EVINCE : a neuropsychiatric desktop expert system for the diagnosis of dementia

EVINCE

A Neuropsychiatric Desktop Expert System for the Diagnosis of Dementia

EVINCE

A NEUROPSYCHIATRIC DESKTOP EXPERT SYSTEM
FOR THE
DIAGNOSIS OF DEMENTIA

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Rijksuniversiteit Limburg te Maastricht,
op gezag van de Rector Magnificus, Prof. Mr. M.J. Cohen,
volgens het besluit van het College van Dekanen,
in het openbaar te verdedigen op
donderdag, 5 maart 1992 om 16.00 uur

door

Leonard André Plugge

geboren op 18 mei 1956 te Middelburg

PROMOTORES:

Prof. Dr. J. Jolles
Prof. Dr. Ir. A. Hasman

BEOORDELINGSCOMMISSIE:

Prof. Dr. H.G. Schmidt (voorzitter)
Prof. Dr. Ir. J.H. van Bemmel (Erasmus Universiteit Rotterdam)
Dr. H.F.A. Diesfeldt (Stichting Verpleeghuizen Nederland)
Prof. Dr. F. Sturmans
Prof. Dr. F.H.M. Verhey

Druk: Datawyse Maastricht / Krips Repro Meppel

ISBN 90 5291 072 3

The research described in this thesis was conducted at the department of Neuropsychology & Psychobiology of the University of Limburg, and is part of the "Aging Program" of the University of Limburg. The research was supported by a research grant from the Medical Technology Assessment Program of the Dutch Ministry of Welfare, Public Health and Culture (W.V.C.) (TA-87-19) (CRO-236509).

"Van alle dingen is de mens de maat, van de zijnde dat zij zijn, van de niet-zijnde dat zij niet zijn."

Protagoras (480-411 v.C.)

I EXPERT SYSTEMS AND PSYCHIATRY

1 INTRODUCTION

Neuropsychiatry is a multidisciplinary specialism with its origin in psychiatry and covering the grey areas between neurology, psychiatry and psychology. As with other multidisciplinary specialisms (for example, Artificial Intelligence. See paragraph 2.) it is difficult to formulate a precise definition of neuropsychiatry, and it is probably best described by what the focus of attention is. In an editorial Yudofski and Hales (1989) state that "A prominent focus of neuropsychiatry is the assessment and treatment of patients with psychiatric illnesses or symptoms associated with brain lesions or dysfunction".[1]

Although the multidisciplinary approach of neuropsychiatry is an advantage in studying mental disorders, for example dementia, it also poses the problem of diversity. Each discipline brings its own manner of examination, terminology and criteria into neuropsychiatry, which task it is to compare, combine and integrate these findings into diagnoses that offer clinicians a basis for treatment.

One obvious solution to this problem is to give specialists training in all these related disciplines. However, there is also an obvious -though difficult to measure- limit as to how much knowledge from several specialisms can be crammed into one person. Furthermore, the amount of specialists that can be expected to be trained will be limited, which in turn will limit other clinicians to request aid or to refer their patients.

Another solution to this problem is to combine specialized knowledge from two or more disciplines into a medium that can aid clinicians in using knowledge that is not part of their specialism. Since late 1960, such a potential medium is offered by Artificial Intelligence (AI) in the form of Expert Systems (ES). As already many ESs have been developed for medicine, we wil limit the scope of the discussion by focussing on the ESs that were developed for psychiatry.

The first part of this chapter will be devoted to a brief discussion on the origin of AI and how this led to ES research. The second part elaborates on what ESs are and how they generally work. Also, four types of knowledge representation used in ESs will be discussed. The remainder of this text will review the results of the psychiatric ESs that were developed during the last 15 years. In the final part we will try to enumerate the problems that were encountered by their developers and try to formulate some criteria and goals for the development of a psychiatric ES on the domain of dementia diagnostics.

2 The Origin of Expert Systems

In the 1950's a new field of research called Artificial Intelligence (AI) emerged as an interdisciplinary science of cognitive psychology and computer science.* The definition of AI has been a topic of debate ever since J. McCarthy and M. Minsky used the name for their project at MIT in 1959. This was mainly due to the lack of a good definition for 'intelligence' and to the pretentious tone of the name. This lack of definition made it also difficult for AI research to claim credits for new techniques, ideas and applications, as evidenced by the so called Lighthill Report[2] in 1973, which stated that much of the results at that time should be attributed to sciences in the category "Advanced Automation" and "Computer-Based Studies of the Central Nervous System". This report hampered British AI research for almost a decade, until the Alvey Report[3], which also introduced the name Intelligent Knowledge-Based Systems (IKBS) in 1982. However, the name IKBS was never accepted to the same extent as AI, and did not terminate the definition debate (For example, see the discussion in The Knowledge Engineering Review, started by M. Lam)[4,5,6,7,8,9,10]. Therefore, we will continue using the name AI in this paper, together with D. Waterman's definition which states that the goal of AI is "...to develop computer programs that (...) solve problems in a way that would be considered intelligent if done by a human."[11].

During the first decade of AI, this goal was pursued by scientists trying to develop general problem solvers (GPS). Although this effort failed, it did bring about one major finding: problem-solving is more a matter of knowledge than of inference techniques. AI researchers had severely underestimated the amount of (common) knowledge humans have and use for even the simplest problem to be solved. It was also clear that the software and hardware at that time were inadequate to solve this problem. This situation has not changed much since then. The solution to this problem was to drastically reduce the scope of problem solvers, and to develop special-purpose problem solvers with very limited but highly specialized knowledge, i.e. Expert Systems (ES).

3 What are Expert Systems?

Just like AI research has its debate about intelligence, ES research has its share of debate about defining what an Expert System actually is. One definition sais that an ES is "a computer system that achieves high levels of performance in task areas that, for human beings, require years of special education and training."[12] Although this definition applies equally well for any knowledgebase program or system, most other attemps to define what an ES is also allude to the requirement of human knowledge and skills that is represented in the ES and lead to a performance that is equivalent to that of a human.

---

* The term 'Artificial Intelligence' was first suggested by John McCarthy, co-founder of the MIT Artificial Intelligence Project in 1959.

However, the level of performance and the task area are not the only criteria by which to decide whether a computer program is an ES. Several other features are commonly considered essential: the ability to explain how a solution was reached, an intelligent search capability (or at least no blind search). reasoning with different levels of certainty (as opposed to a binary 'yes' or 'no'), a separation between inference mechanism and knowledge, and -last but not least- symbolic reasoning. The origin of the discussion about these criteria can be traced back to the dispute about the definition of AI. In a reply to Paul M. Churchland and Patricia Smith Churchland's article[13], Searle suggested a separation of AI in a 'hard' and a 'weak' approach[14]. The 'hard' AI was thought to develop programs that model human behavior as exactly as possible, while the 'weak' AI was believed to be interested only in developing programs capable of simulating human performance, i.e., produce a performance equal or better than humans, irrespective of the means used. Although this division can be useful to make a distinction between, for example, systems that are designed to provide a psychological model and systems that should provide correct answers, in practice, the distinction between modelling and simulating is less clear than Searl seems to suggest. This is especially true for the 'weak' AI, where researchers often use probem solving methods that are thought to resemble human problem solving, for reasons of speed. Thus, even 'weak' AI researchers use models of human problem solving. The recent development to use AI techniques in 'conventional' programs has blurred this distinction even further. For example, on-line spelling checkers, proof reading programs, and installation programs that try to establish an optimal configuration for computer memory all try to solve problems which were formerly performed by (proficient) humans. These programs even use fallible rules-of-thumb to solve their problems.

Thus, although it is difficult to provide unambiguous criteria to distinguish ESs from conventional computer programs, one could say that ESs focus exclusively on tasks (presently) performed by human experts, and that ESs are built in a such way that a user is able to ask (check) what the program 'knows', what it has done, why it has done that, and what it is about to do.* Furthermore, ESs typically cover task domains that are at the edge of what can be done efficiently using conventional programming aids.

The architecture of ESs is definitively different from regular programs. Most ESs consist of three (more or less) separate parts: 1) a knowledge base, 2) an inference engine, and 3) a user/explanation interface. (See Figure 1)
Although large differences can exist between ESs, in general the functioning of the inference engine and the user interface depends heavily on the organization and representation of the knowledge used. Based on ideas on how humans organize their knowledge, several representational methods have been developed. Three of the most common used methods are: 1) rules, 2) semantic nets and 3) frames. We will briefly discuss each one of them.

* However, sometimes ESs are developed without explanation facility to protect the implemented knowledge. (Peter van Lith, Lithp Systems BV. Personal communication.)

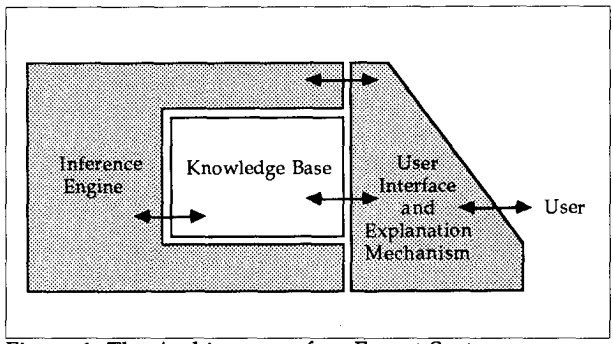

Figure 1. The Architecture of an Expert System.

3.1 Rule Based Knowledge Representation

In rule based ESs the main part of the knowledge is represented in rules of the form IF: <antecedent> THEN: <consequent>. The number of antecedents and consequents may vary, and the IF: part of a rule may combine several antecedents using booleans like AND or OR. During a run, the inference engine will try to match the available facts with the antecedents, and if a match is found the rule is said to 'fire', or to be 'executed'. The consequent of a rule can be a new fact that in turn can cause one or more rules to be executed, i.e. a chain of rules is executed. (See Table 1)

```
Fact: weather = heavy_clouds

Rule 1:
    IF:  weather = sunny
  THEN:  chance_of_rain = none

Rule 2:
    IF:  weather = heavy_clouds
  THEN:  chance_of_rain = high

New fact: chance_of_rain = high

Rule 3:
    IF:  chance_of_rain = high
  THEN:  take_a_raincoat = wise

New fact: take_a_raincoat = wise
```

Table 1 Rule Based Inference.

Reasoning of the ES can thus be traced backward from the last rule, up to the first rule that fired. Furthermore, the explanation mechanism can tell the user why a certain fact was not asserted, by showing which antecedent(s) did not match any of the facts known to the inference engine.

The facts used by the inference engine can either be derived from the systems knowledge or by asking the required information from the user.

One inference method is called 'forward chaining', because the direction of inference starts with a fact, for example "weather = heavy_clouds" (See Table 1), and checks which rule's antecedent matches this fact. When such a rule is found, then that rule's consequent is used for further matching.

Another approach is 'backward chaining', where the inference engine uses the consequent of a rule as a goal to be proven, for example, the fact "take_a_raincoat = wise" in Table 1. It will trace the rules backward to prove that the hypothesis "take_a_raincoat = wise" is true. Rule 3 can

prove the hypothesis, if the antecedent "chance_of_rain = high" can be proven, which in turn can be proven using Rule 2, if it can be established that the "weather = heavy_clouds".
Although this proved to be a very powerful and flexible way of processing knowledge, with an intuitive resemblance to human problem solving, it lacked a good organisation of the factual (or conceptual) knowledge which is treated as an unorganized collection of data. Additionally the organization of the rules (rule base) in a true rule based ES is also cumbersome, as all the rules are listed sequentially and tested one after the other until one is found that can be executed. Such a system can easily get cought in a never ending loop of asserting and retracting facts. Other methods of knowledge representation have been developed, such as the logic based language Prolog, or object oriented languages such as Smalltalk. However each of these representational methods has its own advantages and drawbacks, and none of them has a satisfactory answer to problems like the representation and use of uncertainty (i.e. uncertainty management) and handling of conflicting evidence (i.e. conflict resolution).

3.2 Semantic Nets

Semantic Nets offer a model for the organization of conceptual data through the use of meaningful connections. Each connection has a direction and a label identifying the nature of the relation. For example, the concept 'car' can have an 'is_a' relation to 'vehicle', i.e., 'car' is an instance of 'vehicle'. An other link from 'car' is called 'has_part' and points to the concept 'engine'.
A semantic net is a useful technique when it is important that the ES is able to find associations with its available knowledge. However, semantic nets are less useful for the representation of rule-like knowledge as used in rule based ESs.

3.3 Frame Based Knowledge Representation

CAR Frame	
speed:	45 km/h
fuel_level:	20 l
(if fuel_level = 0	
then speed = 0 km/h)	
gear_level:	if: speed > 10
	and speed ≤ 50
	then: gear_level = 2
	if: speed > 50
	and speed ≤ 80
	then: gear_level = 3

Table 2 Frame Based Representation.

A representation method incorporating both former methods uses frames to depict knowledge. A frame[15] is a kind of fill-in form with labels and slots (See Table 2). A label is a property of a concept, while the slot contains its present value. Sometimes special functions are attached to these slots called 'demons'. In Table 2 the slot "fuel_level" contains a value, but also has a demon attached to make sure that the car's speed will be 0 km/h when there is no fuel, i.e. demons can control the relation between two or more slots. Instead of a value, a slot could also hold a rule or function to determine its own value. Furthermore, frames can be used to hierarchically organize knowledge. For example, in

Table 2 the CAR frame could represent knowledge about all cars, while another frame could represent a specific car. This child frame can then inherit attributes from its parent and have additional attributes specific only to itself.

3.4 Hybrid Systems and Expert System Shells

The above mentioned methods are only three main stream representational techniques. Other methods exist[16], but most present day ESs are hybrid systems combining the strengths of rules, frames and semantic nets. Furthermore, many of today's ESs are developed using so called ES shells. An ES shell is a computer program with the aforementioned inference engine and explanation mechanism, but without any knowledge. However, they do incorporate a specific representational structure for the knowledge base. Thus, an ES shell is only suitable for task domains where the knowledge used can be mapped onto the representation technique used in the shell.

Many ES shells, or the ideas behind them, are related to the first medical ES called MYCIN, developed at the Stanford University[17]. After MYCIN was developed, it was essentially stripped of its knowledge, leaving a shell called E(ssential)MYCIN. After EMYCIN many other shells have been developed based on the same methodology. The increasing computational power of mini- and microcomputers has fostered this development even further.

4 Psychiatric Expert Systems

Medicine was one of the first domains for which ESs were developed, and since MYCIN many more have been developed. However, psychiatry has never been a major topic of interest for ES developers. During the last two decades four ESs were developed on topics within psychiatry: HEADMED[18], BLUE-BOX[19], an unnamed psychiatric version of Pathfinder[20], and Methuselah[21].

4.1 HEADMED

The ES HEADMED was developed with the ES shell EMYCIN and its task was to advise clinicians about psychopharmacology, and to function as a tutorial. As Brooks and Heiser stated in their article, one of the aims of building HEADMED was to see whether a rule-based control structure, derived from MYCIN, could be transferred to a new knowledge domain[22]. Although the project was abandoned before it was completed[23], it did point out that one of the main problems for medical ESs is to obtain large amounts of information from the user that are needed for diagnosis and treatment recommendation.

4.2 BLUE-BOX

BLUE-BOX was developed to advise physicians on the selection of an appropriate treatment for patients suffering from a depression. Like HEADMED, BLUE-BOX was implemented in EMYCIN. It used information about the patient's symptoms and medical history to generate a treatment recommendation, i.e., the kind of drug, dosage, administration and side effects. The aim of developing BLUE-BOX was first of all to investigate the knowledge engineering problem. Like HEADMED, BLUE-BOX was never submitted to validation tests.

4.3 The Psychiatric version of Pathfinder

One of the first psychiatric ESs that was submitted to a validation test was developed by Feinberg and Lindsay using a new version of Pathfinder[24], which was originally an ES for the interpretation of lymph node tissue examination. This Pathfinder version was designed to distinguish endogenous and nonendogenous depression. The ES was tested in diagnosing 51 patients. The results showed that, of the 42 patients it was able to diagnose, 76% received a correct classification. Unfortunately the authors provided very scanty information about the architecture, the implemented knowledge and the source of the standard diagnoses with which the ES was compared.

4.4 Methuselah

This ES focuses on geriatric diagnoses, more specifically on dementia, primary degenerative dementia, multiinfarct dementia and major depression. The ES gathered its information on-line during a consultation session using seminatural language. Methusaleh was evaluated using the medical charts of 45 geriatric patients, and by comparing the ES's diagnoses with "independent clinical judgement"[21]. The results showed that in 38 cases at least one of the diagnoses generated by Methuselah was in agreement with the clinical diagnoses. However, no detailed information was presented about those diagnoses, nor about the clinicians who functioned as a source for independent judgments.

5 Discussion

There are several lessons to be learned from the psychiatric ESs developed during the last 2 decades.

Firstly, as Morelli et al. already stated several years ago: "the development and utilization of expert systems in the mental health field has lagged behind their use in other domains"[25]. Unfortunately this situation has not changed since. In the same article, Morelli et al. suggest that this slow penetration of ES in psychiatry is due to the limitations in the nosology of mental disorders. They further note that, due to the lack of hard physiological criteria, psychiatry bases much of its diagnoses on behavioral criteria which they consider far more difficult to implement in an ES than, for example, the underlying pathogenesis. The use of such descriptive criteria was also

regarded as an obstacle by Werner[21] who described psychiatry, or more specifically geriatric psychiatry, as an ill structured task domain. However, we disagree with their view that ESs for psychiatric diagnostics are too difficult and that it would be more productive to develop ESs for patient treatment and management. Once again we may quote Morelli et al. that "important breakthroughs in this area will likely depend upon the adoption of a more widely accepted and more detailed descriptive basis for diagnostic decisions"[25] In other words, part of the problem lies in the standardization and exactness of the diagnostic criteria, and not in its descriptive nature. Apart from the fact that a good understanding of diagnostics is required to be able to formulate appropriate treatment and management recommendations, none of the authors discusses the possibilities ESs offer in improving diagnostic criteria and consensus. ESs can form an important means for testing and comparing formalized diagnostic criteria. For example, if a set of criteria is changed in favor of another one which is supposed to make a better decision possible, they could both be tested in an ES to see whether the result is what was expected.

Secondly, another important lesson to be learned from the aforementioned projects is that the destiny of ESs that were mainly developed to investigate certain AI related problems are doomed to end without the prospect of producing a field application. After reporting the problems and solutions the developers (or their institution) loose interest in the domain and focus on a different (AI related) problem. For an ES to become useful, it seems that the focus of attention should be the problem domain, and not some specific AI related problem such as a new search algorithm. Furthermore, thorough evaluation studies are required to establish the expertise of the ES and to show both the weak and the strong points of the system.

Thirdly, as reported by several developers, it is deemed of utmost importance to increase the availability of ESs. First of all, this means that ESs should be available at the clinicians desktop. In this respect it clearly does not suffice to have a terminal on a desk connected to a minicomputer or mainframe, i.e., time-shared computers, because such systems often show "unpredictable fluctuations of response time"[21].

Fourthly, ESs should be built such that they offer the possibility to decrease the amount of time needed from the clinician to enter the desired information. One solution would be to integrate the system with a data base. However, a more obvious improvement would be to provide the user with fill-in forms and a list of the possible answers to select from. As with databases, these forms could be filled in by the clinician or a trained medical assistant after the medical examination. All previously reviewed ESs use semi natural language data input, which is very time consuming and certainly not error proof.

Fifthly, as stated before, neuropsychiatry is a medical field which uses information from many different disciplines, like neurology and psychology. Consequently, a neuropsychiatric ES should include relevant expertise from related disciplines, for example, neurology and psychology. Of the 4 previously discussed ESs, only Methuselah comes close to that prerequisite, even though its developer (Werner[21]) places it in the domain of geriatric psychiatry. Such a discipline centered approach can be found in most other ESs. For example, the few neurologic ESs, such as NEUREX[26] on the domain of neurological localization and NEUROLOGIST-I[27] for the

localization of lesions within the central nervous system, all operate within the boundaries of neurology.

6 Concluding remarks

In order to build a neuropsychiatric ES while avoiding the problems previously mentioned, the following aims should be pursued:

1. The ES should be built to aid clinicians in making a diagnosis in a specific field of neuropsychiatry, and not (at least not solely) as a means to test some specific AI related problem as a goal in itself.
2. The ES should encompass knowledge from different fields, preferably with the aid of one expert that is proficient in the related fields. Thus it can be prevented that conflicting expertise, for example, due to different schools of thought, will impede the development of the ES at an early stage. A confrontation between the expertise of the ES and other experts will be more fruitful after the ES has shown what its implemented knowledge can accomplish.
3. Although the aim of the ES should be to reach a proficiency equal or better than the average clinicians working in the specified field, it should not be developed with the aim to provide an exact model of the human expert problem solving behaviour. Any system that could presently be developed would lack many of the capabilities of the human expert. Instead, the aim should be to extract a protocol from the human expert that is useful for implementation and delivers the desired performance.
4. The ES should be easily available for clinicians, preferably in the same way that other instruments are available, i.e., as a personal tool. Thus the ES should be developed with the aim to use it on a personal computer with a financially attractive configuration.
5. There is no need for an ES to be able to communicate with its user through a seminatural language interface. Such interfaces still lack the sofistication actually needed and place unnecessary high demands on the available hardware and software. A much simpler and effective approach would be to offer the user standard fill-in forms with lists to choose the appropriate values. This is comparable with the interface found in many database applications.
6. The ES should be developed such that it can easily be expanded, augmented, repaired, changed, and maintained. For that purpose a modular system seems the best approach.
7. The ES should be tested in phases. After informal tests during the development, a formal test should be set up using other cases than those used during development. The formal test should have a design that is comparable with that used to assess the agreement between human raters.
   Furthermore, the level of performance required from the system must be decided on before the experiments starts. We would like to stress the importance of a thorough and rigorous evaluation of the ES. As the ES is developed to be given a responsible role in a difficult medical field, it is only natural that its performance is

tested in the same rigorous manner as humans are before they are allowed to use their skills in daily routine.
Because the evaluation of ESs deserves a more elaborated discussion than is possible here, the next chapter will be devoted to this subject.

Although these aims will not guarantee that a useful ES will emerge, it will make it likely that previous problems will be avoided and that -in contrast to many ESs developed sofar- the resulting ES will go beyond the stage of a research or demonstration prototype.

7 References

1. Yudofsky SC, Hales RE. The reemergence of neuropsychiatry: definition and direction. J Neuropsychiatry, 1989, 1; 1: 1-6
2. Lighthill J. Artificial Intelligence: a general survey. In: Flowers BH (Ed.) Artificial Intelligence: a Paper Symposium. London Science Research Council, 1973, pp 1-21.
3. Department of Industry. A programme for advanced information technology: the report of the Alvey Committee. London: Her Majesty's Stationery Office, 1982.
4. Lam, M. Lighthill 17 years on. The Knowledge Engineering Review, 1990, 5; 4: 265-276.
5. Michie D. Lighthill 17 years on: end of a shotgun divorce. The Knowledge Engineering Review, 1990, 5; 4: 277-284.
6. Jackson P. Reply to Lam. The Knowledge Engineering Review, 1990, 5; 4: 285.
7. Wilks Y. AI and Anglo-Saxon Attitudes: a response to Martin Lam. The Knowledge Engineering Review, 1990, 5; 4: 285-288.
8. McCarthy J. Lessons from the Lighthill Flap. The Knowledge Engineering Review, 1990, 5; 4: 288-290.
9. Sparck Jones K. Re Lam: Lighthill 17 years on. The Knowledge Engineering Review, 1990, 5; 4: 290.
10. Lam M. A Rejoinder. The Knowledge Engineering Review, 1990, 5; 4: 290-293.
11. Waterman D. A Guide to Expert Systems. Addison-Wesley, Reading, Massachusetts, 1986.
12. Hayes-Roth F, Waterman DA, Lenat D. (Eds.) Building Expert Systems. Addison-Wesley, Reading, MA, 1983.
13. Churchland PM, Smith Churchland P. Could a Machine Think? Scientific American. January 1990: 26-31.
14. Searle JR. Is the Brain's Mind a Computer Program? Scientific American. January 1990: 20-25.

15. Minsky M. A framework for representing knowledge. In: P.H. Winston (Ed.) The Psychology of Computer Vision. New York, MacGraw-Hill, 1975, pp. 211-277.
16. Tanimoto SL, The Elements of Artificial Intelligence. Computer Science Press, Rockville, Maryland, 1987, p. 130.
17. Buchanan B, Shortliffe EH. (Eds.) Rule-Based Expert Systems. The MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley, New York, 1984.
18. Heiser JF, Brooks RE. Design considerations for a clinical pharmacology advisor. Proceedings of the 2nd Annual Conference on Computer Applications in Medical Care, Washington, November 1978, pp. 278-285.
19. Mulsant B, Servan-Schreiber D. Knowledge engineering: a daily activity on a hospital ward. Computers and Biomedical Research, 1984; 17: 71-91.
20. Feinberg M, Lindsay RK. Expert Systems in Psychiatry. Psychopharmacology Bulletin. 1986, 22; 1: 311-316.
21. Werner G. Methuselah: An Expert System for Diagnosis in Geriatric Psychiatry. Computers and Biomedical Research, 1987; 20: 477-488.
22. Brooks RE, Heiser JF. Transferability of a rule-based control structure to a new knowledge domain. Proceeding of the 3rd Annual Conference on Computer Applications in Medical Care. Washington, 1979, pp. 56-63.
23. Servan-Schreiber D. Artificial Intelligence and Psychiatry. The Journal of Nervous and Mental Disease. 1986, 174; 4: 191-202.
24. Horvitz EJ, Heckerman DE, Nathwani BN, Fagan LM. Diagnostic strategies in the hypothesis-directed PATHFINDER system. Proceedings of the 1st Conference on Artificial Intelligence Applications. IEEE Computer Society, 1984, pp. 630-636.
25. Morelli RA, Bronzino JD, Goethe JW. Expert Systems in Psychiatry. Journal of Medical Systems. 1987, 11; 2/3: 157-168.
26. Reggia J. A production rule system for neurological localization. Proceedings of the Second Annual Symposium on Computer Applications in Medical Care. Piscataway, New Jersey, IEEE Press, November 1978, pp 254-260.
27. Xiang Z, Srihari SN, Shapiro SC, Chutkow JG. Analogical and propositional representations of structure in neurological diagnosis. Proceedings of the First Conference on Artificial Intelligence Applications, IEEE Computer Society, December 1984.

II EVALUATION OF MEDICAL EXPERT SYSTEMS

1 INTRODUCTION

Although expert systems (ESs) are being developed during more than 2 decades, the evaluation of such systems has received much less attention by developers than topics such as inference techniques, knowledge acquisition and knowledge representation. In 1987, Lundsgaarde reported that only about 10% of the ESs developed until then were subjected to evaluation tests.[1] As stated in the previous chapter, this is probably due to the specific interest of ES developers in AI topics. Furthermore, Wyatt and Spiegelhalter[2] suggest that the sparsity of evaluation studies can also be explained by the lack of paradigms for ES evaluation, making developers uncertain about what to test, and which methods to employ. Although this problem applies to the evaluation of any ES, it is especially true for ESs in the domain of medicine, where it can be very difficult to assess the truth of a judgment. Although there is general agreement that medical ESs, like any other new technology in medical care, should be formally tested before being released for practical medical use, none of the existing evaluation methods seem to fit the problem of ES evaluation very well. For example, as noted by Wasson et al.[3], Wyatt[4], and Wyatt and Spiegelhalter[2], there is a close analogy between the laboratory testing phase of ESs and the safety and dose-finding phases of drug development. However, Wyatt and Spiegelhalter[5] also point out that there are important differences, for example, the difficulty to obtain gold standards for comparison. Another analogy exists between the evaluation of conventional software and ESs.[6] However, this analogy faces the problem that ES decisions are not only computational difficult -if not impossible- to prove, but also that an ES's knowledge is based on fallible human expertise. Furthermore, even though ESs are comparable to conventional software in many respects, there is the danger that they will not be treated as such by the potential users, as the phrases Artificial Intelligence and Expert System suggest that such systems possess certain human capabilities. As Cole[7] notes in an article on tort liability for such systems, this can lead to an over reliance on the ES, so that the user will neglect his own judgement, intuition, and abstract reasoning. In the remainder of this chapter we will discuss what the phases and topics of ES evaluation are, the evaluation methods that can be used, and by which criteria ESs are to be judged.

2 STAGES AND TOPICS OF EXPERT SYSTEM EVALUATION

In contrast to O'Leary et al.[8], who regard evaluation as an experimental field stage next to validation and verification, we adhere to the interpretation of O'Keefe et al.[9] (p.81) that evaluation is "a broader area seeking to assess an expert system's overall value". In this view, ES validation, i.e., assessing whether the system does what it is supposed to do, and knowledge base verification, i.e., assessing whether the system's knowledge is correct, are both part of ES evaluation.

What the topics of ES evaluation are is very much dependent on the proposed structure of the life cycle of ES evaluation, i.e. from the very first ideas up to the final use in medical practice. For example, the five-task framework proposed by Buchanan et al.[10] includes only one task (the final) where testing is explicitly mentioned. The same applies for the spiral life cycle model proposed by Boehm[11] that ends the moment the ES becomes operational. Wyatt[4], on the other hand, proposed a cycle resembling the drug development cycle, which includes several test phases, up to the final stage where an accredited decision support system is available on the market. As exemplified by these three life cycle proposals, there are three principal phases in the ES life cycle: the initial development, laboratory, and field testing phase. This is reflected in the three-stage evaluation method proposed by Wyatt and Spiegelhalter[2], a modification of Wyatt's drug development paradigm, consisting of a definition or prototype development phase, a laboratory testing phase, and a field testing phase. As we believe that the proposed modification of the drug paradigm by Wyatt and Spiegelhalter is the most useful and complete at present, we will use it as a starting point to combine their ideas with other topics and evaluation methods. In our discussion we will focus on Stage II, the laboratory evaluation, because many ESs have either not reached this stage or have been tested unsatisfactorily. The other two stages will mainly be discussed to suggest which types of tests do not belong at Stage II.

2.1 Stage I: Prototype development and evaluation

Although Wyatt and Spiegelhalter did not mention it in their article, the main topic of evaluation during Stage I, after defining the ES's domain, is the verification of the knowledge base. Verification deals with checking the knowledge base for correctness, consistency, and completeness. Several methods have been proposed for knowledge base verification. For example, the decision tables proposed by Suwa et al.[12], and the categorical inferences by Quinlan.[13] However, most of these verification methods are designed to verify only the parameter values of production rules. Recently, Liu et al.[14] proposed the use of numerical Petri nets, based on Zisman's[15] earlier suggestion to use Petri nets to model relationships in production systems. Especially the proposal by Liu et al. to use numerical Petri nets (NPNs) seems very promising, as the use of these nets can be automated relatively easy. Unfortunately, it is too early to expect experience reports about the use of NPNs. The main problem with verification methods is that they are meant for a specific knowledge representation design and need (substantial) modification to be useful for systems with a hybrid knowledge representation. The use of demons (see also I 3.3), for example, is not represented in the NPNs of Liu et al. This is probably one of the reasons why knowledge base verification usually receives even less attention than knowledge base validation. Fieschi[16] calls the methods for this part of ES evaluation 'static methods', as they do not require running the system. However, there is an exception. An alternative is to employ 'face verification', i.e., tracing the rulebase during numerous trial runs, and examining the rules manually. Although this has the disadvantage that it is informal, laboriously and error prone, it is probably the method used most often.

The possibilities to connect the system with other existing information systems is

also an important subject, for example an interface to a patient database. However, the diversity of systems used and the lack of standardization severely impaires the possibilities for integration of ESs with other information systems, unless the ES was specifically designed for a certain configuration. However, care should be taken that the ES's architecture is open enough to build interfaces for communication with other information systems.

System speed is another important topic during this stage. From the beginning, developers should try to optimize the system such that long periods of user inactivity is avoided, and that the total time needed for a consultation is minimized.[17] The transparency of the human-computer interface (HCI) and its sensitivity to human errors are also important in that respect. A HCI is transparent when it is easy to use and when it requires minimal instruction for operation, i.e., the HCI should guide the user in the operation of the system. The sensitivity to human errors of the HCI can be reduced by offering good information, checking answers for typing mistakes and limiting the number of open-ended questions. In this respect, Boden's remark that "A little common sense and a little natural language ability might be a very dangerous thing"[18(p.224)] is very important. Both transparency and human error sensitivity can be tested through informal evaluations by asking clinicians who are unexperienced computer users to operate the system with minimal instruction.[19] The testing of the HCI should not be postponed until the system is completed, because then it is much more difficult to change the design. However, HCI evaluation should neither be limited to this first stage, the system's architecture should allow for some modification of the HCI at later stages. Unfortunately, the use of a commercial expert system shell can severely limit the degrees of freedom in HCI design, because some aspect of the HCI will be determined by the shell and not by the developer. In that case, testing the HCI also means that the suitability of the shell is tested. It should be stressed that, in the end, the ease of use of an ES will be as important as the correctness of the ES's conclusions.

An often forgotten topic of evaluation is the system's hardware requirement. Until the advent of the personal computer most systems were developed using mini- or mainframe computers. And even today, with powerful personal computers readily available, there are ES developers who think, overhastily, that systems designed on and for a personal computer must be inferior to those designed and running on a mainframe or powerful workstation. However, the growth of the power of present personal computers has virtually closed the gap between personal computers and workstations. Due to the almost equivalent power of personal computers and workstations, the much less expensive personal computer has a clear advantage, as many institutions that could benefit from ES technology do not have the financial resources to invest in large computers. In this respect, the maintenance of the final system is of importance. The knowledge base should be designed such that it can easily be updated and modified. This can be tested during the subsequent evaluation stages when the knowledge base is extended and modified. However, it should be remembered that knowledge base maintenance is probably not done by the initial developers, but by some central distributing company or organization.

Although not explicitly mentioned by any of the authors previously referred to,

there is general consensus that if developers consider that the system performed well on the informal tests during the first phase, and there is reason to believe that the knowledge base is correct, consistent, and complete, then it is time for a formal evaluation of the system.

2.2 Stage II: Formal laboratory evaluation

The main topic of Stage II is validation of the ES, i.e. the quality or assumed expertise of the system. What should be validated is partly determined by the goal of the system, i.e. whether it was intended to simulate or model the human expert. As Fieschi[16] notes, systems which *simulate* the reasoning of an expert should be judged on their capacity to reach the same conclusions, while systems that *model* an expert should additionally be judged on their inference method and reasoning chain. As most ES developers would not claim that their system is a (realistic or complete) model of the human expert, we will limit our validation discussion to the systems that simulate human expertise. However, we would like to stress that there is a large grey area between simulating and modelling human behaviour. Furthermore, before we elaborate on how a simulation system can be validated, we would like to point out the possibility to perform this validation phase in two steps.

2.2.1 Internal and External Validation

Although an ES could be built with the purpose to be used by a specific institution, i.e. a custom designed system, the intention of most developers will be to deliver a system that can and will be used by other institutions. It is therefore of the utmost importance that the criteria used and the conclusions reached by the system will be acceptable outside the place of development. One way to accomplish this is to use knowledge that is widely agreed upon, i.e., consensus knowledge, such as the reports by consensus meetings.[20] Additionally, it will be necessary to validate the ES by comparing it with experts outside the institution of development. However, as Bachant and McDermott[21] note, expectations should not be too high during the first few years of development. Consequently, immediately comparing a newly developed ES with experts outside the institution of development should be avoided. A solution to avoid this problem is to validate the system in two steps, where the first step would be to compare the system with the expert(s) who served as domain expert(s) (the *internal* comparison), and the second step to compare the system with experts from outside the development institution (the *external* comparison). If the system fails to meet the expectations of the first test, then there is no need to perform the -usually more laborious and expensive- second test, while the chances of success will be higher if the first step was completed successfully. Furthermore, performing both an internal and an external comparision is useful to check whether the existence of schools of thought is causing an ES to fail.

2.2.2 The Comparison Data

How such a validation must be performed is still a much debated topic. As evidenced by our previous discussion on the use of human experts for comparison, we do not consider (at present) equivalence tests between two systems on the same problem domain a serious method for validation. At present (medical) ESs are still too immature to be used as serious validation candidates. Medical ESs should first of all show that they can accomplish what they were designed for: perform better than the average human clinician who has to deal with the same problem domain.

The most common method of comparison is to use historic data, i.e., patient records, and feed these into the system. The advantage of this method is that the data are readily available and that the test can be performed quickly. This is important to control the cost of validation.[9]

A second possibility is to set up a prospective experiment, using the data from newly entered patients that are to be judged by the ES and the human expert at the same time. However, this method is very time consuming and expensive, especially when the type of patient required is relatively rare. An additional problem -which is not restricted to prospective studies- is the difficulty to assess whether the collected data are complete and correct. Because not all clinical findings can be objectively verified, for example through autopsy, there is no definite solution to this problem.

A third option is to let experts synthesize cases along with the appropriate diagnosis. However, as O'Keefe[9(p.83)] notes, "this is problematic because any set of synthesized cases is unlikely to represent a well-stratified sample". Apart from the questionable quality of such cases, this method is also time consuming and expensive, because it demands much time and effort from the medical experts.

Although the second option is probably methodologically the best because it ensures a natural selection of cases that have to be dealt with by the human clinician, for this stage (II) of evaluaton, there are important advantages in favor of the first method, i.e., the use of historic data. By using historic data a time consuming and expensive method in an early phase of development can be avoided. Furthermore, these data can more easily be checked for completeness, and -although not with absolute certainty- for their correctness before they are are used. The prospective approach would be more suitable for stage III, when there is more certainty that the system performs well enough to be tested outside laboratory settings.

2.2.3 The Number of Cases and Case Selection

The next question is how many cases, and which cases should be used in the comparisons, i.e., the internal and external comparisons.

As to which cases should be used for comparison, it will be clear that only those cases can be used that did not take part in any of the previous phases of system development. Two options for case selection have been proposed: selected and randomized patient cases.[9,22] As each method has its pros and cons, we suggest to make a practical choice between the two options. For the internal comparison a stratified, for example by complexity or diagnosis, randomized case selection seems the most appropriate, because the cases to be used will probably be derived from the

same institution where the system was developed. A stratified randomization will avoid the problem that the domain expert (the human expert who cooperated in the acquisition of the knowledge for the system) will have to choose the cases for comparison. However, to test the sensitivity and specificity of the system, some selection must be allowed to assure that there are both cases with a positive diagnosis and cases with a negative diagnosis. Only when there are not enough cases available the option of selected cases should be chosen.
For the external comparison, the stratified randomization would still be the best option. However, here there is an additional problem. If the system is to be compared with experts from many locations, then it will practically be very difficult -if not impossible- to have all those clinicians diagnose many cases. In such circumstances, only a small number of cases can be used, which eliminates the use of randomized cases. A methodologically better approach in that situation is to use a carefully selected sample of cases. This can affect our confidence in the test results, but as O'Keefe et al.[9(p.83)] note "the law of large numbers simply does not apply here. The issue is not the *number* of test cases, it is the *coverage* of the test cases". Additionally, the problem of the number of test cases to be used is not restricted to the selected sample cases, but also to the randomized sample cases. As Van Bemmel[22(p.193)] notes "the number of degrees of freedom is so large that (...) one would need many millions of cases to test a sizable expert system". The number of available patient records alone, makes this impossible.

2.2.4 Stress Testing

Another topic of validation is 'stress testing', i.e. testing whether the system produces errors to certain combinations of data.[6] Although this is a very important topic, we believe that the possibilities for stress testing are very limited during this stage of system evaluation. To find such errors either synthesized cases based on hypotheses about possible errors are required , which is unlikely to succeed due to the large number of possible errors, or all possible combinations of parameters have to be present in the cases, which is impossible. The best option at this stage would be to select cases that are considered difficult by most experts, but then again this is not quite what is meant by stress testing. The third system evaluation stage, when the system is tested in field situations, is probably more suitable for stress testing, because at that point more, and more realistic, unforeseen cases will have to be judged by the system.

2.2.5 The Golden Standard

For both internal and external comparison there is the problem of the golden standard, or: how should *any* diagnosis made be judged for correctness? For the internal comparison we would like to simplify the problem by stating that the domain expert is to be regarded as the golden standard. This can be justified by the fact that the domain expert has also served as the model after which the ES was build. Consequently, the ES is supposed to produce results comparable to its model.

Incidentally, this approach can even be used when there is more than one domain expert, except that the test may then produce more than one result.

For the external comparison the use of an existing or constructed golden standard is imperative, since we are supposed to compare the system with many clinicians who's diagnoses can differ in varying degrees from the domain expert. Probably the best solution is to have cases that contain post mortem data that confirm the clinical diagnoses. However, such cases are difficult to obtain. Furthermore, as stated before, not all clinical diagnoses can -or even have to- be confirmed by post mortem data. In such cases some authority in the domain must serve as a golden standard. We suggest that a committee be formed consisting of experts who are internationally considered as experts in the domain. When the expertise of several disciplines is involved, then the choice of experts should cover these disciplines. For example, in dementia diagnostics it is recognized that the opinion of psychiatrists, neurologists and psychologists should be taken into account.[23,24] Furthermore, this expert committee should be able to form an opinion through joint discussion to obtain an optimal result. Such a committee provides an upper limit for system performance. To obtain a lower limit, other clinicans with which to compare the system should be selected at random, or at the least form a rough average sample. Thus, the upper and lower limit yield the *acceptable performance range* mentioned by O'Keefe et al.[9], where the upper limit represents the desired system performance, while the lower limit represents the threshold for rejection or acceptance of the system. To be more specific, if the system performs at par with the *average* clinician, then the system should be rejected, because it was supposed to perform at expert level. However, it should be noted that unless an ES performs at par with the best clinician, "...it can be expected to have a negative influence on some of them", because the clinician could be deceived by the suggestions of the system.[16(p. 98)]

2.2.6 Obtaining and Judging the Diagnoses

To acquire the diagnoses from the clinicians for comparison with the ES, the clinicians and the ES should have the same data available in a structured way. This could lead to a 'checklist effect', i.e. a positive effect on decision making because more complete and structured data than normally is the case are available. Furthermore, because the quality and quantity of the information given to the clinicians in this setup, is likely to be superior to the information that is usually available to them, it is plausible to assume that this makes the test more demanding of the ES, because the clinicians are better informed than usual.[2] Furthermore, the clinicians should be allowed to state their diagnoses in a way that does not restrict the terminology they use. This is especially important when clinicians from several disciplines participate in the study.

To judge the diagnoses from both the ES and the clinicians, O'Keefe et al.[9] suggest to

use a kind of Turing test.* In this test the diagnoses of both the clinicians and the ES are being judged by other clinicians, while being blinded for the source of the diagnoses. Although blinded experiments are common practice for the drug evaluation paradigm, in this case it has a serious flaw, because it assumes that the answers of the clinicians and the ES are linguistically indistinguishable. At least for the present, this assumption must be considered unrealistic, unless the diagnoses from both sources were transformed such to make identification impossible. However, it is unlikely that such a liguistic tranformation can be done reliably. A better approach would be to categorize all diagnoses and use some statistical measurement of agreement, for example, Cohen's Kappa[25]. However, in some cases, for example in the external validation with many clinicians and only few cases, the judgement will probably be done qualitatively with only some descriptive statistics as proportions and averages.

2.3 Stage III: Field evaluation

After an ES proves to be successful during the previous laboratory tests, the system will be ready to be tested in field circumstances. This stage is probably the most laborious and expensive due to the number of participants involved and the time needed for the evaluation. The nature of a field evaluation study is that a quasi experiment is performed in a situation that includes independent environmental variables (confounding variables) that were not present in the laboratory tests, such as attitude towards computers or time pressure. Furthermore, unwanted experimental effects can occur in a field study, for example, the problem of (observer) interference with the observed clinicians: the so called 'Hawthorn effect'[26]. Another effect is more specific to the use of ESs in the experiment, i.e., the carry-over-effect. This is the educational effect that the ES can have on the clinicians who use the system. Contrary to the drug evaluation paradigm it is impossible to use a placebo. However, as Wyatt and Spiegelhalter[2] suggest this effect can be partly controlled for by alternating periods where the clinicians work with and without the ES, and by measuring the clinicians performance prior to the experiment. For these experimental problems educational research methods can prove very helpful.

As mentioned before, the most suitable experiment for this stage is a prospective study. However, before such a lengthy study is started an other retrospective study is possible. Such a study would be performed by external clinicians, using their own patient records, preferably patients of whom the clinical data are supplemented, if relevant, with data from an autopsy. In this way several topics can be examined at the

* In order to show that a computer is capable of imitating human intelligence, A.M. Turing designed a test in which a subject behind a teletype console can communicate with a second teletype console in another room. The second teletype console can either be operated by a human or a computer. The subject's task is to find out whether he is communicating with a computer or a human being by asking questions. If the subject is unable to tell who he is communicating with while he has been communicating with a computer, then the computer passed the test.

same time: the validity of the system using patient records containing data that were collected in a different way, the stability of the system, and the effectiveness of the HCI. Furthermore, during this experiment the trial run-in period mentioned by Wyatt and Spiegelhalter[2] can be performed.

3 Concluding remarks

As we have stated before, we have focussed on the methodological issues of the Stage II evaluation, a stage that has been underrated by most developers. Consequently, most systems did not even come close to the field evaluation of Stage III (one exception is INTERNIST-I[27]). The depreciation or underestimated difficulty of ES evaluation is reflected in many of the reports about new ESs, which give only scanty details on, for example, the evaluation methods, the cases used, and the people involved. Therefore, when the goal of ES development is to produce an ES for practical use, future reports on ES development should pay at least as much attention -and maybe even more- to evaluation studies as to knowledge acquisition or system architecture. Like medical drugs, or psychological tests, ES evaluation is essential in order to prove their usefulness. In this respect, a policy change of the publication journals involved might be required, because ES evaluation experiments will typically be interdisciplinary studies, involving both knowledge engineers and domain experts.

Given the number of tests and the people involved, we can also conclude that the amount of time to be invested in ES evaluation is gravely underestimated. Previously knowledge acquisition was seen as the bottleneck in ES development. Now it seems that there is another bottleneck, one that is probably even more costly than the knowledge acquisition phase. The European project for Common Standards for Quantitative Electrocardiography provides good impression about how expensive and complicated the evaluation of information systems is.[28] Therefore, it seems obvious that Stage III evaluations cannot be performed without the financial support of institutions that can take over the financial and commercial responsibility of the system after it has proven its value in the laboratory. Given this commercial interest, and the multitude of difficulties of ES evaluation, it is important that an independent technology assessment body is formed to judge (among others) decision support systems, since there is no principal difference between the evaluation of ESs, new drugs, or even the examination of medical students.[2,29]

4 REFERENCES

1. Lundsgaarde HP. Evaluating medical expert systems. Social Science and Medicine. 1987, 24: 805-819.
2. Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Medical Informatics. 1990, 15; 3: 205-217.
3. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: application and methodological standards. New England Journal of Medicine, 1985; 313: 793-799.
4. Wyatt J. The evaluation of clinical decision support systems: a discussion of the methodology used in the ACORN project. Lecture Notes in Medical Informatics, Proc. AIME '87, Marseille, 1987; 33: 15-24.
5. Wyatt J, Spiegelhalter D. Preface. Medical Informatics. 1990, 15; 3:183-184.
6. Smeets RPAM, Talmon JL, O'Moore R. General methodology and problems in assessment of decision support systems. In: O'Moore R, Bengtsson S, Bryant JR, Bryden JS (Eds.), Lecture Notes in Medical Informatics, Medical Informatics Europe '90, Springer-Verlag, Berlin, 1990; 40: 225-230.
7. Cole GS. Tort Liability for Artificial Intelligence and Expert Systems. Computer/Law Journal, 1990, X; 2: 127-231
8. O'Leary T, Goul M, Moffit KE, Radwan AE. Validating Expert Systems. IEEE Expert, June 1990: 51-58.
9. O'Keefe RM, Balci O, Smith EP. Validating Expert System Performance. IEEE Expert, Winter 1987: 81-90.
10. Buchanan BG, et al. Constructing an Expert System. In: Hayes-Roth F, Waterman D, Lenat D. (Eds.) Building Expert Systems. Addison-Wesley, Reading, Mass., 1983: 127-167.
11. Boehm BW. A spiral model of software development and enhancement. ACM Software Engineering Notes, March, 1986.
12. Suwa M, Scott AC, Shortliffe EH. An approach to verifying completeness and consistency in a rule-based expert system. AI Magazine, 1982, 3; 3: 16-21.
13. Quinlan JR. Internal consistency in plausible reasoning systems. New Generation Computing. 1985; 3: 157-180.
14. Liu NK, Dillon T. An approach towards the verification of expert systems using numerical Petri nets. International Journal of Intelligent Systems. 1991; 6: 255-276.
15. Zisman MD. Use of production system for modeling asynchronous, concurrent process. In: Waterman DA (Ed.), Pattern-Directed Inference Systems. Academic Press, Inc., New York, 1978.
16. Fieschi M. Towards validation of expert systems as medical decision aids. Int J Biomed Comput, 1990; 26: 93-108.

17. Werner G. Methuselah: An Expert System for Diagnosis in Geriatric Psychiatry. Computers and Biomedical Research, 1987; 20: 477-488.
18. Boden M. Artificial Intelligence and "Natural Man". In: Bernold T, Albers G. (Eds.) Artificial Intelligence: towards practical applications. Elseviers Science Publishers B.V., North-Holland, 1985: 221-227.
19. Edmonds E. Human-computer interface evaluation: not user-friendliness but design for operation. Medical Informatics, 1990, 15; 3: 253-260.
20. McKhann G, Drachman D, Folstein M, et al. Clinical Diagnosis of Alzheimer's Disease: Report of the NINCDS-ADRDA Work Group under auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology, 1984; 34: 939-44
21. Bachant J, McDermott J. R1 Revisited: Four years in the trenches. AI Magazine, 1984, 5; 3: 21-32.
22. Van Bemmel JH. Formalization of Medical knowledge. Methods of Information in Medicine. 1986, 25; 3: 191-193.
23. Dutch Consensus Development Conference: Diagnosis of the Dementia Syndrome. Utrecht, C.B.O. National Organization for Quality Assurance in Hospitals, 1985.
24. Stuart CY, Hales RE: The Reemergence of Neuropsychiatry: Definition and Direction. Journal of Neuropsychiatry and Clinical Neurosciences. 1989, 1; 1: 1-6.
25. Cohen J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70; 4: 213-220.
26. Roethligsburger FJ, Dickson WJ. Management and the Worker. Harvard University Press, Cambridge, MA, 1939.
27. Miller RA, Pople Jr.HE, Myers JD. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. New England Journal of Medicine, 1982, 307; 8: 468-476.
28. Willems JL, Arnaud P, Van Bemmel JH, Degani R, Macfarlane PW, Zywietz Chr, (for the CSE Working Party). Common Standards for Quantitative Electrocardiography: Goals and Main Results. Methods of Information in Medicine, 1990, 29; 4: 263-271.
29. Dowie J. The evaluation of decision aids: the role of the decision owner. Med. Inform., 1990, 15; 3: 219-228.

III ACQUAINT*

1 INTRODUCTION

In 1987 the Department of Neuropsychology and Psychobiology of the University of Limburg started the development of an expert system with the name 'Evince', for the differential diagnosis of dementia.[1] The goal of developing Evince was to integrate and formalize international criteria and examination procedures, and to aid non-expert clinicians in diagnosing dementia and its possible causes.

As the expert system was also meant to be used in small health care institutions, the decision was made to use a micro computer based expert system shell. Furthermore, the shell had to be able to manage uncertainty, have a quasi natural language type of knowledge representation for easier understanding by the user, and a reasonably good explanation facility. After a careful comparison of the shells available early 1987, we decided to use the expert system shell Acquaint**.[2] In the remainder of this paper we will review the most important features of Acquaint, and report some of our experiences with this shell.

2 KNOWLEDGE REPRESENTATION

In common Artificial Intelligence terms, the knowledge of an expert system is divided into declarative and procedural knowledge, where the declarative knowledge is comprised of data and rules, while procedural knowledge stands for the procedural programming language routines. In Acquaint knowledge is represented in the form of frames of which there are 6 types, i.e., RuleBase, Context, Rule, Concept, Function, and Form frames. Each frame consists of a name for identification and one or more slots with an accompanying value facet. With the exception of the Function frames, these frames comprise the usual declarative knowledge. However, the Function frame encompasses both declarative and procedural knowledge. Therefore, in the remainder of this thesis we will label knowledge consisting of rules as Procedural Knowledge (PK) and knowledge about data -including the Function frame- as Conceptual Knowledge (CK).[3]

---

* Parts of this chapter were published in Plugge LA. Acquaint. Expert Systems, 1990, 7; 4: 243-245.

** Acquaint is an expert system development tool for the IBM compatible personal computer developed by Lithp Systems BV, Purmerend The Netherlands. The version reviewed in this chapter was release 3.25.

2.1 Procedural Knowledge

Procedural knowledge in Acquaint is hierarchically ordered in three frames. The highest ordered frame is the RuleBase frame, which defines the knowledge modules that are used in the application. A knowledge module is a file containing PK and/or CK. Knowledge modules are for the convenience of the developer to modularize the knowledge base, and have no special meaning to the user, or the system. The second frame type is the Context frame, which is used to structure the problem domain into a hierarchical tree of topics to be investigated. Each Context definition is a rule to group

```
(DefContext ALCOHOL-ABUSE
        IN:  PHYSIOLOGY
 COMMENT:    There is reason to investigate the
             possibility of alcohol abuse when one of
             the following statements is true:
        IF:  $OR  (PresentAlcoholUse / 7) >> 3
                  (FormerAlcoholUse / 7)  >> 3
                  AF = 'increased
                  OT = 'increased
                  PT = 'increased
                  Gamma-GT = 'increased)
```

Table 1. Definition of a Context.

subordinate rules and other Contexts, and has the function of a meta rule. Subordinated rules are simply enumerated under the Context to which they belong, while a subcontext has an IN-slot which mentions its supercontext. For example, the context 'ALCOHOL-ABUSE' in Table 1 is triggered when the patient regularly consumes more than 3 glasses of alcohol per day, or when one of the other parameters has the value 'increased'. The operator '>>' ('moving toward the higher bound') is one of the available quasi fuzzy operators, and uses the number 3 to determine an interval with a 20% higher and lower bound and fits the amount of alcohol the patient consumes within these bounds. The degree of fit is expressed as a certainty factor, in this case the higher the alcohol consumption the higher the fit and certainty factor. (Due to the small interval in this example the operator '>' would be sufficient in most cases. However, as the amount is calculated from the weekly use it thus able to differentiate between small deviations upward.) The main advantage of this contextual representation of PK is its modular structure, which makes it relatively easy to add or rearrange knowledge by moving an entire context, without the need to change individual rules. During the development of Evince this proved to be a very efficient way to implement new knowledge.

The third frame type is the Rule frame, the basic element of the PK. (See Table 2) The IF slot of the Rule frame allows the use of several rule properties. For example, the properties FORWARD or BACKWARD inform the system whether a rule can be used in forward or backward reasoning. Furthermore, the usual booleans, i.e., AND, OR,

(DefRule <Rule Name> IF: <Rule Property> <Boolean><condition> <condition> .. THEN: <action> ELSE: <action> GOAL: <hypothesis>)

Table 2. Definition of a Rule.

and NOT can be used, with additional directions on how to use the certainties of the antecedents of a rule. For example, the boolean OR* forces the inference engine to evaluate all premises after which the combined certainty of these premises is calculated (how this is done is discussed below). Unfortunately, Acquaint does not allow combinations of the AND and OR boolean within one antecedent part of a rule. Although this is not uncommon (for example, in NEXPERT only AND rules are allowed), it does impair readability to some extent because it requires the use of several rules. One additional quasi-boolean is MAYBE, which uses the certainty of its premises only when they are above a certainty of .2, otherwise they are ignored. This boolean has proven itself very useful to check the patient's data for atypical symptoms.

The conditions in the IF slot can test the value of concepts with a large set of operators, like: 'equal to' (=), 'smaller than' (<), 'TRUE', 'KNOWN', and quasi fuzzy operators, like: 'approximately equal to' (≈), and the previously discussed 'moving toward the higher bound' (>>).

The consequent part of rules in Acquaint offers more possibilities then can be found in most other shells, i.e., the usual THEN part, but also an ELSE and a GOAL part. The THEN and the ELSE slots usually consist of a certainty factor (CF), ranging between -100 and 100, with the name of a concept and the value to be asserted. However, a whole range of other actions can be performed, like message transmission, or window handling. In contrast to the THEN statement, the ELSE statement will only be activated when the IF statement is *not* true. When the truth of the IF statement is undetermined, i.e. a CF between -20 and 20, then neither the THEN nor the ELSE statement will be executed.

The GOAL part is a new feature of Acquaint's latest release and is best compared with the IF slot of a rule. After the IF statement of a rule is proved to be true, then the GOAL slot is treated as a second IF statement to be proved. However, in contrast to the IF-statement, the GOAL-statement orders the inference engine to examine the entire rulebase, and not just the present context. The concept mentioned in this slot is treated by the inference engine as new hypothesis to be proved, just as the usual IF statements. The GOAL part of a rule gives the inference engine the powerful capability to focus on a specific diagnosis given that the conditions in the IF statement are true.

The type of boolean in the antecedent part of a rule determines how the CF of a rule is calculated, but generally the AND boolean returns the lowest CF, while the OR boolean takes the highest CF. The CF in the consequent part of the rule denotes the maximum certainty with which the conclusion can de drawn. The algorithm to calculate CFs is the same as the one implemented in MYCIN (See Table 3).[4,5]

```
Given CF1 and CF2:
If both are positive:
   CF = CF1 + Cf2 (1- CF1)
If both are negative:
   CF = 0 - (|CF1| + |CF2| (1 - |CF1|))
If one is negative:
   CF = (CF1 + CF2) / (1 - MIN(|CF1|,|CF2|))
```

Table 3. Certainty Factor Calculation.

2.2 Conceptual Knowledge

There are three types of CK frames: the Concept, the Function, and the Form frame. The Concept is the most fundamental CK frame in Acquaint. The number of available slots in a Concept frame is fixed, i.e. you can not define your own slots. Nevertheless, there is still sufficient flexibility with regard to the use and implementation of the slots. Table 4 shows an example of a Concept definition in Evince. The EXPECT slot can handle several types of answers, like VALUE (choose one value), VALUES (choose one or more values), RANGE, and CONVERT (converts numbers in text and visa versa). Although Acquaint offers the possibility to define new EXPECT functions to proces answers, we have found the available functions sufficient for our purpose.

```
(DefConcept EEG
      FACT: Electroencephalogram
    PROMPT: Which deviations have been found in the EEG
   COMMENT: If the deviation you found is not included in the list, then choose 'un-
            known'.
    EXPECT: VALUES  none unknown diffuse_slowing focal_pathology
                    increased_Beta-activity epileptic_activity
       VAL: (<value><certainty>)
   DEFAULT: none
     CLASS: <parent concept>
      PROP: ASK)
```

Table 4. Definition of a Concept.

The PROP (property) slot informs the system whether it is allowed to ask the value (ASK), whether it has to recalculate the concept value (CALC), and whether the value of the concept is a CF or a value-CF pair. The CLASS slot handles the inheritance of slot values.

The use of Function frames is difficult to explain, because they can serve almost any kind of purpose. However, their main purpose is to serve concepts and rules, for example, as procedures for calculation or message transmission.

The Form frame serves as a fill-in form with dynamic links to the concepts used in

it, and is used to circumvent tedious queries by providing the user a form with default values which can be edited. The frame itself does not contain conceptual or procedural knowledge, it merely manages the display of the CK and tells the system whether or not the data should be used instantly or not. In Evince forms are used, for example, to enter administrative data and the results of laboratory blood tests.

2.3 Reasoning

The inference engine (IE) traces the context tree depth first, and compiles a priority list (an agenda) of the rules for each context. This agenda is setup for each context when it is visited. The priority of a rule is based on several criteria, for example, the priority of an OR-rule is higher than an AND-rule. The rules on the agenda are then traced by forward chaining, until all rules have been tried. The rules that did not fire are then evaluated by backward chaining. Only during backward chaining is the IE allowed to ask questions about concepts. This switching between forward and backward chaining can be suppressed by setting a global reasoning switch to backward, i.e., only backward reasoning is allowed, or forward. Additionally, local forward and backward reasoning switches can be put in individual rules. When a rule fires, the total amount of certainty of the IF part of that rule is taken as a percentage of the certainty in the rule's conclusion. For example, if a rule fires with 80% certainty, and a concept in the THEN part has a CF of 90, then that conclusion will receive a CF of 72. The IE raises or lowers the certainty of a concept when an other rule has the same concept in its THEN part, by combining the previous CF with the new CF using the formulas from Table 3.

Reasoning continues until the last context within the root context has been traced. A context will be checked only once. To check a context for a second time, the IE must be explicitly instructed to do so. Apart from hierachically ordered contexts, Acquaint offers the possibility to use isolated contexts, i.e. contexts that are no part of the root context. This feature makes it possible to have two separate knowledge bases within one system.

3 Interface

Due to the character based user interface Acquaint appears rather dull. Nevertheless, although Acquaint is visually less impressive than competing shells with graphic user interfaces, it has the advantage that it does not overwhelm the user with a labyrinth of windows.

3.1 Development Interface

Acquaint does not offer a very friendly development interface. The knowledge base has to be edited with a Lisp oriented editor, or some other straight ASCII editor.* However, the explain and tracing options in Acquaint are good and the compiler has a sensitive debugger that provides informative error messages, which makes it relatively easy to trace faults. Acquaint contains a large number example knowledge bases which explain many of Acquaint's features. Although it is not required to be a Lisp programmer to implement these examples, we have found that at least some basic knowledge of Lisp is certainly of advantage.

3.2 User Interface

Acquaint provides the user with several menu driven options to query the system during a consultation. For example, the user can ask why a question was asked, how values were inferred, and which concepts are known. A trace option allows the user to get a detailed picture of the reasoning chain of the IE, i.e., which rules or contexts fired or failed and in what sequence. The communication is done via specific windows, like INFO windows for message transmission, and the STATUS window for query and answering, and to display the present hypothesis and context under consideration. The use of function keys and other keys is very consistent, and their amount is very limited, which makes it easy to remember them. The answers entered by the user are checked by a simple spelling checker by matching it with the possible answers, making the system very forgiving for typos and spelling errors. We have found that the end user needs very little instruction to operate a ready to run expert system developed with Acquaint.

3.3 System Performance

One serious drawback of Acquaint is that the muLISP** interpreter can not address extended or expanded memory. Nevertheless, Acquaint loads the entire knowledge base into its standard memory. This limits the size of the knowledge base considerably. Lithp Systems has acknowledged this problem and developed a new version (4.0) of Acquaint that uses virtual memory. However, this version has not been released yet.

The speed of run time versions, considering that the system is implemented in LISP, is reasonably fast when used on a personal computer with micro processor of the 808X family. For developers a personal computer with a 80286 or 80386 micro processor is recommended.

---

* This does not apply for Micro Acquaint, a beginners version of Acquaint, which is completely menu driven.

** muLISP is a trademark of the Soft Warehouse.

4 Conclusion

Although Acquaint has some drawbacks (i.e., the size of the knowledge base, the lack of a graphical user interface, and the limited options for development), it does present a powerful tool for professional expert system development for personal computers. Acquaint is financially attractive, because Lithp System does not charge runtime royalties. Other strong points of Acquaint are its clear user interface, the contextual structure for the procedural knowledge, its reasoning with certainty factors and its open architecture. The manual is well written and the example files provide a good source of information on the available options and behavior of the system.

5 References

1. Plugge LA, Verhey FRJ, Jolles J. A Desk-Top Expert System for the Differential Diagnosis of Dementia: An evaluation study. Intl J Technology Assessment in Health Care, 1990, 6; 1: 147-156.
2. Acquaint, Expert Systems, News, 1987, 4; 2: 123.
3. Ronteltap CFM. De rol van kennis in fysiotherapeutische diagnostiek. Thesis, Maastricht, The Netherlands, 1990.
4. Buchanan BG, Shortliffe EH. A Model of Inexact Reasoning in Medicine. In: Buchanan BG, Shortliffe EH (Eds.) Rule-Based Expert Systems, Reading, Addison-Wesley Publishing Company, 1985, pp. 233-262.
5. Acquaint, User's manual for Acquaint and Acquaint-Light. Lithp Systems BV, Purmerend, Holland, Release 3.00, 1987: 1-22.

IV DEVELOPMENT OF EVINCE: AN EXPERT SYSTEM FOR THE DIAGNOSIS OF DEMENTIA

1 INTRODUCTION

From the beginning of expert system (ES) research, medicine has been a major field of interest as a source of human expertise, in order to demonstrate that Artificial Intelligence (AI) techniques made it possible to simulate generally acknowledged intelligent human behaviour with the aid of a computer. As general human intelligence was (and still is) far too complex to be simulated by a computer, ES research focussed on fields of human expertise that were considered to encompass relatively small, highly specialized and well defined domains. Medicine, with its specialized disciplines was judged to be one of them. Consequently, some of the first and best known ESs contained medical expertise, e.g., INTERNIST[1], CASNET[2] and MYCIN[3]. Since then there has been a steady increase of applications on a wide range of medical subjects. The *CRI Directory of Expert Systems*[4] -using a broad definition of ESs- lists 145 medical applications in all kinds of developmental stages. However, only a few medical ESs moved beyond the level of a research (demonstration) prototype. Waterman[5] lists 46 medical applications that are in the research or demonstration stage, and only 7 in the field or production stage.

There are several possible causes for this gap between development, and field or experimental utilization. One cause is the interest of the ES researchers in developing new AI techniques, rather than in the practical use of an application. This emphasis on AI techniques is reflected in the scanty attention given to performance evaluations. A second reason for the discrepancy, is the use of expensive hardware. This is due to the high computational performance which is required for symbolic reasoning. This hardware is usually not available outside the laboratory, which makes prototypes less suitable for field testing. The fact that the performance of new hardware steadily increases does not reduce the problem, because the performance demands increase likewise. A third reason is the misconception that medical knowledge is equally well defined as the problems it addresses. Although the amount of medical knowledge -including the topics of consensus- is large, there is still an abundance of uncertain knowledge and disagreement, both within and between disciplines. This uncertainty and lack of consensus poses great difficulties in developing and evaluating medical ESs, as was shown by Buchanan and Shortliffe when they developed MYCIN.[6] Furthermore, although medical specialists commonly practice in a monodisciplinary manner, they also have -and are expected to have- knowledge of related disciplines in order to be able to judge whether or not to refer the patient, or to consult a member from another discipline. However, the knowledge of the ESs developed so far is usually restricted to one discipline in order to simplify the problem and to reduce the amount of conflicting knowledge. Unfortunately, such simplification does not only hamper the quality of ESs, it also reduces one of the most valuable aspects of (medical) ESs: their potential to spread expertise among a large audience of clinicians within related disciplines. Moreover, ESs can play an important role in making knowledge

accessible for discussion because ES development requires the explicitation of the knowledge involved. Therefore, ES development should not be restricted to clearly defined areas of expertise or one specialism on a domain.

With this in mind the ES EVINCE for the domain of dementia diagnostics was developed. EVINCE was implemented on a simple personal computer (IBM PC/XT compatible) using a commercial expert system development tool (ESDT), and contains multi-disciplinary expertise from such diverse disciplines as neurology, psychiatry, and neuropsychology. In two experiments the performance of EVINCE was compared with the neuropsychiatric domain expert, three prominent experts from three disciplines and 85 clinicians from 5 disciplines. The first experiment revealed a moderate to high level of agreement between EVINCE and the domain expert on the major diagnoses (See chapter V).[7] In the second experiment it was shown that the improved and expanded version of EVINCE produced better results than the average clinician, and performed at par with the three experts (See chapter VII).[8] On the basis of the results of these evaluations, Evince can be considered to be one of the few medical ESs that have left not only the prototype or demonstration stage, but also the laboratory stage.

As the evaluations have shown that the performance of Evince was at an expert level and that such a system can be implemented on a simple personal computer, it was deemed of importance to provide a more detailed description for development of ESs in similar domains. The description will concern itself with the problem domain, the materials and methods used during development, and the architecture of the system.

2 Domain definition

Dementia diagnostics is a typical example of a medical domain where knowledge from different disciplines, notably, neurology and psychiatry, has to be integrated in order to make a reliable diagnosis. This multi-disciplinary approach is due to the nature of the dementia syndrome, which manifests itself through both physiological and behavioral changes in the patient. In recent years, the knowledge of dementia diagnostics and causes of dementia has increased steadily, and has led to an increased level of consensus on the diagnostic criteria. Nevertheless, this increased knowledge seems to have penetrated daily practice only partly and appears to be difficult to apply for the average clinician due to the multidisciplinary approach required.[9,10,11] Furthermore, differences were found between disciplines in their diagnoses. For example, neurologists and psychiatrists differed significantly in their etiological diagnoses when presented with the same patient case descriptions.[11] However, it should be remembered that dementia is only a small part of the medical problems encountered by the disciplines mentioned before. Although one can expect that a clinician is informed about new developments in his or her own field, it is unrealistic to expect the same level of expertise for related disciplines.

There are over 50 different known causes of dementia, of which three causes are the most prominent: Dementia of Alzheimer Type (DAT), Multiple-infarct Dementia

(MID), and depression-induced dementia.[12] Of these, DAT is the most difficult to identify, because it is predominantly characterized by exclusion criteria, i.e., if non of the other causes can be found responsible, then DAT is possible or probable. However, the definite diagnosis DAT can only be made on the basis of both clinical and post-mortem findings. An additional complicating factor is that several diagnoses may coexist, for example, the diagnosis DAT and vascular problems can be found simultaneously, producing the diagnosis MIX.[13] Thus, the search for the dementing cause asks for a complete examination of all possible causes, while postponing the decision for the diagnosis DAT until last.

The task for the clinician, and thus for the ES, is threefold: firstly, a decision has to be made on the question whether or not a dementia syndrome is present, i.e., whether the cluster of symptoms and physiological measures is characteristic for dementia. Secondly, a variety of neurological signs and other somatic and behavioral indices has to be searched for, and thirdly, a decision about the etiology of the dementia syndrome has to be made. The first task meant that the expert system would have to be able to exclude some of the other syndromes which resemble dementia, and be able to report this if it could not achieve that. That is, the ES should have some knowledge outside the domain. The second task implied that the ES should gather information about the etiology and the third task meant that the ES had to decide which cause was most likely, whether there were alternative causes, contradictions, or several causes acting together.

3 Materials and Methods

3.1 Starting-points

There were five factors with important consequences for the development method* chosen.

The first factor was the decision to build a small prototype within a short time for feasibility assessment. This excluded the use of an elaborated knowledge acquisition approach, like the (in 1987) emerging 'Knowledge Acquisition Documentation and Structuring' method (KADS), which requiers "a full analysis of data (...) before any design and implementation".[14(p24)]

The second factor was the decision to develop and implement the ES on a simple IBM compatible personal computer. Thus, avoiding the need to port the system from a workstation, minicomputer or mainframe, with the problem of scaling the system down for personal computer use.

The third factor affecting the development method was the decision to use a commercial expert system development tool (ESDT) to achieve a further reduction of development time. A further requirement was that the ESDT would be suitable for use

* To avoid confusion, we use Jackson's[15] distinction between 'method', i.e., "a way of doing something", and 'methodology', i.e., "the study of method". Here a method is discussed, not a methodology.

on a personal computer. A drawback of such an approach is that it could force a reduction of the problem to fit the ESDT chosen. However, at the start of the project there were several commercial ESDTs available, offering the opportunity to choose the one best suited for the problem at hand.

The fourth factor was the domain of dementia diagnostics. Although the choice of the development method and the knowledge acquisition are discussed separately, in reality they were very much intertwined. Analysis of the domain showed that a structured clinical protocol was used which was partly documented, i.e. a checklist and the diagnostic criteria to be used, and partly a mental protocol. The latter part, i.e., the mental protocol, consisted of rather discrete steps about when to apply which criteria, which data should be used, and what procedure should be used in applying the criteria. Each of these decision steps produced their own diagnoses and/or - conclusions, terminated by a final evaluation by the clinician to make a concluding report with diagnoses.

The fifth factor was the decision to use a top-down development approach and to design a modular system. A modular design would make it easier to expand (breadthwise development) and refine (depthwise development) the system at later stages.

3.2 Development Method

Given the aforementioned starting points, an iterative prototyping Life-Cycle Model (LCM) using incremental implementation seemed to fit these starting points best.[15,16,17] Figure 1 gives a graphical representation of the development method used. The development method consisted of 6 consecutive stages, proceeding in a top-down fashion, each with the possibility to return to a previous stage. Usually, the early stages of top-down development deal with many unknowns, because there is no knowledge about the final model.[15] In such a case a bottom-up method is more useful. However, the situation is very different for expert system development where the system is not designed from scratch, but extracted from the domain expert's mind. Although human experts cannot always explain in detail how they handle problems in their domain, most of the time they do have a general idea, i.e. a mental protocol, of their approach and the information involved. It is the task of the knowledge engineer to extract this protocol and to convert it into a computer model. This means that expert system development centers around explicitating the expert's (mental) model. For this reason it was thought that a top-down development method was more appropriate than a bottom-up approach. In our case it meant using the expert's knowledge to implement the available paper and mental protocol used by the department of Neuropsychology & Psychobiology, which in turn was based on international criteria and procedures.

During the first three stages and the beginning of the fourth stage the method used consisted of knowledge acquisition, i.e. making an inventory of the expert's knowledge about the domain, the tasks and procedures. In the fourth stage one task was analyzed in depth, implemented, and tested. Depending on the results of the test, a next task was selected, or one or more of the preceding stages were revisited. For

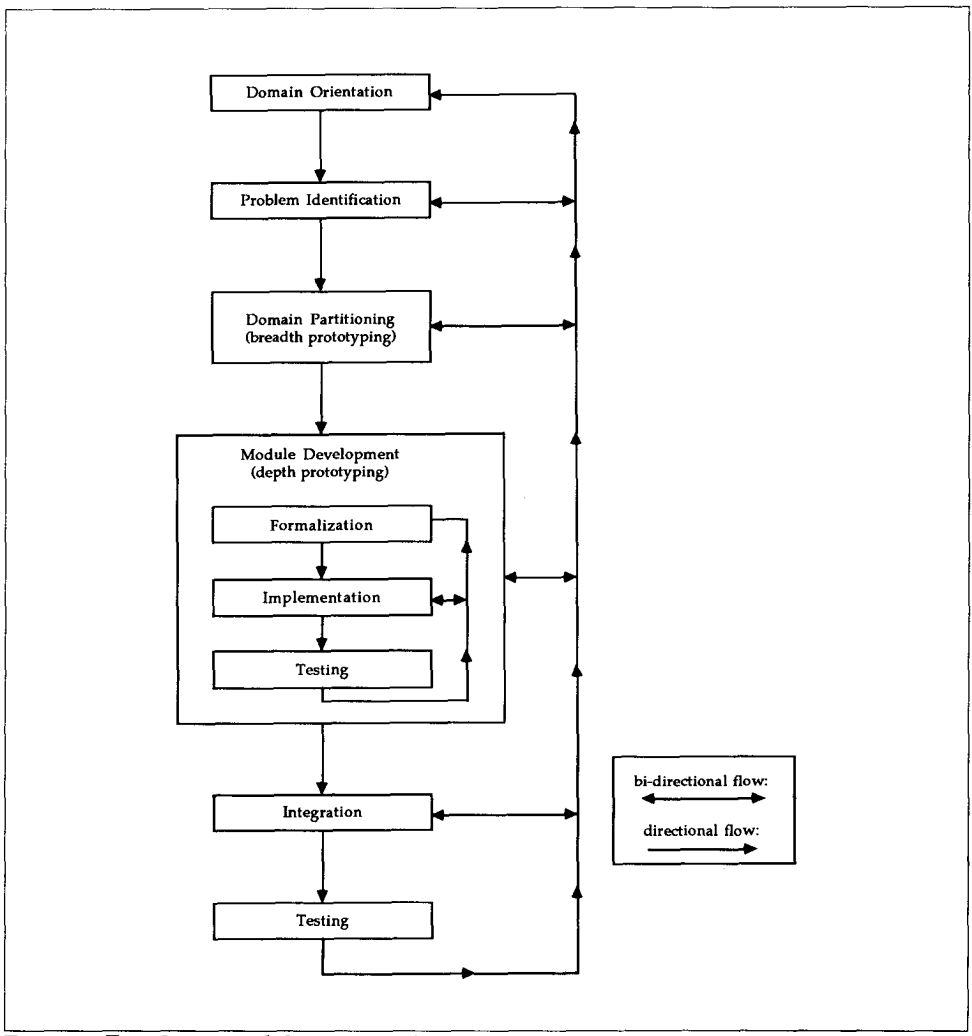

Figure 1. Top-down modular prototyping.

example, one conclusion of a module test may be that essential information was missing, i.e. another task should have been performed first. After the test of a module showed that it was successful, it is integrated into the system. If the partitioning at the third stage was done accurately this would pose no problems. In the last stage the whole system was tested, to see how the modules behave together.

3.3 Knowledge Acquisition

A neuropsychiatrist* of the Maastricht Academic Hospital was asked to cooperate in the project, because he was specialized in the diagnosis of dementia. He was medical coordinator of the Maastricht Memory Clinic (MMC), a department specialized in early and differential diagnosis of dementia. Having a neuropsychiatrist cooperating in the project had the advantage that both neurology and psychiatry, two essential disciplines for the diagnosis of dementia, were covered. Another advantage for knowledge acquisition was the availability of a standardized protocol consisting of a documented checklist and the diagnostic criteria to be used. This protocol was based on international guidelines for the diagnosis of dementia as detailed in the *Diagnostic and Statistical Manual, DSM-III-R* and the report of the NINCDS-ADRDA consensus work group and proved very succesfull in discriminating non-dementing patients from demented, even in the more early stages.[18,19,20,21] During knowledge acquisition the role of the domain expert would be to explain the checklist and the international guidelines used. This approach had the advantage that the knowledge would be internationally acceptable, while avoiding the need to use more than one domain expert.

To let the knowledge engineer familiarize himself with the subject, terminology, procedures and the people involved, knowledge acquisition started from the moment the project was initiated. To that end the knowledge engineer visited several consultations and patient examinations.** This included visiting several weekly conferences about the patients involved. Each session attended was concluded by an interview. Furthermore, the neuropsychiatrist and neuropsychologists involved provided the necessary literature concerning the international criteria used. The fact that the knowledge engineer was a cognitive psychologist was a clear advantage during the whole project. After this initial period, the actual knowledge acquisition started, proceeding in a top-down fashion, using the available checklist and the international criteria as a guideline. As dementia diagnostics consisted of three tasks (See paragraph 2 above), the first task, i.e., deciding on the syndrome dementia, would be implemented first, followed by the tasks to find possible causes.

Although several knowledge acquisition methods exist[22], only a few were considered appropriate in this domain: observation, (structured) interview and document analysis. For several reasons (ethical, privacy), consultations and medical examinations were observed and only discussed afterwards. The possibility to set up simulation consultations was considered but rejected, because it would be very time consuming while the information gain was questionable. Knowledge acquisition proceeded top-down, i.e., from a general view of what dementia diagnostics means, upto specific information on, for example, the effects of a deviating level of vitamin B_{12}. Each interview session was tape-recorded and the typewritten version was given to the domain expert for comment and error checking.

---

* F.R.J. Verhey jr., M.D.

** In each case the patient's consent was asked.

3.4 The Expert System Development Tool (ESDT)

For the development of the ES the ESDT Acquaint was chosen. Acquaint is a Lisp based hybrid system, using frames with a fixed format to represent conceptual and procedural knowledge*. The central organizational property of Acquaint is the rulebase. Rules are arranged in hierarchical contexts, controling subject related rules. Each context is itself a rule, except that it has rules and subordinate contexts instead of a THEN-part (For a more elaborate discussion see Plugge[23]). In our development method each context is a module, i.e., a domain expert's task or topic of investigation.

Acquaint offers much less facilities for the structuring of conceptual knowledge than for the procedural knowledge. The main structuring facility is inheritance, which organizes concepts hierarchically. In our system this feature was used only for concepts carrying similar types of information, for example, laboratory blood test results. The procedural knowledge, i.e. the rules and contexts, was used as the main method for structuring, the organization of the conceptual knowledge was merely used for the convenience of the knowledge engineer. Acquaint's explanation mechanism provided the user with sufficient facilities to examine conceptual information in coherence with each context, obliviating the necessity for a separate structure for conceptual information. In our implementation the conceptual part of the knowledge base was treated as a kind of blackboard, from which each context and/or rule retrieved the necessary information and returned its findings.

Another feature of Acquaint is the use of MYCIN like certainty factors, which was important in our implementation to represent certainties about diagnoses. For example, by consensus the clinical diagnosis DAT is made only with the labels 'possible' or 'probable'. Although the use of certainty factors is not the best method to represent certainty, it is relatively easy to understand for the user in comparison with, for example, the Bayesian method.[24] (For a more elaborated discussion of certainty factors, see Chapter III paragraph 2.1 and 2.3)

---

* In traditional AI terminology *procedural* knowledge refers to the procedural programming laguage, while *declarative* knowledge refers to the rules and facts. However, the Function frame in Acquaint encompasses both declarative and procedural knowledge. Therefore, in the remainder of this thesis we will label knowledge consisting of rules as *procedural knowledge* and knowledge about data -including the Function frame- as *conceptual knowledge.*

4 The Architecture of Evince

The architecture of Evince will be discussed from two perspectives: 1) the knowledge representation design, and 2) the system design. This distinction is made, because the knowledge of Evince was to be used in an interactive and a batch mode. In the former mode the user would interactively go through a consultation session, while in the latter mode the system would examine one or more patient cases independently.

4.1 Conceptual and Procedural Knowledge representation

As mentioned before, the procedural knowledge was partitioned into different levels (modules), coinciding with the examination tasks performed by the neuropsychiatrist. These tasks and sub-tasks were represented in socalled *contexts*, while more detailed knowledge was represented in rules within these contexts. This resulted in a hierarchical tree, that is traced depth first by Acquaint's inference engine (see Figure 2, page 41). Each context has the capability to decide whether or not it will be worthwhile to check its subordinate rules and contexts by testing its premises. If a context's premises are not met, the tree will be pruned at that point, and another context at the same level will be checked. However, in our implementation not all contexts use selection criteria, because some tasks are performed routinely (e.g., the contexts MEDICATION and LAB-TESTS), or because the use of selection criteria would be the same as performing the task itself (e.g., the context ALCOHOL-ABUSE). An example of three contexts using selection criteria is given in Table 1, page 42.
In Table 1, the context DEMENTIA is visited only when the Global Deterioration Scale score is greater than 1 (a score of 1 means there is nothing wrong). If this is not true, then the contexts DEPRESSION-DEMENTIA, M.I.D., ALZHEIMER, and Mix are not visited. The contexts AMNESIA and MILD-COGNITIVE-DISORDER are at the same level as the DEMENTIA context, but will only be visited if one or more diagnoses are not true. Note, that the two bottom contexts will neither be visited when the inference engine is unable to establish that the required diagnoses were true, i.e. when they are known to be 'unknown'.

Although each context was developed relatively independent of each other there was one exception: the Evaluation context. Because the Evaluation context is used to process the findings (conclusions) from the previous contexts, it was developed in parallel with all the other contexts. It is also the largest context without subcontexts, because only the resulting diagnoses from the previous contexts are considered. However, in a few rules more specific concepts are checked to produce more detailed information. (See Table 2, page 43)

In contrast to contexts which are checked sequentially, rules are assigned a priority by the inference engine's agenda. This scheduling is performed separately for each context. The priority of a rule depends on the amount of information available, the amount of certainty of competing conclusions, and the amount of information required to prove a rule. For example, an OR-rule will receive a higher priority than an AND-rule, because in an OR-rule only one true premise is required to make the

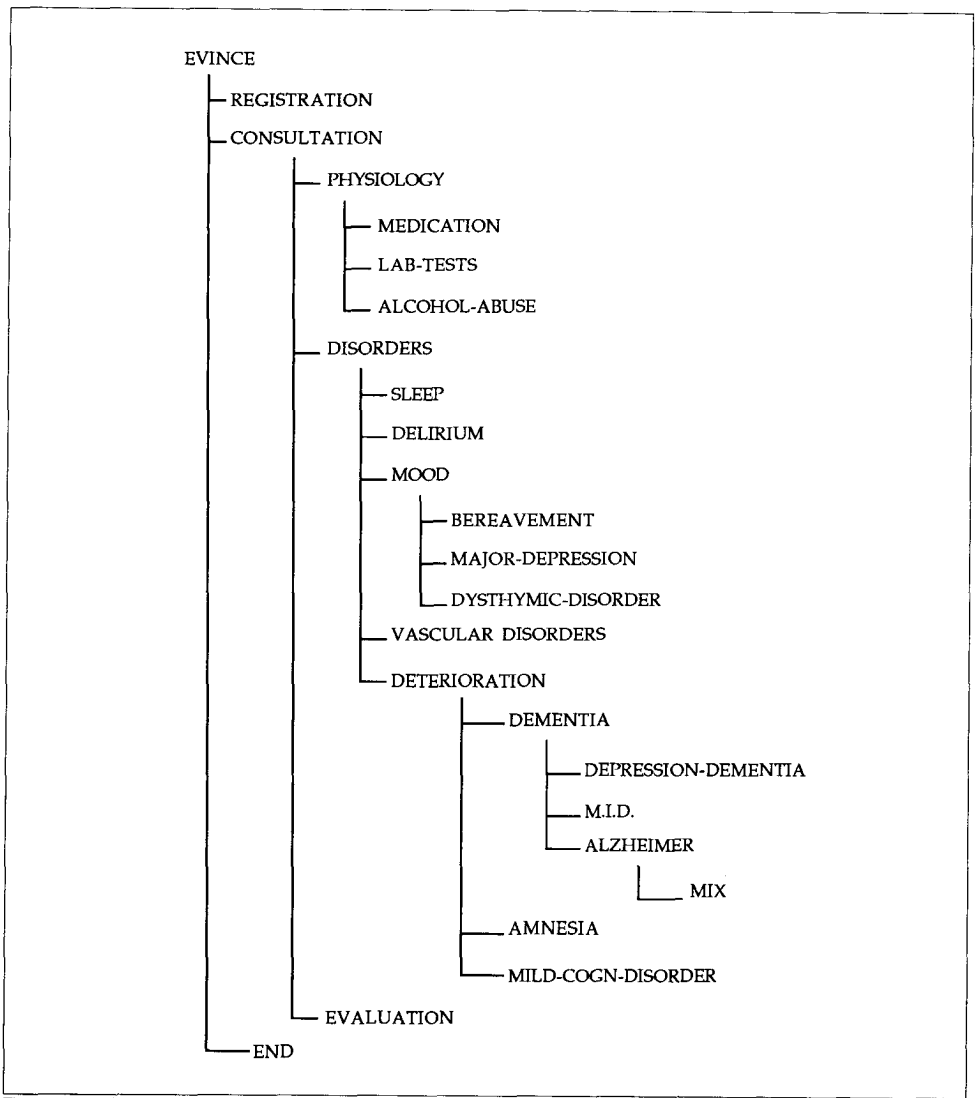

Figure 2. Context Level Decision Tree.

```
Context DEMENTIA
 IN: DETERIORATION
 IF: GDS > 1

 Context DEPRESSION-DEMENTIA        (IN: DEMENTIA)
 Context M.I.D.                     (IN: DEMENTIA)
 Context ALZHEIMER                  (IN: DEMENTIA)
     Context MIX                    (IN: ALZHEIMER)

Context AMNESIA
 IN: DETERIORATION
 IF: $AND GDS > 1
          $NOTTRUE (Diagnosis Demential Syndrome))

Context MILD-COGN-DISORDER
 IN: DETERIORATION
 IF: $AND GDS > 1
          $NOTTRUE (Diagnosis Demential Syndrome)
          $NOTTRUE (Diagnosis Amnestic Syndrome))
```

Note:
GDS = Global Deterioration Scale
M.I.D. = Multiple-infarct Dementia
MIX = Multiple-infarct Dementia with Alzheimer's Disease.

Table 1 Example of Context Level Knowledge.

rule succeed, while in an AND-rule all premises must be true. Because it takes less effort to prove an OR-rule, it is given a higher priority.

Furthermore, rules can be treated as data by using them as an additional premise in another rule. For example, the two main rules used for the diagnosis dementia (See Table 3, page 44) are actually one rule with a nested OR (This is why the rule *R-B-Criterium* does not need a THEN-part).

When an OR-rule checks its premises until one is found to be true, this could mean that not all premises will be examined. The inference engine can be forced to evaluate all premises by using the OR* prefix. Like an ordinary OR-rule, one true premise is required for the rule to succeed. However, the evidence from the other premises is used to possibly increase the certainty of the conclusion. Apart from the reason to increase the certainty of conclusions, this strategy was also used to make sure that related information was asked from the user within the same context. Thus, erratic questioning behaviour by the system can be avoided.

Although Acquaint's frame based representation of conceptual knowledge is restricted to the predefined slots, there was sufficient flexibility to store very different types of information. This was an important property, because the diagnosis of dementia requires the system to consider a variety of data. For example, the concept *EEG* stores high level qualitative information, while the concept *ComputeHamilton* is actually a hidden rule calling the function '+' to evaluate several other concepts. (See

```
(DefRule R-Delirium-III
    IF: $FORWARD
        (Diagnosis Delirium)
    THEN:
    MSG-INFO (According to the DSM-III-R criteria (See pp. 100-103) $FullName
             has a delirium. \(p= ^A$R-Delirium-III ^A\) However, in order to examine
             the possibility of a demential syndrome, a delirium must be precluded))
(DefRule R-DeliriumAlcohol
    IF: $FORWARD
        $AND* (Diagnosis Delirium)
             (Diagnosis Alcohol Abuse)
    THEN:
    MSG-INFO (The delirium could be caused by excessive use of alcohol.
             \(p= ^A$R-DeliriumAlcohol ^A\) ))

(DefRule R-DeliriumMedicine
    IF: $FORWARD
        $AND* (Diagnosis Delirium)
             DrugDisorder = '("confusional states")
    THEN:
    MSG-INFO (The delirium could be medication induced.
             \(p= ^A$R-DeliriumMedicine ^A\))
    MSG-INFO (Confusional states can be induced by the following drugs\:
             PrintKeyList(MedicationClasses "confusional states")))
```

Note:

DefRule:	define rule
$FORWARD:	use this rule only in forward mode, i.e., do not ask questions if the value of the concept is unknown.
MSG-INFO:	to send a message to the screen.
PrintKeyList:	A functions that prints items that are associated with a certain key.
p=<rulename>:	Print a certainty factor.
$FullName:	Prints the patient's name.

Table 2 Example of Rules from the Evaluation Context.

Table 4, page 45) On the other hand, the concept *Diagnosis* is a collection of diagnostic instances gathered during the consultation process, each carrying the certainty with which it was asserted or rejected.

Although not directly transparent to the user, the concept class diagnosis is used to make a distinction between diagnostic conclusions and information requested from the user. By storing all the conclusions as instances of the concept Diagnosis the rules in the Evaluation context simply had to check these instances and their relevant combinations.

```
(DefRule R-B-Criterium
  IF:     $OR*  AbstractThinking
                JudgementDisorder
                CharacterChange
                HighCorticalFunct = 'afasia
                HighCorticalFunct = 'agnosia
                HighCorticalFunct = 'apraxia
                HighCorticalFunct = '|constructive apraxia|)

(DefRule R-Dementia
  IF:     $AND  PastMemory >= 3
                RecentMemory >= 3
                ShortTermMemory >= 3
                R-B-Criterium
                Functioning > 3
  THEN:  -90 (Diagnosis Mild Cognitive Disorder)
         -90 (Diagnosis Amnestic Syndrome)
          90 (Diagnosis Demential Syndrome)
  ELSE:  -90 (Diagnosis Demential Syndrome)
         -90 (Diagnosis Medication Induced Dementia)
         -90 (Diagnosis Epileptic Induced Dementia)
         -70 (Diagnosis Depression Induced Dementia)
         -90 (Diagnosis Multi Infarct Dementia)
         -90 (Diagnosis Dementia Alzheimer Type)
         -90 (Diagnosis MIX))
```

Note:
Rules are predeeded by **R-**, e.g. *R-Dementia*.
The figures preceeding the conclusions are the maximum or minimum certainty with which a conclusion is asserted.

Table 3. Rules for the Diagnosis Dementia.

```
(DefConcept EEG
  FACT:      Electro Encephalogram
  PROMPT:    EEG findings
  COMMENT:   Choose 'unknown' if your finding is not listed below.
  EXPECT:    VALUES  none unknown diffuse_lowered focal_pathology
                     increased_Beta-activity epileptic_activity
  DEFAULT:   none
  CLASS:     LOG (an artificial class for filing purposes)
  PROP:      ASK)

(DefConcept ComputeHamilton
  FACT: Score of the Hamilton Depression Rating Scale
  IF: Sex = 'male
  DO: Mood + Guilt + Suicide + SleepStartDisorder + SleepThruDisorder +
      SleepEndDisorder + Activities + Restraining + Agitation + PsychologicalFears
      + PhysicalFears + GastroIntestSympt + GenPhysicalSympt + Hypochondria +
      BodyWeight + DisorderInsight + LossOfLibido
  IF: Sex = 'female
  DO: Mood + Guilt + Suicide + SleepStartDisorder + SleepThruDisorder +
      SleepEndDisorder + Activities + Restraining + Agitation + PsychologicalFears
      + PhysicalFears + GastroIntestSympt + GenPhysicalSympt + Hypochondria +
      BodyWeight + DisorderInsight + FemGenitalDisorder
  PROP: ASKNOT)

(DefConcept Diagnosis
  EXPECT: GROUP
  CLASS:  LOG
  PROP:   ASKNOT)
  % Example Instances %
  (DefConcept (Diagnosis Amnestic Syndrome))
  (DefConcept (Diagnosis Delirium))
  (DefConcept (Diagnosis Mild Cognitive Disorder))
  (DefConcept (Diagnosis Depression Induced Dementia))
  (DefConcept (Diagnosis Dysthymic Disorder))
  (DefConcept (Diagnosis Major Depressive Episode))
  (DefConcept (Diagnosis MIX))
  (DefConcept (Diagnosis Multi Infarct Dementia))
  (DefConcept (Diagnosis Complicated Bereavement))
```

Note:
FACT = The print name of the concept; PROMPT = The question to be asked; COMMENT = Explanation; EXPECT = Values to expect; CLASS = The concept's class PROP: = Properties, e.g., ASK, ASKNOT; DO = equivalent to THEN.

Table 4 Four Examples of Conceptual Knowledge.

4.2 System Design

Although an interactive consultation is efficient, because not all tasks have to be completed given sufficient evidence, it is clearly less efficient when the data of several patients have to be entered. Therefor the system was expanded to provide a way to consult the system in batch mode. Even at this level of system design the modular approach proved itself very useful. As the ESDT Acquaint can address only 640 kByte of memory, the system designed so far had to be split in several modules to be loaded when necessary (See Appendix I). The present system's design is shown in Figure 3. The Manager Module offers the user a choice between Batch Mode and Interactive Consultation. If Batch Mode is chosen, the user must choose one or more patient data files to be processed by the Consultation Module. For both consultation modes the same modules are used, i.e., the Consultation Module and the Report Module, which contain the actual knowledge base. The respective modules are sent a message whether they should work in Batch Mode or in Interactive Mode. In Batch Mode the number of files to process determines how many times a consultation is repeated. In both modes, the report is stored on file, and can be viewed or printed from the Manager Module.

Although the use of the Batch Mode Consultation decreases the consultation time, it has the disadvantage that the system cannot be asked any questions. Furthermore, there is a certain risk that the system will not arrive at a (correct) diagnosis due to missing data, because it is not allowed to ask questions during Batch Mode. Presently the data of the patients are stored in separate files, but a database is being designed.

Because the knowledge of Evince is dispersed over several modules, it is difficult to say 'how much' knowledge Evince actually contains. Furthermore, the complete knowledge is separated in 4 different types, of which the rules and concepts are the central part. Table 5 lists the amount of knowledge type per module.

Table 5. Number of Knowledge Frames

	Manager	Database	Consult	Report
Forms	3	14	6	1
Contexts	5	4	25	3
Rules	19	14	73	36
Concepts	10	126	129	32
Formulas	15	13	99	6

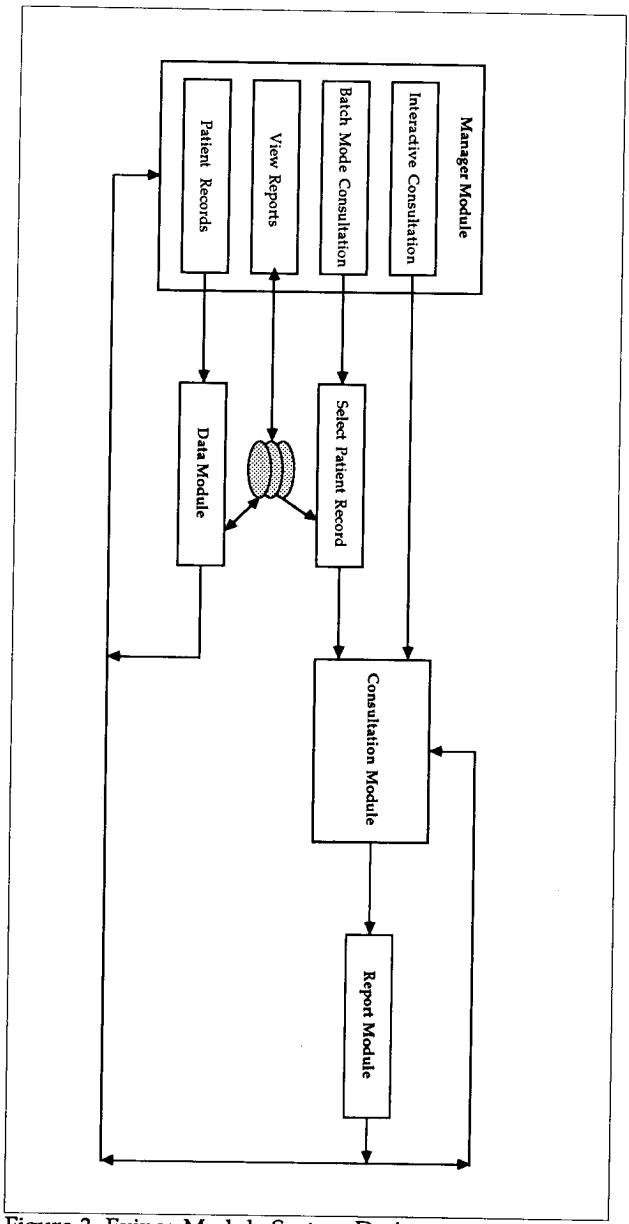

Figure 3. Evince Module System Design.

5 Concluding Remarks

During the course of development two experiments were set up to test Evince. The result of these tests showed that the knowledge base could compete with top experts in its field and produce better results than the average clinician.[7,8] Moreover, the results showed that Evince provides a valuable tool for an interdisciplinary approach in dementia diagnostics. Thus the risk of discipline biased diagnoses, as found in one of our experiments, can be avoided.[8]

Also, the top-down development method and the modular design have proven to be very useful. A top-down development approach guarantees that the knowledge documentation and structuring is done prior to the implementation, because it requires that a complete model is used. However, it does not require that knowledge acquisition is detailed on all levels prior to the implementation, as is required by the KADS method.[14] This top-down approach is further supported by the modular design, which makes it possible to implement parts at different levels of detail, that can be enhanced at later stages. Furthermore, a modular system can be expanded without changing the structure, by simply adding other modules. For example, with the modular approach a therapeutic module can be added to the system and even have an architecture that is very different from Evince.

The implementation of international guidelines for dementia diagnostics makes it unlikely that the design of the present diagnostic system will have to be changed radically within the near future. This is a clear advantage of using the domain expert as a mediator for the available consensus knowledge, and not as the main source of knowledge. Nevertheless, it remains likely that the present system will be subject to some small changes over time, due to new methods, the introduction of new findings and new criteria. It is at that point that the modularity of the system will prove itself a valuable approach for system maintenance.

With reference to the remarks made in the introduction about causes hindering ESs research to reach an operational stage, it seems legitimate to state that the method used to develop Evince has made it possible to avoid these pitfalls. Also, it should be stressed that reports on this development have been quite different from what is (still) common practice in ES research. Many studies present only some percentages about correct or incorrect diagnoses (c.f. Rienhoff et al.[25]). In contrast, the reports about Evince emphasized a careful description of the evaluation methods and results, reporting about the patients used, the experts, the data, and methods of analysis. As ESs are primarily meant for practical use, we think that it should become common practice that ES research reports pay more attention to the topics just mentioned. In our opinion this will not only enhance the development of new techniques, but also have a positive spin-off for the domain of interest.

6 References

1. Myers JD, Pople HE. INTERNIST: a consultative diagnostic program in internal medicine. Proceedings of the First Annual Symposium on Computer Applications in Medical Care, 1977: 52.
2. Kulikowski CA, Weiss S. Strategies of data base utilization in sequential pattern recognition. Proceedings of the IEEE Decision and Control Conference, 11th Symposium on Adaptive Processes, 1972: 103-105.
3. Shortliffe EH, Axline SG, Buchanan BG, et al. An artificial intelligence program to advise physicians regarding antimicrobial therapy. Computers and Biomedical Research, 1973; 6: 544-560.
4. Smart J, Langeland-Knudsen J. The Cri Directory of Expert Systems. Learned Information Europe Ltd., Oxford, 1986.
5. Waterman DA. A guide to Expert Systems. Addison-Wesley Publishing Company, Reading Ma., 1986: 272-288.
6. Buchanan B, Shortliffe EH. (Eds) Rule-Based Expert Systems. The MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley, New York, 1984.
7. Plugge LA, Verhey FRJ, Jolles J. A desktop expert system for the differential diagnosis of dementia: an evaluation study. Int J Technology Assessment in Health Care, 1990, 6; 1: 147-156.
8. Plugge LA, Verhey FRJ, Jolles J. Differential diagnosis of dementia: a comparison between the expert system Evince and Clinicians. Journal of Neuropsychiatry, 1991, 3; 3; 1-7.
9. Garcia CA, Reding MJ, Blass JP. Overdiagnosis of dementia. Journal of the American Geriatrics Society. 1981, 29; 9: 407-410.
10. Kukull WA, Larson EB, Reifler BV, Lampe TH, Yerby M, Hughes J. Interrater reliability of Alzheimer's disease diagnosis. Neurology, 1990; 40: 257-260.
11. Plugge LA, Verhey FRJ, Van Everdingen JJE, Jolles J. Differential diagnosis of dementia: an experimental study into intra- and interdiscipline agreement. Journal of Geriatric Psychiatry and Neurology, 1991, 4; 2: 90-97.
12. Marsden CD. Neurological causes of dementia other than Alzheimer's disease. In: Kay & Burrows (Eds.) Handbook of Studies on Psychiatry and Old Age. Elseviers Science Publishers BV, Amsterdam, 1984: 145-67.
13. Marsden CD. Assessment of dementia. In: Frederiks (Ed.) Handbook of Clinical Neurology. Volume 2 (46), Neurobehavioural Disorders, Elseviers Science Publishers BV, Amsterdam, 1985, pp 221-232.
14. Hickman FR, Killin JL, Land L, Mulhall T, Porter D, Taylor R. Analysis for knowledge-based systems: a practical guide to the KADS methodology. Ellis Horwood Ltd, Chichester.

15. Jackson MA. System development. Prentice/Hall International, Englewood Cliffs, NJ, 1983.
16. Yourdon E, Constantine LL. Structured design. Prentice/Hall, Englewood Cliffs, NJ, 1979.
17. Van Vliet JC. Software engineering. Stenfert Kroese, Leiden, 1988.
18. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. Organic Mental Syndromes and Disorders. 3rd Edition, Revised. Washington DC, American Psychiatric Association, 1987.
19. McKhann G, Drachman D, Folstein M, et al.: Clinical Diagnosis of Alzheimer's Disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology. 1984, 34: pp 939-44.
20. Verhey FRJ, Vreeling FW, Jolles J. DSM-III and NINCDS/ADRDA criteria for dementia and Alzheimer's disease: impact of diagnostic procedures on daily practice. In: Wurtman RJ, Corkin SH, Growdon JH. et al. (Eds.) Alzheimer's Disease: advances in basic research and therapies. Proceedings of the Fifth Meeting of the International Study Group on the Pharmacology of Memory Disorders Associated with Aging. Zürich, 1989: 419-423.
21. Jolles J. Early diagnosis of dementia: possible contributions of neuropsychology. In: Traber J, Gispen WH, Senile Dementia of the Alzheimer Type, Springer-Verlag, Berlin, 1985: 84-100.
22. Neale IM. First generation expert systems: a review of knowledge acquisition methodologies. The Knowledge Engineering Review, :105-145
23. Plugge LA, Acquaint. Expert Systems, 1990, 7; 4: 243-245.
24. Tonn BE, Goeltz RT. Psychological validity of uncertainty combining rules in expert systems. Expert Systems, 1990, 7; 2: 94-100.
25. Rienhoff O, Piccolo U, Schneider B. (Eds.) Expert Systems and decision support in medicine. 36, 33rd Annual Meeting of the GMDS EFMI Special Topic Meeting, Springer-Verlag, Berlin, 1988.

V A DESKTOP EXPERT SYSTEM FOR THE DIFFERENTIAL DIAGNOSIS OF DEMENTIA.*

1 INTRODUCTION

An expert system (ES) is a computer program designed for specific fields of expertise in which it attains a performance equal or better than that of a human expert. More specifically, medical ESs can be regarded as computer software products with a medical data base that are designed to assist physicians and medical personnel in diagnosis, therapy, and related tasks in medical care.[1] As the use of medical ESs is expected to produce profound changes in health care, it is deemed of utmost importance to know the potential of this new technology.[1,2,3] For instance, it is of relevance to know whether medical ESs are able to make similar - or even better - diagnoses than a physician, and whether medical ESs could be of value in improving health care. In addition, it is important to acquire information about the possible benefits and risks of this new technology.[4]

The medical ESs that have been developed up till now, share some characteristics. Firstly, they have been developed for areas in which the expert knowledge is fairly well defined, such as CENTAUR (interpretation of pulmonary tests) and AI/MM (renal physiology).[5] Secondly, a common feature of most medical ESs is that the amount of data used is relatively small. Thirdly , hardly any medical ES developed so far has left the prototypical or demonstration stage and there is a paucity of studies in which the medical ES is compared with the physician, i.e. 'domain expert'. Fourthly, there is a very rapid increase in the use of personal computers in medicine, but until now no reports have been published about their potential use for medical ESs. Thus it is of interest to know whether it is possible to develop a medical ES for more complex problems such as those in areas where physiological, behavioral and psychosocial data have to be combined, whether such an medical ES can be developed on a - cheap - personal computer and whether such a medical ES yields conclusions that are similar to those of the physician/domain expert. The present paper applies these questions to the field of neuropsychiatry and the increasingly important topic of the differential diagnosis of dementia.

In the first part of this paper a description will be given of the development of EVINCE-I. The second part presents the assessment of the medical ES EVINCE-I versus the domain expert in diagnosing 29 patients.

---

* This chapter was published as: Plugge LA, Verhey FRJ, Jolles J. A desktop expert system for the differential diagnosis of dementia: an evaluation study. Int J. Techn Assessment in Health Care, 1990, 6; 1: 147-156.

2 Development of Evince

2.1 Differential Diagnosis of Demential Syndromes

Dementia is a concept that still has very controversial definitions, although most experts agree that it is better regarded as a syndrome than as a nosological entity. Some essential features of the demential syndrome are: memory impairment, intellectual deterioration and changes in personality.[6] Dementia can have many different causes: up to 50 different causes are enumerated by Marsden[7]. The two main causes of dementia are Dementia of Alzheimer Type (DAT), with a prevalence of 39-50% of the total number of cases, and multi-infarct dementia (MID), with a prevalence ranging from 13-30%. This leaves approximately 20-48% accounted for by the remaining causes.[7,8,9] These broad ranges can be explained by the difficulty in defining the concepts DAT and MID. Some researchers regard the overlap between DAT and MID as a special category, usually referred to as Mix. Moreover, some include depression-induced dementia, i.e. pseudodementia[9], which has a prevalence of approximately 9%, whereas others exclude this condition from their figures.

Given the succinct description presented above, the task of the ES in making a diagnosis in the individual patient is threefold:

1. deciding on the diagnosis 'dementia',
2. searching for neurological signs and other somatic indices,
3. interpreting these signs to decide on the aetiology of dementia.

This means that the ES should have knowledge of common disorders that present themselves as dementia, e.g. depression, and disorders that preclude the diagnosis of dementia, e.g. delirium. After stating the diagnosis as being dementia, the ES should be able to differentiate between the possible causes of dementia. Since dementia is caused in 52-80% of all cases by DAT and MID, and because depression is a major disorder that can present itself as dementia, it was decided to centre the domain around these three causes. The possibility of adding some or all of the other causes in a later version of the ES was left open; their choice would depend on the data used for diagnosing the three main causes. This decision was made to economize on the amount of data needed in the present version of the ES.

Beside making valid diagnoses, i.e. a significant agreement between the ES and the domain expert, it was deemed important that both the ES and the domain expert express their certainty about any diagnosis made. Such a requirement would enable us to make a more precise comparison between the ES and the domain expert. Furthermore the ES should be able to reproduce the process of medical decision making (MDM) which takes place in the domain expert, for later evaluation of the rules and data used. This would simplify comparison of the differences and similarities found.

2.2 The Choice of the Expert System Shell

The most common structure for ESs is a computer program consisting of a knowledge base, an inference engine to infer new facts with the help of the rule base and data available and a user interface to regulate the communication between the user and the inference engine.[10]

ES-shells consist of the above-mentioned units, excluding the knowledge base, and one of the major decisions to be made in the development of an ES is the choice of the appropriate ES-shell. This is because the shell determines the type of knowledge representation and the inferences possible. For the development of EVINCE-I the expert system tool ACQUAINT was chosen.[11,12] Knowledge in ACQUAINT is represented in frames, comparable to a database structure, which makes it relatively easy to add or delete knowledge. Reasoning in ACQUAINT can be both hypothesis and data directed, i.e. forward and backward, and includes the use of certainty factors* as a measure of confidence. Rules can be used at different levels of reasoning, i.e. rules controlling rules, and are organized in easy readable IF-THEN statements. Further more, ACQUAINT is operational on the widespread standard IBM-compatible PC with 512 K-RAM and two floppy drives. Another important feature is that the end-user is supplied with a run-time copy of the ES and not the whole tool, which makes it financially very attractive.

2.3 Defining the Knowledge Base

After the initial preparations of forming a general picture of the domain, a start was made by further defining the problem. This was done by drawing a detailed picture of the MDM strategy used by the domain expert, a neuropsychiatrist who is a member of the department of Neuropsychology and Psychobiology at the University of Limburg. This expert was chosen, because of the method he used, which is based on international guidelines for the diagnosis of dementia as detailed in DSM-III-R[13] and the report of the NINCDS-ADRDA consensus work group[14]. Thus the knowledge to be implemented would be widely acceptable. The domain expert was therefore interviewed about these guidelines, and the protocols resulting from them were examined. This resulted in a decision tree which reflects the search strategy used by the domain expert. The decision tree is organized into contexts which each control a body of rules and data concerning the subject to be investigated. The data used by EVINCE-I can be found in Table 1 (see the next page) and the diagnosis in Table 2. The global decision procedure will be discussed in the next chapter.

* For a more detailed description of the use and calculation of certainty factors the reader is referred to references 5, 11 and chapter III, paragraph 2.1.

Subject:	Contents:
Identification data:	age, name, I.D.-number, date of birth, consultation date, sex, hand preference, occupation, social-economic status, educational level.
Haematology:	BSE, Hb, Ht-percentage, Leucocyte
Chemistry:	Kalium, Na, Ca, P, Glucose, B_1, B_{12}, AF, Gamma-GT, OT, PT, T4, TSH.
Auxilliary tests:	EEG, CT-scan, ECG.
Anamnesis:	weight, alcohol use (former and present), appetite, smoking habit (former and present), intoxication, occupational functioning, social functioning.
Cognitive functions:	memory, abstract thinking, higher cortical functions (aphasia, apraxia, agnosia, constructive apraxia), consiousness, attention, judgement, orientation, perception.
Motor functions:	speech, motorial activity.
Personality:	character, personality, social activities, disorder-insight.
Complaint patterns:	onset, deterioration course, deterioration pace, physical complaints.
Sleep:	somnolence, sleep-start disorder, sleep-through disorder, sleep-end disorder, nocturnal confusion.
Mood:	lability, mood fluctuations, daytime fluctuations, guilt feelings, suicidal thoughts, inhibitions, agitation, fears with psychological or somatic manifestations (general and gastro-intestinal), libido, hypochondria, depersonalization and depersonalization type, paranoiac symptoms, symptoms of compulsion, Hamilton depression score.
Atherosclerosis:	anamnestic hypertension, TIA, CVA, atherosclerosis, focal neurological symptoms, focal neurological signs, Hachinski score[15].

Note: The list above represents the data gathered during medical examination, not the variables and functions used for window management, file variables, etc.

Table 1. Patient characteristics used by EVINCE-I

* Demential syndrome (senile/pre-senile).
* Primary degenerative dementia (senile/pre-senile).
* Multiple infarct dementia.
* Major Depression.
* Delirium.
* Cognitive Disorders.
* Hyperthyreosis and Hypothyreosis
* Diabetes.
* Alcoholism.
* Electrolyte disorder.
* Inflammation.
* Vitamin-B-insufficiency.
* Medication.
* Epilepsy.

Table 2. Diagnostic knowledge used by EVINCE-I

2.4 Description of the Decision Procedures.

The root node, or top-level context, is called 'Initiate' and contains an introduction to the system. The next context contains a few rules to decide whether the problem of investigation is part of the knowledge domain. If it fits the domain, or if the user is not sure, the actual consultation will begin. The actual consultation begins by collecting general data about the patient, such as date of birth, age, sex, et cetera, in the context 'Patient Data', and data from auxiliary investigations, such as blood and urine tests, in the context 'Preliminary Data'. In this latter context EVINCE-I will check the data for possible deviations which could yield any preliminary diagnosis. These findings are then used during the examinations to follow. After that EVINCE-I will try to find out the patient's state of consciousness. If the system is sure that the patient's consciousness is clouded, i.e. one or more symptoms fit the criteria for delirium, it will terminate the consultation. The system will then leap to the 'Report' context and make a final report. The decision to abstain from further examination is taken on the basis that the consciousness should not be clouded, which is one of the DSM-III-R criteria for dementia[13].

If the patient's consciousness is not clouded, EVINCE's will try to determine whether this particular case is an instance of dementia without making a decision about the etiology. EVINCE-I does so by using the remaining DSM-III-R[13] criteria for dementia. If dementia is not diagnosed, EVINCE-I will skip any context concerning dementia and jump to the context which determines whether depression is the cause of any of the problems encountered. In the 'Report' context it collects the diagnoses made thus far, checks them for mutual implications, and presents its findings.

If dementia is diagnosed, the system will try to determine the cause of the demential

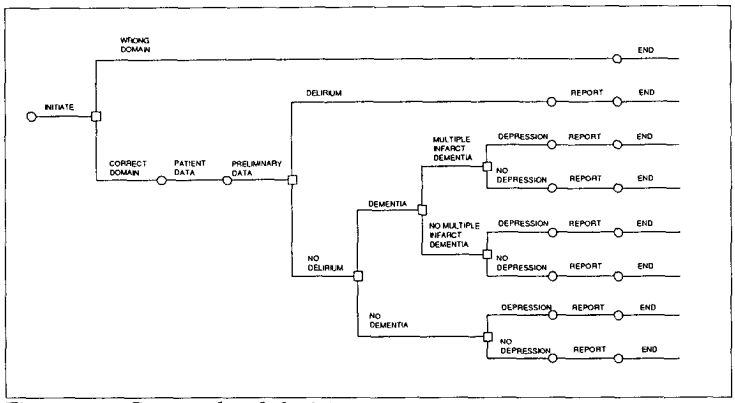

Figure 1. Context-level decision tree.

syndrome. This is done in a special way, related to the nature of the main cause: DAT. Since DAT is initially diagnosed by using exclusion criteria the diagnosis DAT is delayed until all other causes have been excluded. Therefore EVINCE-I will first check the data for any vascular problems that could lead to the diagnosis of MID, the next major cause of dementia. Independent of the outcome EVINCE-I will check the data for signs of depression. Finally a last check is performed in the 'Report' context to see if any of the preceding diagnoses can be used to exclude the diagnosis DAT, e.g. a vitamin B_{12} deficiency.

When all these steps have been taken, and DAT can still not be excluded, the ES will use some of the preliminary data to adjust the certainty factor of the final diagnosis DAT. The certainty factor is raised, for example, when the EEG pattern shows deceleration in the alpha-rhythm activity. The certainty factor is decreased if the EEG shows any signs of focal pathology. However, none of these data can exclude or confirm DAT by itself, they are used as the NINCDS-ADRDA work group[14] propose: for adjustment of the diagnostic confidence.

The final report that EVINCE-I produces is made in the 'Report' context. This context reflects the evaluative activity of the domain expert and the reformulation of the findings in legible standard phrases. This context, or child context, will write complete standard reports in future versions.

3 A Comparison between Evince and the Domain Expert

3.1 Introduction

As stated above, it was deemed important to know whether a desk-top ES is powerful enough to make valid diagnoses. The performance of EVINCE-I was therefore compared to that of the domain expert in an experiment in which the following two hypotheses were tested:

1. expert and EVINCE-I agree on their diagnoses,
2. expert and EVINCE-I agree on the relative certainties of these diagnoses.

Furthermore the number of false positive and false negative diagnoses made by EVINCE-I was compared to that of the expert, in order to see whether a specific type of mistake prevailed in any diagnostic category.

3.2 Methods

From the patient records available from the Maastricht Memory Clinic, 19 patients were drawn who had been diagnosed with at least 50% certainty as having dementia, according to the estimation of the domain expert, irrespective of its cause or other diagnoses. These cases were taken to see whether EVINCE-I would make any false-negative dementia diagnoses. To check if EVINCE-I would make any false-positive dementia diagnosis, 10 other patients were drawn who were not diagnosed as having dementia. Most of these patients suffered clinically from a (mild) depression. The mean age of the first group was 73 (sd = 7.5) and that of the second group was 47 (sd = 14.6). None of the applied diagnostic criteria uses age as a distinguishing factor, they are reported here for the sake of completeness. All patients had been diagnosed by one domain expert and none of the cases had been used in test runs when the system was being developed.

The expert received a form with a short instruction which asked him to write down the diagnoses of each patient in key words with a certainty value. This value could vary between 0 (unknown) to 100 (absolutely certain). This certainty scale was used because EVINCE-I gave certainties within the same range. The data of the 29 patients given to EVINCE-I were entered by the knowledge engineer. All data requested by EVINCE-I were present in the patient records.

3.3 Results

The main diagnoses were: dementia (DEM), dementia of the Alzheimer type (DAT), multiple infarct dementia (MID), depression (DEP) and a residual category (RES). Since the RES category contained too great a variety of diagnoses for a meaningful comparison, this category was dropped from the statistical analysis. Each patient received one or more of the diagnoses mentioned above. The categories of possible diagnosis made by the human expert and the system can be found in Table 3. First it was examined to what extent EVINCE-I and the domain expert were in agreement for all combinations of diagnosis made (the categories 1 through 8 in table 3), by calculating the kappa[16]. The results can be seen in Table 4, test 1.

Cat Test	Diagnosis DEM	DAT	MID	DEP
1	x	x	-	-
2	x	x	-	x
3	x	-	x	-
4	x	-	x	x
5	x	-	-	-
6	x	-	-	x
7	-	-	-	-
8	-	-	-	x

Note: DEM=dementia; DAT=Dementia of Alzheimer's Type; MID=multi-infarct dementia; DEP=depression; x=present; -=absent.

Table 3. Diagnostic categories.

Test	Categories	Obs.	Exp.	Kappa	Z-score
1	1, 2, 3, 4, 5, 6, 7, 8	0.69	0.18	0.62	7.19
2	(1+2),(3+4),(5+6),(7+8)	0.90	0.29	0.85	7.13
3	(1+2+3+4+5+6),(7+8)	1.00	0.55	1.00	4.89
4	(1+2),(3+4+5+6),(7+8)	0.93	0.33	0.90	6.82
5	(1+2+5+6),(3+4),(7+8)	0.93	0.34	0.90	6.78

All Z-scores are significant: $p < 0.0001$
See Table 3 for the meaning of the categories.
Note: Obs. is the observed agreement and Exp. is the agreement expected by chance.

Table 4. Interrater agreement per diagnostic category

The overall level of agreement between the two raters was moderate. The reason that the overall agreement was not high can be found in the categories (7+8) and (3+4), which were supposed to make a distinction between patients diagnosed as depressed or not (See Table 3), but failed to do so adequately. The amount of agreement between the raters increased when the diagnosis DEP was not taken into consideration, as can be seen in test 2 of Table 4.

In test 3, we analyzed the agreement between the two raters on the diagnosis DEM, i.e. categories (1+2+3+4+5+6) and (7+8). This is an important diagnosis to test for, since the diagnosis DAT and MID are dependent on it. Table 4, test 3, shows that both raters showed a perfect agreement on this diagnosis. Since MID and DAT are considered to be mutually exclusive diagnoses in this system, we tested the agreement for the diagnosis DEM with DAT, i.e. categories (1+2), (3+4+5+6) and (7+8), and DEM with MID, i.e. categories (1+2+5+6), (3+4) and (7+8), separately. In both cases, the raters showed a high degree of agreement on their diagnoses. (See Table 4, tests 4 and 5)

The results indicate that the agreement between EVINCE-I and the domain expert is "high" to "perfect" when corrected for the diagnosis DEP. Clearly depression is the only diagnosis that is not handled correctly by EVINCE-I.

After this analysis the ratings, i.e., the certainties about a the diagnoses, were transformed, with normal rounding, to an 11 point (0-10) scale in order to test whether the certainties of EVINCE-I and domain expert agree in direction and relative magnitude. We used the same diagnosis as above, but examined them separately since we were only interested in the individual certainties of the diagnoses. For each diagnosis we computed the Spearman rank correlation coefficient.

Category	Spearman's corr:	DF:	t	Prob:
DEM	0.916	27	11.86	$p < 0.001$
DAT	0.821	27	7.47	$p < 0.001$
MID	0.849	27	8.35	$p < 0.001$
DEP	0.560	27	3.51	$p < 0.002$
N = 29				two sided

Table 5. Interrater agreement using a 11-point certainty scale.

Table 5 shows that the raters displayed a high degree of correlation in their certainty about the diagnostic subjects, except for the diagnosis DEP, which showed a moderate correlation.

Although the RES category was dropped from the statistical analysis, we can still find some clues to the ability of EVINCE-I to diagnose rarer causes of dementia. For example, one patient was diagnosed by EVINCE-I as suffering from MID, but had in

fact Multiple Sclerosis. Since the system knew nothing about MS, MID is a logical near miss. Another patient's diagnosis contained the comment "cave vitamin B_{12} deficiency" from the domain expert. EVINCE-I picked this possible cause up and made the same note in its report.

In two cases, the psychiatrist noted a deviation in the function of the thyroid gland, and EVINCE-I managed to note one of these two cases. Alcohol as a probable cause for dementia was noted in two cases by both the expert and EVINCE-I.

3.4 Discussion

Given the results presented the tests confirm our notion that EVINCE-I can be considered a reasonably good replication of human expertise on the subject matter. The results further show that an ES makes it easier to trace problems in the domain knowledge. Furthermore, the fact that EVINCE-I uses not more than 122 concepts*, which is substantially less than the total data gathered by the human expert, and still is able to produce these results, shows that an ES can be used as a means to assess the contribution of the data used by the expert. Further development of EVINCE-I and a test of these possibilities is currently being performed.

The fact that the correlation between the domain expert and EVINCE-I was high for the categories DEM, DAT and MID, provides evidence that the knowledge represented in EVINCE-I is sufficiently sophisticated to yield valid diagnoses within these three categories. EVINCE-I also seemed to do a good job for the more rarely occurring causes of dementia, when it had the knowledge to detect them. One case with a deviating thyroid gland functioning was missed by EVINCE-I because only the value of the TSH hormone deviated (0.6 mU/l) and not the T4 hormone (11.6 pmol/l), while the rules which assess the functioning of the thyroid glad require a deviation of both hormone levels.**

The only category EVINCE-I performed less well on, is that of depression. In examining the reasons for the discrepancy between EVINCE-I and the domain expert, two possible causes are relevant:

1. EVINCE-I might be relying too heavily on the interpretation of the Hamilton Depression Rating Scale (HDRS)[17]. In an evaluative discussion the domain expert stated that the HDRS-score is used as a measure of severity and not as a diagnostic criterium.
2. Another cause for the low correlation may be due to a rule in EVINCE-I that decides whether or not to investigate depression. It does so only if a patient appears to be, or feel, depressed. This might be the pitfall for patients who do not

---

* The term 'concept' refers to any object that is not a rule, context, form or formula. For example, the concept 'HELP' has no relation to the patient data, but to the user of the program. It is however included in the concept count like many other of such control concepts. As the second version of Evince was devided over 4 modules, the number of concepts is different from the first version which consisted of one large program.

** According to the patient's records a treatment for the thyroid disfunction was administered and three years later both hormone levels were within the normal range.

show obvious signs of depression.
These causes explain why EVINCE-I made 4 false negative and 2 false positive diagnoses in this category. The first cause indicates that the system does not have sufficient criteria, while the second cause indicates that the system rejects investigating the possibility of depression too often. However, there is yet a further possibility, which is typical for the problems in this domain. Since a psychiatric diagnosis (especially depression) can hardly ever be determined with absolute certainty, it could be that the domain expert, and not EVINCE-I, made the wrong diagnoses. An experiment involving several human experts and an improved version of EVINCE-I has been performed to examine this possibility. The data of that experiment are being analyzed now.

Furthermore the domain currently covered by EVINCE-I is expanded to include amnestic and dysthymic disorders. A future version of EVINCE-I will also have to be expanded to incorparate more of the over 50 possible causes of dementia than those presently available (See Table 2). This could lead to a system that provides a standard protocol for clinicians and for researchers who need to select patients with dementia.

4 Concluding Remarks

As stated in the introduction, the EVINCE-I project set out to assess the possibility of developing an ES for the differential diagnosis of demential syndromes. We further wanted to assess the possibility of whether such an ES could be implemented on a personal computer and whether it would yield conclusions that are similar to those of the physician. The results of the assessment presented above strongly confirm our belief that the use of such an ES is feasible and offers promising leads for use in future assessments. An ES on the subject of demential syndromes could - for example - be a tremendous help in [1] mobilizing expert knowledge and make this knowledge accessible, [2] assessing the contribution of data to the reliability of the diagnosis, [3] examining the efficiency of the organization, procedures and decisions in gathering data (i.e. time required, costs involved, inconvenience for the patient, et cetera) [4] investigating the effect of missing data when re-diagnosing patients.

An ES used as a model system could thus offer not only the opportunity to assist in making reliable diagnoses, but also provide a new technique to assess a broad range of subjects in the domain of expertise. At the moment an adapted and expanded version of the ES, EVINCE-II, is being used to assess the topics mentioned above.

5 REFERENCES.

1. Potthoff, P., Rothemund, M., Schwefel, D. et al. Expert Systems in Medicine. International Journal of Technology Assessment in Health Care, Cambridge University Press, 1988, 4, 121-33.
2. Maxmen, J.S. Long-term trends in health care: The post-physician era reconsidered. In D. Schwefel (Ed.), Indicators and trends in health and health care, Berlin: Springer Verlag, 1987, pp 109-115.
3. Schwartz, W.B. Medicine and the computer: The promise and problems of change. New England J of Medicine, 1970, 283, 1257-64.
4. Gurstein, M. Social impacts of selected artificial intelligence applications - The Canadian context. Futures, 1985, 652-71.
5. Waterman DA. A Guide To Expert Systems. Reading, Addison-Wesley P.C., 1986.
6. Jolles J. Cognitive, emotional and behavioral dysfunctions in aging and dementia. In: Swaab DF, Fliers E, Mirmiran M, et al. (Eds), Progress in Brain Research, Amsterdam, Elsevier Science Publishers BV, 1986; 70, 15-39.
7. Marsden CD. Neurological Causes of Dementia other than Alzheimer's Disease. In: Kay & Burrows (Eds), Handbook of Studies on Psychiatry and Old Age, Amsterdam, Elseviers Science Publishers BV, 1984, 145-67.
8. Haase GR. Diseases Presenting as Dementia. In: Wells ChE. (Ed), Dementia, Philadelphia: Davis Comp., 1971, 2nd ed., 27-67.
9. Marsden CD. Assessment of dementia. In: Frederiks (Ed.), Handbook of Clinical Neurology, Vol.2 (46), Neurobehavioural Disorders. Amsterdam, Elseviers Science Publishers BV, 1985, pp 221-232.
10. Hayes-Roth F, Waterman DA, Lenat D. Building expert systems. Reading, Addison-Wesley P.C., 1983.
11. ACQUAINT: User's manual for Acquaint and Acquaint-Light. Manual, Purmerend, Lithp Systems BV, 1987.
12. Tello ER. ACQUAINT. BYTE, Peterborough, McGraw-Hill Inc., 1987, 7, 265-72.
13. American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders. Third Edition, Revised, Washington DC, American Psychiatric Association, 1987.
14. McKhann G, Drachman D, Folstein M, et al. Clinical Diagnosis of Alzheimer's Disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology, 1984, 34, 939-44.
15. Hachinski VC, Illif ZE, Zilka E et al. Cerebral bloodflow in dementia. Archives of Neurology, 1975; 32: 632-637.

16. Schouten JA. Statistical measurement of interobserver agreement. Doctoral thesis, Utrecht, 1985.
17. Hamilton M. Development of a Rating Scale for Primary Depressive Illness. Brit J soc clin Psychol, 1967, 6, 278-96.

VI DIFFERENTIAL DIAGNOSIS OF DEMENTIA: AN EXPERIMENTAL STUDY INTO INTRA- AND INTER-DISCIPLINE AGREEMENT.*

1 INTRODUCTION.

During the past decades, the number of old people in western countries has increased both in absolute numbers and as a proportion of the total population of these countries. Increasingly, interest has been focussed on diseases associated with aging. Dementing conditions are prominent in this respect: recent epidemiological studies indicate that at least 5% of all subjects aged 65 or older suffer from a moderate or severe dementia, and between 2.6% and 52.7% for mild dementia, while the percentage with moderate or severe dementia increases to 20% for subjects over 85 years of age.[1,2,3,4,5]

The need for accurate diagnosis and classification has led to the organization of consensus meetings in different countries.[6] Criteria have been formulated based on the discussions of these consensus meetings. For instance, the criteria published by McKhann et al.[7] after the 1984 meeting organized by the NINCDS-ADRDA are still a cornerstone in the classification of Alzheimer's disease, which accounts for 50 to 80% of all cases of dementia. Unfortunately, no concrete information is available on the actual level of consensus among clinicians involved in the differential diagnosis and classification of dementia. In addition, quantitative information about the influence of these consensus meetings on physician behavior in daily practice is lacking. Furthermore, in the majority of health care institutions, the diagnosis and classification of dementia is done by specialists of one discipline. Since dementia is a condition which is characterized by neurological, psychiatric, psychosocial and somatic aspects, the final diagnosis and classification could be biased by the nature of the medical specialization of the physician. For this reason, it is recommended that a patient thought to suffer from dementia should be examined by a multidisciplinary team of specialists.[8] The purpose of the present study was to provide quantitative information about these points.

The first aim of the study was to assess and compare the level of agreement within several medical disciplines routinely involved in diagnosing dementia. The second aim was to determine whether there were any differences in the use of diagnostic classes between the disciplines. The hypothesis was that such differences exist because of the specialized knowledge of each discipline, which causes specialists to be especially sensitive to certain kinds of information and not for others. The third aim of the study was to assess the influence of a consensus meeting on this level of agreement. To this end, we examined the diagnoses made by clinicians on the basis of ten standard cases.

---

* This chapter was published as: Plugge LA, Verhey FRJ, Van Everdingen JJE, Jolles J. Differential diagnosis of dementia: an experimental study into intra- and interdiscipline agreement. J. Geriatric Psychiatry and Neurology, 1991, 4; 2: 90-97.

2 METHODS

2.1 Subjects

In close cooperation with the Dutch National Organization for Quality Assurance in Hospitals (CBO), an inquiry was organized to gather information on the effect of a consensus meeting on the differential diagnosis of dementia. The aim of the CBO concensus meetings is to improve the quality of differential diagnosis and classification by the formulation of practical guidelines through joint discussion between clinicians from several disciplines.[9] Interest in the subject was so great that not all applicants could attend the meeting. Four hundred and fifty-eight people were registered, representing six disciplines, i.e., neurologists, psychiatrists, general physicians, nursing home physicians, psychologists, and a residual group from other disciplines, such as research and hospital management.

A number of participants considered themselves too inexperienced with the subject and returned the questionnaire; 127 participants completed the preconsensus meeting questionnaire. Of these respondents, 90 completed the postconsensus questionnaire. The data of these 90 respondents was analyzed (Table 1).

Discipline	Participants n (%)	Respondents n (%)
neurologists	100 (22)	24 (27)
psychiatrists	57 (13)	13 (14)
general physicians	74 (16)	12 (13)
nursing home physicians	133 (29)	26 (29)
psychologists	66 (14)	10 (11)
Other	28 (6)	5 (6)
Total	458 (100)	90 (100)

Table 1. Number of Participants and Respondents in Absolute Figures and as a Percentage of Total.

2.2 Materials

Two sets of five case reports, each selected from the records of the Maastricht Memory Clinic, were prepared such that the two sets of five patient case descriptions were comparable in the severity and complexity of the memory disorder. The sequence of the cases in each set was randomized in order to avoid possible order effects (Table 2).

Each case description was presented as a standard letter and contained all information necessary to make a diagnosis in accordance with standard research criteria recommended by the *DSM-III-R* and the NINCDS-ADRDA work group.[7,10] The information was incorporated in the following paragraphs: introduction, medical

Survey	Case	Sex	Age	Diagnosis
1	1	F	74	Moderate dementia, probable DAT, with depressive symptoms
	2	F	80	Moderate dementia, probable DAT; neuroleptic-induced parkinsonism
	3	M	78	Severe dementia, MID; neuroleptic-induced parkinsonism
	4	F	66	Mild dementia, possible DAT or depression-induced dementia
	5	F	71	Slight cognitive deficit (no dementia), history of CVA, adjustment disorder with depressive symptoms
2	6	M	62	Slight cognitive deficit (no dementia), history of TIA or mood disorder
	7	F	80	Severe dementia, probable DAT
	8	F	86	Mild dementia with depression, MID
	9	M	72	Mild dementia, possible DAT, MID, depression.
	10	M	67	Moderate dementia, MID

Note: DAT=Dementia of Alzheimer's type; CVA=cerebrovascular accident; MID=Multi-infarct dementia; TIA=transient ischemic attack.

Table 2. Summary of Diagnoses, Based on the Diagnostic Report With Neurological, Psychiatric and Neuropsychological Information From a Multidisciplinary Expert Committee.

history, history as reported by the patient, history as reported by a partner or close member of the family, psychiatric and neurological history, medical history, medication data, intoxication data, psychosocial data, daily functioning, physical examination, neurological examination, psychiatric examination, blood examination, neuropsychological examination, additional examination (i.e. computed tomographic scan, chest radiograph, electroencephalogram, electrocardiogram).* Two inquiry booklets were assembled containing: an instruction page, a questionnaire, and one set of five case descriptions. Booklet 1 contained the cases 1 through 5 and booklet 2 the cases 6 through 10.

* To guarantee the patient's privacy, any information that might identify the patient was either changed or omitted.

2.3 Inquiry Procedure

One week before the consensus meeting, each of the 458 registered participants was sent a copy of booklet 1 with an introductory letter informing the participant that the questionnaire was concerned with the current status of the clinical diagnosis of dementia. No indication was given that a second questionnaire would be presented after the consensus meeting. To guarantee the privacy of the participants the questionnaires were labeled with numbers assigned by the CBO administration office. In the instruction to booklet 1 each participant was asked to answer questions concerning his or her age, years of medical experience, discipline, job location, hours per week spent on the differential diagnosis and classification of patients with dementia, and whether or not she or he had read the consensus meeting handbook. Finally, after examining the case descriptions, they were asked to write down their diagnoses. In both the intruction form and questionnaire it was indicated that making more than one diagnosis was allowed, by using the Dutch plural/singular style "diagnose(n)". The phrases 'etiologic' and 'syndrome' were not used, to avoid making a direct reference to the consensus meeting and thereby possibly influencing diagnostic usage. The participant was asked to hand in the questionnaire at the registration desk on arrival at the consensus meeting. Of the 458 participants, 127 people handed in their form.

One week after the consensus meeting, booklet 2 was sent to those who had returned the first questionnaire. This time, the questionnaire contained questions about whether or not they had read the consensus meeting handbook, the amount of time spent on reading it, and the diagnoses concerning the cases 6 through 10. A reminder was sent 2 weeks later. The reminder also included a list of the response percentages for each discipline to encourage participants from disciplines with a low response rate to return their form. Of the 127 people who replied to the premeeting questionnaire, 90 responded to the postmeeting inquiry.

Since both the premeeting and postmeeting responses were needed to measure an effect, the participants who had only responded to the premeeting questionnaire were dropped from the analysis. Thus, data from 90 respondents were available for further analysis.

2.4 Classification of Diagnostic Judgements

Since the consensus meeting was also concerned with diagnostic terminology, and the questionnaire data were also to be used for a qualitative analysis of consensus on terminology at a later stage, no instruction was given to the participants as to which terminology or classification should be used. The only instruction was to include all relevant diagnoses in keywords.

To perform a meaningful quantitative analysis of these diagnoses, we constructed a classification system according to the following principles: Diagnoses on the syndrome level and the etiologic level were coded separately. On the syndrome level, the possibilities were: dementia; cognitive disturbances not labeled as dementia; no cognitive disturbances; no statement about cognitive functioning. The etiologic level was classified according to the following causes: primary neurodegenerative;

cerebrovascular; neurological other than cerebrovascular; internal, such as endocrine and/or metabolism; drug induced; depression induced; related to psychosocial factors. Although the term depression is usually adhered to in a syndrome way, for instance in *DSM-III-R*, it was used here in an etiologic sense, i.e., as a possible cause for dementia or cognitive deterioration (cf. "depression-induced dementia"), to avoid use of the term pseudodementia. Diagnostic statements were thus classified at the syndrome and etiologic level.

3 RESULTS

3.1 Subject Characteristics

The number of responses were compared with the actual percentage of participants present during the consensus meeting (Table 1). No significant difference was found between the response percentage per discipline and the actual percentage per discipline present during the consensus meeting (Pearson chi-square = 2.024, $df = 5$, $p > .8$). Because the nonrespondents did not return their form, no additional comparisons could be made. Since discipline category 6 did not represent a coherent clinical discipline, this category was dropped from further analysis, leaving 85 respondents.

The data were then checked for differences between disciplines concerning age, general medical experience, experience in diagnosing dementia, and the reported time spent in reading the handbook (Table 3). This was done to exclude these factors as possible causes of differences.

An analysis of variance showed that there was no difference between the disciplines concerning their age (n = 82, $F = 0.693$, $p > .5$, 3 missing cases), or concerning the number of years of experience in health care (n = 83, $F = 1.843$, $p > .1$, 2 missing cases).

Discipline	N	Mean Age, yr	Mean Healt Care Experience, yr	Mean Hours Reading, hr	Experience in Skewness of Distribution
neurologists	24	43.8	17.1	2.19	0.587
psychiatrists	13	40.3	11.9	2.50	0.529
general physicians	12	39.8	12.1	2.75	-0.422
nursing home physicians	26	40.1	13.1	2.90	0.521
psychologists	10	43.7	13.9	3.39	-1.156

Table 3. Discipline Characteristics.

There was no significant difference between the reported time spent in reading the consensus meeting handbook (ANOVA: n = 83, $F = 1.843$, $p > .1$, 2 missing cases).

To measure the level of experience in diagnosing dementia we used information about the time spent per week on this subject. The answers of the respondents were divided in two range categories: 0 to 3 hours and 4 hours or more per week. A Pearson chi-square test revealed that the discipline category and the time spent on dementia diagnostics were significantly associated (chi-square = 13.762, $df = 4$, $p < .008$). The skewness of the distributions (Table 3) showed that this significant association was due to the psychologists and general physicians who spent more time on dementia diagnostics than the other clinicians.

3.2 Levels of Consensus

As a measure of the degree of agreement, the proportion of the number of pairs for which there was agreement to the possible pairs of assignment was used. The formula to compute this proportion of agreement (S)* was derived from the method used to compute the K coefficient of agreement for k raters and n objects.[11] For each discipline the level of agreement S for each patient on the syndrome diagnoses was calculated. The results are shown in Figure 1 (see the next page).

To test for differences in premeeting and postmeeting consensus for the syndrome diagnoses, a (two sided) sign test was used to compare the level of agreement S of each discipline for the 5 case report pairs, making a total of 30 pairs. This revealed that there was no significant difference between the premeeting and postmeeting level of consensus ($p > .5$). The same was done to test for differences within each discipline. None of the disciplines, however, demonstrated a significant difference between premeeting and postmeeting agreement. The same procedure was followed for the diagnosis at an etiologic level. The results are shown in Figure 2 (see the next page).

The (two sided) Sign Test for differences between premeeting and postmeeting agreement on the etiologic diagnoses did not show significant differences ($p > .9$).

---

* The formula to calculate S for each patient is as follows:

$$S = \frac{1}{k\,(k-1)} \sum_{j=1}^{m} n_j\,(n_j - 1)$$

where k is the number of raters, m is the number of categories, and n_j is the number of raters in category j. The level of consensus (S) for a patient is thus: the proportion of the number of pairs for which there is agreement to the possible pairs of assignments.
For example: k=4 raters judge a patient using m=5 categories. If 2 raters chose category 4 and the two other raters chose category 1 and 2 respectively, then n_1=1, n_2=1, n_3=0 and n_4=2. Using the summation part of the formula above this yields 2. Multiplying 2 with the first part, i.e. 1/12, of the formula yields 2/12, i.e. an agreement of .167.
If all 4 raters would have chosen the same category, then the proportion would have been 12/12, i.e. an agreement of 1.

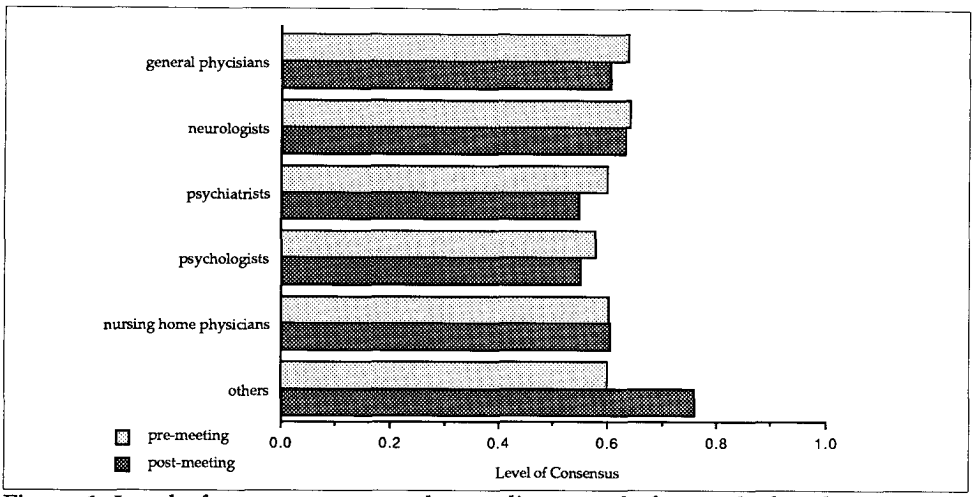

Figure 1. Level of consensus on syndrome diagnoses before and after the consensus meeting.

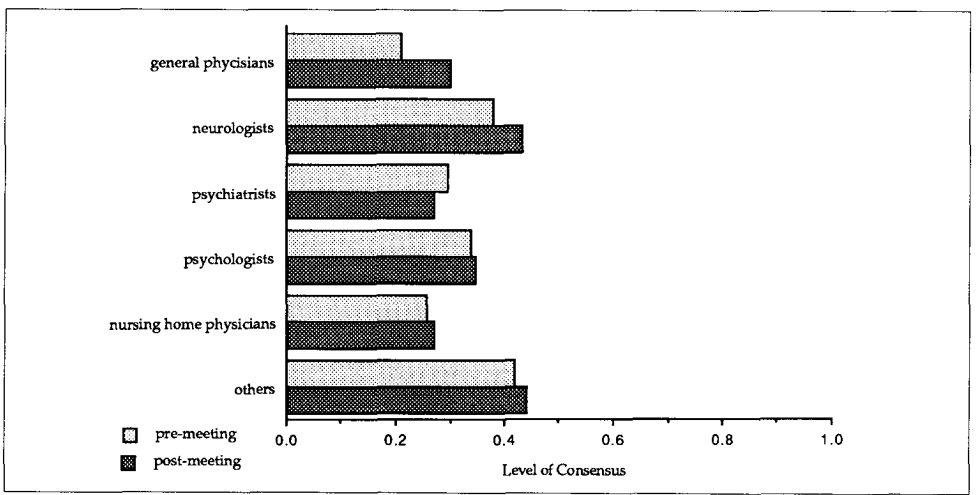

Figure 2. Level of consensus on etiologic diagnoses before and after the consensus meeting.

Differences within each discipline were again tested. None of the disciplines, however, showed a significant difference between premeeting and postmeeting agreement.

A (two sided) sign test showed that the level of consensus for syndrome diagnoses was significantly higher than for the etiology diagnoses ($p < .001$). Since 66% (59 out of 90) of the respondents had (sometimes partly) read the consensus meeting handbook before the meeting and before they filled in the first questionnaire, we analyzed the level of agreement between the readers and non-readers: no significant differences were found.

In summary, we found: (1) no significant differences existed between the disciplines concerning health care experience, age, or time invested in reading the syllabus; (2) general physicians and psychologists spent significantly more time on dementia diagnostics than physicians of the other disciplines; (3) there was no significant difference between or within the discipline categories on premeeting and postmeeting level of consensus for both syndrome and etiologic diagnoses; and (4) there was significantly more consensus on syndrome than on etiologic diagnoses.

3.3 Differences in Diagnostic Classification

We were also interested in the difference between disciplines concerning the nature of their diagnosis. The hypothesis was that some disciplines would show a special interest in particular information, resulting in a preference for certain diagnoses. The miscellaneous discipline category was dropped from the analysis, since that group could not meaningfully be compared with a discipline in daily practice.

The Waller-Duncan K-ratio t test was used to analyze the data.[12] This test makes it possible to compare means of several groups at the same time, while minimizing the Bayes risk under additive loss. This means that a correction is made, using Bayes Theorem, for the increased chance of finding significant differences due to multiple comparisons.* For the syndrome diagnoses, the results showed that psychiatrists and nursing home physicians differed from neurologists in that they made a diagnosis without further specification significantly more often. (Table 4, see the next page)

* A Waller grouping expresses the ordered differences found by the Waller-Duncan K-ratio t-test. For example, in Table 4, the mean score of the neurologists (letter B) differs from that of the psychiatrists (letter A), but not from the mean score of the psychologists (letter A and B).

Discipline	N	Mean	Waller Grouping
psychiatrists	13	1.923	A
nursing home physicians	26	1.846	A
general physicians	12	1.500	A B
psychologists	10	1.200	A B
neurologists	24	0.625	B

K ratio=100, df=80, MSE=1.632, F=3.25, T=2.102
(Means with the same letter are not significantly different)

Table 4. Differences in the use of syndrome diagnoses without specification

For the diagnosis Alzheimer's type dementia the situation was almost reversed. This diagnosis was made significantly more often by neurologists than by any of the other physicians, with the largest difference being between the neurologists and the psychiatrists (Table 5).

Discipline	N	Mean	Waller Grouping
neurologists	24	3.667	A
psychologists	10	2.400	B
general physicians	12	2.250	B
nursing home physicians	26	2.154	B
psychiatrists	13	1.846	B

K ratio=100, df=80, MSE=2.638, F=3.951 T=2.068
(Means with the same letter are not significantly different)

Table 5. Differences in the use of the Diagnosis Alzheimer's Type Dementia.

Diagnoses with medication and/or intoxication as etiology were also used differently by physicians of the various disciplines. General physicians and nursing home physicians made this diagnosis significantly more often than psychologists and psychiatrists, while the neurologists held a middle position (Table 6).

Discipline	N	Mean	Waller Grouping
general physicians	12	1.333	A
nursing home physicians	26	1.154	A
neurologists	24	0.792	A B
psychologists	10	0.400	B
psychiatrists	13	0.385	B

K ratio=100, df=80, MSE=0.794, F=3.179 T=2.159
(Means with the same letter are not significantly different)

Table 6. Differences in the use of the diagnosis Medication-/Intoxication

The diagnosis depression was used significantly more often by psychiatrists than by neurologists, while the other disciplines occupied a middle position (Table 7).
For the diagnoses cerebrovascular, neurologic, somatic, and psychosocial disorders, no significant differences were found between the disciplines. However, none of the psychologists made diagnoses in the somatic category.

Discipline	N	Mean	Waller Grouping
psychiatrists	13	2.923	A
general physicians	12	2.500	A B
nursing home physicians	26	2.192	A B
psychologists	10	2.100	A B
neurologists	24	1.708	B

K ratio=100, df=80, MSE=1.523, F=2.262 T=2.332
(Means with the same letter are not significantly different)

Table 7. Differences in the use of the diagnosis depression

4 Discussion

The present study addressed questions concerning (1) the level of agreement or consensus between medical disciplines in differential diagnosis of dementia; (2) the differences in the use of diagnostic classes on the syndrome and etiologic level; and (3) the effect of a consensus meeting in obtaining consensus. With respect to the first question, the quantitative data demonstrate that it was significantly more difficult to reach agreement on an etiologic than on a syndrome level. However, it should be noted that the differences between the syndrome and etiologic consensus can be partly explained by the number of possible categories to choose from. For each syndrome, there are several etiologic explanations, thus making it possible, by chance alone, to have more different diagnoses where etiology is concerned. Moreover, the data clearly show that the disciplines did not differ in their level of agreement, but that the agreement within each discipline was based on different diagnoses on both the syndrome and etiologic level. This finding is important because it is the first time that quantitative differences in differential diagnostics between disciplines have been reported. Furthermore, it gives additional impetus to the necessity of improving diagnostic competence in the field of dementia. Apart from the organization of consensus meetings, continuing education seems obligatory.

There were striking differences between the disciplines with regard to the second question. It is apparent that the various disciplines focus upon different aspects of the dementia syndrome and its multidimensional facets. Psychiatrists and nursing home physicians gave diagnoses at the syndrome level significantly more often than neurologists did, i.e., diagnoses without further specification of the etiology. However, when psychiatrists gave a cause for cognitive dysfunctions, they used the diagnosis "depression" significantly more often than neurologists. Neurologists, however, used the diagnosis "Alzheimer's disease" significantly more often than physicians of the other disciplines. General physicians and nursing home physicians considered medication or intoxication as a cause of cognitive disorders significantly more often than psychiatrists and psychologists did.

These findings are of importance because of their probable cause: it is likely that the diagnoses primarily reflect the medical specialization of the diagnostician, and thus his or her medical knowledge and experience. For example, depression is of special interest to psychiatrists, while neurologists are more interested in Alzheimer's disease. A similar finding was reported by Lopez et al.[13], who also found that neurologists and psychiatrists varied in their interpretation of standardized patient data, e.g. "(...) differing interpretation of the significance of symptoms, differing importance given to comorbid conditions, or differing interpretations of the diagnostic criteria"[(p.1521)].

It should be noted that no opinion is given in this study as to which diagnoses were correct: as mentioned above, the different specialists appear to focus upon different aspects of the multidimensional condition. A follow-up study addresses this question more specifically. Furthermore, it should be mentioned that the disciplines did not differ on all etiologic diagnoses. The etiologic diagnoses on which there was agreement were: cerebral vascular disorders, neurologic disorders other than cerebrovascular, somatic disorders (with the exception of the psychologists who never

used this diagnosis) and psychosocial disorders. The differences concerned the frequency of the use of syndrome diagnoses without further specification, and the frequency of the use of the following etiologic diagnoses: Alzheimer's disease, medication and intoxication, and depression.

With respect to the third question addressed by this study: a significant difference between the level of diagnostic agreement before and after the consensus meeting could not be established. It should be stressed that this does not mean that consensus meetings are without value. It is possible that this particular consensus meeting was not effective, or that the time interval of 2 weeks (1 week before and 1 week after the meeting) might have been too short to change the use of diagnostic criteria and jargon. However, it does not seem probable that the lack of significant effects was caused by the fact that several respondents had already read part of the consensus handbook, because no significant differences were found between those respondents who read the syllabus before the meeting and those who did not.

There are several implications of the findings presented here. The differences reported in this study confirm the notion that a multidisciplinary committee is needed to avoid the diagnosis and classification of dementia from being biased by the medical specialization of the diagnostician. The fact that clinicians within the same discipline tend to agree with each other on diagnostic classes makes it imperative that physicians from other disciplines are asked for a second opinion to reduce the chance of overlooking or underestimating evidence which might suggest other diagnoses. This means that the present referral policy should be changed such that a multidisciplinary approach will become regular practice.[14] Incidentally, the reemergence of neuropsychiatry as a specialty between neurology and psychiatry might also be of relevance in this respect.[15] An additional implication is that continuing education in the field of dementia might be necessary. It is especially important to teach medical specialists about those aspects of the multifaceted conditions of dementia that are not usually taken into consideration in their own specialty. This notion will be investigated in a follow-up study which will examine the effect of using the expert system EVINCE as a diagnostic aid in education and clinical practice.[16] Lastly, the fact that the consensus meeting had no measurable effect does not mean that consensus meetings are without value. As stated before, the lack of significant results could be because of the short time interval between the meeting and the second questionnaire, a long-term follow-up might reveal changes in consensus, if it is possible to control interfering factors. Consensus meetings undoubtedly play an important role in the integration of disciplinary expertise about complex medical controversies. This, in turn might increase the knowledge of the participants so that they are capable of formulating clinically sound criteria for the assessment of the multidimensional condition of dementia, and subsequently improve their own clinical management.

5 References

1. Henderson AS: The coming epidemic of dementia. Australian and New Zealand Journal of Psychiatry, 1983; 17: 117-127.
2. Bergmann K, Kay DWK, Foster EM, et al.: A follow-up study of randomly selected community residents to assess the effects of chronic brain syndrome and vascular disease. In: New Prospects in the Study of Mental Disorders in Old Age. Proceedings of the Vth World Congress of Psychiatry, Mexico. Amsterdam, Exerpta Medica, 1971, pp 856-865.
3. Kaneko Z: Epidemiological studies on mental disorders of the aged in Japan. In Proceedings of the 8th International Congress of Gerontology: Vol. 1, Abstracts of Symposia and Lectures. Washington D.C., International Association of Gerontology, 1969, pp 284-287.
4. Kaneko Z: Care in Japan. In: Howells JG. (Ed), Modern Perspectives in the Psychiatry of Old Age. New York, Brunner/Mazel, 1975, pp 519-530.
5. Committee Health Care Future Scenarios (Stuurgroep Toekomstscenario's Gezondheidszorg): Ouder worden in de toekomst 1984-2000. Utrecht, Jan van Arkel, 1985.
6. Consensus Conference on Dementia: Differential diagnosis of Dementing Diseases. Journal American Medical Association. 1987, 258: 3411-6.
7. McKhann G, Drachman D, Folstein M, et al.: Clinical Diagnosis of Alzheimer's Disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology. 1984, 34: 939-44.
8. Health Counsel, (Gezondheidsraad): Psychogeriaterische ziektebeelden. 's-Gravenhage, Gezondheidsraad, 1988.
9. Dutch National Organization for Quality Assurance in Hospitals, (Centraal Begeleidingsorgaan voor de Intercollegiale Toetsing, C.B.O.): CONSENSUS: informatie over de opzet van door het C.B.O. georganiseerde consensusbijeenkomsten. Utrecht, National Organization for Quality Assurance in Hospitals, 1985.
10. American Psychiatric Association: Diagnostic and statistical manual of mental disorders. Organic Mental Syndromes and Disorders. 3rd Edition, Revised. Washington DC, American Psychiatric Association, 1987.
11. Siegel S, Castellan NJ: Nonparametric Statistics for the Behavioral Sciences. 2nd Edition, New York, McGraw-Hill Book Company, 1988.
12. SAS Institute Inc: SAS/STAT Guide for Personal Computers. 6th Edition, Cary, NC, SAS Institute Inc, 1985.
13. Lopez OL, Swihart AA, Becker JT, et al. Reliability of NINCDS-ADRDA criteria for the diagnosis of Alzheimer's disease. Neurology, 1990; 40: 1517-1522.

14. Dutch Consensus Development Conference: Diagnosis of the Dementia Syndrome. Utrecht, C.B.O. National Organization for Quality Assurance in Hospitals, 1985.
15. Stuart CY, Hales RE: The Reemergence of Neuropsychiatry: Definition and Direction. Journal of Neuropsychiatry and Clinical Neurosciences. 1989, 1; 1: 1-6.
16. Plugge LA, Verhey FRJ, Jolles J: A Desk-Top Expert System for the Differential Diagnosis of Dementia: An evaluation study. International Journal of Technology Assessment in Health Care. 1990, 6; 1: 147-156.

VII DIFFERENTIAL DIAGNOSIS OF DEMENTIA: A COMPARISON BETWEEN THE EXPERT SYSTEM EVINCE AND CLINICIANS.*

1 INTRODUCTION

The diagnosis of dementia and dementing diseases is based on neurological and psychiatric findings, but is usually made by physicians from one discipline. In a previous study, we showed that such a monodisciplinary approach had a significant effect on the type of diagnoses made.[1] We compared the diagnoses of neurologists, psychiatrists, nursing home physicians, general physicians and psychologists, and found that neurologists made the diagnosis of Alzheimer's Disease more frequently than clinicians of any of the other disciplines. Consistent with, for example, DSM-III-R guidelines, psychiatrists used the diagnosis of depression more often than neurologists, while the other disciplines took a middle position.[2] Furthermore, psychiatrists and nursing home physicians more often made a syndrome diagnosis without an etiologic diagnosis than did neurologists.

These findings are consistent with the results reported by Hoffman[3] and by Verhey et al.[4], who all found that a thorough multidisciplinary neuropsychiatric examination resulted in a therapeutically important alteration in the referring diagnosis (41% and 45% respectively). The reemergence of neuropsychiatry is an important step towards a solution for such discipline-related diagnostic biases. However, there are more complicating factors in the diagnosis of dementia. Although international criteria have been developed to improve the consensus on the definitions of dementia, these criteria are only slowly being applied, if at all, in daily practice. Furthermore, there is a lack of consensus on the selection and weighing of data, and on the examination procedure. For instance, in the NINCDS/ADRDA criteria no consensus could be reached about the selection of the neuropsychological methods to be used.[5]

One way to improve dementia diagnostics and classification is to use computer controlled medical protocols to gather the data.[6] However, these programs are usually sophisticated data bases with a disadvantage: the data set requested by the program is always the same, irrespective of patient characteristics. Moreover, these programs do not incorporate knowledge on data integration, detection of data inconsistencies or data relevance.

In a review of computer-based decision aids, Morelli et al.[7] compared five prominent decision-making paradigms: data bank analysis, statistical pattern recognition, Bayesian analysis, the logical flow chart, and knowledge-based expert systems. In their conclusions they stated that "the expert system approach appears to

---

* This chapter was published in: Plugge LA, Verhey FRJ, Jolles J. Differential Diagnosis of Dementia: A Comparison Between the Expert System Evince and Clinicians. Journal of Neuropsychiatry and Clinical Neurosciences, 1991; 3, 4: 398-404.

be the most promising. Its main strengths are 1) the ability to incorporate different kinds of knowledge into the decision-making process, 2) the ability to mimic the way humans reason about a problem, 3) the ability to explain and justify the system's conclusions...".[7(p. 166)]

Based on these considerations the neuropsychiatric expert system Evince was developed. Evince is based on international rules and criteria for dementia diagnostics as described in the DSM-III-R and proposed by the NINCDS-ADRDA Work Group.[2,5] A neuropsychiatrist (the domain expert) -the second author of this paper- provided the expertise in applying these rules and criteria. A first evaluation showed that the diagnoses produced by Evince showed a high level of agreement with those made by the domain expert.[8]

After this first evaluation Evince was developed further. To test this expanded version of Evince, an evaluation experiment was set up involving a multidisciplinary committee of three expert clinicians and 85 clinicians from 5 different disciplines. The multidisciplinary expert committee (MEC) provided diagnoses as a reference for comparison with the diagnoses of Evince and of the 85 clinicians. The hypothesis was that Evince would make more correct diagnoses (i.e., the number of diagnoses that are in agreement with those of the MEC) than the average clinician taking part in the experiment. Thus, the performance levels of the experts and the 85 clinicians would provide an upper and a lower limit to decide whether the expert system's performance was acceptable.

2 METHODS

2.1 Subjects

The subjects were participants in a consensus meeting on the differential diagnosis of dementia organized by the Dutch National Organization for Quality Assurance in Hospitals (CBO) in the Netherlands, in November 1988. Each of the 458 registered participants was asked to cooperate in an inquiry concerning the present state of dementia diagnostics. As the inquiry data were to be used in a more extensive study on the use of classification in dementia diagnostics, the inquiry was devided in two parts - one before and one after the consensus meeting. Of the 458 participants, 127 people handed in their first form, and of these 127, 90 filled out and returned their second inquiry form.

Based on information provided by the participants for the registration office of the CBO, 85 respondents represented 5 disciplinary categories: 1) neurologists, 2) psychiatrists, 3) general physicians, 4) nursing home physicians, and 5) psychologists, leaving a residual category of five respondents that was dropped from the analysis. The data from the 85 respondents were used in this study.

2.2 Materials

Ten case descriptions were selected from the patient records of the Maastricht Memory Clinic. The cases were selected so that both classical and more complex cases were present, and with different levels of severity of the cognitive -or memory-disorder.

Each case description was formulated in terms of a standard medical report and contained all information necessary to make a diagnosis according to standard research criteria recommended by the DSM-III-R and the NINCDS/ADRDA work group.[2,5] This information was incorporated in the following paragraphs: 1) introduction, 2) past history, 3) anamnesis, 4) anamnesis as reported by a partner or a close member of the family, 5) psychiatric and neurological history, 6) medical history, 7) medication data, 8) intoxication data, 9) psychosocial data, 10) daily functioning, 11) physical examination, 12) neurological examination, 13) psychiatric examination, 14) blood examination, 15) neuropsychological examination, 16) additional examination (e.g., CT-scan, chest x-ray, EEG, or ECG). To guarantee the patient's privacy, any information that might identify the patient was either changed or omitted. Each participant was asked to answer questions concerning his or her age, years of medical experience, discipline, nature of medical practice, hours per week spent on differential diagnosis and classification of patients suspected of suffering from dementia. Finally, after examining the case descriptions, the participants were asked to write down their diagnoses on the form.

2.3 Multidisciplinary Expert Committee.

To establish a reference for comparison of the diagnoses, an independent multidisciplinary committee of three expert clinicians was established, consisting of a neurologist, a psychiatrist, and a psychologist. Each member of the MEC received the same 10 patient case description and was asked to study the data and formulate diagnoses. The MEC was then given the opportunity to discuss these diagnoses in a joint conference to reach a consensus on the final diagnoses. The MEC was asked to state the diagnoses at both the syndrome and the etiologic level. The conference lasted aproximately 4 hours. The MEC reached a consensus for all patients, except for the etiologic diagnoses of patients number 4 and 6. (See Tabel 1 on the next page.)

2.4 Classification of Diagnostic Judgement

The consensus meeting was also concerned with diagnostic terminology; the inquiry data were going to be used for qualitative analysis of agreement on terminology at a later stage. No instruction was given to the participants as to which terminology or classification they should use. The only instruction given was to include all relevant diagnoses in key words.

To perform a meaningful quantitative analysis of these diagnoses, we constructed a classification system according to the following principles: Diagnoses on the syndrome level and the etiologic level were coded separately. On the syndrome level, the possibilities included 1) dementia, 2) cognitive disturbances not termed dementia,

			MEC		Evince	
Case	Sex	Age	Syndrome	Etiology	Syndrome	Etiology
1	F	74	Moderate dementia, depressive symptoms	Probable AD	Mild dementia, mild depression	AD
2	F	80	Moderate dementia	Probable AD, neuroleptic-induced parkinsonism	Moderate dementia	AD
3	M	78	Severe dementia.	MID, neuroleptic-induced parkinsonism	Moderate dementia	MID
4	F	66	Mild dementia	Possible AD, or Major Depression	Mild dementia	Major depression (possibly medication-induced), bereavement
5	F	71	Slight cognitive deficit (no dementia)	History of CVA, adjustment disorder with depressive symptoms	Cognitive deficit (no dementia)	Vascular problems, dysthymic disorder
6	M	62	Slight cognitive deficit (no dementia)	History of TIA, or mood disorder	Cognitive deficit (no dementia)	Vascular problems
7	F	80	Severe dementia	Probable AD	Severe dementia	AD
8	F	86	Mild dementia with depression	MID	Mild dementia	MID
9	M	72	Mild dementia	Possible AD, MID, depression	Mild dementia	AD, depression
10	M	67	Moderate dementia	MID	Moderate dementia	MID

Note: AD=Alzheimer's disease; CVA=cerebrovascular accident; MID=Multi-infarct dementia; TIA=transient ischemic attack.

Table 1. Summary of diagnoses made by the Multidisciplinary Expert Committee and the Expert System Evince.

3) no cognitive disturbances, and 4) no statement about cognitive functioning. The etiologic level was classified according to the following causes: 1) primary neurodegenerative, 2) cerebrovascular, 3) neurological other than 2, 4) internal, such as endocrine and/or metabolism, 5) drug induced, 6) depression-induced, 7) related to psychosocial factors. Although the term *depression* usually is used to mean a syndrome (for instance, in DSM-III-R), it is used here in an etiologic sense, i.e. as a possible cause for dementia or cognitive deterioration (cf., "depression-induced dementia"). This makes it possible to avoid the term *pseudodementia*. Thus, the diagnostic statements of the 85 clinicians, the MEC and Evince were classified at the syndrome and etiologic level.

3 The Expert System Evince.

Evince was developed with the expert system tool Acquaint.[9] The minimum requirements for Evince are as follows: an IBM PC-compatible microcomputer (preferably an AT), with a minimum of 520 Kbytes of RAM (640 Kbytes of RAM is recommended), two floppy drives (a hard-disk drive is recommended), and MS-DOS or PC-DOS version 2.1 or later. Evince can be used with a monochrome or a color monitor.

Evince is a package consisting of the actual program, i.e., the inference engine and user interface, and three knowledge modules. In Module 1 the user can decide to consult Evince in batch mode (i.e., let Evince diagnose the preentered data of one or more patients) or in interactive mode (i.e., the system asks questions and the user provides the answers). The Modules 2 and 3 consist of the procedural knowledge depicted in Figure 1. These two modules embody 28 contexts, 110 rules and 129 concepts. Additionally, these two modules use 143 formulas, i.e., functions which perform calculations, window and file management, etc. Because the knowledge modules are separate from the actual program, they can also be stored separately (e.g. on a network) to prevent unauthorized use.

Knowledge in Evince is represented in rule and concept frames. Concept frames represent the knowledge of patient data, while the rules embody the procedural knowledge. When the value of a concept is unknown to the system and required by a rule, it will either be inferred whenever possible or requested from the user. A rule consists of an IF and a THEN part. The IF part compares the concept value with the test value, and triggers the THEN part of the rule when the values match. Each concept value has a certainty attached to it; for example "maybe" = 50, and "unknown" = 0. These values range between -100 (absolutely false) and 100 (absolutely true) and are used to assess with how much certainty a conclusion can be drawn. Additionally, each conclusion has a "certainty" value that informs the system how certain that conclusion is when the premises are absolutely true. Consequently, the lower the certainty of the premises, the lower the certainty of the conclusion.

As stated in the introduction, Evince was developed further on the basis of the results of a previous experiment. The knowledge base was extended to include knowledge about disorders that can be related to medication.[10] The diagnostic

capabilities on the subject of depression were refined with DSM-III-R criteria, and were extended to include the diagnosis "dysthymic disorder". The knowledge about dementia was reorganized to investigate the level of deterioration, which then would lead to the diagnosis of dementia or amnesic syndrome. This resulted in a hierarchical examination protocol consisting of diagnostic topics called "contexts" (See Figure 1, page 43).

Each context governs a set of rules and sometimes one or more subcontexts (or child contexts). A context is in fact a higher order rule that establishes whether it is worth checking its subordinate rules and contexts. If the system considers a context irrelevant, then all the subordinate rules and contexts will also be considered irrelevant. For example, the context "Mood" tries to establish whether there is reason to assume that the patient is depressed. If this is not the case, then all child contexts will be ignored. However, some contexts will always be examined because they concern information that is considered relevant (i.e., the contexts "Registration", "Medication", and "Lab-Tests"). An other context that always is used is the "Evaluation" context. In this context, all the diagnostic information gathered from the previous contexts is collected, checked for inconsistencies and mutual consequences, and finally transformed into natural language statements. These statements are then printed in the form of a report. The last context is called "End" and gives the user the opportunity to ask the system how the presented conclusions were reached. It should be noted that such questions can also be asked during the actual (interactive) consultation.

4 Results

Table 2 (see the next page) shows the number and percentage of respondents compared with those present during the consensus meeting. None of the disciplines within the respondents was over- or under-represented in comparison with the number of participants per discipline.

The MEC reached a consensus on the syndrome diagnosis for all 10 patients. However, the MEC did not reach a consensus on the etiologic diagnosis for the patients 4 and 6. Two etiologic diagnoses were given for these patient (Table 1). A diagnosis made by the clinicians or Evince was considered "correct" if the diagnosis was made by the MEC (Table 1). Thus, 10 points could be scored for the syndrome, and 10 points for the etiologic diagnoses. The diagnoses were not compared for the level of severity of the demential syndrome, or for the use of the classifications "possible" and "probable" for Alzheimer's disease because these frequently were omitted by the 85 clinicians. Comparison of the syndrome diagnoses of the 85 clinicians with the MEC revealed that they had a mean±SD of 7.6±1.4 correct diagnoses (n=85; range 5-10). No differences were found between the disciplines (See Table 3).

discipline category	No. of participants		No. of respondents	
neurologists	100	(22%)	24	(27%)
psychiatrists	57	(13%)	13	(14%)
general physicians	74	(16%)	12	(13%)
nursing home physicians	133	(29%)	26	(29%)
psychologists	66	(14%)	10	(11%)
Other*	28	(6%)	5	(6%)
Total	458	(100%)	90	(100%)

* The residual category Others was dropped from the analyses, leaving a total of 85 respondents.

Table 2. Number of participants and respondents in absolute figures and percentage of total.

Discipline category	N	Mean	sd	range
neurologists	24	7.833	1.31	5-10
psychologists	10	7.667	1.37	5-10
general physicians	12	7.654	1.33	5-10
nursing home physicians	26	7.300	1.83	5-10
psychiatrists	13	7.100	1.44	5-9

Table 3. Mean number of syndrome diagnoses in agreement with the MEC.

Discipline category	N	Mean	sd	Waller Grouping		
neurologists	24	6.208	1.38	A		
psychologists	10	5.900	1.52	A	B	
general physicians	12	4.917	1.38		B	C
nursing home physicians	26	4.885	1.68		B	C
psychiatrists	13	4.615	1.94			C

K-ratio=100, df=80, MSE=2.518, F=3.53, T=2.11

Table 4. Mean number of etiologic diagnoses in agreement with the MEC.

Evince diagnosed all 10 case descriptions correctly, i.e., in agreement with the MEC. However, when the level of severity of the demential syndrome was taken into account, it was revealed that Evince considered patients 1 and 3 to be less severely demented than did the MEC (i.e., mild vs. moderate). Furthermore, Evince did not provide a level of severity for cognitive deficits (patients 5 and 6) because it was not designed to do so. (See the two right columns of Table 1.) On the etiologic level, the 85 clinicians reached a mean of 5.3±1.7 correct diagnoses (n=85; range 1-8). Evince, however, made all 10 etiologic diagnoses in agreement with the expert committee (Table 1). However, Evince did not use the classification "possible" and "probable", as did the expert committee. (The present new version incorporates rules to enable the use of such a distinction.) Furthermore, in contrast to the MEC panel, Evince did not make the diagnosis multi-infarct dementia for patient number 9. The MEC decided that this diagnosis should not be excluded completely, given the finding of a small hypodensity on the CT-scan. In accordance with international consensus, Evince did not consider the CT-scan finding alone sufficient to make the diagnosis multi-infarct dementia, since the patient's history and examination did not reveal a cerebrovascular accident.

With respect to the etiologic diagnoses, a significant difference was found between the disciplines concerning the number of etiologic diagnoses that agreed with the expert committee. Neurologists had significantly more correct etiologic diagnoses than clinicians of the other disciplines, except for psychologists (see Table 4).

5 Discussion.

The results showed that the average clinician made fewer etiologic diagnoses than syndrome diagnoses that were in agreement with the MEC. This can be explained partly by the fact that there are fewer choices in the latter. Thus, it is possible to have a lower score for the etiologic diagnoses by chance. However, the diagnostic performance of Evince was not affected by these differences in chance. Furthermore, the average score of the clinicians for both syndrome and etiologic diagnoses was considerably lower than the score of Evince. Although the diagnosis of Alzheimer's type dementia is predominantly made on the basis of exclusion criteria, there are positive criteria, such as an insidious onset and the absence of a clouded consciousness. Therefore, it was possible for Evince to reject the diagnosis Alzheimer's type dementia, even when none of the other diagnoses were applicable, i.e., Alzheimer's type dementia was not treated as a 'waste bucket' diagnosis.

Another important finding is the disciplinary difference found in the number of diagnoses that were in agreement with the MEC, specifically the difference between psychiatrists and neurologists. Both neurologists and psychologists had significantly more etiologic diagnoses in agreement with the MEC than did psychiatrists. Our previous study[1] on interdisciplinary differences revealed that psychiatrists more often made a syndrome diagnosis without an etiologic diagnosis, in contrast to the neurologists. This difference between neurologists and psychiatrists was also seen in the present study. The possibility that this difference is due to level of experience

and/or involvement in dementia diagnostics could not be established. There was no significant difference between disciplines concerning their health care experience. A significant difference was found concerning the time spent on dementia diagnostics (chi-square=13.76, df=5, p<.008); psychologists and general physicians spent more time on dementia diagnostics than the other disciplines. However, this does not explain the discrepancy between neurologists and psychiatrists. A more plausible cause of interdisciplinary difference would be the fact that the 85 clinicians were not given the opportunity to discuss the cases to develop multidisciplinary consensus. However, this monodisciplinary approach does not deviate from what is common practice at present. The question as to whether the difference between neurologists and psychiatrists might be due to the nature of the medical specialization and to the clinicians' experience, is the subject of a forthcoming study.

The number of patients used in this evaluation, can be considered relatively small in comparison with the (few) other ES evaluations. However, compared with medical comparison studies, this number is quite acceptable, particularly given the number of observers involved.[11,12,13] In order to obtain the voluntary cooperation of clinicians outside the research institution, the number of patients to be diagnosed must be kept to a methodologically acceptable minimum.

Although most clinicians are familiar with written case reports as an alternative to seeing the real patient, it is possible that this has had a negative influence on their results. However, it should be noted that the standardized patient information was given to both the 85 clinicians and the MEC panel. As Lopez et al. have remarked, this ensures a reduction of variance stemming from the patients and the clinicians.[14] Another possible cause for the low etiologic performance of the clinicians might be the rather low response rate. Although 85 (67%) out of 127 clinicians responded to both the first and the second inquiry, they account for only 18% of the total number of participants. However, it should be noted that the participants who responded both times considered themselves competent, while others returned an empty form with the remark that they were too inexperienced. A more plausible cause for the low average score can be found in a studies by Chimowitz et al.[15],Voytovich et al.[16] and Friedlander et al.[17,18] who found that diagnostic errors are related to reasoning mechanism such as "premature closure" (i.e. premature diagnostic conclusions) and "anchoring" (i.e. adhere to an initial hypothesis despite subsequent contradicting or inconsistent evidence).

Although the diagnoses made by the MEC could not be compared with postmortem and/or long-term follow-up data, it is thought that these diagnoses were reliable, because the three clinicians involved were recognized experts in their discipline, and they had ample opportunity to discuss each case thoroughly. As mentioned in the introduction, the overall results of this experiment are in agreement with those described by Hoffman[3] and Verhey et al.[4], who assessed referral diagnoses of behavioral disorders with a multidisciplinary team.

The aim of comparing Evince with human clinicians was to assess the performance of the system, not to show that an expert system can replace the human clinician. Although expert systems are able to mimic human reasoning, they (still) lack important human capabilities (such as intuition), as well as the vast amount of world

knowledge. However, the results warrant the conclusion that the implemented neurological, psychiatric and psychological knowledge was successfully applied by Evince with the material presented, and that the system can assist clinicians in diagnosing dementia. As Hoffman observed: "it is clear that the techniques of neuropsychiatric diagnosis have currently advanced to the point where a major limitation exists in their knowledge of application...".[3(p.967)] However, as Teitelbaum noted, we cannot expect clinicians to be experts in all disciplines.[19] Nevertheless, they should be able to recognize situations that demand referral and collaboration. A computer-based decision aid like Evince could help to achieve this. It provides the human clinician the opportunity to apply up-to-date knowledge of internationally approved standard criteria, and it can help to make the clinician aware of the decision process involved and provide a useful check for completeness of the necessary information. Furthermore, it offers the clinician a tool to integrate expert diagnostic information from different disciplines.

Although it is unlikely that expert systems will replace the human clinician, due to the limitations noted earlier, it is to be expected that, in the near future, expert systems will play an important role in assisting medical diagnostics.[20,21] However, before expert systems are given such an important function in medical diagnostics, their performance should be thoroughly tested, both in laboratory and in field situations. Therefore, collaborative studies have been initiated to compare the diagnoses made by Evince both retrospectively and prospectively with clinical and post mortem diagnoses, and to test the system in field situations. Finally, the authors welcome other proposals for collaborative studies, particularly from outside the Netherlands.

6 REFERENCES.

1. Plugge LA, Verhey FRJ, Van Everdingen JJE, Jolles J. Differential diagnosis of dementia: an experimental study into intra- and interdiscipline agreement. Journal of Geriatric Psychiatry and Neurology, 1991, 4; 2: 90-97.
2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 3rd Edition, Revised. Washington, DC, American Psychiatric Association, 1987, p.106.
3. Hoffman RS. Diagnostic errors in the evaluation of behavioral disorders. JAMA, 1982, 248; 8: 964-967.
4. Verhey FRJ, Vreeling FW, Jolles J. DSM-III and NINCDS/ADRDA criteria for dementia and Alzheimer's disease: impact of diagnostic procedures on daily practice. In: Wurtman J. (Ed.): Alzheimer's Disease. Proceedings of the Fifth Meeting of the International Study Group on the Pharmacology of Memory Disorders Associated With Aging, Zürich. Cambridge, MA, Center for Brain Sciences and Metabolism Charitable Trust, 1989; pp 419-423.
5. McKhann G, Drachman D, Folstein M, et al. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under auspices of Department of Health and Human Services Task Force on Alzheimer's

Disease. Neurology, 1984; 34: 939-944.

6. Roth M, Tym E, Mountjoy CQ, Huppert FA, Hendrie H, Verma S, Goddard R. CAMDEX: A standardised instrument for the diagnosis of Mental Disorder in the Elderly with special reference to the early detection of dementia. Br J Psychiatry, 1986, 149: 698-709.
7. Morelli RA, Bronzino JD, Goethe JW. Expert systems in psychiatry: a review. Journal of Medical Systems, 1987, 11; 2/3: 157-168.
8. Plugge LA, Verhey FRJ, Jolles J. A desk-top expert system for the differential diagnosis of dementia: an evaluation study. International Journal of Technology Assessment in Health Care, 1990, 6; 1: 147-156.
9. ACQUAINT: User's manual for Acquaint and Acquaint-Light. Manual, Purmerend, the Netherlands, Lithp Systems BV, 1987.
10. World Health Organization. Drugs for the elderly. Copenhagen, World Health Organization Regional Office for Europe, 1985.
11. Shinar D, Gross R, Hier DB, et al. Interobserver reliability in the interpretation of computed tomographic scans of stroke patients. Archives of Neurology, 1987; 44: 149-155.
12. Lee D, Fox A, Vineula F, et al. Interobserver variation in computed tomography of the brain. Archives of Neurology, 1987; 44: 30-31.
13. Burke WJ, Miller PhM, Rubin EH, et al. Reliability of the Washington University Clinical Dementia Rating. Archives of Neurology, 1988; 45: 31-32.
14. Lopez OL, Swihart AA, Becker JT, et al: Reliability of NINCDS-ADRDA clinical criteria for the diagnosis of Alzheimer's disease. Neurology, 1990, 40: 1517-1522.
15. Chimowitz MI, Logigian EL, Caplan LR. The accuracy of bedside neurologic diagnoses. Annals of Neurology, 1990; 28: 78-85.
16. Voytovich AE, Rippey RM, Suffredini A. Premature conclusions in diagnostic reasoning. Journal of Medical Education, 1985; 60: 302-307.
17. Friedlander ML, Phillips SD. Preventing anchoring errors in clinical judgement. J Consult Clin Psychol, 1984; 52: 366-371.
18. Tversky A, Kahneman D. Judgement under uncertainty: heuristics and biases. Science 1974; 185: 1124-1131.
19. Teitelbaum ML. Toward Better Integration of Medical and Psychiatric Care. JAMA, 1982, 248; 8: 977.
20. Maxmen JS. Long-Term Trends in Health Care: The Post-Physician Era Reconsidered. In: D. Schwefel (Ed.), Indicators and Trends in Health and Health Care, Springer-Verlag, Heidelberg, 1987, pp. 109-115.
21. Potthoff P, Rothemund M, Schwefel D, Engelbrecht R, Van Eimeren W. Expert Systems in Medicine: Possible Future Effects. International Journal of Technology Assessment in Health Care, 1988; 4: 121-133.

VIII DISCUSSION.

1 INTRODUCTION.

In the preceding chapters a description was given of the development and the phased laboratory evaluation of a neuropsychiatric expert system for the diagnosis of dementia, along with an assessment of the agreement among clinicians and a comparison of their judgements with a panel of human experts. The result of the first evaluation showed that the expert system performed at a level comparable to the performance of the domain expert whose knowledge was used to develop the system (Chapter V). The second experiment showed that the diagnosis of dementia is hampered by discipline related disagreements between clinicians (Chapter VI). In the third experiment the expert system showed capable of an interdisciplinary performance comparable to that of a panel of three external (from other institutions) human experts from different disciplines and better than that of 85 clinicians from several disciplines (Chapter VII). In this last experiment it was also shown that Evince can handle not only the classic cases, but also notoriously difficult cases, such as those where a choice has to be made between an early stage of dementia or benign memory complaints. This is especially important, because clinicians will benefit more from a system that is able to aid in diagnosing difficult cases than in diagnosing easier classic cases. Additionally, Evince proved capable of handling the notoriously difficult problem of a depression coinciding with memory complaints.

The implication of these results is that neither the multi-disciplinary approach, nor - as suggested by Morelli et al. and Werner[1,2]- the lack of hard physiological criteria and the predominant use of descriptive criteria, is necessarily an obstacle for the development of an expert system. The positive results also show that the performance of a medical expert system using relatively inexpensive commercial software and a simple desk-top IBM PC-XT compatible computer should not a priori be dismissed as inferior to systems developed with the aid of expensive soft- and hardware using megabytes of memory and powerful processors.

However, not unlike its human counterparts, Evince is not perfect. Several aspects of the results, our approach and the problems encountered deserve additional discussion which is provided in the remainder of this chapter.

2 TWO ALTERNATIVES TO THE EXPERT SYSTEM APPROACH

In chapter I a short overview was given of expert systems in general and previous attempts to use this tool in psychiatry. Chapter I did not discuss other possible ways of providing the clinician with computerized support, because there is a large array of such support programs, ranging from straightforward billing programs up to AI applications, and because this study was aimed at the development and evaluation of an expert system. However, the expert system approach for medical decision support is not free from criticism, and has led to the development of alternative decision

2.1 The Catalyst Model

One major criticism of the ES approach is that it resembles the classical Greek oracles like the one in the city of Delphi where the god Apollo could be asked for advice.[3] This metaphor depicts the ES as a kind of 'black box' that treats the physician as an ignorant person who is required to transfer all of his available knowledge to the system and wait for the machine to speak. The solution proposed by Miller and Masarie is a "catalyst model" that "can facilitate the physician overcoming the difficult limiting step."[3(p.2)] In this model the physician decides which diagnostic steps to take and which steps require diagnostic support. The computer program keeps track of the "physicians deliberations by providing decision support tools for selected steps in the physician's diagnostic reasoning."[3(p.2)] Although this catalyst approach seems intuitive, there are some problems. The Greek Oracle allegory as a characterization of expert systems, has two important flaws. Firstly, Greek Oracles typically gave cryptic and ambiguous answers and did not provide an explanation of their answer. Even though it can not be excluded that an answer from an expert system, nor any other decision aiding system, can be ambiguous, this is normally not the case with expert systems. Secondly, Greek Oracles did not require the consulting person to be knowledgeable on the problem matter. ESs on the other hand do require that the user has professional knowledge of the subject at hand, albeit that it does not require the user to be an expert. In one of their studies Bankowitz et al. explicitly state that physicians must be familiar with the limitations of the system, in their case the Quick Medical Reference (QMR), and must be capable of overriding inappropriate suggestions.[4] This capability applies equally for users of expert systems.

An important problem with the Catalyst model is that it requires the physician to know when he should consult the system, i.e. when his knowledge is inadequate. This means that the system will not be consulted when the physician erroneously thinks he made a correct decision. Due to discipline related biases as reported in chapter VI, such a situation is bound to occur.

2.2 The Critiquing Approach

A second alternative approach is the use of a critiquing system. In this approach the system requests the physician to enter his plan (or diagnosis) and reasons for these plans, which are then commented on by the system. This technique resembles the catalyst approach in that it uses only the data given on the physician's initiative. Therefore, such a system only evaluates the data considered relevant by the physician.[5] As shown in this study, disciplinary differences alone are sufficient reason to ask for more data than those considered relevant by the consulted physician. Furthermore, the amount of knowledge that a critiquing model needs seems seriously underestimated, as it must be able to keep track of the physician's deliberations, which entails more than simply knowing which step the physician is going to take next. It means that the system must have knowledge of the physician's policy and rationale, and be able to trace the physician's deliberations backward to check previous assumptions and decisions.

3 EVALUATION: PROBLEMS ENCOUNTERED AND SOLUTIONS CHOSEN

At the beginning (early 1987) of the project described in this study, evaluation of decision-support systems was only scarcely debated. However, during the last few years the subject of decision-support systems has gained much more interest, as evidenced by recently published special issues of, for example, Methods of Information in Medicine (1990, vol 29, nr 4) and Medical Informatics (1990, vol 15, nr 3). It is also encouraging to see that ideas about evaluation methods that were developed independent from each other have very much in common and complement each other, for example, the method suggested by Wyatt and Spiegelhalter[6] and the one used in this study. In the following sections we will elaborate on some aspects of the expert system evaluation method used in this study.

3.1 Level of performance

In previous evaluation studies of medical expert systems the developers tried to assess the system's performance by comparing it with expert clinicians, or with clinicians without providing information as to how experienced these clinicians were. However, as expert systems are generally not meant to be used by human experts, it is of the utmost importance to find out what the minimum performance of a decision-support system must be in order to be an aid to the non-expert. To the best of our knowledge, this is the first study where the performance of an expert system was explicitly compared to that of both experts and domain-competent non-experts in order to establish an upper and lower limit of performance for determining the potential use of the system. How well the 85 clinicians performed in comparison with the expert panel and, consequently, with Evince, is shown in Figures 1 and 2 on the next page.
Even on the syndrome diagnoses, where the 85 clinicians showed a higher level of agreement with the expert panel, the performance of Evince is better than that of the majority of the clinicians (possible reasons for this proficiency will be discussed further on in this chapter).

3.2 The patient cases used in the experiments

In this study cases from the institution that developed the expert system were used for the evaluation experiments. Although this is common practice at the moment, it would be better to have a more diverse source. For this reason, the development of a reference library containing standardized cases from different countries would be desirable. However, the development of such a reference library is a major project that demands large scale cooperation to establish some standardization before such a library can be developed, a task which was far beyond the possibilities within this study.[7]

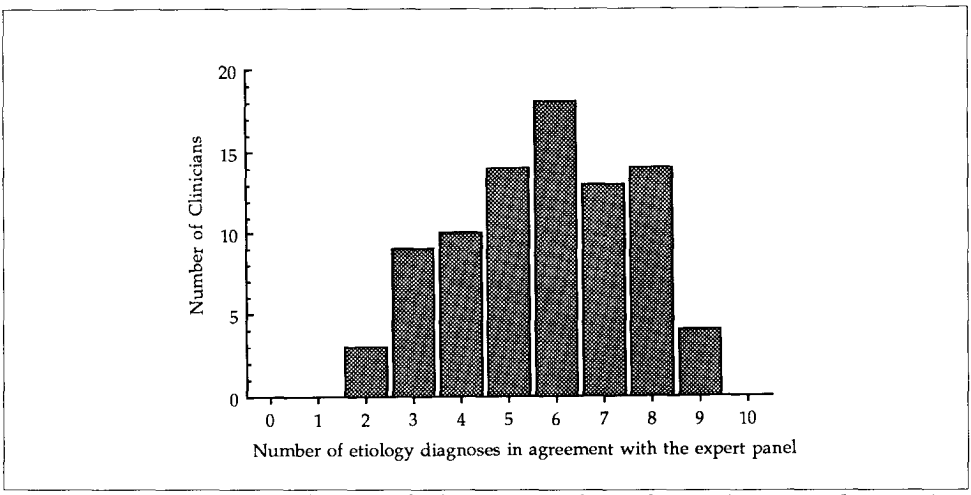

Figure 1. Frequency Distribution of Clinicians with Etiology Diagnoses that are in Agreement with the Expert Panel.

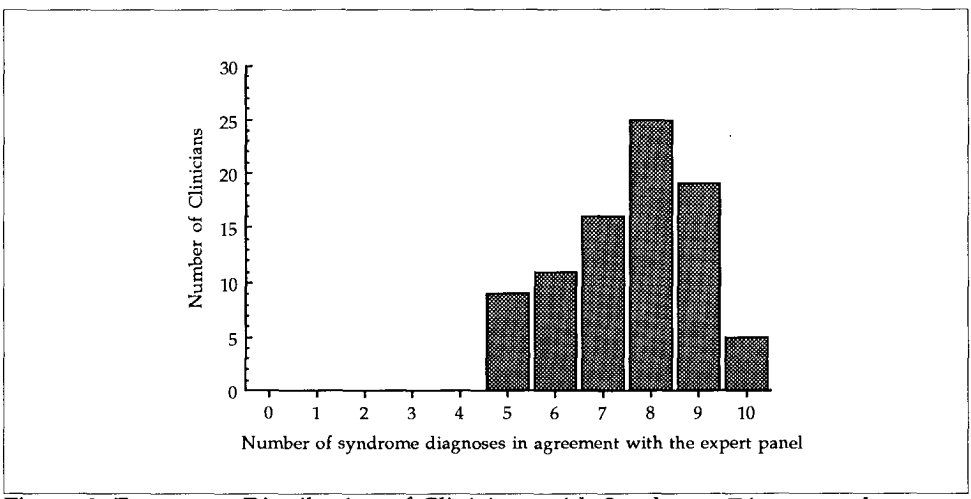

Figure 2. Frequency Distribution of Clinicians with Syndrome Diagnoses that are in Agreement with the Expert Panel.

The number of cases, and the variation in the cases, is another major problem. Due to the large number of parameters used in dementia diagnostics and in the expert system it is impossible to cover every possible combination of these parameters by creating or selecting patient cases. Furthermore, the number of cases that could be

used in the second evaluation experiment was limited due to the number of people that were asked to voluntary cooperate in diagnosing these cases, i.e., the participants of the consensus meeting. The higher the number of cases used, the less likely it would be that a large number of people would have cooperated due to the amount of time required to judge these cases. Given these limiting factors, the second evaluation experiment used stratified random cases as a practical alternative during the laboratory testing stage. In this second evaluation experiment the cases were stratified according to severity of the cognitive disorder, i.e. two cases of severe dementia, two cases with a cognitive deficit (no dementia), three moderate and three mild cases of dementia. Furthermore, cases with the following etiologies were selected: Alzheimer's Dementia, Multi-infarct Dementia and depression. The individual selection of the cases was random. The aforementioned stratification was chosen to provide a plausible representation of cases that can be encountered in medical practice. As we quoted before (Chapter II, paragraph 2.2.3) from O'Keefe et al., "The issue is not the number of cases, it is the *coverage* of the test cases".[8(p.83)] Given this axiom, the practical problem of the number of clinicians involved, the care taken in the selection of the cases and the number of cases used in other medical comparison and evaluation studies, the number of cases used in this study is acceptable.[5,9,10,11,12,13]

3.3 Comparison of the diagnoses

Another problem concerns the comparison of the diagnoses provided by the expert system, the clinicians and the expert panel. In this study the comparison was performed by the developers (not the panel of three external experts). In the ideal situation it would have been desirable to have an independent institution perform the comparison if it were possible to make that third party unaware about which diagnoses were produced by a human and which by the system. Unfortunately, the limited prose capabilities of our system (and, for that matter, almost any other system) made it very implausible that such a third party could have been successfully blinded for the source of the diagnoses. Given these problems, we share Shortliffe's opinion that researchers should be able to evaluate their own system up to the laboratory stage, more so, because it is common practice for researchers in other fields to perform their own analyses on the data they gathered.[14] Furthermore, this approach does not preclude the possibility to have independent raters at a later stage in the field evaluation.

4 Possible Reasons for the Proficiency of Evince

Compared to other medical expert systems the performance of Evince is remarkably good, and it is tempting to ascribe this to the system. However, it is important to see whether there are other factors that can explain the positive results. Prior to this discussion we would like to stress that much of the following discussion is speculative. Further experimentation will be necessary to test the suggested hypotheses.

4.1 Distinguishing between clinicians

As noted before, this is -to the best of our knowledge- the first study where an explicit difference was made between expert clinicians and domain competent non-experts. The reason for this differentiation is that the expert system is supposed to -at least- approach the proficiency of human experts and surpass -at least- the average (domain competent) non-expert in order to be beneficial to this group. Apart from the fact that it is more realistic to compare the system's performance with both experts and the average clinician, it provides a less demanding lower boundary of performance. This means that Evince's proficiency can partly be explained by the fact that a difference was made between experts and domain competent non-experts. Consequently, it is likely that other medical expert systems will show a better performance if they were reevaluated using this method.

4.2 The multidisciplinary versus the monodisciplinary approach

The multidisciplinary knowledge of Evince could be another reason for its proficiency because Evince was compared with clinicians who diagnosed the patient cases individually (except for the expert panel). The finding in Chapter VI that clinicians from different disciplines focus on different etiologies, makes it plausible that the interpretation of the patient data was discipline dependent. Consequently, it is very well possible that the performance of the 85 clinicians would have been better if they were given the chance to discuss the cases with colleagues from other disciplines. However, this possibility also emphasizes one of the themes of this study, namely, that the diagnosis of dementia requires a multidisciplinary approach. This is reflected in the knowledge represented in Evince which is derived from international criteria such as provided by the DSM-III-R, the NINCDS-ADRDA work group, the Hamilton Depression Rating Scale, the Hachinski Ischaemic Index, etc. Unfortunately, such a multidisciplinary approach is far from common practice because it is time consuming and requires clinicians from several disciplines to be available for a joint conference. In this respect the data about the performance of the clinicians provide a realistic picture of daily practice. It is exactly at this point that Evince could provide the individual clinician assistance: interpretation of data attained through additional examinations, using knowledge from other relevant disciplines. We would like to stress that Evince cannot aid, for example, a neurologist by conducting a complete psychiatric examination, administering and judging a CT-scan, or by analyzing a blood sample. Clinician will still be dependent on colleagues from other disciplines to conduct certain examinations. However, Evince can help the clinician to interpret data resulting from such examination requests. A field evaluation in which clinicians with and without the aid of Evince are compared will have to be performed to test this assumption. Additionally, it will be interesting to see how well a team of randomly selected clinicians perform in comparison to an expert panel, when they are given the chance to discuss patient cases.

4.3 The diagnostic spectrum

The number of possible diagnoses, i.e., the diagnostic spectrum, is another potential source of influence on the results. As the number of diagnoses that is available to Evince is smaller than the number available to the average clinician who's knowledge covers more than dementia, it is possible that this caused the large diversity among the 85 clinicians and, consequently, a lower performance in comparison with Evince. However, there are a few objections to make against this argument. Firstly, like the expert panel, the clinicians were informed that the inquiry was about the present status of dementia diagnostics. Furthermore, the patient case descriptions all provided a paragraph about the patient's cognitive complaints. Thus, the clinicians were appropriately primed to the problem domain. Secondly, if the diagnostic spectrum was a problem, i.e., if diagnoses outside the domain were given, then it is likely that this also would have occurred within the expert panel, which it did not. Thirdly, the only problem with the diagnostic spectrum that did occur was within the limits of the domain of dementia. For example, further analysis of the data by Verhey et al.[15] showed that clinicians displayed less agreement about the diagnoses of patients with low levels of deterioration (as measured by the Global Deterioration Scale score[16]), than for patients with a high level of deterioration. However, clinicians are commonly expected to be able to deal with these types of patients. If the clinicians would have had problems eliminating diagnoses outside the problem domain, then the present status of dementia diagnostics in daily practice would have been even more serious than expected. The finding that the diagnostic spectrum of dementia itself is difficult for the average clinician was conform our expectations, and exactly the reason why Evince was developed.

4.4 Human errors

A much more obvious source of errors leading to a better performance of Evince are typical human errors such as the incorrect use of criteria, neglecting criteria, overlooking data, incorrect interpretation of data, applying an incorrect procedure in analyzing the data, and anchoring (see also Chapter VII, paragraph 5). Which errors are made most often and which have the largest impact cannot be determined on the basis of our data as we were unable to record the diagnostic reasoning of the 85 clinicians. A completely different study would be required to cover that subject.

4.5 The sample of 85 clinicians

The problem with the sample of 85 clinicians is whether it is a true representation of all clinicians involved in dementia diagnostics. There were at least three decisions made by the clinicians that influenced the composition of the sample: 1) the clinician's decision to take part in the consensus meeting, 2) the decision to take part in the first inquiry, and 3) the decision to take part in the second inquiry. The clinicians could have made the first decision because they either felt certain or uncertain about their diagnostic knowledge concerning dementia, or simply because they acknowledged that there is a need for more consensus about dementia diagnostics. It should be

stressed that the meeting was not a course in dementia diagnostics, but a conference where criteria, procedures and recent findings were debated, led by a panel of experts from all over the country. This makes it plausible that the attending clinicians were knowledgeable on the subject matter. In other words, *if* there was a sampling bias, then it is likely that this bias resulted in a more favourable diagnostic performance. However, it cannot be totally excluded that some attenders were not sufficiently qualified given the fact that a few inquiry forms were returned with a notice from the clinician that he did not feel competent enough to participate. As for the latter two decisions, i.e., the decision to participate in the first and the second inquiry, we can only gues what the reasons of the participants could have been, for example, to help science, or to test their competence.

Unfortunately, the problem of selection bias cannot be ruled out in any experiment that relies on the use of volunteers. This problem is bound to reoccur in field testing, when institutions and individuals volunteer to cooperate in such an experiment. The best solution would be to have an exhaustive sample, i.e., testing the entire population. However, even if that were possible, this still leaves the question open whether the results will be representative for clinical practice in other countries.[17] The second best solution is to compare known population characteristics (if available) with the characteristics of the sample to test whether the sample can be considered to be a good represention of the population.

4.6 The knowledge representation

The type of knowledge representation, i.e, rule based, could be considered another possible reason for the good performance of Evince, meaning that rule based system would provide a better representation of the medical knowledge in this domain. However, this seems to us a poor argument for the explanation of the performance of Evince. Stating that a rule based knowledge representation leads to a better performance of this expert system is like stating that programming language X leads to better programs than programming language Y. The decision to use a rule based system was based on the assumption that this type of knowledge representation would be easier to understand for the potential user, not because the knowledge would be easier to represent. Furthermore, Acquaint is not a pure rule based expert system shell, but a hybrid system.

More important than *how* knowledge is represented, is *what* is represented. If the knowledge of the system is inadequate, then the performance of the system will be inadequate, no matter how it was represented internally. This does not mean that knowledge representation is not an important issue in expert system development. In this domain the use of frames and rules were convenient, because part of the documented knowledge, for example, the DSM-III-R, was available in a rule-like form, while the frames provided a good tool for representing the diagnostic concepts.

Like programming languages, types of knowledge representation should be judged on their adequacy to solve the problem. An adequate type of knowledge representation can lead to a system with a good performance, but it is certainly not a guarantee for succes.

5 Field Evaluation

Although Evince passed the formal laboratory tests with results well above average, a field evaluation still has to be performed. Preparations for such a study were made, but revealed several new problems left to be worked out. For example, one of the problems to perform a field experiment is to find a suitable institution for the experiments. Not every institution offers suitable opportunities for a field study, such as sufficient and consistent availability of time and personnel, and complete (or reasonably complete) patient records. Another problem encountered was that the available patient records did not contain all the information required by Evince, because the data were not gathered by the institution according to a standard procedure. Furthermore, some of the data in the patient records were gathered using clinical tests that were different from those used by Evince. This means that some of the concepts had a different definition, making it difficult to decide whether the results could be used by Evince. Temporal data were a problem for methodological reasons. In some cases the available data were derived from examinations that were performed at different moments and at different institutions, causing recent and relative old data to be intertwined in the records. Although Evince can reason with incomplete data, it limits the performance of the system, and devaluates its utility. Evince was designed to use international criteria and procedures for dementia diagnostics, to help the clinician to standardize the examination protocol and to use the data derived using this protocol, not to interpret arbitrary data from any available patient record.

The alternative would be to perform a prospective study using our standard examination procedure. However, such an experiment requires a large effort from the cooperating institutions and is expensive, because diagnostic procedures customary in the institution would have to be performed in parallel with the experimental approach in order to avoid disruption of the daily routine in the institution.[18] This problem illustrates the important dilemma between acceptability of costs and scientific credibility as recently noted by Shortliffe.[14] Still, we believe that a field experiment should be performed because there are many parameters involved in the routine use of an ES that are difficult to test in a laboratory, like user characteristics, geographical differences, organizational routine, familiarity with information systems, etc.

However, despite these problems there were a few preliminary positive findings. As was expected, there were hardly any problems in operating the system. It turned out that the users became rapidly familiar with the available menus and function keys. This ease of use was noted many times on other occasions when clinicians were allowed to freely use the system. Even clinicians who were not familiar with computers were able to operate the system after some brief instructions.

6 Integration of Expert Systems with other information systems

One of the most important future issues to be solved is the integration of ESs with other information systems.[19] The most central part of these information systems is undoubtedly the patient database. Integration of information systems would mean that the same database can be used for different purposes.[20] Unfortunately, the information in databases is still not standardized in such a way that it can be used by different information systems. Different institution use different coding systems or free text, administer different tests, use different checklists, and even have different opinions about the concepts used.[5] This means that existing databases cannot be used by, for example, an ES that was developed by another institution without modifying the database's content. This incompatibility poses an important limitation to the widespread use of ESs and also indicates that there are serious shortcomings in the development of medical consensus and its subsequent use in daily practice.

The best solution to this problem is a rigorous standardization of the data used by information systems, in particular the items stored in databases. If this can be accomplished, then an ES (or any other information system) can be built without the need to augment or reorganize existing information. Without standardized information it will be unlikely that ESs will or even *can* be used on a large scale.

7 Self Learning Medical Expert Systems

One of the many important properties ESs lack in comparison with humans, is the ability to learn, i.e., machine learning. Although progress has been made in this respect, the learning ability of computer systems is very limited.[21] Apart from the question whether it is possible to design a system that is capable of learning, there is the question whether it is desirable. As mentioned before, one of the advantages of ESs is their consistent use of diagnostic criteria, without respect to the location of their use. However, a self learning system discovers relations and creates rules on the basis of the cases presented and is allowed to use these rules after acceptance by the user(s). Consequently, this leads to systems that differ from location to location, for example, due to cultural and geographical differences between patient populations. This would result in a situation resembling the present with geographic, inter-disciplinary and even intra-disciplinary differences in the use of clinical concepts, diagnostic criteria and procedures. To avoid such local dependency self learning ESs would have to be provided with material, either in the form of raw data from specially developed reference databases, or in the form of explicit rules from professional and/or scientific forums. Given these problems, it seems more beneficial to have ESs that do not learn by themselves, at least not when they are being used in daily practice. Another consequence of a self learning, or locally educated system is that it will be difficult to evaluate, because its knowledge changes continuously, i.e., the system becomes unpredictable. It seems unlikely that clinicians will accept an ES that behaves differently depending on which patients have been examined with the system.

However, a self learning mechanism can prove useful for the development of ESs

and for research as a tool for automated discovery, much like experiments and statistics are being used.[22] For example, machine learning could be used to discover relations between data, or to assess the importance of symptoms and signs. In this respect, the use of artificial neural networks is very promising, although such systems are truely black boxes because they cannot provide any explanation for the decisions taken.[23,24,25]

8 The possibilities and limitations of Evince and Expert Systems in general.

At present computerized equipment and computer software have penetrated health care in almost every medical discipline and at many stages of patient care. Information systems are replacing traditional paper patient records, other programs calculate drug dosage or monitor vital functions of patients in intensive care. Undoubtedly the computer will continue to pervade health care. In this respect ES technology is only one of the latest examples of computer technology to enter health care and, similar to the introduction of computerized billing applications during the 1960s, medical personnel will increasingly rely on computerized decision aids. Although there are still many technical and methodological difficulties in developing and evaluating computerized decision aids (even in such seemingly simple applications as computerized Admission-Discharge-Transfer or billing systems[26]) and although there are important questions about the legal implications of computerized health care, there are also important advantages.[27,28,29]

For ESs in general, and Evince in specific, the main advantage lies in their ability to integrate knowledge from several disciplines, their consistent approach to diagnostics, and their ability to use all available data and recent diagnostic knowledge. In addition, the use of an ES makes the diagnostic results more comparable between institutions, even more than in the case where paper protocols are used, because not only the tests and procedures are standardized, but also the interpretation of the resulting data. In contrast, the average clinician's knowledge is not always up to date, sometimes shows inconsistent reasoning, and does not always use the diagnostic criteria consistently (c.f. Chimowitz et al.[30]). This is especially true for a problem domain where the required knowledge comes from several disciplines while the individual clinician is specialized in one discipline, as is usually the case in dementia diagnostics.[31] This does not mean that clinicians can or even should be replaced by ESs. Apart from the (important) question whether it is desirable to have a computer taking care of the interaction with the patients, ESs lack too many properties and too much knowledge to be able to replace the human clinician. Therefore, Evince could never replace the clinician. Furthermore, Evince cannot perform the actual medical examinations, these will still have to be administered by human clinicians. Unfortunately, this is also the weak spot of Evince: if the clinician makes mistakes in the medical examination than it is very likely that the resulting diagnoses will also be erroneous, i.e. 'garbage in, garbage out'.

However, an ES like Evince can play an important *additional* role in medical diagnosis, and apart from being an important aid in diagnostics, there are also other positive

effects. For example, the use of Evince -or any ES, for that matter- can produce a checklist effect, which helps the clinician not to forget essential topics of examination. Also the carry-over effect can be benificial, asuming that errors of the ES do not pass unnoticed. This means that the clinician starts learning how the ES performs the diagnostics and adopts the criteria and approach of the system. A possible consequence of this learning mechanism is that the clinician no longer needs the ES after using it for a longer period, or, more dangerously, thinks he no longer needs the ES. Another possible advantage of ESs (and more generally computerized decision aids) is that it requires a clear definition of knowledge. As Van Bemmel noted, this endeavor by itself might reduce the need for decision support methods.[32]

Next to these important advantages we would like to point out a possible danger of uniform diagnostic protocols due to a widespread use of ESs. To some extent, the use of alternative procedures can provide insight in other and new aspects of a disease. Therefore, a uniform approach can be a serious limitation to the development of new diagnostic procedures. However, care should be taken that this argument is not used to justify a panacea of protocols, because the limitation of a uniform procedure can be overcome by having an ES that uses different knowledge on the same domain, and by the fact that clinical research institutions will still have to rely heavily on human resources. This means that ESs can be compared with other standardized medical technologies. The use of non standardized methods would then be restricted to research institutions for experimentation.

When approximately 41% to 45% of the patients who are examined for memory complaints receive a wrong or partially wrong diagnosis (as reported by Hoffman[33] and Verhey et al.[34]), then this will result in inappropriate therapy or referral. Consequently, the use of an ES could have an important impact in the improvement of health care and a concomitant reduction of health care costs.

Next to the role ESs can play directly in health care, they can also have an important indirect impact through their use in research, where there is a strong need for a uniform means of measurement, for example, to select patients for a clinical trial. Up till now, the procedure was to develop a standard selection procedure which was then checked for inter- and intrarater reliability, because the interpretation of the data gathered is subject to variation. With the use of ESs only the inter- and intrarater variability due to differences in the actual examination will have to be checked. This could result in an important reduction of the costs and an increased reliability of the experimental results.

Given the previously mentioned carry-over effect, ESs can also play a role in medical education, although they will have to be adapted to provide the student user much more (educational) information than is usually the case in ESs developed for routine medical practice.[35] A major advantage of ESs as an educational tool is that the expert knowledge is more explicitly available and that the student is able to manipulate the data to see what the effect is on the diagnosis. However, as in routine medical practice, the finesses of medical education will still require the aid of human clinicians. Finally, the use of ESs in education could help to avoid that potential users view such systems (and other information systems) as if they were Greek oracles and help them to understand both the benefits and the risks of ESs, or AI products in general[36].

9 CLOSING COMMENTARY

As Hellman[37] noted, "Probably the three major societal forces which will impact which health professionals will be in demand in the year 2000 are the graying of America, the economics of health care, and the over supply of MDs and their changing role". Without doubt, this statement is applicable to the entire western civilization, including the Netherlands. Hellman further notes, that this will create a demand for professionals to take care of this increasing part of our population, i.e., the elderly, and an increased demand on financial resources. Although the total cost of the health care sector in the Netherlands has stabilized in recent years, there is a clear change of policy to emphasize the "limits to growth" in health care.[38] This means that health care will have to be made more efficient, for example by more accurate diagnoses to reduce the occurrence of incorrect referrals and to increase the adequacy of therapies. However, due to the limited financial resources, it is unlikely that this efficiency will be realized by employing more (expensive) health care professionals, which means that less people will have to take care of more patients. ES, and information technology in general, will undoubtedly play an important role in the endeavor to solve these problems in the (near) future.

10 REFERENCES

1. Morelli RA, Bronzino JD, Goethe JW. Expert Systems in Psychiatry. Journal of Medical Systems. 1987, 11; 2/3: 157-168.
2. Werner G. Methuselah: An Expert System for Diagnosis in Geriatric Psychiatry. Computers and Biomedical Research, 1987; 20: 477-488.
3. Miller RA, Masarie FE jr. The Demise of the "Greek Oracle" Model for Medical Diagnostic Systems. Editorial. Methods of Information in Medicine, 1990, 29; 1: 1-2.
4. Bankowitz RA, McNeil MA, Challinor SM, Parker RC, Kapoor WN, Miller RA. A computer-assisted medical diagnostic consultation service: Implementation and prospective evaluation of a prototype. Annals of Internal Medicine, 1989, 110; 10: 824-832.
5. Van der Lei, J. Critiquing based on computer-stored medical records. Thesis. Rotterdam, 1991.
6. Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Medical Informatics. 1990, 15; 3: 205-217.
7. Willems JL, Arnaud P, Van Bemmel JH, Degani R, Macfarlane PW, Zywietz Chr, (for the CSE Working Party). Common Standards for Quantitative Electrocardiography: Goals and Main Results. Methods of Information in Medicine, 1990, 29; 4: 263-271.
8. O'Keefe RM, Balci O, Smith EP. Validating Expert System Performance. IEEE Expert, Winter 1987: 81-90.
9. Shinar D, Gross R, Hier DB, et al. Interobserver reliability in the interpretation of computed tomographic scans of stroke patients. Archives of Neurology, 1987; 44: 149-155.
10. Lee D, Fox A, Vineula F, et al. Interobserver variation in computed tomography of the brain. Archives of Neurology, 1987; 44: 30-31.
11. Burke WJ, Miller PhM, Rubin EH, et al. Reliability of the Washington University Clinical Dementia Rating. Archives of Neurology, 1988; 45: 31-32.
12. Rothschild MA, Swett HA, Fisher PR, Weltin GG, Miller PL. Exploring subjective vs. objective issues in the validation of computer-based critiquing advice. Computer Methods and Programs in Biomedicine. 1990, 31: 11-18.
13. Kors JA, Sittig AC, Van Bemmel JH. The Delphi Method to Validate Diagnostic Knowledge in Computerized ECG Interpretation. Methods of Information in Medicine. 1990, 29; 1: 44-50.
14. Wyatt J. Stanford AI in Medicine Workshop, March 1990. Methods of Information in Medicine. 1991, 30; 1: 65-67.
15. Verhey FRJ, Plugge LA, J.J.E. van Everdingen, Jolles J. Verschillende disciplines, verschillende diagnosen? - een enquête onder de deelnemers van de consensusvergadering over dementie. Tijdschrift voor Gerontologie en

Geriatrie, 1991; 22: 187-194.

16. Reisberg B, Ferris SH, de Leon MJ, et al. The Global Deterioration Scale (GDS): an instrument for the assessment of primary degenerative dementia (PDD). American Journal of Psychiatry, 1982; 139: 1136-1139.
17. Nolan J, McNair P, Brender J. Factors influencing the transferability of medical decision support systems. International Journal of Biomedical Computing, 1991, 27: 7-26.
18. Weaver RR. Assessment and diffusion of computerized decision support systems. Int. J. of Technology Assessment in Health Care, 1991, 7; 1: 42-50.
19. Kwa HY, Van der Lei J, Kors JA. Expert systems integrated with information systems. Comput Methods Programs Biomed, 1987, 25; 3: 327-32.
20. Devries C, Degoulet P, Jeunemaitre X, Sauquet D, Morice V, et al. Integrating management and expertise in a computerized system for hypertensive patients. Nephrol Dial Transplant, 1987, 2; 5: 327-331.
21. Rosenbloom PS, Newell A. The chunking of goal hierarchies: A generalized model of practice. In: Michalski R, Carbonell J, Mitchell T. (Eds.), Machine learning: An artificial intelligence approach. Vol. 2, Los Altos, CA, Morgan-Kaufman, 1985.
22. Buchanan BG. Can machine learning offer anything to expert systems? Machine Learning, 1989, 4; 3/4: 251-254.
23. Plugge LA. Possibilities and limitations of neural networks. Psychologie en Computers, 1990, 7; 4: 107-116.
24. Stubbs DF. Three applications of neurocomputing in biomedical research. Neurocomputing. 1990; 2: 61-66.
25. Hart A, Wyatt J. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. Medical Informatics, 1990, 15; 3: 229-236.
26. O'Desky RIO, Ball MJ, Ball EE. Computers in Health Care for the 21st Century. Methods of Information in Medicine, 1990, 29; 2: 158-161.
27. Brannigan VM, Dayhoff RE. Liability for personal injuries caused by defective medical computer programs. Am J Law Med, 1980; 7: 123-167.
28. Brannigan Vm, Dayhoff RE. Medical Informatics: The revolution in law, technology, and medicine. J Leg Med, 1986; 7: 1-54.
29. Miller RA, Schaffner KF, Meisel A. Ethical and legal issues related to the use of computer programs in clinical medicine. Ann Intern Med, 1985; 102: 529-565.
30. Chimowitz MI, Logigian EL, Caplan LR. The accuracy of bedside neurologic diagnoses. Annals of Neurology, 1990; 28: 78-85.
31. Lopez OL, Swihart AA, Becker JT, et al. Reliability of NINCDS-ADRDA clinical criteria for the diagnosis of Alzheimer's disease. Neurology, 1990; 40: 1517-1522.

32. Van Bemmel J. Formalization of Medical Knowledge. (Editorial) Methods of Information in Medicine, 1986, 25; 3: 191-193.
33. Hoffman RS. Diagnostic Errors in the Evaluation of Behavioral Disorders. JAMA, 1982, vol. 248; 8:964-967.
34. Verhey FRJ, Vreeling FW, Jolles J. DSM-III and NINCDS/ADRDA Criteria for Dementia and Alzheimer's Disease: impact of diagnostic procedures on daily practice. In: Wurtman J. (ed): Alzheimer's Disease. Proceedings of the fifth meeting of the international study group on the pharmacology of memory disorders associated with aging. Zürich, 1989; pp 419-4.
35. Ronteltap CFM. De rol van kennis in fysiotherapeutische diagnostiek. Thesis (in Dutch with English summary), Maastricht, Holland, 1990.
36. Dreyfus HL. What Computers Can't Do: A Critique of Artificial Reason. Harper and Row, New York, 1972.
37. Hellman S. Health care in the year 2000 and beyond. STG Bulletin, 1991; 27: 62-73.
38. Rutten F, Banta HD. Health care technologies in the Netherlands. Int. J. of Technology Assessment in Health Care. 1988; 4: 169-170.

SUMMARY

This thesis describes the development of the neuropsychiatric expert system Evince for the differential diagnosis of dementia, and the subsequent experiments in which the expert system was compared with clinicians from several medical disciplines who are involved in dementia diagnostics. An expert system is a computer program that is equally proficient in solving problems, in a specific domain, as a human expert in the same field.

Although many medical expert systems have been developed only a few have been subjected to an extensive formal evaluation. Even less expert systems have been developed for the psychiatric domain, and none for the neuropsychiatric field of dementia diagnostics, because this domain is generally considered too difficult to implement in an expert system, due to the use of descriptive criteria, and a multitude of data with very different origins. Furthermore, many of the expert systems that were developed were meant to be used on a mainframe, mini computer, or workstation, which makes such systems less attractive for many small or medium large institution.

In this thesis it is shown that it is possible to develop an expert system for an IBM-PC compatible desk-top microcomputer on the domain of the neuropsychiatric dementia diagnostics, that produces a performance equal to a multidisciplinary team of expert clinicians and better than the average clinician.

Chapter I gives a short introduction in the history of artificial intelligence and expert system research. Furthermore, an introduction is given into the previous efforts to develop expert system within the domain of psychiatry, and the lessons that can be learned from those earlier attempts. Chapter II identifies the lack of thorough expert system evaluations studies as one of the reasons why expert systems have been so slow in penetrating daily practice. Additionally some of the most recent ideas about expert system evaluation are discussed. An adapted three stage model from Wyatt and Spiegelhalter* for expert system evaluation is proposed, and some of the problems and possibilities of this model are discussed. In chapter III the expert system development tool Acquaint that was used for the development of Evince is reviewed. Chapter IV continues with a description of the development method used, i.e., the incremental top-down development of Evince. Furthermore, in this chapter the sources of knowledge, the knowledge acquisition, and the architecture of Evince are discussed. The following chapters contain a description of the experiments performed to assess the performance of Evince.

Chapter V is a report about the first experiment in which Evince was tested in diagnosing 19 patients with varying stages of dementia and 10 patients showing other disorders except dementia. It is shown that EVINCE-I and the human expert are in perfect agreement on the diagnosis dementia and correlate highly on the diagnoses dementia of the Alzheimer type and multiple infarct dementia. The results showed that the approach followed for the development of Evince was successful. Therefore,

* Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Medical Informatics. 1990, 15; 3: 205-217.

Evince was developed further on the basis of these results. This revised and expanded version of Evince was to be tested by comparing it with clinicians from several disciplines from outside the institution where Evince was developed.

In order to gain more insight in the performance of the average clinicians involved in dementia diagnostics, an experiment described in chapter VI was set up in which 90 clinicians from 6 disciplines diagnosed 10 case descriptions of patients, judged by a multidisciplinary expert committee to suffer from dementia. Five cases were diagnosed before and 5 after a consensus meeting on the diagnosis of dementia. A significant change in the level of agreement between the disciplines could not be established. The analysis did show a significant difference between the disciplines in the use of etiological diagnoses. The results indicated that, in order to avoid possible bias caused by medical specialization, a multidisciplinary approach for this type of patients is recommended.

Finally, in chapter VII an experiment is described in which Evince is compared with 85 clinicians in diagnosing 10 patients suspected of suffering from dementia. A multi-disciplinary expert committee provided a standard diagnosis as reference for comparison. The results show that the syndrome and etiological diagnoses made by Evince were in very close agreement with those of the expert committee and that the diagnostic performance of Evince was better than that of the average clinician.
In the final chapter (VIII) a general discussion of the findings is presented, and the limitations and possible implications of the use of Evince for dementia diagnostics in practice are discussed. Furthermore some topics for future research are addressed.

SAMENVATTING

In dit proefschrift wordt de ontwikkeling beschreven van een neuropsychiatrisch expertsysteem (Evince) voor de differentiële diagnostiek van dementie, en de experimenten waarin het expertsysteem werd vergeleken met clinici uit verschillende medische disciplines die betrokken zijn bij dementiediagnostiek. Een expertsysteem is een computerprogramma dat, bij het oplossen van problemen in bepaald domein, prestaties levert die vergelijkbaar zijn met die van een menselijke expert op dat terrein. Hoewel in de loop der jaren veel medische expertsystemen zijn ontwikkeld, werden slechts enkele daarvan onderworpen aan een grondige formele evaluatie. Nog minder expertsystemen werden ontwikkeld voor de psychiatrie en geen op het gebied van de neuropsychiatrische dementiediagnostiek, omdat dit terrein te gecompliceerd geacht werd voor implementatie in een computer programma, onder andere door het gebruik van descriptieve criteria en de grote verscheidenheid van de vereiste gegevens. Verder waren veel van de ontwikkelde expertsystemen bedoeld voor gebruik op zogenaamde 'mainframes', mini-computers of werkstations, waardoor het gebruik van dergelijke programma's minder interessant is voor kleine(re) instellingen.
In dit proefschrift wordt aangetoond dat het mogelijk is een expertsysteem voor een IBM-compatibele microcomputer te ontwikkelen op het gebied van de neuropsychiatrische dementiediagnostiek, dat resultaten levert die vergelijkbaar zijn met die van een multidisciplinair team van experts, en beter dan de prestaties van de gemiddelde clinicus.

In hoofdstuk I wordt een korte introductie gegeven in de geschiedenis van het wetenschappelijk onderzoek van kunstmatige intelligentie en expertsystemen. Verder krijgt de lezer een kort overzicht van eerdere pogingen om expertsystemen op het gebied van de neuropsychiatrie te ontwikkelen en de lessen die daaruit geleerd kunnen worden.
In Hoofdstuk II wordt ingegaan op het gebrek aan grondige expertsysteemevaluaties en wijst het gebrek aan dergelijke evaluaties aan als één van de oorzaken waarom expertsystemen zo langzaam in de dagelijkse praktijk doordringen. Bovendien worden enkele van de meest recente vooraanstaande ideeën over expertsysteemevaluatie besproken. Er wordt een aangepast model van Wyatt en Spiegelhalter* met drie stadia geïntroduceerd, waarna de mogelijkheden en problemen van dat model worden besproken.
Hoofdstuk III beschrijft het expertsysteem-ontwikkelingsgereedschap Acquaint dat gebruikt is voor het bouwen van Evince. Hoofdstuk IV gaat verder met een beschrijving van de gebruikte ontwikkelingsmethode, i.e. de incrementele en van het algemene in de bijzonderheden afdalende methode. Bovendien worden in dit hoofdstuk de kennisacquisitie, de kennisbronnen en de architectuur van Evince besproken. De daaropvolgende hoofdstukken bevatten een beschrijving van de experimenten die werden uitgevoerd om de prestatie van Evince vast te stellen.

* Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Medical Informatics. 1990, 15; 3: 205-217.

Hoofdstuk V bevat de rapportage over het eerste experiment waarin Evince vergeleken werd met een menselijke expert in het diagnostiseren van 19 patiënten in verschillende dementiestadia en 10 patiënten met andere stoornissen, maar geen dementie. Aangetoond wordt dat Evince en de menselijke expert volledige overeenstemming vertonen over de diagnose dementie en een hoge mate van overeenkomst over de diagnose dementie van het type Alzheimer en multiple infarct dementie. Deze resultaten laten zien dat de gevolgde ontwikkelingsmethode succesvol was. Op basis van deze resultaten werd Evince verder ontwikkeld.
Deze aangepaste en uitgebreide versie van Evince werd getoetst in een vergelijking met clinici uit verschillende disciplines die afkomstig waren van buiten het instituut waar Evince werd ontwikkeld. Om meer inzicht te krijgen in de prestaties van de gemiddelde clinicus die te maken heeft met dementiediagnostiek, werd een experiment opgezet dat beschreven wordt in hoofdstuk VI, waarin 90 clinici uit 6 disciplines ieder 10 patiënten diagnostiseerden aan de hand van een casusbeschrijving. Vijf van de casusbeschrijvingen werden vóór en vijf na een consensusbijeenkomst over dementiediagnostiek beoordeeld. Op basis van de beschikbare gegevens kon geen belangrijke verandering in het niveau van overeenstemming tussen de disciplines worden vastgesteld als gevolg van de consensusbijeenkomst. De analyses lieten echter wel een significant verschil zien tussen de disciplines in het gebruik van etiologische diagnosen. De resultaten gaven aan dat een multidisciplinaire aanpak bij dementiediagnostiek is aan te bevelen, om te voorkomen dat de uiteindelijke diagnose gekleurd wordt door een medische specialisatie.
Tenslotte wordt in hoofdstuk VII een experiment beschreven waarin Evince vergeleken wordt met 85 clinici in het diagnostiseren van 10 patiënten waarvan vermoed werd dat ze leden aan een vorm van dementie. De diagnosen van een multidisciplinair forum bestaande uit drie experts op het gebied van dementiediagnostiek fungeerden als referentie voor de vergelijking. De resultaten van die vergelijking lieten zien dat de syndroom en etiologische diagnosen van Evince nauw overeen kwamen met die van het forum en beter waren dan die van de gemiddelde clinicus.
Het laatste hoofdstuk (VIII) bevat een algemene discussie over de resultaten, beperkingen van Evince en de mogelijke implicaties van het gebruik van Evince bij dementiediagnostiek, en enkele onderwerpen voor toekomstig onderzoek.

DANKWOORD

Hoewel dit proefschrift op mijn naam staat, heeft het mede dankzij vele anderen uiteindelijk gestalte gekregen. Zonder volledigheid na te willen streven wil ik van de gelegenheid gebruik maken een aantal van de betrokkenen te bedanken.

Allereerst Jelle Jolles, die mij uitnodigde om in zijn vakgroep dit onderzoek te komen doen. In de afgelopen jaren ben ik je alleen maar meer gaan waarderen en respecteren, zowel als 'werkbaas' als in de persoonlijke omgang. Ondanks je drukke werkzaamheden voor onze vakgroep was je altijd aanspreekbaar. Ik had mij geen betere begeleider kunnen wensen.

Sommige mensen leer je altijd te laat kennen, zoals Arie Hasman, en dat neem ik mezelf kwalijk. De besprekingen met jou, Arie, en jouw oordeel waren erg belangrijk voor mij. Ik hoop dat we in de toekomst vaker met elkaar kunnen samenwerken.

Frans Verhey jr, jou ben ik speciale dank verschuldigd voor je enthousiaste medewerking, je heldere uitleg, en het geduld waarmee je mijn vragen beantwoordde. Het feit dat wij in de afgelopen jaren vrienden zijn geworden zegt wat dat betreft genoeg.

Alle 127 respondenten van de CBO consensusbijeenkomst, de 3 leden van het expertpanel (Dr. M.F.A. Diesfeldt, Dr. J.A.M. Frederiks en Prof. Dr. W. van Tilburg) en in het bijzonder Dr. J.J.E. van Everdingen ben ik eveneens dank verschuldigd. Bij dezen.

Peter, Paul en Jeroen, jullie wil ik tegelijk bedanken voor al die stimulerende discussies - al dan niet met de voeten op tafel - over ons werk en allerlei andere belangwekkende onderwerpen.

De firma Lithp (spreek uit als: *'lisp'*) Systems en in het bijzonder Peter van Lith (spreek uit als: *'lit'*) bedank ik voor de service die zij al die tijd hebben verleend.

Ma, jou wil ik bedanken voor de steun die ik - lang voordat ik aan dit proefschrift begon - kreeg om dat te doen wat ik graag wilde, ook al had je soms je twijfels.

Tot slot wil ik José en Lennart bedanken voor hun geduld met mij het afgelopen jaar.

CURRICULUM VITAE

Leo Plugge werd op 18 mei 1956 geboren te Middelburg. In 1972 voltooide hij de MAVO aan de Scholengemeenschap Scheldemond te Vlissingen en in 1974 de HAVO-top van de Rijkspedagogische Academie (RPA) te Middelburg. Aan diezelfde RPA werd vervolgens in 1977 de akte van bekwaamheid als volledig bevoegd onderwijzer behaald, waarna hij te Vlissingen werkzaam was in het onderwijs. In 1980 startte hij met de studie psychologie (oude stijl) aan de Rijksuniversiteit Utrecht. In 1981 werd het propaedeutisch examen behaald, in 1983 het kandidaatsexamen met als bijvakken fysiologie, sociologie en filosofie, en in 1986 het doctoraal examen, met als hoofdvak psychologische functieleer, als uitgebreide nevenrichting psychometrie, statistiek en modelvorming en als bijvak onderwijskunde. Het accent van zijn opleiding in het na-kandidaats lag op het terrein van de artificiële intelligentie. Een gedeelte van de opleiding op dat terrein werd gevolgd aan de Universiteit van Amsterdam. Aansluitend op het behalen van de bul trad hij in tijdelijke dienst van de Rijksuniversiteit Limburg, bij prof. dr. J.Jolles, leider van het Deelproject 'Hersenen en Gedrag' (Hoofdproject Veroudering). In het kader van Technology Assessment werd, met subsidie van het ministerie van Welzijn, Volksgezondheid en Cultuur, onderzoek verricht naar de ontwikkeling en evaluatie van een neuropsychiatrisch expertsysteem, waarvan dit proefschrift de weerslag is. In de toekomst zal hij zich verder bezig houden met expertsystemen, artificiële neurale netwerken en de integratie van die twee.

Appendix I

Hardware and software requirements.

Version 3.25 of Acquaint requires a IBM compatible PC microcomputer with a minimum of 520 Kbytes of RAM (620 Kbytes of RAM is recommended) and a hard-disk drive. Although Evince can be used on a microcomputer with the Intel 8088 or 8086 microprocessor, a 80286 is recommended.
Evince can be used with a monochrome or a color monitor.

The complete Evince program package consists of:

EVINCE.EXE	225.920	Kbyte:	the shell program
EVINCE.TXT	9.724	Kbyte:	the text source for the shell
MANAGER.RB	21.397	Kbyte:	the over head rulebase
MANAGER.RBE	3.567	Kbyte:	the text source
DATABASE.RB	73.553	Kbyte:	the database rulebase
DATABASE.RBE	29.098	Kbyte:	the text source
CONSULT.RB	92.893	Kbyte:	the batch/interactive consult rulebase
CONSULT.RBE	26.830	Kbyte:	the text source
REPORT.RB	33.334	Kbyte:	the report rulebase
REPORT.RBE	8.636	Kbyte:	the text source
CONFIG.LOG	68	KByte:	for configuration storage.
Total:	524.988	KByte	