare computerprogrammatuur - onconventionele architectuur, waarbij er een scheiding wordt aangebracht tussen de door het systeem gebruikte kennis en de procedures waarin deze kennis gebruikt wordt om bepaalde doelstellingen te realiseren. Traditioneel vindt toepassing van expertsystemen vooral plaats bij klassificatie-vraagstukken, zoals medische diagnostiek. Het HEPAR-systeem werd met behulp van Rotterdamse patiëntengegevens geëvalueerd. Ook werd aandacht geschonken aan de vraag in hoeverre het ontbreken van gegevens van invloed is op de uitkomsten. Geconcludeerd wordt dat het HEPAR-systeem op basis van de ter beschikking staande patiëntengegevens tot uitspraken komt welke goed overeenstemmen met de actuele situatie van de patiënt.

In hoofdstuk 5 wordt de evaluatie van het in Kopenhagen ontwikkelde COMIK-algoritme voor de diagnostiek van geelzucht besproken. Op basis van eenvoudige informatie welke na anamnese, lichamelijk onderzoek, en eenvoudige biochemie beschikbaar is kent dit algoritme kansen toe aan de 4 hoofdgroepen van oorzaken van geelzucht, te weten acute niet obstructieve oorzaken (bijvoorbeeld virale hepatitis), chronische niet obstructieve oorzaken (bijvoorbeeld levercirrhose), benigne obstructieve oorzaken (bijvoorbeeld galstenen) en maligne obstructieve oorzaken (bijvoorbeeld pancreascarcinoom).

Het meest opvallende aan het COMIK-algoritme is de presentatie naar de gebruiker toe. De Kopenhaagse groep ontwierp daarvoor een papier- en potloodversie van het model, waarbij iedere arts op eenvoudige wijze enige berekeningen kon uitvoeren welke tot de uiteindelijke kanstoekenning leidden. Het algoritme is inmiddels in een aantal landen uitgetest, waarbij bleek dat de bruikbaarheid gehandhaafd bleef. De evaluatie van dit hulpmiddel met Rotterdamse patiëntengegevens wordt beschreven. Op basis van deze evaluatie wordt geconcludeerd dat de bruikbaarheid van dit systeem ook in Nederland behouden blijft. Om deze uitkomsten te realiseren dienden er vooraf wel een aantal aanpassingen aan het COMIK-algoritme plaats te vinden. Deze aanpassingen werden noodzakelijk nadat gebleken was dat er op een aantal punten belangrijke verschillen tussen de ziekenhuizen in Kopenhagen en Rotterdam bestonden. Dit betrof aanpassingen voor verschillen in de relatieve incidenties van de diagnostische groepen, en verschillen in normaalwaarden welke voor de biochemische parameters gehanteerd werden. De aangebrachte aanpassingen zijn niet incidenteel, maar altijd actueel bij overdracht van een diagnostisch hulpmiddel tussen ziekenhuizen onderling.

In hoofdstuk 6 wordt de evaluatie van een stroomdiagram voor de diagnostiek van geelzucht besproken. Stroomdiagrammen worden vaak toegepast om problemen te structureren, zowel in een diagnostische context, maar ook vaak daarbuiten. De voorgeschiedenis van dit stroomdiagram is analoog aan die van het COMIK-algoritme. Ook de toepassing, een globale vierwegsklassificatie van geelzucht, en de wijze waarop evaluatie plaatsvindt waren analoog. De resultaten van dit hulpmiddel blijven wat achter bij die van het COMIK-algoritme. Een mogelijke verklaring hiervoor is het feit dat het bij dit hulpmiddel niet mogelijk was om correcties aan te brengen voor de Rotterdamse relatieve incidenties van de diagnostische categorieën, hetgeen bij het algoritme wél mogelijk was.

In hoofdstuk 7 worden de aangelegde gevensverzamelingen gebruikt voor de ontwikkeling en evaluatie van een diagnostisch model voor geelzucht. Voor het afleiden van het model werd de grootste (N=214) database gebruikt. In de analyse werden variabelen op basis van hun bijdrage aan de diagnostische klassificatie geselecteerd. De variabelen die in het uiteindelijk ontwikkelde model werden opgenomen zijn vergeleken met de variabelen van het COMIK-algoritme. Hierbij blijkt dat een aanzienlijk aantal variabelen ook voorkomt in het COMIK-algoritme. Zeker als men let op het grote aantal kandidaatvariabelen in Rotterdam, de verschillen in de karakteristieken van de patiëntenpopulaties op beide lokaties, en de relatief geringe omvang van deze gegevensverzameling (214 versus 1002 patiënten in Kopenhagen)

moet geconstateerd worden dat de overeenkomende variabelen kennelijk zeer 'harde' informatie met betrekking tot de diagnostiek van geelzucht in zich dragen. Een evaluatie van het model met gegevens uit de N=100 database, laat zien dat de prestaties van het lokale model goed overeenkomen met die van het Kopenhaagse. Men heeft dus de keuze om òfwel een elders ontwikkeld model na zorgvuldige aanpassing te gebruiken, òfwel zelf een model te ontwikkelen op basis van een lokale patiëntenpopulatie.

In hoofdstuk 8 wordt een vergelijking gemaakt tussen de door clinici aangevraagde diagnostische tests (diagnostisch test-gedrag), en de door een formeel model aanbevolen tests. Bij deze vergelijking wordt het COMIK-algoritme gebruikt om kansuitspraken te doen over patiënten. Op basis van deze kansuitspraken én een expliciet voorstel voor testgebruik is het mogelijk om op formele gronden bepaalde tests te selecteren. Op basis van deze analyse wordt geconcludeerd dat het inderdaad goed mogelijk is om de door clinici aangevraagde tests (het testgedrag) te voorspellen. Dit impliceert dat clinici goed in staat zijn om relevante tests te selecteren. Het is echter wel aannemelijk dat het geobserveerde testgedrag in een aantal opzichten redundant was: diverse tests met overlappende informatie worden simultaan gebruikt. Ingegaan wordt op de mogelijkheden om de met dit experiment verkregen informatie te gebruiken in een educatieve context.

Hoewel het in de eerdere hoofdstukken beschreven COMIK-algoritme door de auteurs voorbestemd was om als 'papier en potlood' hulpmiddel door het leven te gaan, heeft het bij routine gebruik het praktische nadeel dat er toch snel (menselijke) fouten in de berekeningen sluipen. Zeker bij herhaalde toepassing op een serie patiënten, zoals bij de Rotterdamse evaluatie, vallen deze vergissingen snel op. Het ligt dan ook voor hand de bewerkingen waar mogelijk te automatiseren. In hoofdstuk 9 wordt deze gecomputeriseerde versie van het COMIK-algoritme, COMIP: een **COM**puter **I**cterus **P**rogramma, besproken. Geheel in overeenstemming met de ervaringen, opgedaan bij de Rotterdamse evaluatie van het COMIK-algoritme, biedt het ontwikkelde computerprogramma de gebruiker de gelegenheid om het programma *voor* gebruik aan de lokale situatie aan te passen. Het betreft hier aanpassingen voor de biochemie en aanpassingen voor de relatieve incidenties van de relevante ziektecategorieën. Ook werd in de mogelijkheid voorzien de bijdrage van de in het model gebruikte parameters aan de uiteindelijke kanstoekenning te visualiseren. Door deze aanpassingen is het programma vooral te zien als een prototype voor op kansrekening gebaseerde diagnostische hulpmiddelen.

In het laatste hoofdstuk wordt nader ingegaan op de elementen die als een rode draad door de voorgaande hoofdstukken heenlopen. Hierbij komen de problemen rondom de concepten 'diagnose' en 'gouden standaard' aan de orde. Ook wordt ingegaan op technieken voor evaluatie van de kwaliteit van diagnostische systemen. Er wordt een visie gepresenteerd op de doelgroep, en het verwachte gebruik van de diagnostische hulpmiddelen. Tenslotte wordt het toekomstperspectief besproken.

Curriculum Vitae

Robert Segaar was born on September 8, 1958. In 1976 he started his study medicine at Leyden University. During his study he worked as a student-assistant in the departments of Immunohematology and Neonatology of Leyden University Medical Center. In 1983 he graduated in medicine. In 1984 and 1985 he was a resident in the Pediatric Surgery department of Sophias' Children Hospital in Rotterdam. From 1985 to 1989 he worked as a researcher at the department of Public Health and Social Medicine, Erasmus University Rotterdam. From 1989 to 1990 he worked as a researcher for the Center for Clinical Decision Analysis at Erasmus University Rotterdam. Since 1990 he is a secretary for the Health Council of the Netherlands in The Hague, and a researcher for the institute of Medical Technology Assessment at Erasmus University Rotterdam.

Acknowledgements

A thesis is never is the work of an individual. Therefore, I wish to acknowledge the contributions of others in this final chapter. First of all, I want to express my gratitude to professor Dik Habbema and professor Paul Wilson. In my first meetings with them, they must at times have been puzzled by my bewildered ideas. I admire their capacity to transpose those ideas into structure and direction. Both established conditions within their departments that allowed completion of this study. Dik Habbema masters the technique of turning theory into practice. Beside that, he has a well-developed sense for omissions hidden in written material. He also had to use this while editing my drafts. Paul Wilson, as experienced clinician, made great contributions to the medical contents of this thesis. He was my link to clinical reality. I also took advantage of his knowledge to improve the language in this thesis.

The department of Public Health and Social Medicine, where work on this thesis started, brings forth many unusual researchers. Only few departments concentrate such disparate topics within one organization. I would like to thank all personnel from the department for their comments on my work, and specially Diederik Dippel and Jan Kardaun.

I am indebted to the personnel of the department of Internal Medicine II and the Medical Registration of Dijkzigt Hospital. They were helpful in collecting the data and in addition provided comment on my activities.

In the last year spent on this thesis, I benefitted from my contacts with the Center for Clinical Decision Making. This involved matters like accommodation and access to computing facilities, but also, more importantly, the opportunity to take part in many interesting discussions with its researchers. Professor Benbassat, visiting the center, also offered useful comments.

The information provided within this thesis is understood best in the context of related research commenced elsewhere. Although most chapters provide particulars on these links, a recapitulation is justified, as an explicit acknowledgement to excellent work done elsewhere.

In 1984 the Danish 'Computer Ikterus' (COMIK) group in Copenhagen published their first paper in a series on formal diagnostic aids for jaundice. This publication and the ones following, were a starting point for my research activities. The members of the COMIK Group, especially Jørgen Hilden and

Axel Malchow-Møller, became actually involved. Their feedback on my work was of great value. During his visits to the Netherlands Jørgen often contributed to my understanding of the problems faced. The hospitality of the COMIK group during my stays in Copenhagen was greatly appreciated.

The other line of research, was started by Peter Lucas from the Center for Mathematics and Computer Science in Amsterdam, and Roel Janssens of the department Gastroenterology of the Leyden Academic Hospital. They were the 'authors' the HEPAR expert system.

Finally, I want to direct my gratitude to the medical faculty of the Erasmus University for funding this project from their 'poolplaats' budget.

Rotterdam, January 1991.