Validating Medical Knowledge Based Systems

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Daalen, Cornelia van

Validating medical knowledge based systems / Cornelia van Daalen. - Delft: Delft University of Technology, Faculty of Mechanical Engineering and Marine Technology. - Ill.
Proefschrift Technische Universiteit Delft. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN 90-370-0089-4
Trefw.: PLEXUS (computerprogramma) / expertsystemen ; gezondheidszorg.

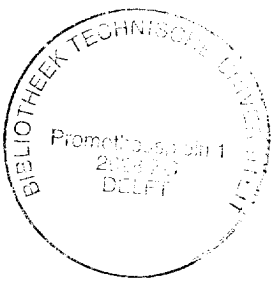
Validating Medical Knowledge Based Systems

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. drs. P.A. Schenck,
in het openbaar te verdedigen ten overstaan van een commissie
aangewezen door het College van Dekanen
op dinsdag 24 augustus 1993 te 14.00 uur,
door

Cornelia van Daalen

geboren te Delft,
werktuigkundig ingenieur.



Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. ir. H.G. Stassen,
Prof. dr. ir. E. Backer.

The hardware for the field evaluation of PLEXUS was provided by Apple Computer B.V.

Contents

1 Introduction

The research described in this thesis concerns the validation of medical knowledge based systems in general, and of the medical knowledge based system PLEXUS in particular. First, the background of the research will be described in Section 1.1. This initiated the development of the medical knowledge based system PLEXUS which is mentioned in Section 1.2. The necessary background information about medical knowledge based systems in general is provided in Section 1.3. A brief historical overview of validation research is given in Section 1.4. During the development of PLEXUS it became clear that validation issues have received little attention in the literature. This has led to the problem formulation as described in Section 1.5. The outline of this thesis is presented in Section 1.6.

1.1. Background of the research

The project originates from a research program which was started in 1968 at the Laboratory for Measurement and Control; it concerned the development and evaluation of externally powered prostheses and orthoses for the arm (Stassen, 1989). The project was carried out in cooperation with the rehabilitation centre 'De Hoogstraat' in Utrecht. Some of the patients who use the orthoses are patients with a paralysed arm due to a nerve injury in the area between the neck and the arm, i.e. a brachial plexus injury.

One of the objectives of the rehabilitation centre was to decrease the total rehabilitation time for patients with a brachial plexus injury, without diminishing the quality of the treatment. This necessitated insight in the rehabilitation process of brachial plexus injuries. A retrospective study of 136 patient files from the rehabilitation centre 'De Hoogstraat' showed that the diagnosis of brachial plexus injuries is often neglected by the referring hospital and revealed that it is necessary to propagate the knowledge of the possibilities for treatment (Jaspers, 1986).

An earlier approach at investigating the rehabilitation of patients with a spinal cord injury had resulted in a quantitative model of the rehabilitation process which provides a prognosis of the results of the treatment on the basis of previous experiences (Stassen *et al.*, 1980). For brachial plexus injuries a similar

system theoretic approach was firstly studied (Jaspers *et al.*, 1982). A major impediment to the applicability of data based methods in the domain of brachial plexus injuries was the lack of a large reliable set of patient data (Jaspers, 1990).

In general, the data based approach is applicable to problems requiring dynamic models, such as treatment and prognosis, or to diagnostic problems that require a case to be classified within a limited number of categories. If the number of diagnostic categories becomes large, data based classification systems are generally less suitable (Jaspers, 1990). Furthermore, in these kinds of systems the model output is fitted to the data by adjusting the model parameters. Neither the model structure nor the parameters have a meaningful interpretation, which leads to a lack of transparency.

In this domain, there are insufficient data to apply a data based approach and there is insufficient knowledge to build a precise enough deterministic model based on physiological data which can describe the behaviour. However, there is a number of experts in the domain of brachial plexus injuries. After studying the possibilities for the application of a knowledge based system, it was decided to use this approach for representing knowledge about brachial plexus injuries in the computer. Knowledge based systems allow uncertain, imprecise and expert knowledge (for instance, rules of thumb) to be represented in the computer. The way in which this may be done will be explained in Section 1.3.

1.2. PLEXUS

As a result, the knowledge based system PLEXUS has been developed. The aim of the system is to assist neurologists, neurosurgeons, orthopaedic surgeons, rehabilitation physicians and traumatologists in the diagnosis and treatment planning of brachial plexus injuries (Jaspers *et al.*, 1989; Jaspers, 1990; van Daalen *et al.*, 1993). It is meant for physicians who are not specialised in these injuries. The system has been developed in cooperation with the departments of Neurosurgery of the Leiden University Hospital and the 'De Wever' Hospital in Heerlen.

In order to obtain advice from the system, the physician enters patient data into the system by means of a graphical user interface. On the basis of these patient specific data, and the general knowledge about brachial plexus injuries which is stored in the computer, the system will suggest a diagnosis and a treatment plan to the physician. The system will be discussed in detail in Chapter 3. In order to provide the necessary background knowledge for this thesis, the subject of medical knowledge based systems will be introduced below.

1.3. Medical knowledge based systems

Knowledge based systems are information systems which manipulate knowledge, rather than manipulating signals as is done when using algorithmic or statistical methods. The way in which this knowledge may be represented in the computer will be explained below in Section 1.3.1. Knowledge based systems contain domain specific knowledge instead of the comparatively domain-free methods derived from areas such as computer science or mathematics (Jackson, 1986). This enables application in domains where the knowledge available is not precise enough to be able to, for instance, develop physiological models and for which there is not enough data to allow the implementation of statistical methods. The group of knowledge based systems which have received most attention are expert systems.

Expert systems are knowledge based systems which solve problems or provide advice at a level which is comparable to a specialist in the domain. These large domain-specific programs first came to be known as consultation programs, for they fit the image of an expert-specialist who is asked to provide advice about some difficult problem. By the late 1970's they became known as expert systems (Clancey and Shortliffe, 1984). Medical expert systems are based on symbolic models of disease entities and their relationships to patient factors and clinical manifestations. One of the most well-known expert systems is the system MYCIN (Shortliffe, 1976) which contains knowledge of infectious diseases.

In these systems, there is usually a division between the knowledge itself and the way in which this knowledge is manipulated. Most expert systems contain at least the following components:

- a knowledge base, in which the domain specific knowledge is represented,
- an inference engine, which manipulates the knowledge contained in the knowledge base,
- a human-machine interface (or user interface), allowing the user to interact with the system.

When conducting a consultation with a knowledge based system, a user may either volunteer patient specific information or the knowledge based system may request patient specific information from the system user. The inference engine, which is the reasoning mechanism, will then use this patient specific information and the general domain specific knowledge which is represented in the knowledge base, and will draw conclusions regarding a specific patient. The conclusions are shown to the user on the computer screen.

In this thesis, the term knowledge based system will be used, as it includes expert systems and allows a broader category of systems, i.e. not only systems which perform diagnosis or provide therapy recommendations, but any kind of information system incorporating symbolic knowledge representation. The issue

of knowledge representation will be mentioned below. To provide the necessary background for this thesis, two well-known methods of knowledge representation: production rules and object based methods, will be briefly explained.

1.3.1. KNOWLEDGE REPRESENTATION

A model of the domain knowledge is represented in the computer using a knowledge representation formalism. One method of knowledge representation which became popular after the development of MYCIN (Shortliffe, 1976) is the production rule formalism.

Production rules. Production rules allow the representation of heuristic knowledge. Systems which use production rules are called rule based knowledge based systems. A production rule is an if-then rule, relating conditions to actions. All knowledge in a domain may be represented in this way. The inference engine will match the patient specific information against these production rules and will draw conclusions. Since certain statements which appear in the condition of one rule will usually also appear in the conclusion of other rules, chains of rules are applied by the inference engine. One of the features of knowledge based systems is the ability to deal with uncertain knowledge. Uncertainty can be incorporated in the production rules by, for instance, attaching a number between 0 and 1 to the actions, indicating to which extent the action holds in a certain situation. When a chain of rules is applied, the final conclusion will include a certainty factor which is derived from the combination of the certainty factors of the rules involved in the chain.

Object based methods. Object oriented programming languages have become increasingly popular. Whereas production rules are very suitable for representing heuristic knowledge (for example, rules of thumb) and shallow knowledge, object oriented formalisms are more suited to representing structural knowledge, such as anatomical knowledge. One way of describing objects is through the use of frames (Minsky, 1975). Using this formalism, all knowledge concerning a certain concept is combined into one unit called a 'frame'. The information is grouped in terms of a record of 'slots' and 'fillers'. With a special slot filled by the name of the object and other slots being filled with the values of various common attributes which are associated with such an object. Frames can be organised in a taxonomic hierarchy of classes and subclasses. The fundamental idea is that properties in the higher levels of the frame system are fixed, insofar as they represent things which are typically true about the object. The lower levels have slots that must be filled with actual data (Jackson, 1986). The frame languages

support a reasoning mechanism called inheritance. The values of slots of more general frames are propagated to more specific ones.

The general inference method supported by frame languages is inheritance. Other ways of reasoning about objects, such as finding the values of certain slots, must be programmed using procedures for the deduction of information.

Most current knowledge based systems are not limited to one knowledge representation formalism. Furthermore, these systems often combine conventional and knowledge based programming techniques. Many knowledge based systems are developed using a knowledge based system shell. A knowledge based system shell is a program which contains a reasoning mechanism and an empty knowledge base, into which the domain specific knowledge can be entered using the knowledge representation formalisms that are supported by the reasoning mechanism.

More comprehensive discussions concerning knowledge representation may be found in, for example, Jackson (1986), Lucas and van der Gaag (1988) and Steels (1990).

1.4. Validation

After the first knowledge based systems had been developed, the developers wanted to prove that these systems possessed expert problem solving capacity (see, for example, Yu *et al.*, 1979). This is usually done by comparing the system to a number of experts in the domain. Evaluation methods which enabled this comparison to be made were the first knowledge based system evaluation methods to be described.

It was also recognised that it is possible and relatively easy to perform checks on knowledge bases, which are more elaborate than the syntactic checks which are used in conventional programs. Rule based systems are particularly suited to these checks, and methods have been described for checking rule bases for completeness and consistency (Nguyen *et al.*, 1987). This entails, for instance, checks for missing rules, cycles and redundant rules. Ginsberg (1987) described a method which allowed rule based systems to be analysed over complete inference chains. These methods for investigating completeness and consistency are called verification methods. They are relatively low cost methods for investigating and improving knowledge based systems.

The aim of most medical knowledge based systems is to improve patient care. In order to investigate whether this objective is achieved, an evaluation of the human-machine system in the target environment is required. Only a limited number of knowledge based systems have achieved the level of development

which is necessary to perform a clinical evaluation. Therefore, few clinical evaluations have been reported (see, for example, Bankowitz *et al.*, 1989).

A further evaluation procedure is the testing of a knowledge base with actual or generated test cases. This is often called dynamic validation. The use of generated test cases has not received much attention in the literature (Shwe, 1989). Knowledge based systems are often developed for domains in which there are not enough test cases for an adequate validation of the system. Using generated test cases, it is possible to directly address those aspects of the system which require investigation.

The validation of knowledge based systems should proceed in parallel with the design and development of a system. Although the literature on this subject is very diverse, three general validation procedures have been encountered in the literature: verification, dynamic validation and evaluation. These procedures allow different aspects of the system to be validated. In general verification will be performed first, followed by dynamic validation. After thorough verification and dynamic validation, a laboratory evaluation will be carried out. A clinical evaluation of the knowledge based system is only performed after a laboratory evaluation has shown the system to be safe and potentially useful.

1.5. Problem definition

During the course of the development, PLEXUS underwent preliminary validation at various stages. This involved testing the system with retrospective actual test cases and generated test cases, and studies in which system output was compared to the diagnoses and treatment plans provided by the experts who were involved in the development of the system.

Rather than being a research prototype, PLEXUS is aimed at actual use. Therefore, thorough formal validation of the system is of the utmost importance. It is necessary for the developer to ensure that the system fulfils its intended goals. Furthermore, potential users will probably not accept a system which has not been thoroughly validated.

Although it was recognised early on that these systems require validation, it has become apparent from the literature on knowledge based systems that the representation of knowledge has long been one of the major topics of research and validation of medical knowledge based systems only received little attention until recently. Furthermore, the literature on the subject of validation is very diverse. This implies that in order to be able to validate PLEXUS it is necessary to study the broader context of validation in general and then to use this information to determine the validation methods which can be applied to PLEXUS.

These validation methods should then be used in the validation of PLEXUS. On the basis of these studies it should be possible to identify possible problem areas, to suggest ways of solving these problems, and to draw conclusions regarding the applicability of PLEXUS in actual practice. Furthermore, since this investigation can be seen as a case study in validation, general recommendations concerning the validation of medical knowledge based systems should also result. The research described in this thesis thus concerns the validation of medical knowledge based systems in general and of the medical knowledge based system PLEXUS in particular.

1.6. Outline of the thesis

A survey of validation literature was performed in order to provide a basis for this research. The review of literature on performance evaluation of medical knowledge based systems is described in Chapter 2. In this context, the term performance is related to the quality of the human-machine system, rather than implying technical performance measures.

Chapter 2 includes both laboratory evaluation as well as clinical evaluation of medical knowledge based systems. In this chapter, a general framework for performance evaluation of medical knowledge based systems is introduced.

The architecture of the knowledge based system PLEXUS is discussed in Chapter 3. The system is compared to other neurological knowledge based systems and to other knowledge based systems in the domain of nerve injuries in the neck. Preliminary validation studies of the knowledge based system are described. Based on the positive results which were achieved in these studies, a laboratory evaluation of the system was performed in cooperation with independent experts from different countries, and a clinical evaluation of the human-machine system was carried out in a number of hospitals in The Netherlands.

The setup and results of the laboratory evaluation are discussed in Chapter 4, and the setup and results of the clinical evaluation are described in Chapter 5. Both evaluation studies were performed according to the general framework for performance evaluation which was already introduced in Chapter 2. In addition, the clinical evaluation involved a study of the usability and acceptance of the system. A general investigation into the acceptance of knowledge based systems was also conducted. It was decided to address both physicians and process-operators in this study, since knowledge based systems in medicine as well as knowledge based assistance in supervisory control (Sassen, 1993) are topics of investigation at the Laboratory for Measurement and Control. The study of the

attitudes of physicians and process-operators towards knowledge based systems was performed in cooperation with J.M.A. Sassen, and is described in Chapter 6.

The conclusions which can be drawn from the evaluation studies that were performed are discussed in Chapter 7. An analysis of the results of the investigations has led to general recommendations regarding the design and validation of medical knowledge based systems. These recommendations are also mentioned in the last chapter.

2 A review of literature on performance evaluation of medical knowledge based systems

Due to the multidisciplinary nature of knowledge based system design and development, the literature on the validation of knowledge based systems is very diverse. The terminology is not precisely defined and the procedures which are used vary from author to author. In this chapter, the terminology which will be used is first defined. Most of the validation literature involves empirical evaluation of the performance of a knowledge based system. In this context the term performance is related to the quality of the accomplishments of the human-machine system, rather than implying technical performance measures. The studies which are described in the literature can be divided into laboratory evaluations and field evaluations. The aspects of importance in the design of a performance evaluation are summarised in a framework for evaluation design. This framework includes the choice of a goal for evaluation, evaluation setup, analysis of the results and threats to the validity of a study. This framework will be introduced, after which the information found in the literature on laboratory and field performance evaluation is discussed and compared along the lines of the framework. Many different evaluation setups and methods of analysing the results have been encountered. Most investigators are very positive after a laboratory investigation, however, quite often no further evaluations of the systems, such as field evaluation, are reported. The discussion on the evaluation methods found in the literature has led to recommendations for performing evaluation studies of knowledge based systems. From the review it becomes clear that empirical performance evaluation is only a limited part of the validation process, and that knowledge based system validation should be a continual process which should proceed in parallel with the design and development of a system.

2.1. Introduction

The validation of medical knowledge based systems has gained interest in recent years. Most knowledge based systems are no longer only research prototypes, but are aimed at actual use. To be able to achieve actual use, it is necessary for the developer to ensure that the system fulfils its intended goals. Furthermore, potential users will probably not accept a system which has not been thoroughly validated.

The literature on the subject is very diverse. This is enhanced by the inherent multidisciplinary nature of knowledge based system design and validation. The aims of this review are to investigate performance evaluation methods which have been proposed in the literature, to survey actual performance evaluation studies which have been described, to integrate the information from the literature, and to propose a framework for the evaluation of the performance of knowledge based systems. This study is

being carried out to provide a basis for the evaluation of the medical knowledge based system PLEXUS (Jaspers, 1990).

The terminology which is used in the literature is not precisely defined. Therefore it is necessary to define the terms as they will be used in this review. The main concepts used in the literature are verification, validation and evaluation.

A distinction is generally made between the following two different procedures:
- determining whether the system has been built right,
- determining whether the right system has been built.

The first procedure is usually called verification, and the second is usually called validation (see, for example, Gupta, 1991). Most authors agree that these two aspects have to be investigated, however, after having stated these general descriptions, the actual working definitions which are used for these procedures are very diverse.

2.1.1. TERMINOLOGY USED IN THE LITERATURE

Two general categories of definitions will be described. Definitions related to the development life cycle, and definitions related to a prototype system or to a final product. A discussion of the differences between these two categories of definitions, and of actual validation methods which have been described in the literature will lead to the definitions which will be used in this review.

2.1.1.1. Life cycle related definitions

The definitions given by Lydiard (1992) are an example of what can be found in the literature, and they are close to those used within the software engineering community. The definitions are related to the development life cycle.

Verification: verification is an activity which should ensure that the product of one phase of the life-cycle is consistent with itself and with the source from which it has been derived. As such, verification should be carried out at the end of each phase of development.

Validation: validation is an activity which should ensure that the product at the end of each phase of the development process complies with the software requirements it was intended to satisfy. Validation is usually achieved through testing.

Evaluation: evaluation is a feature of both verification and validation, and it concerns the assessment of the quantitative and qualitative characteristics of the KBS application through comparison with required standards.

Like Lydiard (1992), Green and Keyes (1987) relate verification to the life cycle, and define it as showing that the specification or code fully and exclusively implements the requirements of the superior specification.

2.1.1.2. Knowledge based system related definitions

There are also authors who do not relate the definitions to every phase of the development life cycle, but to a prototype system or to a final product. There are, however, still many differences of opinion concerning the terminology which is used. Some examples of these differences are demonstrated below.

Verification has been defined as authentication that the formulated problem contains the actual problem in its entirety and is sufficiently well structured to permit the derivation of a sufficiently credible solution (O'Leary *et al.*, 1990). Nykänen (1990) defines verification as the act of checking correctness according to specifications. According to Fieschi (1990) verification is a static method which does not require running the system.

The term validation is also interpreted in various ways. Validation has been defined as the process of assuring that the knowledge and advice is accurate, complete and consistent (Miller and Sittig, 1990), or as the comparison of quality measures with a frame of reference (Nykänen, 1990). Shwe *et al.* (1989) define validation as the process of proving or showing to a satisfactory degree that the behaviour of an expert system is correct with respect to the specifications of the system.

The relation between the terms verification, validation and evaluation also has different interpretations. Fieschi (1990) divides evaluation into verification, and test and validation. According to O'Leary *et al.* (1990) verification is a part of validation, and according to Shwe *et al.* (1989) verification is only part of static validation. O'Keefe *et al.* (1987) state that validation is part of evaluation.

Since there is no consensus as to the terminology which is used in the evaluation literature, Laurent (1992) proposes to solve the definition question by using the term validation as the general term for defining the whole set of activities the goal of which is to contribute to guarantee (up to a certain extent) the quality and the reliability of a knowledge based system.

Validation: a validation process is a process which attempts to determine whether a knowledge based system does or does not satisfy one of its specifications. Validation is the sum of all validation processes.

Laurent (1992) divides validation into two kinds of processes, objective and interpretative validation. Interpretative validation is referred to as evaluation and is defined as below.

Evaluation: an interpretative validation process is a validation process which attempts to determine whether a knowledge based system does or does not satisfy one of its pseudo-formal specifications. Evaluation is the sum of all these processes. A pseudo-formal specification comes from the approximate translation of a non-formalisable validation concept.

Objective validation is referred to as verification and is defined as below.

Verification: an objective validation process is a validation process which attempts to determine whether a knowledge based system does or does not satisfy one of its purely formal specifications. Objective validation is the sum of all these processes.

2.1.1.3. Differences between definitions

A few of the differences which exist in the definitions, which were discussed above, will be highlighted. Laurent (1992) defines the term validation to include the complete field. Some authors define the term evaluation to denote the complete field (O'Keefe *et al.*, 1987), and others use verification and validation (Lydiard, 1992).

Laurent (1992) divides the complete field into objective and interpretative investigation, yet others (Fieschi, 1990) divide the field into static (without running the system) and dynamic (running the system) methods. Shwe *et al.* (1989) differentiate between procedures involving independent experts and procedures for proving correctness against specifications. The procedures which aim at proving correctness with respect to specifications are then further divided into static and dynamic methods.

It may thus be concluded that different terms are used to denote the complete field. Furthermore, different criteria are used to subdivide the field. Another way to try and solve the terminology problems is to look at actual methods which have been applied and are described in the literature.

2.1.1.4. Actual validation methods

Various tools and methodologies have been applied for validation purposes. Looking at the practical research which has been carried out in this area, the

activities usually relate to a product rather than the complete development life cycle. A number of different approaches can be distinguished. Firstly, there is a category of activities which determines a number of objective requirements which the knowledge in the knowledge base has to adhere to, and uses static tools to determine whether the domain model has been correctly implemented. This is usually called verification. Verification used in this way looks only at the software, and not, for example, at the specifications, therefore it cannot be used at any phase of the development life cycle. This is therefore consistent with only part of the definition given by Lydiard (1992), and part of the objective validation definition given by Laurent (1992).

There is also a category of activities which aims at investigating whether the domain model is correct. This is investigated by using test cases and testing the implementation of the model. This is also part of the objective validation given by Laurent (1992) and part of the validation definition given Lydiard (1992).

The final category of activities consists of investigations which aim at comparing the behaviour of the knowledge based system to experts in the domain and to potential users, and approaches which investigate the human-machine system in the field. These approaches are all interpretative validation procedures and belong to validation in the perspective of Lydiard (1992), and to evaluation in the perspective of Laurent (1992).

2.1.2. TERMINOLOGY TO BE USED IN THIS REVIEW

Both groups of definitions discussed above (Lydiard, 1992 and Laurent, 1992) have certain elements which should be incorporated in the definitions which will be used in this chapter. The definitions proposed by Lydiard (1992) emphasize the phases of the life cycle of software development and the definitions proposed by Laurent (1992) recognise interpretative and objective elements.

The complete area will be called validation, consistent with Laurent (1992). This will be divided into interpretative validation (evaluation), and into objective validation (verification and development validation). The definitions will be related to the life cycle, rather than to the prototype or final product. Since the whole field is called validation, and development validation is only part of it, the term validation is used in two different ways, so there is one term missing in this field. Thus, it can be seen that a problem arises.

Validation: a validation process is a process which attempts to determine whether at each phase of the life cycle the product complies with one of its requirements. Validation is the sum of all validation processes.

Evaluation: an evaluation process is a process which should ensure that the product at the end of each phase of the development process complies with one of the pseudo-formal requirements which it was intended to satisfy.

Verification can be seen to be consistent to the definition of verification given by Lydiard (1992).

Verification: is an activity which should ensure that the product of one phase of the life-cycle is consistent with itself and with the source from which it has been derived. As such, verification should be carried out at the end of each phase of development.

Development (or dynamic) validation: is an activity which should ensure that the product at the end of each phase of the development process complies with one of the formal requirements it was intended to satisfy.

Validation issues are related to the complete development life-cycle. With respect to the knowledge based system itself, validation does not only include investigation of the knowledge base of the system, but is equally important to, for instance, the interaction which takes place and the inference procedures which are used.

At the end of the development life cycle, after thorough verification, development validation and evaluation have been carried out, two different empirical evaluation processes may be distinguished. These are often categorised into two kinds of study (Wyatt and Spiegelhalter, 1990), laboratory testing and field evaluation. The laboratory testing phase is necessary to further investigate whether the system is safe and whether it has at least the potential to benefit patients.

The following definitions will be used in this review:

Laboratory evaluation: empirical evaluation of the knowledge based system in the laboratory environment.

In the literature, this procedure is very often assumed to be a development validation procedure. In agreement with Shwe *et al.* (1989) who distinguish validation and evaluation, however, development validation and laboratory evaluation may be thought of as different activities. Development validation will be considered to be a procedure which attempts to find as many errors as possible, and which may be automated to a certain extent using, for example, graph theory or script (patient) generators. Whereas a laboratory evaluation is a time consuming process which is not applied in as many iterations as is dynamic validation. However, formal laboratory evaluation is also necessary in order to evaluate, for instance, the system's problem solving performance level and the potential usefulness of the human-machine system. Therefore, laboratory evaluation will be carried out once development validation of a knowledge based system has been completed.

Field evaluation: empirical evaluation of the human-machine system in the target environment. Field evaluation encompasses investigation of a large number of aspects. The specific issues addressed depend on the nature of the system, on the domain, and on the clinical role of the system. These issues include: the investigation of the impact of the system on physician actions, on patient care, and on health care processes, a cost benefit analysis, the examination of subjective reactions, and the investigation of system use (Miller, 1986).

Design and validation are closely related, as validation should be part of the development life cycle. The need to design quality into software from the beginning of the development life cycle is now generally accepted (Fox, 1993). There is an awareness that soundness of design, clarity of specification and integrity of implementation are issues which must be taken seriously in safety critical fields such as medicine (Fox, 1993). The formulation of a comprehensive design theory will also facilitate the incorporation of validation procedures into the development life cycle.

The validation process should be continual (Gaschnig, 1983). The relevant form of the validation will depend on such issues as system scale (Lundsgaarde, 1987), the system's maturity and goals, and on the character of the domain (Miller, 1986). The aspects which are being validated will evolve during the development period. Validation will become increasingly formal as a developing system begins to achieve real-world implementation (Gaschnig *et al.*, 1983). Before a laboratory evaluation is carried out, the system should have been thoroughly verified and validated. Laboratory evaluation should provide additional evidence that the system is safe and potentially useful. In some domains it may be advisable to use the system in parallel with the

present method of performing the task for a period of time, prior to performing a formal field trial. After a successful field evaluation, the design and validation of the system will be not be finished, as follow-up studies will be necessary to investigate the large-scale usefulness of the system, and the maintenance of the system will have to be addressed (Gaschnig *et al.*, 1983).

2.1.3. SUBJECT OF THE REVIEW

The validation of a knowledge based system is an extensive process which requires many aspects to be considered, since difficult processes can not be validated by a single criterion or number (Gaschnig *et al.*, 1983). Therefore, the literature is of a very diverse nature. Since there are many different aspects which may be investigated, and the complete field is too extensive to cover in this survey, it is necessary to make a choice as to the areas of validation which will be covered. One of the main aims of a knowledge based system is usually to improve the quality of care. In order to determine whether this is indeed the case, a performance evaluation can be carried out. Furthermore, most of the validation literature concerning medical knowledge based systems covers performance evaluation. Although, according to Lundsgaarde (in 1987), approximately 90% of all medical knowledge based systems have not undergone independent performance evaluation in controlled or real-time clinical environments. Most work has been done on the subject of evaluating the laboratory performance of knowledge based systems, where the emphasis has been on investigation of the accuracy of the output generated by the knowledge based system itself. Hardly any of the literature addresses human-machine interaction, cost-benefit and impact on health care in general. This review will concentrate on laboratory and field evaluation studies which are carried out after verification, development validation and evaluation have been performed.

Further information about verification and validation of knowledge based systems may, for example, be found in Nguyen *et al.* (1987), Stachowitz and Combs (1987), Green and Keyes (1987), Shwe *et al.* (1989), Laurent (1992), Lydiard (1992), Preece and Shinghal (1992), and Meseguer (1992). Many papers, most of which regard non-medical applications, have been brought together in the book edited by Gupta (1991) and in the work edited by Ayel and Laurent (1991). Verification and development validation of knowledge based systems are extremely important topics which have not received enough attention in the literature.

2.2. Performance measurements

This review will concentrate on the literature concerning empirical performance evaluation studies of medical knowledge based systems, which are carried out after verification, development validation and evaluation have been carried out. Empirical performance evaluations are either laboratory or field evaluations which are directed at investigating the quality of the accomplishments of the human-machine system. This is therefore only part of empirical evaluation and does not include, for instance, usability and acceptability.

Table 2.1 shows a global categorisation of performance evaluation literature on the subject of medical knowledge based systems. The evaluation procedures, laboratory and field evaluation are shown horizontally in the table. Vertically, the literature is divided into those articles which report actual evaluation studies which have been carried out, and into those which discuss evaluation theoretically. All literature has been classified into the categories which correspond best to the major topics of discussion. Most actual evaluation studies also discuss the topic theoretically, however, these papers have been classified into the relevant application categories only. Not all articles which are related to specific applications concern actual knowledge based systems, however, they are all computer programs (e.g. Bayesian systems) which aim at assisting physicians in solving difficult problems in their domain.

In the next sections, literature on performance evaluation will be reviewed. This will consist of a survey of methods which may be used for performance evaluation, aimed at investigating whether the system can be used in actual practice, rather than more informal evaluation which is concerned with improving the system performance.

First, a number of items will be identified which are of importance in the design of an evaluation study. These items have been brought together into a framework of choices to be made when designing an evaluation. All items in the framework will be discussed separately, consisting of a short discussion of the information which has been found in the literature about that particular topic, followed by a synthesis of the item, based on the information from the literature. This procedure has been carried out both for laboratory and field evaluations of medical knowledge based systems. At the end of the chapter the complete synthesis is summarised. This summary can be seen as a discussion model for performance evaluation.

Table 2.1. Global categorisation of performance evaluation literature.

	theory	application	
laboratory evaluation	Chandrasekaran Hilden Indurkhya Lundsgaarde O'Keefe	Adlassnig Aikins Alonso-Betanzos Bernelot Moens Catanzarite François Gorry Haberman Hickam Kingsland Kors McDermott	R.A. Miller (82) Murray Nelson Quaglini Reggia Rothschild Soula Spitzer Wong Yu Zagoria
field evaluation	Nykänen Spiegelhalter Sutton(b) Whitbeck	Adams Bankowitz Kent McDonald Pryor Sutton(a) White	
lab & field evaluation	Gaschnig P.L. Miller (86,90) O'Moore Rossi-Mori Wyatt (90)	Fieschi	

The framework for performance evaluation will first be introduced. Following this, each item in the framework will be discussed in relation to laboratory evaluations of knowledge based systems. After the discussion concerning items of importance for laboratory evaluations, a number of reported laboratory evaluation studies will be compared. Then, the items in the framework will be discussed in relation to field evaluations, and a number of reported field evaluation studies will be compared.

The design of an evaluation study requires many different aspects to be defined precisely, before carrying out the evaluation procedure. The aspects of importance in evaluation design can be categorised into a framework for evaluation design, which is shown below. Most of the items in the framework have been mentioned before by O'Moore *et al.* (1990) in their discussion on the design of an evaluation, but for the purpose of this review, the items have been ordered and modified. The framework will provide the structure for the discussion on performance measurements. All items in the framework will be discussed in detail. The same framework will be used for the discussion of both laboratory and field performance measurements. The framework for evaluation design is shown in Figure 2.1.

- selection of a goal for performance evaluation
    - quantification of the goal for performance evaluation
- evaluation setup
    - selecting test input (selection)
    - consultation
        - specifying who uses the system (human-machine system)
        - specifying physicians to test against
        - specifying a standard of performance
    - comparison (variables, judging)
- analysis of the results
- identifying threats to the validity (e.g. bias and confounding)

Figure 2.1. Framework for performance evaluations.

The items which are mentioned above are also shown in a dataflow diagram (Figure 2.2).

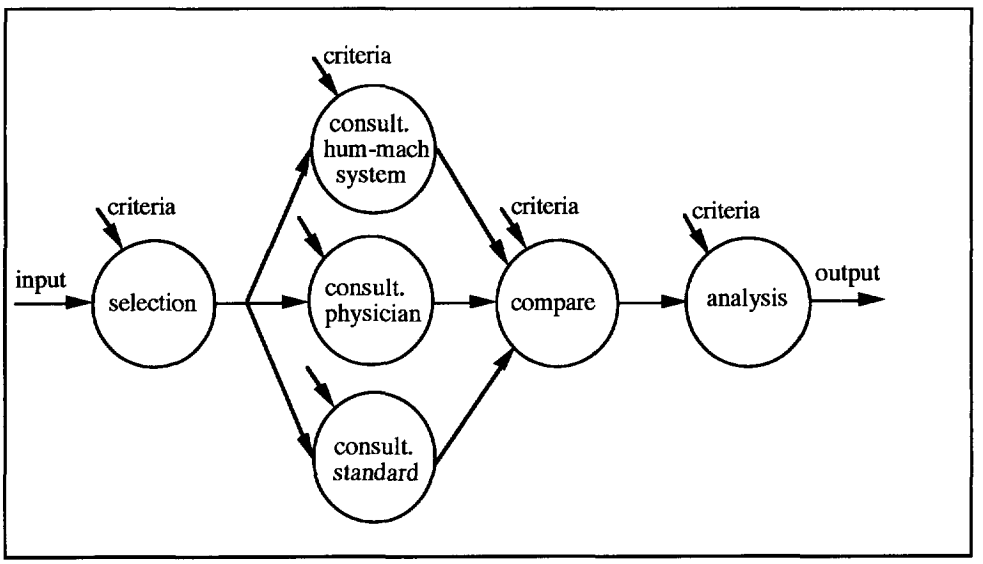


Figure 2.2. Dataflow diagram for the design of performance evaluations.

The dataflow diagram may be interpreted as follows. After a goal or limited number of goals have been chosen for evaluation, the input data will have to be selected. The data usually consist of a number of test cases. The criteria for test case inclusion will have to be decided on during the design of the evaluation. The filtered input will then have to be processed by one or more human-machine systems, various physicians and a standard of performance. This will involve a number of choices, for example, the choice of the persons who will use the knowledge based system, how many physicians will be involved and how to establish a standard. The output which is generated by all three systems during the consultation will have to be compared, and criteria for the comparison will have to be chosen. The output which is meant here is not necessarily the same as the final advice. Following this, the results of the comparison will have to be analysed according to certain criteria, and conclusions will be drawn from the results of the analysis with respect to the goals of the evaluation.

The choices which can be made during the design of a performance evaluation are shown in the dataflow diagram. The choices are represented by the

arrows labelled 'criteria'. Each subsection of this chapter will represent one entry in the flowdiagram, and the possible choices (criteria) for that particular entry will be discussed in the subsection. In each subsection, information from the literature will discussed first. The final paragraph of each subsection will consist of conclusions which may be drawn from the literature. Performance evaluations are subject to a considerable number of threats to the validity of the study, therefore possible sources of bias and confounding and their implications on the evaluation results will also be discussed.

In the literature, a division can be made between laboratory measurements of the knowledge based part of a system and field evaluations of human-machine systems. Therefore, these studies will be described first. Measurements of the performance of knowledge based systems in the laboratory will be discussed (Section 2.3). Following this, performance measurements in the clinical environment will be reviewed (Section 2.4). For both kinds of evaluation the classification framework (Figure 2.1) will be used as a basis for the discussion. The discussion on laboratory performance concerns knowledge based part of the system, rather than the complete human-machine system, since this complies to what is usually found in the medical knowledge based system literature. Human-machine studies will also have to be carried out in the laboratory, however, this has not been found in the literature, therefore these studies will be discussed separately in Section 2.5.

2.3. Laboratory performance evaluation of medical knowledge based systems

2.3.1. SELECTION OF A GOAL FOR PERFORMANCE EVALUATION

Before carrying out a performance evaluation it is necessary to decide on the goals of the evaluation, for otherwise it will not be possible to decide whether the system satisfies all the necessary requirements. There is a number of goals which may be pursued during a laboratory evaluation, examples of goals are potential usefulness, correctness of output, safety, transferability, efficiency, reliability, correctness of reasoning. However, in order to carry out an evaluation, the object of the evaluation must be clearly defined and be formulated into a testable hypothesis. Although this seems straightforward, it is not often seen that a testable hypothesis is stated before an evaluation is carried out.

2.3.1.1. Quantification of the goal
The goals of reported evaluation studies may be stated in the literature, however, they are very rarely quantified. Gaschnig *et al.* (1983) stress that when designing a system, explicit statements of what the measures of the program's success will be and how that failure or success should be evaluated should be included. Requirements could be that the system should diagnose significantly more than a certain amount of the cases correctly, that the system produces significantly better decisions than the potential user, that the system produces a significant fewer number of unsafe decisions than the potential user etc. Actual performance measurements are often an investigation of the output where the goal is not quantified in advance. It may be difficult to quantify a goal in advance. This makes it extremely difficult to empirically test hypotheses. However, it is necessary to quantify a goal in advance to be able to investigate whether the goal is satisfied.

2.3.2. EVALUATION SETUP

The general method which may be chosen is shown the in dataflow diagram above (Figure 2.2). The input is given to the knowledge based system, and the knowledge based system provides an output. The input may also be given to a number of physicians who may be potential users, experts or both. If a definite standard of performance is known, this will also be used. The outputs provided by these three groups may then be compared. Comparison may exist of a direct comparison between outputs, or of an indirect subjective comparison by experts. The indications for a specific experimental setup will be discussed in more detail. The following subsections will consist of a discussion of the criteria for each of the entries of the flow diagram (Figure 2.2).

Choice of experimental unit. Although the choice of experimental unit is not usually specified explicitly in the literature, the experimental unit used for laboratory evaluations of the medical knowledge based system itself often is the patient. The experimental unit can sometimes be data from a patient visit or result.

2.3.2.1. Selection of test input

Level of test input. In laboratory studies, the test input usually consists of a set of retrospective cases. There is some discussion as to the level of difficulty these cases should have. Chandrasekaran (1983) proposes that if the statistical evaluation should represent the system's performance in a real clinical setting the cases should be selected as representative of the target clinical setting, which makes it necessary to have knowledge of the distribution of types of cases. O'Keefe *et al.* (1987) state that the issue is not the number of test cases, it is the coverage of test cases and mention that in a domain where 90% of cases is standard and 10% is difficult, a 90% success rate is not very good. Rossi-Mori *et al.* (1990) distinguish obvious, typical and atypical cases in the test sample, and they stress that it is important to investigate the cases that the user feels are not trivial.

For actual studies, test cases have been taken from a database of cases (Murray *et al.*, 1986; Kors *et al.*, 1990). Cases have been taken from one (Quaglini *et al.*, 1988; Nelson *et al.*, 1985; Hickam, 1985; François *et al.*, 1992) or more hospitals (Fieschi, 1990). For other studies, test cases have been taken from the literature (Catanzarite *et al.*, 1981; Miller *et al.*, 1982).

There are various ways in which the test cases have been chosen. For some studies random samples (Murray *et al.*, 1986; Fieschi, 1990) or stratified random samples (Kors *et al.*, 1990) have been used. In other studies, cases consisted of consecutive patients (Nelson *et al.*, 1985; Kingsland, 1985; François *et al.*, 1992). In a performance evaluation study of MYCIN (Yu *et al.*, 1979) a set of challenging, diverse cases was chosen to test the system.

The choice of test cases depends on the goal of the evaluation. Besides testing cases which are representative of the distribution which may be expected in the target environment, it will usually be necessary to investigate the behaviour on challenging and probe cases. Furthermore, in any evaluation study, the test cases that are used must not have been used before. It is necessary to separate training and test cases. Decision-aids usually perform much better on training data than on a fresh set of data collected as part of a separate study (Wyatt and Spiegelhalter, 1990).

Number of test cases. Many authors state that it is not the number of test cases which is important, but the coverage of cases. However, a minimum number of test cases is necessary to be able to obtain results which are sufficient to be able to conclude about the results of the evaluation. But, there is also a practical limit to the number of cases which can be evaluated, since physician cooperation is needed to carry out the evaluation. Unrealistic time demands on physicians may lead to failure of an evaluation study. Therefore,

measures to try to limit the time demands, such as focusing on fewer variables and designing checklists, may be in order (Gaschnig *et al.*, 1983). Efficient use must be made of the people involved (Miller, 1986). The numbers of test cases involved in reported laboratory evaluations vary from about 10 (Yu *et al.*, 1979; Murray *et al.*, 1986; Rothschild *et al.*, 1990), 30 (Quaglini *et al.*, 1988; Kors *et al.*, 1990; Catanzarite *et al.*, 1981), 100 (Zagoria and Reggia, 1983; Nelson *et al.*, 1985), to 212 (François *et al.*, 1992) and 415 (Hickam *et al.*, 1985).

No mathematical background has been found for the choice of the numbers of test cases involved in the laboratory evaluations. Some estimation of trial size should be carried out in advance. Cohen (1977) provides sample size tables for differences between proportions when applying a normal curve test to the arcsine transformation of the proportions. Effect size (difference to be detected between arcsine transformation of the proportions), power level and significance have to be specified in advance, and the sample size can be looked in the tables. There may be a practical limit to the number of cases which can be used in a certain evaluation. However, some estimation can be made in advance as to whether this number will be sufficient to potentially provide relevant evaluation results. If this is not the case it should be decided whether action should be taken to obtain more cases or accept the influence on the statistical conclusion validity.

2.3.2.2. Consultation

There are various different situations which will have to be compared. A comparison will have to be made between the human-machine system, a number of physicians and a standard of performance.

There are several studies that compare more than one system. For instance, Zagoria and Reggia (1983), who report an evaluation study in which a Bayesian system was not only compared to humans, but also to two other systems.

Specifying who uses the system (human-machine system)

It is usually not mentioned explicitly, but in laboratory testing the test cases are often entered into the system by the knowledge engineer. If the goal of the evaluation is to evaluate the performance of the knowledge based system, and not of the human-machine system, the knowledge engineer should enter the patient data into the system. The idea probably being that when the knowledge engineer enters the data, no errors will be made due to data entry, and therefore only the problem solving part of the knowledge based system will be evaluated.

As was mentioned above, the human-machine system should also undergo laboratory evaluation, during which potential users interact with the system. However, since this has not been described in the literature and requires a somewhat different approach, this will be discussed separately in Section 2.5.

Specifying physicians to test against

Level of expertise. In most performance measurements of medical knowledge based systems, the objective is to compare the knowledge based system to a number of physicians. In the literature, these have either been experts in the domain (Quaglini *et al.*, 1988, Soula *et al.*, 1988), experienced physicians (Kors *et al.*, 1990; Haberman *et al.*, 1985), the case physicians (Hickam, 1985; François *et al.*, 1992), potential users of the system, a combination of experts and potential users (Fieschi, 1990), or physicians ranging in expertise (Zagoria and Reggia, 1983).

The level of expertise of the people involved in the evaluation will depend on the objective of the system. To be able to show potential usefulness of the knowledge based system it is necessary to involve prospective users in the evaluation. If the system is designed as an expert system, both experts and potential users will be involved in the evaluation. It may be necessary to demonstrate the expert problem solving capabilities of a system, also to be able to convince potential users in the field of the quality of the system, and because it is not possible to test a range of potential users.

Number of physicians. In actual evaluation studies, the number of physicians who have been involved in the comparison varies from 1 (Haberman *et al.*, 1985), 2 (Aikins *et al.*, 1983), 3 (Zagoria and Reggia, 1983), 5 (Kors *et al.*, 1990), 6 (Quaglini *et al.*, 1988; Fieschi, 1990), 9 (Yu *et al.*, 1979) to 29 (Murray *et al.*, 1986).

No justification has been found for the choice of the numbers of physicians. The intra- and inter-physician variability which exists will have to be taken into account. The number of physicians should be estimated in advance. If it is not possible in practice to involve this number of physicians in the evaluation, then the influence of the reduction of this number on the external validity of the measurement should be determined.

Specifying a standard of performance
The standard of performance is the objective reference against which will be evaluated. The central question in all evaluation studies is, what is the objective reference by which to evaluate the techniques (van Bemmel, 1988). In some domains the actual diagnosis and optimal treatment plan for a patient may be known, or it may be possible to follow up cases to determine the true cause of their symptoms. Whenever this is possible, the true answer should be used as the standard. In an evaluation of a prognostic system for severe head injury (Murray *et al.*, 1986) system output could be compared with the real outcome after 6 months. In other domains however, the standard may never become known. In these domains there are various ways to approximate a standard, these methods will however introduce errors into the measurement.

Sometimes, the histopathological or surgical diagnosis (Adlassnig and Scheithauer, 1988) is taken to be the optimal diagnosis, or the standard diagnosis is taken from the literature (Catanzarite *et al.*, 1981). In other evaluations, in order to take into account the variability which may exist between experts, the consensus opinion of a number of experts is taken to be the standard (Kingsland, 1985). However, the method used to obtain the standard is not described. One study (Kors *et al.*, 1990) describes a method which may be used to increase agreement. This method is called the Delphi method. The Delphi method is an anonymous feedback method, in which participants are asked their opinion in a first round, in the following rounds they receive all opinions from the previous round without identifying who provided which opinion and are asked their opinion again. It is then expected that the judgements will converge, ideally reaching some sort of consensus. Van Bemmel (1988) states that involvement of human experts in the evaluation requires a feedback procedure to reduce the inter-observer variability.

In cases where the real output is unknown, another possibility is to not create a subjective standard but to, for instance, have a panel of expert judges analyse the results of the knowledge based system and physicians (implicit standard). This last possibility will be discussed in section 2.3.2.3.

In domains where an actual standard of performance is available, the actual standard should always be used. However, since in most medical domains there is no standard of performance, another solution should be found. A consensus analysis would seem the best method to provide a partial solution to this problem. Intra- and interexpert variability will have to be taken into account when this standard output is obtained from experts. If it is not possible to directly compare the outputs generated by physicians and the

knowledge based system, the standard does not have to be made explicit. The comparison may be carried out by a panel of judges. However, this will also introduce problems of intra- and interexpert variability.

2.3.2.3. Comparison

Variables to be compared. The variables which are to be measured depend on the objectives of the system and the goal of the evaluation procedure. In most actual evaluation studies, final system output has been measured (Fieschi, 1990; Adlassnig and Scheithauer, 1989; Murray *et al.*, 1986; Haberman *et al.*, 1985). Some researchers also consider the structure of reasoning of the knowledge based system. Chandrasekaran (1983) proposes to evaluate the efficiency with which conclusions are reached, by using a thinking aloud protocol obtained from experts. In the evaluation of ANEMIA, Quaglini *et al.* (1988) use questionnaires in order to investigate the physicians' diagnostic reasoning.

Measurement of only final system output does not take into account the micro-structure of problem-solving behaviour which can be important in permitting the extrapolation from representative cases to conclude about the overall competence of the system (Gaschnig *et al.*, 1983). In order to be able to extrapolate to overall competence, the subconclusions the system makes must be correct, therefore, subsystem analysis should also be carried out.

Fieschi (1990) has carried out a sensitivity study by tightening and widening the limits around different interpretation zones in certain rules, changes in interpretation of low, normal etc., and investigating the difference in output.

Depending upon the goals of the evaluation, the variables which are to be compared have to be chosen. Analysis of the output of the complete system, subsystem analysis and change in output in a sensitivity analysis seem to be measurements which are very important. Gaschnig *et al.* (1983) point out that it is pertinent to carry out a sensitivity investigation. There are often many variables involved, therefore it is probably difficult to carry out this kind of study, although efforts should be made to do a sensitivity analysis. The development validation phases, however, may be more appropriate for this purpose than the laboratory evaluation.

It is difficult to assess whether an attempt at measuring the reasoning would add to the evaluation procedure. There is a number of problems which arise when trying to approximate reasoning. Firstly, it is not known whether the methods described in the literature indeed measure reasoning. A system's subconclusions may be measured, but although the subconclusions must be correct, the subconclusions do not necessarily have to resemble the experts'

think aloud protocols. Even if it is possible to measure reasoning, experts may reason in different ways, which makes determining the reasoning an even more difficult problem. Furthermore, the potential user may not be interested in the system resembling expert reasoning, but may want the reasoning to resemble the way he would himself reason had he known the answer. There are also systems for which it is not desirable to have the reasoning resemble the expert's reasoning, for instance, if the system is concerned with time critical decisions. However, intermediate conclusions should be correct. Detailed subsystem analysis should be carried out during dynamic validation of a system.

Judging the results. Depending upon the nature of the domain it may or may not be possible to objectively determine the correctness of the output of the system. In domains where there is only one correct answer which the system should have given, the result may be interpreted directly (Adlassnig and Scheithauer, 1989; Kingsland, 1985; Nelson *et al.*, 1985; Murray *et al.*, 1986; Reggia, 1985). Even when direct comparison is possible, it is often necessary to specify a correctness rule (Aikins *et al.*, 1983; Kingsland, 1985; Miller *et al.*, 1982). In some evaluations, only the first in a list of possible diagnoses is taken into account (François *et al.*, 1992), and in other evaluations the output is interpreted as being correct when the diagnosis is included in a list of possible answers (Nelson *et al.*, 1985). The correctness rule will strongly influence the results of the evaluation.

In cases where the results may be interpreted directly, and where retrospective test cases are used, the term predictive validation is used by O'Keefe *et al.* (1987). However, there are also domains for which it is not possible to carry out a direct comparison. This may occur when multiple answers are expected or when the answer is neither completely correct nor completely incorrect. The results will have to be analysed by a number of experts. These experts must not know the origin of the output, i.e. whether it has been produced by the knowledge based system or by a physician. This is called a blind evaluation (Hickam *et al.*, 1985; Quaglini *et al.*, 1988). Using experts to judge the results again introduces subjectivity into the measurement and raises the reliability and consensus problems described in Section 2.3.2.2. Therefore the inter- and intra-expert variability will have to be investigated.

When a panel is involved in judging the results, the Delphi method (Section 2.3.2.2) may be by used to obtain a judgement. Another approach (Mirkin, 1979) which may be interesting, but has not been seen in the evaluation of medical knowledge based systems is to take a weighted average of opinions, where each participant has a particular weight according to their competence on the subject. The weights may be established in various ways. For instance,

each participant is asked to rate the competence of all other participants. A matrix of competence grades assigned to the experts can then be obtained. When taking the positive eigenvector of this matrix, the vector entries will be the weights which are allotted to the participants. Another method for finding weights is a method which evaluates the competence of experts with respect to the level of consistency of their evaluations with those of the majority. A matrix of experts' opinions, regarding the objects to be evaluated, is drawn up. Its transpose and the matrix are multiplied and the positive eigenvector of this multiplication will again consist of the weights to be used.

2.3.3. ANALYSIS OF THE RESULTS

There is a number of statistical methods which can be used for the analysis of the results. Indurkhya and Weiss (1989) describe various models for measuring performance of medical knowledge based systems. Each method has its own particular advantages and drawbacks. In any evaluation study, various methods of analysis will have to be used, and the results have to be studied from different points of view. Examples of most of the methods described in this section may be found in Appendix 1. It is not possible to present an exhaustive view of all the methods which are described in the literature on the evaluation of medical knowledge based systems, however, the methods of analysis which are used most often will be discussed below.

The analysis of the results can be divided into two different parts. Firstly, calculating measures of performance and secondly, hypothesis testing. Both parts will be described in some detail.

Calculating measures of performance
A number of general approaches for calculating measures of performance has been found in the literature. The first category, consisting of error rate methods, is found to be used most often, whereas the second approach, consisting of confidence level methods, is used in addition to the first by some researchers. The latter kind can only be used for systems which provide either probabilities or certainty factors with their output. The final category of methods for performance measurement is based upon a calculation of agreement.

Error rate methods

Basic error rate method. The basic error rate method is based on the notion that the output given by a system is either right or wrong. The error rate of the system is the number of incorrect cases divided by the total number of cases, and the accuracy is the number of correct cases divided by the total number of cases. An illustration of this may be seen in Example 1 of Appendix 1. The error rate method has been used by Catanzarite *et al.* (1981), Kingsland (1985), Wong *et al.* (1990) and Miller *et al.* (1982).

In addition to being either correct or incorrect, an alternative is to incorporate a class for cases which are partially correct. This does mean, however, that a scoring scheme will have to introduced to be able to classify a case as partially correct, and the scoring scheme will usually be domain dependent. Percentages of cases which are correct, partially correct and incorrect may then be calculated. This does not have to be limited to one degree of partial correctness, the scoring scheme could involve varying degrees of partial correctness.

If there is a possibility of more than one answer per case, and for the system to give more than one answer per case. Then, instead of using correctness of the complete case, another possibility is to define the accuracy as the number of correct system answers divided by the total number of answers of the standard. However, this should be coupled with some measure of how precise the system was, for instance, the positive predictive value, which is defined as the total number of correct answers divided by the total number of answers given by the system (Indurkhya and Weiss, 1989).

Positive negative correctness model. The positive negative correctness model is an extension to the error rate model. There are four categories which the output given by the system may fall into, these are: true positive, true negative, false positive and false negative. Regard the most simple form for which there is one diagnosis which is either present or absent. The true positive situation occurs when both standard and system agree that a certain diagnosis is present, the answer is true negative if both system and standard agree that the diagnosis is absent. The answer is false positive if the system regards the diagnosis to be present whereas according to the standard it is absent, and the false negative situation occurs when the system regards a diagnosis to be absent whereas according the standard it is present. The positive negative correctness is always carried out with respect to a certain diagnosis. A number of additional metrics, beside those mentioned above, may now be calculated. These are shown in Example 2 of Appendix 1. The positive negative correctness model has been used in the study carried out by François *et al.* (1992). The most important metrics are sensitivity and

specificity. Sensitivity is the number of true positive answers divided by the number of answers for which the diagnosis was actually present. This is often called the true positive ratio. Specificity is defined as the number of true negative answers divided by the number of answers for which the diagnosis was actually absent. The metric (1 - specificity) is often called the false positive ratio. The cost or risk of making a false positive or false negative judgement may also be taken into account, by multiplying these conclusions with a cost factor (Indurkhya and Weiss, 1989).

If the system output contains more than one diagnostic category per patient, whereas only one answer can be true, then a correctness rule will have to determine the true positive situation. The two rules which are described in Adlassnig and Scheithauer (1989), are firstly that an answer is true positive if the standard diagnosis is among those given by the system, or that an answer is only true positive when the standard diagnosis corresponds to the top diagnosis as given by the system.

In the last situation, there may be systems which incorporate some kind of certainty factor in the answers. In these cases, if a certainty factor threshold is introduced, the list of possible diagnoses may be reduced. Diagnoses with certainty factors of less than 0.5 may for instance not be included in the list of possible diagnoses. If the threshold is increased to 0.8, then fewer diagnoses will be included in the list. The sensitivity and specificity of the system may be calculated for different values of the threshold (Example 3 of Appendix 1). A curve of sensitivity against (1-specificity) may then be drawn for the various threshold values. This kind of curve is called a Receiver Operating Characteristic (ROC) curve. The area under the curve may serve as a performance measure. The ROC method has been used by Adlassnig and Scheithauer (1989) by varying an internal threshold, and by Bernelot Moens and van der Korst (1991). The ROC method was used by de Dombal and Horrocks (1978) not for reducing the list of possible diagnoses, but by changing the threshold for the probability of the final answer to be present. If, for example, the probability of appendicitis being present is over 50%, and 50% is taken as the threshold, then system is considered to have made a prediction for appendicitis.

Confidence level methods

If the system gives probabilities, weights or certainty factors, then a number of additional methods is available for calculating performance parameters. These methods take into account the weights given by the system.

Accuracy coefficient. An accuracy coefficient has been proposed by Zagoria and Reggia (1983), which has been modified from Shapiro (1977), who uses a

logarithmic coefficient. The coefficient allows the accuracy of correct answers with a higher weight to be better than the accuracy of correct answers with a lower weight. (Example 4 of Appendix 1). The accuracy coefficient used by Reggia has been criticised by Nykänen *et al.* (1990) for encouraging overconfident diagnostic statements. Another modified accuracy coefficient is defined by Bernelot Moens and van der Korst (1991), this coefficient is also described in Example 4 of Appendix 1. According to Bernelot Moens and van der Korst (1991), in their study this accuracy coefficient was heavily influenced by the large number of correct predictions of low probability made for absent diagnoses.

Distance metrics (Indurkhya and Weiss, 1989). The n diagnostic categories given may be represented in an n dimensional space. Each axis is the certainty factor of a diagnosis, and the squared distance between the correct answer and the system's answer may then be calculated for each case. The average squared distance over all cases will then be the performance measure. An example of this method is shown in Example 4 of Appendix 1. This kind of performance measurement using a distance metric is popular among neural net systems (Indurkhya and Weiss, 1989).

Reliability. If the system gives actual probabilities rather than weights or certainty factors, the following method can also be used. It provides an answer to whether the assigned probability means what it should mean. The method has for instance been used by Bernelot Moens and van der Korst (1991). If the average probability of the predictions for the first diagnosis is 0.8, then 80% of these cases in the test population is expected to have the predicted diagnosis. This can then be compared to the observed number of correct diagnoses. This method has been described by Hilden *et al.* (1978) as a reliability measure. Hilden *et al.* (1978) also mention other approaches of reliability analysis.

No standard methodology for calculating measures of performance has been found in the literature. However, in the medical knowledge based system literature, the positive negative correctness model is often found. If there is a list of diagnoses for which the threshold is adaptable, then ROC curves can be drawn up. If certainty factors or probabilities are given then an accuracy coefficient or distance metric can also be calculated. If probabilities are given, as is the case in for instance Bayesian systems, then the reliability of the system can be investigated.

Agreement methods

Rather than using correctness measures, some researchers investigate the agreement between physicians and the knowledge based system. The two approaches which have been encountered most frequently are a calculation of the percentage of outputs for which physicians and knowledge based system agree (Alonso-Betanzos *et al.*, 1989; Aikins *et al.*, 1983), or the Kappa coefficient of agreement (Alonso-Betanzos *et al.*, 1989; Reggia, 1985; Spitzer and Endicott, 1969; McDermott and Hale, 1982; Kors *et al.*, 1989).

Kappa coefficient of agreement. The Kappa coefficient (Cohen, 1968) is often used as a measure of agreement for use with nominal scales. The Kappa coefficient is a chance corrected measure of agreement. An illustration of the calculation of Kappa may be seen in example 5 of Appendix 1. When Kappa equals 1 there is complete agreement, and when Kappa is 0 this equals the agreement expected by chance alone. A weighted form of Kappa (Cohen, 1968) also exists and may alternatively be used.

Hypothesis Testing

Depending on the goal of the evaluation relevant hypotheses have to be tested. One goal of the evaluation may be to investigate the difference between the calculated system and human performance. Depending on the object of the system, another goal of the evaluation may be to show that the system has an expert level of problem solving.

Investigating whether there is a significant difference between the system and a human problem solver. The null hypothesis in this case is: The system performs equally well as the human problem solver. To test whether there is a significant difference, for categorical variables, Hickam *et al.* (1985) have used Chi-square or the Fisher exact test, and François *et al.* (1992) have used Wilcoxon. For variables which are normally distributed the t-test can be used, or an analysis of variance can be carried out (see, for example, Hickam *et al.*, 1985; Yu *et al.*, 1978) to determine whether a significant difference is present.

Investigating whether the system performs equally well as the human expert. The null hypothesis in this case is: System and human expert agreement is less than or equal to a certain amount. This means that the agreement between system and human expert will first have to be calculated. The Kappa coefficient (Cohen, 1968) is often for this. The value of Kappa should be proven significantly more than a certain value. It is unclear, however, which

value this should be, since it is difficult to interpret Kappa. The value which Kappa should significantly exceed should certainly not be zero, because significance above zero does not indicate whether the agreement is enough (Nykänen, 1990). Cohen (1968) states that a substantial value for the lower confidence limit is a more meaningful criterion. Nykänen proposes to investigate the deviation from perfect agreement. To test whether the answer differs significantly from a certain value other than zero, a calculation of the standard error as given by Fleiss (1981) may be used for a test of significance (Example 5 of Appendix 1).

In the preceding analysis, the standard output has been taken to be undisputed. However, when experts are involved in the evaluation, inter and intra-expert variability have to be calculated. The Kappa measure of agreement may be used for this. The influence of the inter- and intra-expert variability on the results of the evaluation will have to be investigated.

2.3.4. THREATS TO THE VALIDITY

There is a number of sources of bias and confounding which can be present in performance measurements of a knowledge based system. These situations must be avoided or taken into account when analysing the results of the evaluation. The possible threats to the validity which are mentioned in the literature most often, will be discussed below.

Pro- and anti computer bias (Chandrasekaran, 1983; Fieschi, 1990; Gaschnig *et al.*, 1983). The evaluation always has to be blind in order to account for this. However, this usually means coding the results which may also introduce biases of errors.

Coding (Chandrasekaran, 1983). In evaluations which involve blind judging of the output, the outputs may have to be coded in order for the source of the output to be unrecognisable to the judges. It must not be possible to distinguish which answers originate from the human-machine system and which answers originate from the unaided physician. The coding will introduce subjectivity into the measurement, since there will undoubtedly be an influence of the people who code the answers, and probably limits the amount of evidence which is available to the judges.

Circularity (Wyatt and Spiegelhalter, 1990). This may be a problem if a decision aid is built and evaluated by the same individual or team, or performs a classification task using the same data and criteria as assessors.

Parochial bias (Kingsland, 1985). The cases used to test the system may not be representative of the complete population. Furthermore, potential users involved in the evaluation may also not be representative of the complete population. Gaschnig *et al.* (1983) also report the bias which may occur when a system operates in a limited domain and receives preselected test cases. On transfer of the system to another location, results may be different.

In laboratory evaluations it is possible to avoid most sources of bias and confounding. In an evaluation possible threats to the validity must be identified first, after which it should be established whether it is possible to avoid these. If this is not possible, the influence on the outcome of the measurement should be described and estimated.

2.3.5. LABORATORY EVALUATION RESULTS

The conclusions which can be drawn from a laboratory evaluation will consist of the results of the analysis viewed in the light of the goals of the evaluation study. A number of domain and system characteristics and certain aspects of the evaluation design will influence the results of an evaluation. With regard to the evaluation design, for example, the correctness rules will have an effect on the outcome of an evaluation. It will make a difference whether only the top answer of a list is taken into account or whether the output merely has to be present somewhere in the list. However, the evaluation design is not the only influence on the results. Various domain and system characteristics will also be of importance. The size of the domain and also the number of different categories which the system categorizes its answers into will be important. The results of two systems even in the same domain, where one system differentiates into ten categories and one differentiates into two categories will be different and it will thus not be possible to compare the results.

2.3.6. RESULTS OF REPORTED LABORATORY EVALUATIONS

Table 1 in Appendix 2 shows a summary of reported evaluations of medical knowledge based systems. The framework which was introduced in Section 2.2 has been used to summarize the evaluations and provides a context for the comparison of the evaluation procedures. Not all items of the framework have been found in all articles, however, those aspects which have been mentioned either explicitly or implicitly have been entered, otherwise the

Table 2.2. Summary of aspects of importance in laboratory evaluations of knowledge based systems.

evaluation framework	conclusions regarding laboratory evaluations
object of evaluation	depending on goal of system show potential improvement on present situation; show expert performance; quantify goals of evaluation
experimental setup	give same input to KBS, physicians and standard; compare outputs
test input	representative (and challenging) retrospective patients; statistically estimate minimum number to be chosen (e.g. using specification of difference worth detecting); if number not possible, estimate influence of practical limitations
human-machine	knowledge engineer enters data
test against	depending on object evaluation potential users, experts; estimate numbers statistically
standard	if true answer exists, use it as a standard; otherwise standard may be obtained by consensus analysis (e.g. Delphi method); take intra- & interexpert variability into account (e.g. Kappa)
comparison	variables measured: system output, subsystem output judging outcome: directly by observation, or if not possible then indirectly involving experts (e.g. Delphi, weighted average), take into account problems of unknown validity and intra- & interexpert variability
analysis	-performance measures (e.g. error rate methods, confidence level methods, agreement) -hypothesis testing; difference with users (e.g. Chi-square, ANOVA, t-test), agreement with experts (e.g. Kappa & z-test)
bias & confounding	some potential bias: pro- anticomputer, circularity, parochial, coding; most can be eliminated

cells have been left empty. The purpose of this table is not to give an exhaustive view of all evaluations which have been carried out, but rather to provide an overview of the variety of methods which have been used, from relatively simple to relatively complex evaluation designs. The table has been drawn up in alphabetical order of the authors. It is interesting to note the many different approaches taken in the evaluation studies, some of which have been carried out over a decade ago.

The table also shows a selection of some of the results which have been obtained in these evaluation studies. Most papers report good results, and many have very promising conclusions. However, many of the papers do not (yet) have follow-up papers. For example, it is striking that the number of reported field evaluations is many times smaller than the number of reported laboratory evaluations.

As has been noted above, it may be seen that various domain and system characteristics, such as the size of the domain, multiple or single answers, certainty factors or no certainties, makes the evaluation designs, methods of analysis and outcomes so different, that a comparison of results of systems in the same domain is very difficult, let alone comparisons of results of systems for different domains.

2.3.7. CONCLUSIONS

A performance evaluation framework has been used for the discussion of the laboratory evaluation of medical knowledge based systems. Each step in the framework has been discussed in the subsections. The final paragraphs of the subsections contain the conclusions which may be drawn from the literature concerning laboratory evaluations. The conclusions which were stated in the subsections have been summarised in Table 2.2. In the table, the framework of steps which have to be taken in the design of an evaluation are shown in the left hand column, and the conclusions which may be drawn from the literature regarding those steps are shown in the right hand column of the table.

2.4. Field performance evaluation

Only few actual field evaluations of medical knowledge based systems have been carried out. Most of the reported field evaluations concern Bayesian systems, which are based on probabilities and not on knowledge based technology. However, in this discussion of performance evaluation they have been included. Although, for example, verification and development validation would not be the same for Bayesian and knowledge based systems, for the performance evaluation most of the methodology is similar.

2.4.1. SELECTION OF A GOAL FOR THE FIELD EVALUATION

The prerequisite to any formal field evaluation is that the patient's safety is assured (Miller and Sittig, 1990). Informal clinical testing is carried out under close supervision of the system's developers and a select group of trial users. It serves as a check on the system's clinical safety and allows the developer to obtain feedback from the physicians on the system's overall performance. Prior to a field evaluation the knowledge based system can be used in parallel (in the background) with the current situation. Field evaluation encompasses investigation of a large number of aspects, including the investigation of the impact of the system on physician actions, on patient care, and on health care processes, a cost benefit analysis, the examination of subjective reactions, and the investigation of system use (Miller, 1986). However, although all the above aspects must be investigated, a limited number of goals must be chosen to investigate during a field study. The objective of a medical knowledge based system is usually to assist physicians with certain tasks, thereby improving final patient outcome. This means that a goal for a field study could be to investigate enhancement of final patient outcome. It will be shown below that this is a measurement which is fraught with difficulties, bias and subjectivity, and is therefore very complicated.

Another kind of field study is described by Kent *et al.* (1985), who investigate the influence on data completeness when a knowledge based system is used, and by McDonald *et al.* (1984) who measure the response rates to reminder messages given to physicians. White *et al.* (1984) investigated the difference in physicians' actions in response to computer generated alerts.

The main goal of a performance evaluation of a decision aid in the clinical environment is to show that the performance of the human-machine system is superior to the performance of the physician without a knowledge based system. Therefore the final goal of the field evaluation could be to study

whether patient outcome improves due to application of a knowledge based system.

2.4.1.1. Quantification of the goal
The quantification of the goal for the field test of performance is more straightforward than in the measurement of technical performance. The final patient outcome should improve significantly due to the application of a knowledge based system. However, in most evaluation studies described in the literature it has been investigated whether the physician produces significantly better decisions in cooperation with a knowledge based system than without the knowledge based system. It should then be exactly defined what is understood by a better decision. However, this is hardly ever stated explicitly in the literature.

2.4.2. EVALUATION SETUP

The default design for any interventional trial is the randomised controlled trial with double-blinding (Wyatt and Spiegelhalter, 1990). However, there is a number of aspects which influences the design which can be used for the evaluation of knowledge based systems. For instance, the user is usually completely free to accept or to reject the recommendation given by the system (Spiegelhalter, 1983).

Spiegelhalter divides studies into experimental, where a controlled trial takes place with balanced allocation to the control and experimental groups, or quasi-experimental where performance is measured before and after introduction of the system. Examples of studies for which historical controls have been used are Murray (1990) and Adams *et al.* (1986).

Ideally a full experimental design is used in the clinical evaluation of the human-machine system. This must take the form of a multi-centre trial, since it is the physician's task which is being studied. However, in actual practice, a number of limitations will often exist, such as time limitations, lack of systems, lack of patient data, or difficulty with randomisation, which will require adaptations to the evaluation design.

Choice of experimental unit. According to Spiegelhalter (1983), the choice of the experimental unit is a difficulty in designing trials of medical decision aids. Spiegelhalter gives a rough categorisation for the choice of the experimental unit. The patient is chosen as the experimental unit if the system provides immediate information and recommendation useful for a particular patient. If the system educates the physician about careful data collection,

clinical judgement and awareness of performance then the physician is chosen as the experimental unit, and if it generates an awareness among a small group of physicians, the group is chosen. Murray (1990) uses neurosurgical units as the basic observational units. McDonald *et al.* (1984) chose practice teams as unit of randomisation.

Usually the object of a knowledge based system is to assist a physician in carrying out a certain task, therefore the physician should be studied. However, since within a certain hospital (or within a certain team) there will be communication between those using and those not using the system, ideally the team or hospital should be chosen as the unit of randomisation. In technical domains, the choice of experimental unit may be more obvious. Adelman (1991) states that in experiments organisational units would be randomly assigned to situations with and without the decision support, and their performance measured when it is stable. The unit of analysis is the performance of the organisational unit.

2.4.2.1. Selection of test input

Task difficulty should be as representative of the operational environment as possible (Adelman, 1991). Prospective patient data will be used in a field evaluation. It will be difficult to determine, however, to which extent the representativeness has been satisfied. In some studies where historical controls are used, the controls consist of prospective cases which have been gathered during a baseline study (Murray, 1990; Adams *et al.*, 1986).

A difference between Bayesian and knowledge based systems which is important in performance evaluation is the fact that Bayesian systems are based on probabilities, which are obtained from data. This means that there is usually no lack of (test) cases, and often many cases are involved in an evaluation, whereas with knowledge based systems this may sometimes prove to be a problem. In the study carried out by Sutton (1989a), 6962 cases were involved, Adams *et al.* (1986) used 16737 cases (4075 during baseline and 12662 during test period, system used by physicians in 3451 cases). Whereas, for instance, an evaluation of the knowledge based system QMR (Bankowitz *et al.*, 1989) involved 31 patients, and Kent *et al.* (1985) used 180 patient visits (system used by physicians for 56 visits) in an evaluation of ONCOCIN.

It is necessary to use a representative sample of patients in the clinical trial. To investigate whether it may be possible to statistically interpret the findings, statistical estimations should be carried out in advance, for example, using sample size tables for differences between proportions (Cohen, 1977).

2.4.2.2. Consultation
A comparison has to be made between various situations, consisting of human-machine system, physicians and a standard of performance. These situations have to be clearly defined prior to the evaluation.

Specifying who uses the system (human-machine system)
In most actual field evaluations, the physician enters the relevant information into the computer, as in the studies carried out by Sutton (1989a), Bankowitz *et al.* (1989) and Kent *et al.* (1985). However, in some studies the data are entered into the computer by research technicians and automatically (McDonald *et al.*, 1984). In the investigation carried out by Adams *et al.* (1986), some of the cases were entered by research assistants and the other cases were entered by the physicians. Some systems obtain their data automatically from a database in which data are routinely stored (White *et al.*, 1984).

The object of the evaluation is to determine the efficacy of the human-machine system when used in the target environment. Therefore, the data should be entered and the system used by those who would also do this if the system were in routine use.

Specifying physicians to test against
The object of the investigation is to test whether the physicians' performance improves through application of a knowledge based system. Therefore, the performance of the human-machine system is compared to the performance of potential users of the system. To form the control and experimental groups, balanced allocation to the situation with and without a knowledge based system may be used (experimental study). However, quite a large number of experimental units may be needed. Therefore some researchers use historical controls (Adams *et al.*, 1986; Murray, 1990). This means that the control data are collected during a period prior to the introduction of the knowledge based system. However, these designs are not as effective in controlling extraneous factors (Adelman, 1991).

The numbers of physicians, groups and hospitals involved in actual field evaluations varies from 27 teams involving 130 physicians (McDonald *et al.*, 1984), 3 centres (Sutton, 1989a), 8 centres and over 250 physicians (Adams *et al.*, 1986), to consultants and ward teams in 2 hospitals (Bankowitz *et al.*, 1989).

Ideally a full experimental design should be used, where the situation with and without a knowledge based system can be studied. However, since there is often a limitation to the number of centres or teams which can be involved in

such a trial, it may not be possible to balance the centre or team using the system with a similar centre or team without a system. Therefore, one of the few possibilities may be to use historical controls.

Specifying a standard of performance
The choice of a standard depends on the variables which are to be measured (Section 2.4.2.3). In most field evaluations described in the literature, human-machine output is measured, rather than patient outcome. In which case a standard for the output has to be chosen using the methods which were mentioned in section 2.3.2.2 on laboratory evaluations.

Bankowitz *et al.* (1989) established a definite diagnosis in 20 of the 31 cases. A diagnosis could be established through histologic, radiographic etc. confirmation, or by clinical means if the patient met certain criteria. Sutton (1989a) uses the diagnosis ultimately assigned by the consultant in charge of the case as the standard. In the investigation carried out by Adams *et al.* (1986), the discharge diagnosis was chosen to be the standard of performance

2.4.2.3. Comparison

Variables to be compared. The variables which are to be measured depend on the goal of the field evaluation study. In most evaluation studies, the output of system and user is compared to the output of the unassisted user, where the output will consist of a decision or a number of decisions to be taken. However, it is not satisfactory to solely measure decisions, because in medical domains for instance, the diagnosis may be correct, but it may for instance delay the treatment of the patients, so that final patient outcome may be worse (Wyatt, 1987). The improvement of final patient outcome is usually an important goal of a knowledge based system.

The object of some systems is also to reduce the number of special investigations or to reduce the use of resources, which should then also be measured. Adams *et al.* (1986) measure the use of resources, including rates of stay in hospital, the number of special investigations and the financial implications.

Kent *et al.* (1985) investigate the improvements in data completeness when a knowledge based system is used. The completeness of the data after use of the knowledge based system is compared to the completeness of the data when the system is not used. In the study performed by White *et al.* (1984) the effect of the system on patient management is studied by means of the distribution of possible system output related actions.

To determine whether the human-machine system has a superior performance to the unassisted physician, the decisions taken by the physicians using the

knowledge based system, and the decisions taken by physicians without the knowledge based system are usually measured. However, as is described above, improvement in diagnosis (decision) does not necessarily imply an improved patient outcome. Therefore, it will also be necessary to measure, for example, final patient outcome and speed. If there are other objectives to the knowledge based system, then the dependent variables which aim to measure the influence of these objectives should also be measured.

Judging the results. The judgement of the results may take place in the same way as described above in Section 2.3.2.3 on laboratory evaluation.

2.4.3. ANALYSIS OF THE RESULTS

The same methods of analysis which were mentioned for measuring laboratory performance may be used for field evaluation studies. Measures of performance which have been used in actual field studies are for instance the error-rate method (Adams *et al.*, 1986), complemented with confidence intervals (Bankowitz *et al.*, 1989), and the positive negative correctness method (Sutton, 1989a).

The choice of experimental unit will also have to be taken into account in the statistical analysis. The discussion in Section 2.3 on laboratory evaluation is centred around the patient as the experimental unit. Adams *et al.* (1986) state that with the doctors as experimental unit some adjustment is necessary to the p-values associated with tests on patient statistics.

The results are dependent on the user for entry of the data and acceptance or rejection of the results. The dependence on the user makes it difficult to generalise the results, unless a cross-section of physicians and institutions have participated (Spiegelhalter, 1983). Furthermore, there are many more possible threats to the validity than there are in a laboratory evaluation, making the interpretation of the results of field trials more difficult.

2.4.4. THREATS TO THE VALIDITY

Possible threats to the validity, in addition to those mentioned in section 2.3.4 on laboratory evaluation, are the following:

Carry-over effect (Wyatt and Spiegelhalter, 1990). This is the possible positive effect on performance due to education of the user by the system.

According to Wyatt and Spiegelhalter (1990) this effect may be compensated by raising the size of the experimental unit, or by quantifying the effect by studying alternating knowledge-based system and control periods. The latter depending on the trial either being multi-centre with randomized or asynchronous periods, or there being no significant changes during the trial period.

Hawthorne effect (Wyatt and Spiegelhalter, 1990). This is the effect by which performance might be expected to improve merely by being seen to measure it in a trial. This effect will be common to both trial and controls. However, according to Spiegelhalter (1983), the Hawthorne effect will tend to decrease any relative benefit of the system. The effect may be quantified by performing a low-profile baseline study (Wyatt and Spiegelhalter, 1990).

Secular trends (Wyatt and Spiegelhalter, 1990). These are changes in the measures of interest which occur during the evaluation period, and which may influence the outcome of the study. They are particularly damaging in studies using historical controls, and which run for a long period of time, or when there are changes in the way a particular task is carried out.

Feedback effect (Spiegelhalter, 1983; Wyatt and Spiegelhalter, 1990). A decision-aid will often make it easier for clinical performance to be monitored, and feedback to the physician may act as a stimulus to improvement.

Checklist effect (Wyatt and Spiegelhalter, 1990). The knowledge based system may encourage a more complete and structured data collection. The discipline imposed by structured data collection may offer a major contribution to clinical insight (Spiegelhalter, 1983). Kent *et al.* (1985) have assessed the influence of a computer-based chemotherapy treatment consultant on the completeness of clinical trial data. Adams *et al.* (1986) have used a design in which there were four different groups, one group which used structured data collection forms, one group which used forms and the diagnostic aid, one using forms and receiving feedback, and one using forms and the diagnostic aid and receiving feedback.

Expert judgement. If experts are involved in the evaluation, this will introduce errors of unknown validity of the judgements, and intra- and interexpert variability. The variabilities may be approximated, however the validity of the judgements will not become known.

Trial size. The trial size will influence the statistical conclusion validity of the evaluation study.

In clinical evaluations of medical knowledge based systems, there are many threats to the validity of the results. Some effects may be eliminated or may be (partly) compensated for. It may depend on the trial design as to whether it is possible to compensate or eliminate these effects. If this is not possible, then the influence on the results of the trial must be estimated.

2.4.5. FIELD EVALUATION RESULTS

The results of the analysis will provide information to allow conclusions to be drawn about the goals of the evaluation. As was mentioned in Section 2.3.5 on laboratory evaluations, there are many aspects which will influence the results of an evaluation, including various domain and system characteristics, and the design of the evaluation.

2.4.6. RESULTS OF REPORTED FIELD PERFORMANCE EVALUATIONS

Only few field evaluations of clinical efficacy have been reported in the literature. Table 2 in Appendix 2 shows a summary of reported field evaluations. The evaluations have been summarised according to the framework for evaluation design. It can be seen that various evaluation designs have been used in these field trials. There are some studies for which the diagnostic accuracy of the human-machine system is higher than the diagnostic accuracy of the unaided physician. The overall results, however, are not as positive as those which were mentioned in the results of laboratory evaluations (Table 1 of Appendix 2). It is not possible, either for laboratory evaluations or for field evaluations, to objectively compare results for different systems and different domains, because domain characteristics, such as domain size and number of categories, system characteristics, and evaluation design have too much influence on the evaluation results.

Table 2.3. Summary of aspects of importance to performance evaluations in the field.

evaluation framework	conclusions regarding field evaluations
object of evaluation	show the human-machine system significantly improves the unaided situation
experimental setup	ideally randomised controlled double blind trial; often quasi experimental design is used; experimental unit: physician, or group of physicians (potential users)
test input	prospective cases; estimate minimum number to be chosen (e.g. using minimum difference worth detecting); if not possible in practice then estimate influence of limitation
human-machine	potential users
test against	potential users, estimate number to be chosen
standard	depending on goal of evaluation and variables measured. if true answer exists, use this as standard; otherwise standard may be obtained through consensus analysis (e.g. Delphi), take intra- & interexpert variability into account (e.g. Kappa)
comparison	variables measured: patient outcome, human-machine output, cost, impact judging results: directly by observation; if not possible then indirectly, involving experts (e.g. Delphi, weighted average), creates problems of unknown validity and intra- & interexpert variability
analysis	difference with unaided users, by hypothesis testing
bias & confounding	many additional potential sources of bias and confounding: carry-over effect, Hawthorne, secular trends, feedback, checklist some may be eliminated

2.4.7. CONCLUSIONS

The performance evaluation framework (Figure 2.1) has been used for the discussion of the clinical evaluation of medical knowledge based systems. Each step in the framework has been discussed in the subsections. The final paragraphs of the subsections contain the conclusions which may be drawn from the literature concerning field evaluations. The conclusions which were stated in the subsections have been summarised in Table 2.3. In the table, the framework of steps which have to be taken in the design of an evaluation are shown in the left hand column, and the conclusions which may be drawn from the literature regarding those steps are shown in the right hand column of the table.

2.5. Laboratory evaluation of human-machine systems

The discussions in Section 2.3 and 2.4 have addressed two kinds of evaluation, laboratory evaluation of the knowledge based system and field evaluation of the human-machine system. However, it appears that in the medical knowledge based system literature, laboratory evaluation of the complete human-machine system has not been mentioned, i.e. investigations where potential users work with the system. The laboratory evaluation is usually directed towards the performance of the knowledge based part of the system by itself. It seems that the evaluation of the complete human-machine system is missing from most evaluation studies, whereas this could provide additional information about the safety and potential usefulness of the knowledge based system. Such an investigation should be performed prior to a field evaluation and will incorporate aspects of both the laboratory investigation of a knowledge based system as well as of the field evaluation of the human-machine system, which were discussed in the previous sections.

The setup is similar to the setup which will be used in a field evaluation. A difference is that retrospective test cases are used, as they are in other laboratory investigations. Furthermore, the variables which may be measured are the same as in other laboratory investigations, since only outputs of the human-machine system can be measured, rather than variables such as final patient outcome. The threats to the validity of the investigation will be similar to those which may be present in a field evaluation. However, due to the fact that the experiments are conducted in the laboratory environment, for example, secular trends will not be present. A summary of the aspects of the

evaluation framework which are important in a laboratory evaluation of the human-machine system can be seen in Table 2.4.

Table 2.4. Summary of aspects of importance to laboratory performance evaluations of the human-machine system.

evaluation framework	conclusions regarding human-machine laboratory study
object of evaluation	show the human-machine system significantly improves the unaided decisions
experimental setup	ideally randomised controlled experiment; experimental unit: physician (potential users)
test input	retrospective cases; estimate minimum number to be chosen (e.g. using minimum difference worth detecting); if not possible in practice then estimate influence of limitation
human-machine	potential users
test against	potential users, estimate number to be chosen
standard	if true answer exists, use this as standard; otherwise standard may be obtained through consensus analysis (e.g. Delphi), take intra- & interexpert variability into account (e.g. Kappa)
comparison	variables measured: human-machine output judging results: directly by observation, if not possible then indirectly, involving experts (e.g. Delphi, weighted average); creates problems of unknown validity and intra- & interexpert variability
analysis	difference with unaided users, by hypothesis testing
bias & confounding	potential sources of bias and confounding: e.g. pro-anticomputer, circularity, parochial, coding, Hawthorne, checklist; some may be eliminated

2.6. Comparison between laboratory and field evaluations

A comparison of laboratory and field evaluations may be seen in the light of the aspects of evaluation which have been discussed. A summary of the differences is shown in Table 2.5. In the left hand column of Table 2.5, the evaluation framework is shown. The second column shows the conclusions regarding the design of laboratory evaluations of the system itself, the third columns shows the design of laboratory evaluations of the human-machine system, and the final column shows the conclusions regarding the design of field evaluations. The differences between the designs can be seen in the table.

Table 2.5. Comparing laboratory and field performance evaluations of medical knowledge based systems.

evaluation framework	laboratory evaluation system	laboratory evaluation human-machine system	field evaluation human-machine system
object of evaluation	depends on goal system; show potentially useful, show expert performance	show improved decisions through use of system	clinical efficacy
experimental setup	compare system and physicians on same cases	experiment	often historical controls experimental unit physician/group
test input	retrospective	retrospective	prospective
human-machine	knowledge engineer	potential users	potential users
test against	depends on goal evaluation; potential users, experts	potential users	potential users
standard	depends on domain; actual standard or group choice	depends on domain; actual standard or group choice	depends on goal, variables; depends on domain; actual standard or group choice
comparison	variables: decision output, subsystem output, sensitivity output judgement: depends on domain; direct or number of experts	variables: decision output judgement: depends on domain; direct or number of experts	variables: patient outcome, decision output, cost, impact judgement: depends on domain; direct or number of experts
analysis	difference with users, agreement with experts	difference with unaided users	difference with unaided users
bias & confounding	some bias & confounding; most can be eliminated	more bias & confounding due to interaction with system	more bias & confounding due to external factors

2.7. Conclusions

Validation should be carried out continually and in parallel with the design of a knowledge based system. Three different validation activities were identified: verification, dynamic validation and evaluation. These activities are important during all phases of the development process. Looking at the practical research which has been carried out in this area, the activities are usually directed at a product rather than at the complete development life cycle. Furthermore, it is usually assumed that the reasoning mechanism has been verified and validated and works according to specifications. It may be concluded that validation research should be directed at the complete development life cycle and should include the complete human-machine system, rather than concentrating on the knowledge that has been represented in the system. Since design and validation are closely related, validation research will also benefit from more structured and formal methods of knowledge based system design.

Verification is the procedure which is least domain dependent and which can greatly improve the knowledge based system at low cost. It is of the utmost importance to apply verification methods from the beginning of the development. Dynamic validation is a process which is aimed at improving the behaviour of the system, and may require test case generation. It is a procedure which has not received much attention in the literature, but which deserves more research and which should become more important, for it appears that empirical evaluation is always restricted in the number of cases and people involved, because of time limitations. Therefore, efforts should be made at developing methods of system improvement and error retrieval methods which place as little time constraints as possible on those involved.

At the end of the development life cycle, after thorough verification, dynamic validation and evaluation have been carried out, two different empirical evaluation processes may be distinguished: laboratory evaluation and field evaluation. In this chapter, a framework (Figure 2.1) has been described for the design of empirical performance evaluation studies. This framework indicates the choices which have to be made when designing a performance evaluation. It is important to decide on a limited number of important goals for the evaluation and then exactly define the evaluation design according to the items in the framework. There are many different goals which need to be investigated, however, in an evaluation study it is necessary to select a number of goals, as it is not possible to study all relevant aspects in one study. The framework is not an exhaustive list of items which should be considered in an

evaluation, but should be seen as a basis for discussion on evaluation of knowledge based systems. The conclusions which can be drawn from the literature concerning laboratory evaluations have been summarised in Table 2.2. There are still many difficulties which arise when performing laboratory evaluations. Therefore, no really satisfactory method of evaluation has been found in the literature.

The same framework for the design of a performance evaluation study has been used as the basis for the comparison of a number of laboratory evaluations which have been described in the literature. As can be seen from this comparison (Table 1 in Appendix 2), many papers report laboratory evaluations, using just as many different methods for performing the evaluation. There are various domain and system dependent factors which determine the choices made in the design of an evaluation and in the methods of analysis used, therefore it is hardly possible to compare the results in one domain, let alone comparing results of systems in different domains.

No laboratory experiments of complete human-machine systems were found in the medical knowledge based system literature. However, in order to investigate the potential usefulness of a knowledge based system it seems important to also perform such an evaluation. A proposal for the design of laboratory evaluations of the human-machine system is shown in Table 2.4.

After the laboratory evaluation, a system should be evaluated further. Depending on the criticality of the system, it may first be used in parallel with people working in the field, and then it may be used in an actual prospective field trial. The framework of evaluation design (Figure 2.1) has also been used for the discussion of field evaluations. Field evaluations suffer from more threats to the validity than laboratory evaluations, and the results will therefore be even more difficult to interpret. A proposal for the design of field evaluation studies is shown in Table 2.3. A summary of reported field evaluations can be seen in Table 2 of Appendix 2. In field evaluations there are also many aspects which will have to be studied, whereas during one field trial only a limited number can be addressed. Many other aspects will have to be investigated besides the performance requirements which have been discussed in this chapter. This will include, for example, cost benefit, impact on physicians, ethical and legal issues. Furthermore, after a field study, the evaluation will not be finished, since the influence of long term use will have to be investigated and aspects of maintenance and updating with their evaluation should receive attention

From this discussion, it becomes clear that validation is indeed a continual process, and only a limited part of this process, performance evaluation, has been regarded in this chapter. Furthermore, dynamic validation is of the utmost importance in the development of knowledge based systems, however, there are no really satisfactory well tested methods for knowledge based system validation.

It can be seen that many investigators are very positive after laboratory evaluations. However, quite often no further evaluations of the systems are reported. Therefore, it is clear that laboratory evaluations of the performance of the system only are just one aspect of evaluation, and the complete human-machine system should also be evaluated at this stage of development. However, the cause may even lie at the very beginning of system development, the problem definition. First, it has to be assured that potential users indeed require assistance in a certain task, so there must be a proven need for a system. Furthermore, there may be another problem which is inherent in the paradigm which is used in conventional expert system design, where the computer is designed to be a machine expert. Woods *et al.* (1987) call this paradigm, which is often used in expert system design, the 'cognitive-tool-as-prosthesis' paradigm, and Miller and Masarie (1990) refer to it as the 'Greek Oracle' model. When looking at the complete human-machine system, both system and user are trying to solve the same problem in parallel (Rossi-Mori *et al.*, 1990; Lipscombe, 1989). The user's task in cooperating with the system is as an interface between system and the environment, which includes evaluating the system's advice. Another approach to system design would be to design the system as an 'instrument' (Woods *et al.*, 1987). From the 'cognitive-tool-as-instrument' perspective, computational technology should be used, not to make or recommend solutions, but to aid the user in the process of reaching a decision (Woods *et al.*, 1987). Such knowledge based systems will require other modes of interaction with the user. The aim of the system is to really provide assistance, rather than both user and system solving the same problem.

The discussion on performance evaluation in this chapter was centred around systems which are designed according to the machine expert paradigm. Application of the 'cognitive-tool-as-instrument' paradigm will undoubtedly influence the validation methods which are appropriate.

3 Computer-assisted diagnosis and treatment planning of brachial plexus injuries. The knowledge based system PLEXUS

The brachial plexus is a network of nerves which is situated in the area between the neck and the arm, and innervates the muscles of the shoulder, arm and hand. Results of a retrospective study of patient files have shown that the localization of brachial plexus injuries and determining the appropriate treatment plan are complex problems, which may potentially benefit from computer assistance. The knowledge based system PLEXUS has been developed to assist physicians in the diagnosis and treatment planning of brachial plexus injuries. PLEXUS uses patient history information and results of neurological, neurophysiological and radiological examinations. The system's graphical user interface is based on a familiar scheme, and does not require previous computing or typing experience. Preliminary evaluation studies of the system's problem solving performance have produced promising results.

3.1. Introduction

The brachial plexus is a very complex network of nerves, which innervates the muscles of the shoulder, arm and hand. Injuries of the brachial plexus most often occur in young men during motorcycle accidents. The diagnosis and management of brachial plexus injuries are reputed to be very difficult and specialist tasks. To investigate whether computer assistance in the domain of brachial plexus injuries could overcome some of the problems associated with the diagnosis and treatment planning of brachial plexus injuries, a retrospective study of patient files was carried out, and a questionnaire was distributed among a number of physicians. These studies are described in Section 3.3. As a result, the knowledge based system PLEXUS was developed. In order to illustrate the features of the knowledge based system PLEXUS, some background knowledge of brachial plexus injuries is necessary. Therefore, the anatomy of the brachial plexus and possible causes of brachial plexus injuries are firstly discussed in Section 3.2.

Various other neurological advisory systems have been mentioned in the literature. Some of these systems are directed towards the central nervous system and others are aimed at localizing peripheral nerve injuries, such as brachial plexus injuries. The different methods of knowledge representation used in these systems will be discussed in Section 3.4.

PLEXUS consists of a diagnostic and a treatment planning module. The architecture of both knowledge modules, the inference (i.e. reasoning) methods which are used, and the design of the graphical user interface will be dealt with

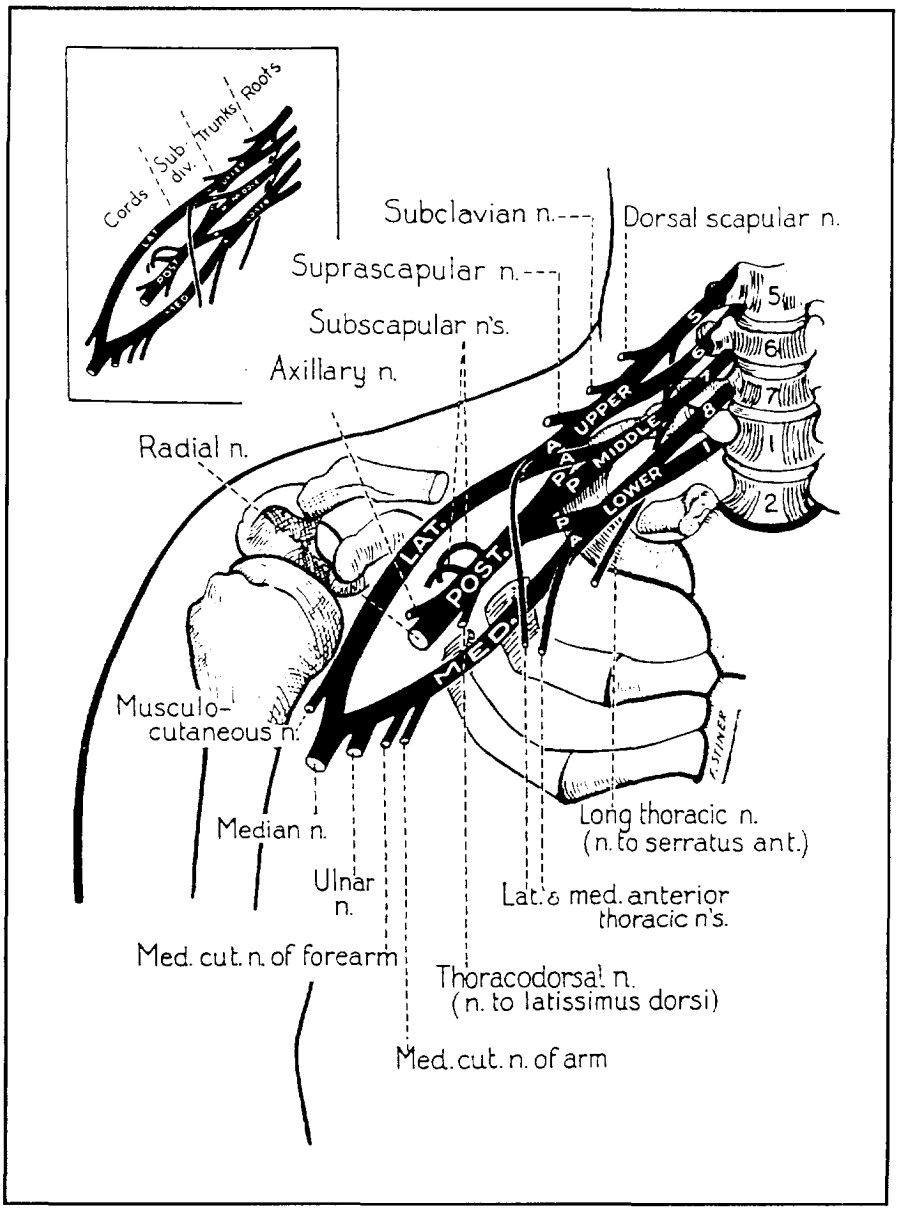


Figure 3.1. Anatomy of the brachial plexus (Haymaker and Woodhall, 1945). Reprinted with permission of the publisher.

in some detail in Section 3.5. The difference between PLEXUS and other knowledge based systems in the field of brachial plexus injuries will also be discussed.

Preliminary evaluation of the problem solving performance of PLEXUS has been carried out. The results of the preliminary performance evaluation studies are described in Section 3.6. The evaluation results were very promising, which encouraged extensive formal evaluation of the system. The formal evaluation of PLEXUS is the topic of discussion in the following chapters.

3.2. The brachial plexus

3.2.1. ANATOMY OF THE BRACHIAL PLEXUS

The brachial plexus is a network of nerves which is situated in the area between the neck and the arm, and which innervates the muscles of the shoulder, arm and hand. In addition, the nerves of the brachial plexus provide the sensory function in the arm and hand, and also carry autonomic fibres which can, for instance, stimulate the sweat glands and constrict the blood vessels.

In order to provide the background which is necessary for the rest of this chapter, the anatomy of the brachial plexus will be mentioned briefly. Detailed discussion may be found in, for example, Kerr (1918), Sunderland (1968) and Leffert (1985). The anatomy of the brachial plexus is shown in Figure 3.1. The brachial plexus generally originates at the five spinal nerves C5, C6, C7, C8 and T1 which leave the spinal cord. These spinal nerves are indicated by the white numbers 5, 6, 7, 8 and 1 in Figure 3.1. The spinal nerves join and divide to form a network of nerves. The spinal nerves are formed by the union of motoric nerve rootlets and sensory nerve rootlets which arise from the spinal cord. The motoric nerve rootlets are situated at the front (ventral) side of the spinal cord and the sensory nerve rootlets are situated at the back (dorsal) side. This can be seen in Figure 3.2.

Part of the brachial plexus is situated above the clavicle, this is called the supraclavicular part of the brachial plexus, and part of the network is situated below the clavicle, this is called the infraclavicular part of the brachial plexus.

Supraclavicularly, C5 and C6 usually join to form the truncus superior (or upper trunk), C7 forms the truncus medius (or middle trunk), and C8 and T1 make up the truncus inferior (or lower trunk). This is shown in Figure 3.1. A number of nerves leaves the plexus supraclavicularly; these are the n.dorsalis scapulae, the n.thoracicus longus, n.suprascapularis and the n.subclavius.

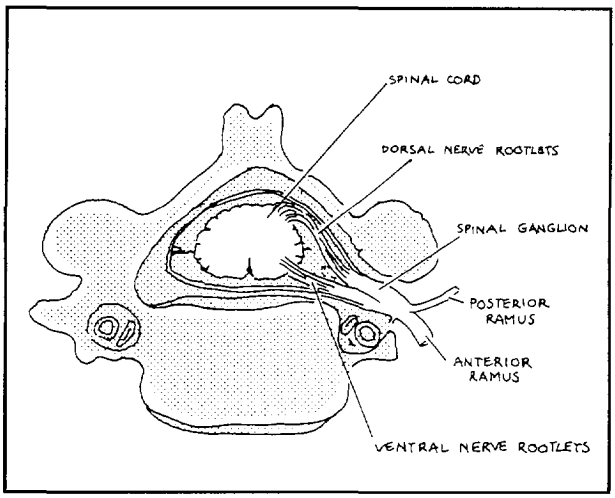


Figure 3.2. Formation of a spinal nerve (Jaspers, 1990).

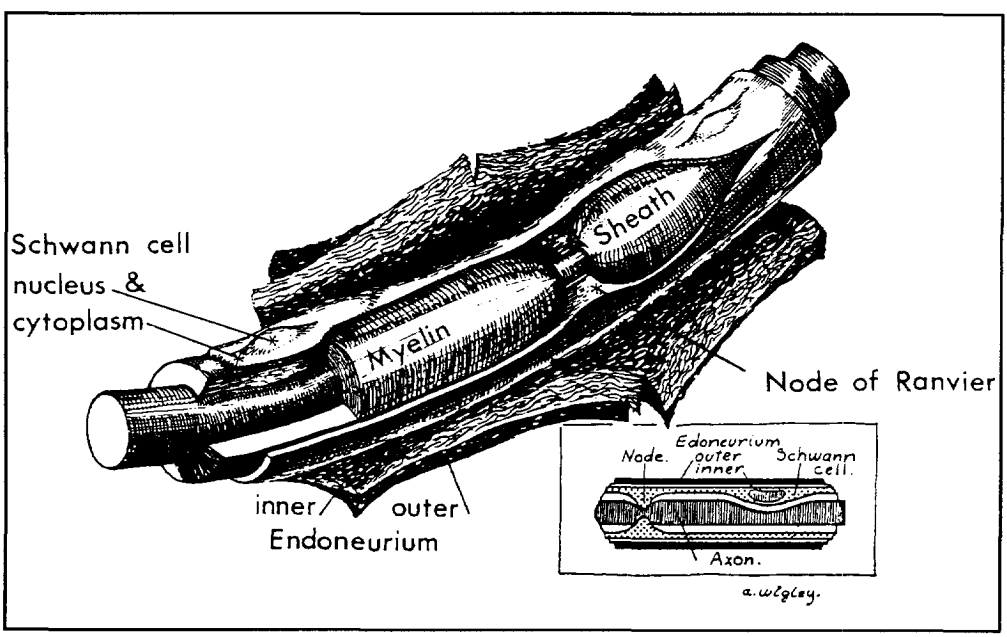


Figure 3.3. Features of a myelinated nerve fibre (Sunderland, 1968). Reprinted with permission of the publisher and the author.

The trunci each divide into an anterior (front) part and a posterior (back) part. This is indicated by the letters A and P in Figure 3.1. The anterior parts of the truncus superior and the truncus medius join and form the fasciculus lateralis (or lateral cord). The posterior parts of all three trunci join to form the fasciculus dorsalis (or posterior cord) and the anterior part of the truncus inferior forms the fasciculus medialis (or medial cord). The fasciculi are situated infraclavicularly, i.e. below the clavicle. A number of nerves directly leaves the fasciculi. The n.pectoralis lateralis leaves the fasciculus lateralis. The n.subscapularis and the n.thoracodorsalis leave the fasciculus dorsalis, and the n.pectoralis medialis leaves the fasciculus medialis. Finally, the fasciculi divide into the peripheral nerves which supply the muscles in the arm. The fasciculus dorsalis divides into the n.axillaris and the n.radialis. The fasciculus lateralis divides into the n.musculocutaneus, and part of the fasciculus lateralis joins with part of the fasciculus medialis to form the n.medianus. The fasciculus medialis also forms the n.ulnaris.

Individual variations of the general anatomy may also occur. The brachial plexus may, for example, be formed by the roots C4 to C8 or T1. This is called a prefixed plexus. Another possibility is a postfixed plexus, which is formed by roots C5 or C6 to T2.

3.2.2. PATHOLOGY OF BRACHIAL PLEXUS INJURIES

Brachial plexus injuries may be characterised according to the locations which are injured and the severity of the injury. The locations which may be injured consist of the anatomic structures which were discussed above, for example spinal nerve C5, truncus superior or fasciculus lateralis. The severity of a brachial plexus injury may be classified according to the structures in the nerve which are affected. A peripheral nerve consists of nerve fibres surrounded by supportive tissues. A nerve fibre consists of an axon which is the prolongation of the nerve cell, a myelin sheath, and Schwann cells whose main function is to form the myelin sheath. An example of a nerve fibre is shown in Figure 3.3. The nerve fibres are surrounded by connective tissue called endoneurium. A funiculus (fascicle) is a bundle of nerve fibres invested by a sheath of connective tissue, the perineurium. The epineurium comprises all the connective tissue outside the perineurium. An illustration of these structures is shown in Figure 3.4.

The severity of a nerve injury may be classified according to the nerve structures which are involved. The classification according to Seddon (1943) is well-known, and is defined as follows:

- Neurapraxia: a lesion in which there is no axonal degeneration.

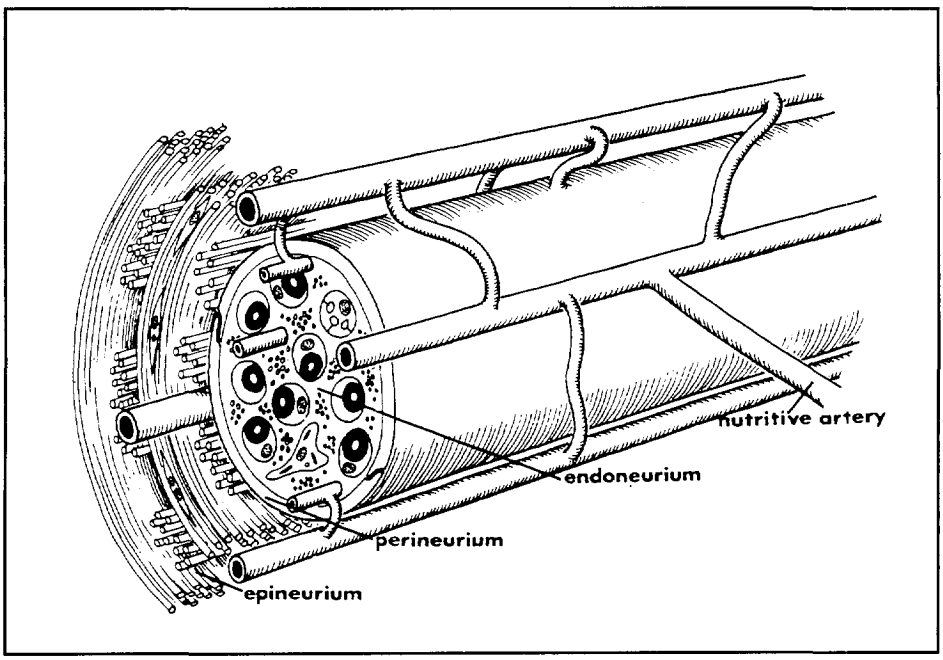


Figure 3.4. Representation of a peripheral nerve (Bischoff, 1975). Reprinted with permission of the publisher.

- Axonotmesis: a lesion characterized by complete interruption of axons, but with preservation of the supporting structures of the nerve -Schwann tubes, endoneurium and perineurium.
- Neurotmesis: is the name given to a lesion of such severity that all essential parts of the nerve are destroyed. The simplest and commonest variety of the condition is that resulting from anatomical division, but interruption of the same kind can occur without any apparent loss of anatomical continuity.

The classification according to Sunderland (1968) is also used very often. This divides the severity of the injury into five degrees:

- First degree damage: interruption of conduction in the affected axons with preservation of the anatomical continuity. Recovery is usually rapid and complete.
- Second degree damage: loss of continuity of axons. Continuity of the endoneurial sheath of the nerve fibres is preserved. Spontaneous recovery will usually be complete, but is delayed due to axon regeneration.
- Third degree damage: loss of continuity of nerve fibres. Endoneurial tube continuity is destroyed. The perineurial sheath and funicular continuity are preserved. Recovery in individual structures takes place more slowly and is usually incomplete.
- Fourth degree damage: loss of continuity of funiculi. Only the epineurium maintains the continuity of the nerve. The onset of recovery is unduly delayed, the course of recovery is grossly irregular and the end result is functionally insignificant.
- Fifth degree damage: lesion in which the affected segment of the plexus has been ruptured or cleanly severed. There is no spontaneous recovery at all.

A further kind of injury is one in which the nerve roots are avulsed (torn away) from the spinal cord, this may affect motoric or sensory rootlets alone, or both. Avulsions are the most serious type of plexus injury.

A common aftermath of trauma to the brachial plexus is fibrosis. This may be localised or diffuse. The fibrocytic reaction occurring in damaged tissues may resolve or progress to permanent scarring. Such fibrosis may constrict nerve fibres and impair their blood supply and, in these ways, delay or prevent regeneration, or delay or prevent the restoration of function in axons that have regenerated (Sunderland, 1982).

Table 3.1. Etiology of brachial plexus injuries (Narakas, 1993).

Cause of brachial plexus injury	number	%
traction injuries	1028	66
secondary compression after trauma	17	1
gun shot or missile injury	24	1.5
iatrogenic injury	40	2.5
lacerations	8	0.5
obstetrical palsy	281	18
post-radiation	88	6
tumours	39	2.5
varia	30	2
total	1555	100

Table 3.2. Etiology of brachial plexus injuries (Slooff, 1993).

Cause of brachial plexus injury	number	%
traumatic:		
- traction/crush - lacerations - gunshot wounds	315	48
- obstetric	240	37
- iatrogenic	27	4
tumours	30	4.5
entrapment syndromes	12	2
irradiation	10	1.5
miscellaneous	21	3
total	655	100

3.2.3. ETIOLOGY OF BRACHIAL PLEXUS INJURIES

The brachial plexus may be damaged by traction, compression, penetration or due to non-traumatic causes. The distribution of injury causes in the patients seen by Narakas (1993) is shown in Table 3.1, and the distribution of injury causes in the patients seen by Slooff (1993) is shown in Table 3.2. Traction is the most frequent type of injury in brachial plexus lesions. These injuries are usually due to a forceful widening of the angle between the shoulder and the neck, or between the upper arm and the trunk (Jaspers, 1990). Illustrations of various injury mechanisms may be seen in Figure 3.5.

Traction injuries most frequently occur during road traffic accidents. Approximately 70% of traumatic brachial plexus injuries are due to traffic accidents, and approximately 70% of the lesions in traffic accidents involve the use of a cycle or motorcycle (Narakas, 1985). In Table 3.1 and Table 3.2 it can be seen that brachial plexus injuries which occur during the delivery of babies, i.e. obstetrical injuries, also constitute a significant percentage. Although the percentage in Table 3.2 may be influenced by the fact that dr. Slooff specializes in obstetrical brachial plexus lesions.

3.2.4. DIAGNOSIS

In order to decide on the appropriate therapy it is necessary to obtain a precise diagnosis, consisting of the exact locations within the brachial plexus, which are injured and of the severity of the injury. It is not possible to directly measure the state of the nerves, therefore indirect measurements are needed. This requires extensive neurological, neurophysiological and radiological examinations, the results of which have to be interpreted and combined with patient history information. The most important data which are needed will be discussed briefly below. For a more complete discussion, see Jaspers (1990).

Patient history information. It is very important to know the exact cause of the injury, as this may provide a clue regarding the extent and severity of the injury. For instance, high velocity injuries are often more severe than injuries which take place at a low velocity. The additional trauma which is present may also provide an indication of the severity of the injury. When patients have sustained multiple additional injuries, they are more likely to have a severe brachial plexus injury.

Neurological examination. The motor function examination is of the utmost importance for determining the locations within the brachial plexus which are injured. The strengths of the muscles in the upper extremity have to be

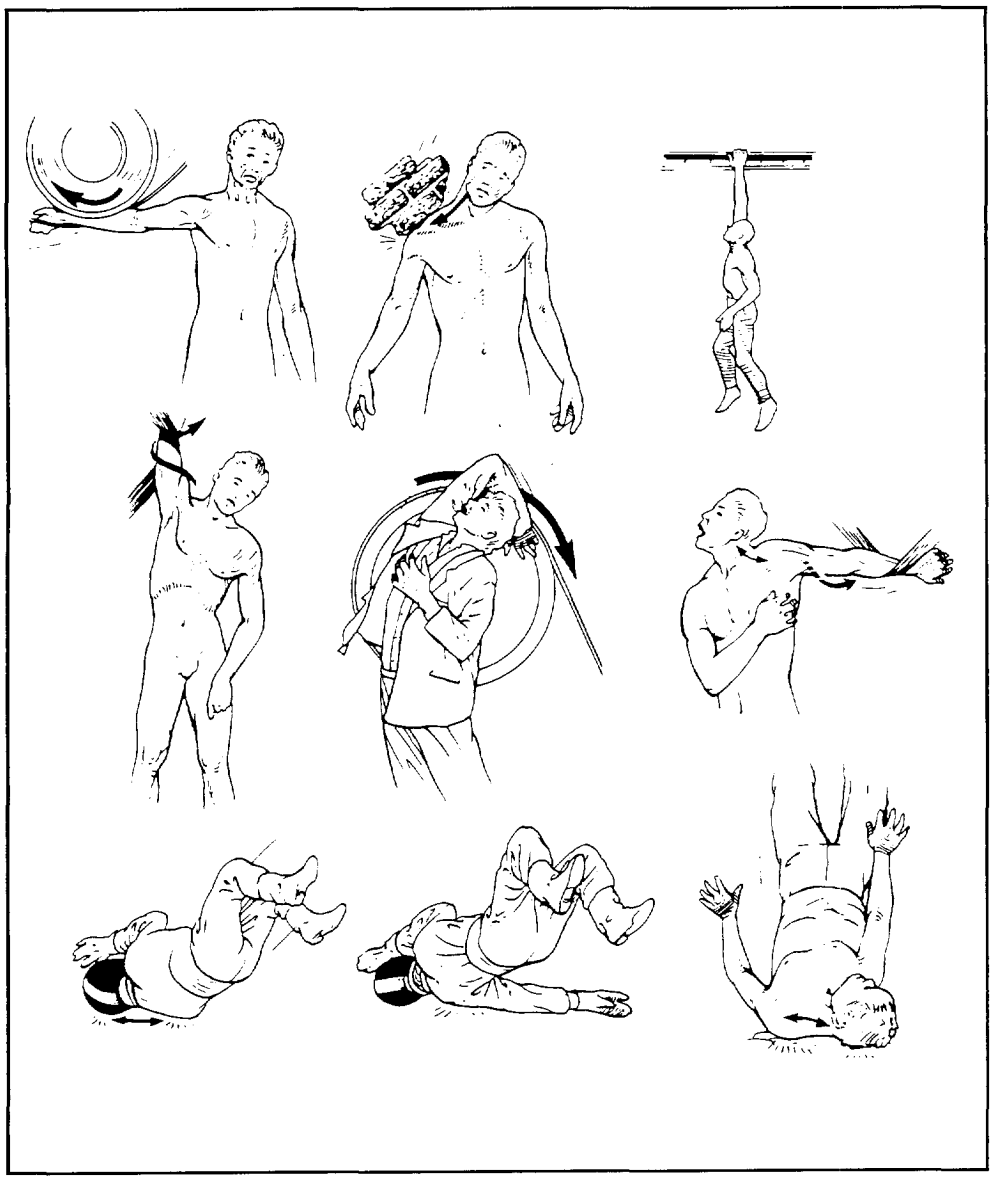


Figure 3.5. Trauma mechanisms resulting in supraclavicular brachial plexus injuries (Coene, 1985). Reprinted with permission of the author.

investigated very thoroughly. There is a number of key muscles which provide especially relevant information, such as the m.rhomboideus, m.levator scapulae, m.supraspinatus, m.infraspinatus and the m.serratus anterior. These muscles are innervated by nerves which leave the spinal nerves or brachial plexus at an early stage and are therefore indicative of the level of the injury.

Sensory examination of the segmental or peripheral nerve innervation areas provides further information. However, due to overlap of neighbouring dermatomes the resulting sensory deficit might be smaller than expected (Thomeer, 1991).

The sign of Tinel-Hoffman may be elicited by percussion of the brachial plexus. A painful feeling indicates the presence of a lesion for which a connection with the spinal cord has been preserved. The radiation of the painful feeling towards a certain area on the arm refers to the location of the lesion.

Another important clinical sign is the presence of Horner's syndrome which is characterised by a small pupil, drooping of the upper eyelid and absence of sweat secretion on the forehead, all on the same side of the body as the brachial plexus injury. These functions are normally innervated by axons from nerve roots of T1. The presence of Horner's syndrome is indicative of root avulsion of at least T1 (Thomeer, 1991).

Neurophysiological examination. An important neurophysiological examination is electromyography (EMG). This provides additional information about the state of the muscles, and allows the testing of muscles which are otherwise inaccessible. The EMG may reveal findings, such as early reinnervation which may not yet be detectable by physical examination. A further neurophysiological examination is measurement of somatosensory evoked potentials (SEP). This examination gives qualitative information about the connection between a sensory nerve and the central nervous system. Sensory nerve action potentials (SNAP) give information about the conduction of a nerve. In case of preganglionic injury (i.e. root avulsion) the sensory axons remain intact, therefore this method allows discrimination between pre- and postganglionic injuries to the plexus. Theoretically, SNAPs are positive in a pure preganglionic injury. Although this examination is used extensively, the interpretation of the findings may be difficult due to, for instance, possible simultaneous presence of both pre- and postganglionic lesions.

Radiological examination. The radiological examination comprises three investigations. Firstly, a plain X-ray of the clavicle, cervical spine, scapula and humerus is made. Any damage to these structures is also indicative of the severity of the lesion.

The second examination involves an investigation of the area where the spinal nerves leave the spinal cord. One of the reasons for performing a cervical

myelography combined with CT-scan is that it may provide information as to whether nerve roots are avulsed. The absence of the rootlets on the radiology makes an avulsion of the root from the spinal cord very probable. Meningeal tears frequently result in cyst formation filled with contrast fluid, called meningoceles, which can be seen on the CT(myelogram). The existence of meningoceles, especially of extensive meningoceles outside the foramen, is a sign of a severe lesion to the roots, and is very suspect for an avulsion.

Sometimes an angiogram of the subclavian artery is necessary. The severity of any vascular injury is also indicative of the severity of the brachial plexus injury.

3.2.5. THERAPY

Jaspers (1990) discusses a general scheme for treating traumatic brachial plexus injuries. Sharp lesions, such as stab wounds, have to be treated surgically immediately. In all other cases, in the early stages when an accurate diagnosis still has to be established, associated injuries, such as fractures, vascular injuries and head trauma are treated. Conservative treatment is started to prevent contractures and to control pain, and orthoses are provided.

When an accurate diagnosis is available, a decision is made as to whether the patient should be considered for neurosurgical repair. The injuries which are treated surgically are usually the more serious injuries which will not recover spontaneously. Depending on the nature of the injury, there are different neurosurgical procedures which may be performed. One such procedure is nerve grafting. A nerve graft is a length of donor nerve which may for instance be taken from the leg, in order to replace the injured part of a nerve. Another surgical procedure is a nerve transfer. In plexal root avulsions (nerve roots which have been torn away from the spinal cord) the only possible means to restore continuity is by coaptation with neighbouring nerves, either from within the brachial plexus, i.e. intraplexal nerve transfer, or from outside the plexus, i.e. extraplexal nerve transfer. It may be necessary to use both nerve grafts and nerve transfers in order to reconstruct a brachial plexus.

When nerves are expected to be in continuity but no recovery occurs, a possible operative procedure is the removal of scarring tissue. This procedure is called a neurolysis. It is indicated only in late cases, and it is a potentially hazardous procedure. After neurosurgery, conservative treatment is again necessary to mobilize joints, to re-educate reinnervating muscles and for psycho-social support.

When the final prognosis is definite, secondary surgery may be considered. This may, for instance, entail transferring certain muscles or tendons, or fixation

of certain joints. A comprehensive discussion about the treatment of brachial plexus injuries may be found in Alnot and Narakas (1989).

3.3. Need for assistance

To determine whether there is a need for assistance in the domain of brachial plexus injuries, two points of view may be distinguished. Firstly, retrospective treatment results can be studied to determine objectively whether patient management could possibly improve if physicians would use a computer advisory system. Secondly, there must be a recognised need for assistance on the part of the physicians who are the potential users of such a system. Reasons for an objective need for assistance will be discussed first, after which the opinion of potential users of a decision support system for brachial plexus injuries will be considered.

3.3.1. OBJECTIVE NEED FOR ASSISTANCE

In order to investigate the difficulties associated with the diagnosis and management of patients with a brachial plexus injury, Jaspers (1990) performed a retrospective study on 136 patients who had been referred to the rehabilitation centre 'De Hoogstraat' in The Netherlands from different hospitals across the country. Of these 136 patients, 93 patients had been admitted to the rehabilitation centre before 1981 and 43 patients were admitted between 1981 and 1985. Jaspers (1990) identified a number of problems, including the following:

- Localization of brachial plexus injuries is a very complex process. This difficulty is due to the complex anatomy of the brachial plexus and to possible anatomic variations.
- In the early stages of the injury, there are often associated injuries which require immediate attention, so that the brachial plexus injury is left unattended.
- Diagnosis is often neglected by the referring clinic, because neurosurgical possibilities may not be known, or because physicians may have a pessimistic view on the results of reconstructive neurosurgical procedures. Of the 43 patients who were admitted to 'De Hoogstraat' between 1981 and 1985 and who were involved in the investigation, 21% had not received any additional diagnostic tests in the referring clinic besides motor and sensory examination. For 47% of the 136 patients who were studied, the referring physician had not recorded a diagnosis, indicating the site, the extent or the severity of the injury, in the patient file.

- Patients are often referred to a rehabilitation centre at a very late stage. For the patients who were admitted to 'De Hoogstraat' after 1981 and who were involved in the investigation, the average time to admission was 12.5 months and the median time to admission was 3 months.
- Only few patients are treated neurosurgically. In the group of patients admitted to 'De Hoogstraat' between 1981 and 1985, the percentage of nerve repairs was 25%.

The full results of the study have been described by Jaspers (1990). From the above, it is clear that there is a need to improve the diagnostics of brachial plexus injuries, to increase an awareness of the possibilities of neurosurgical treatment and of the necessity to refer brachial plexus patients to a specialist centre at an early stage.

3.3.2. SUBJECTIVE NEED FOR ASSISTANCE

In addition, it is necessary to investigate whether there is a need for assistance on the part of potential users of a computer advisory system. Grolman (1989) performed a preliminary study among 67 neurologists in The Netherlands. One of the aims of the study was to investigate the need for assistance. For this investigation, a questionnaire was developed and distributed among the neurologists. Since the sample was not an aselect sample of Dutch neurologists, and only 19 of the 67 questionnaires were both returned and at least partly completed, careful interpretation of the results is required. Some results of the study are shown in Table 3.3.

Table 3.3. Results of a questionnaire distributed among 67 neurologists in The Netherlands.

Number sent	67			
Number returned and completed	19 (28%)			
Question	very/yes	fair/some	poor/no	no answer
System will be used	7 (37%)	4 (21%)	5 (26%)	3 (16%)
Would use system personally	8 (42%)		5 (26%)	6 (32%)
Expect problems for introduction	2 (10%)	7 (37%)	6 (32%)	4 (21%)

In the table, it can be seen that according to 11 of the 19 neurologists a computer program in the domain of brachial plexus injuries will be used in practice if it is available, and 8 out of 19 physicians would personally use the decision support system. Some problems may be expected in introducing decision support systems. Some of the problems which were mentioned are:
- lack of time,
- the physician will have to get used to the system,
- the computer must not interfere with the physician's critical thinking,
- these systems do not work except for small (sometimes trivial) domains,
- difficulties in determining the quality of the knowledge,
- lack of computing experience.

It has to be noted that the letter which accompanied the questionnaire explicitly mentioned that the physician has the final responsibility when using a decision support system and that the conclusions which are drawn by such a system are meant only as advice. This may explain the fact that the question of responsibility and the nature of the conclusions which are often identified as problems to the introduction of knowledge based systems, were not mentioned by these physicians.

Physicians were also asked to indicate any criteria related to the acceptance of decision support systems. Some of the items mentioned, regarded requirements for the system to be:
- easy to use/ user friendly,
- quick,
- reliable,
- of good medical quality,
- of practical use,
- made by well-known and experienced physicians,
- well tested,
- easily accessible,
- able to motivate conclusions.

From the above, it can be seen that these physicians are positively inclined towards a decision support system for the diagnosis and treatment planning of brachial plexus injuries, although there are some physicians who do not think that such a system would be used. There is, however, a number of requirements which will have to be met by such a system, in that it must be of impeccable medical quality and it must not be time consuming to use. Furthermore, such a system must be user friendly and be well validated.

3.4. Neurological advice giving systems

Various neurological advice giving systems have been described in the literature. Some of these systems are meant to assist in the diagnosis of disorders of the central nervous system and others are directed towards assistance in the domain of the peripheral nervous system, which also entails brachial plexus injuries. A discussion on systems for brachial plexus injuries will follow a short description of a number of knowledge based systems for other areas of neurology.

3.4.1. NEUROLOGICAL KNOWLEDGE BASED SYSTEMS

Various knowledge based systems have been developed in the domain of neurology. One of the features that most distinguishes neurological localization from general diagnostic problem solving is its use of spatial knowledge, i.e. neuroanatomy (Reggia *et al.*, 1986). The importance of the (neuro)anatomy makes neurology an interesting domain for developing knowledge based systems. The knowledge representation methods which have been used in neurological knowledge based system have progressed in parallel with other medical knowledge based systems. Some of the first computer systems in medical decision making were based on the Bayes' theorem. A Bayesian system for application in clinical neurology is discussed by Salamon *et al.* (1976), who describe an experiment in computer aided diagnosis of a number of disorders covering both the brain and the spinal cord.

Shortly afterwards, rule based systems became popular among the medical decision making community. Reggia (1978) used the rule based approach for localization of damage to the central nervous system. The system used results of the neurological examination of patients in a coma to categorize these unconscious patients. The main purpose of this work was to evaluate the suitability of the rule based methodology for representing knowledge about neurological localization.

The rule based representation was found to be a poor representation for neurological localization because localization knowledge is conceptually organised in a frame-like fashion and is very context dependent. According to Reggia (1978) a conceptual, visually-oriented representation is used by physicians in localizing damage to the nervous system.

Geometrical methodologies. Catanzarite *et al.* (1981) and Banks and Weimer (1985) used geometric methodologies for representing neuroanatomic knowledge. Catanzarite *et al.* (1981) developed the NEUROLOGIST system for consultation in clinical neurology, which firstly localizes the neurological disease

and then uses these data as well as the mode of disease onset to rapidly focus on a limited number of possibilities which are then sequentially investigated. The anatomic localization submodule consists of a database of horizontal sections through the central nervous system from the spinal cord to the cerebral cortex. At each level, the system generates a convex polygon including all malfunctioning tracts present at that level.

Banks and Weimer (1985) partitioned the nervous system into a hierarchical set of nested cubes. Each cube is divided into 27 smaller cubes until the smallest cubes are reached which are each 3 mm. on a side. In addition to the cubes, the knowledge base contains anatomic objects. The cubes and objects are associated with lists of properties which describe the relationships of the cubes and objects.

These systems are interesting with respect to the graphical possibilities which are provided, however, according to Xiang *et al.* (1985), the analogical geometrical approach has a major disadvantage which does not invalidate but limits the conclusions which can be derived from anatomical analysis. It oversimplifies the real life situation, because it does not provide appropriate levels of abstraction and flexibility.

Propositional methodologies. Propositional representation of the knowledge allows more flexibility and abstraction possibilities. A propositional representation has, for instance, been used by First *et al.* (1982) in the knowledge based system LOCALIZE which uses a network of objects and links to represent the anatomical knowledge. The system is meant to assist physicians in localizing lesions in the peripheral nervous system, and will therefore be discussed in more detail in the next section. However, propositional representation also has limitations (Xiang, 1985) in that not all structural information can be abstracted in the form of propositions, certain geometrical details are lost, and graphics and imaging processing techniques are not supported because they rely on geometrical data.

Reggia *et al.* (1986) describe a system which is intended to be a general framework for neurological localization and diagnosis, and which presently focuses on the problem of neurological localization in the cerebrum, brainstem and cranial nerves. The problem solving knowledge consists primarily of associative knowledge organised in a hierarchical semantic network. This network includes, for example, causal relationships between disorders and manifestations, spatial relationships between anatomical loci, and containment relationships between spatial loci and physiological systems.

An important object of the work concerns the study of plausible reasoning in neurological knowledge based systems. More specifically, one of the aims is to test and extend parsimonious covering theory as an inference method for

knowledge based systems. Parsimonious covering is a method which finds the minimum number of disorders which best explains the manifestations which are present. The method handles simultaneous disorders and is justifiable in terms of past empirical studies of diagnostic reasoning.

Geometrical and propositional representation. The most recent approaches to modelling neuroanatomic knowledge combine geometric and propositional representations (Xiang *et al.*, 1985; Ohe and Kaihara, 1988; Niggeman, 1990). The papers describe representation methodologies which will allow various kinds of inference, rather than being limited to entering signs and obtaining the location of lesions.

Xiang *et al.* (1985) use a semantic network approach for representation of spatial structure and function of the neuroanatomy. A physical entity, each of its physical-spatial properties and its function are all independent concepts, which relate to each other when, in combination, they describe the entity. Analogical or geometrical, propositional and functional knowledge are integrated into a single network.

Ohe and Kaihara (1988) describe a system which uses three levels of anatomical knowledge: topological, functional and geometrical. Topological and functional knowledge are represented using PROLOG, and the geometrical knowledge is described using a special methodology to convert the position of an anatomical object in a diagram into the form of a list.

Niggeman (1990) describes the ANATOM system. This system contains anatomical knowledge in three different representation formalisms. Propositional representation, two-dimensional depictional representation and a three-dimensional model. The communication between the formalisms is mediated by a meta-interpreter. The depictional model allows the most direct presentation and handling of the knowledge because presentation and representation are identical. The depictional representation can be used as a knowledge acquisition tool.

The recent approaches to modelling neurological knowledge use normative models, rather than fault models which are the basis of earlier systems. Aspects of validation and actual use of these systems have not yet been described in the literature.

The implementations mentioned above show that knowledge representation methodologies have progressed from representation of a single kind of anatomical knowledge to the explicit representation of different kinds of anatomical knowledge. Furthermore, the traditional text-based interaction with the computer is being largely replaced by graphical interaction. There is an ever increasing emphasis on visual information, which has to be facilitated in the knowledge based system.

3.4.2. KNOWLEDGE BASED SYSTEMS FOR BRACHIAL PLEXUS INJURIES

There are various knowledge based systems which aim at (or also aim at) assisting in the domain of brachial plexus injuries. Various approaches have been described in the literature. The four different methods which can be identified are:

- the statistical approach (Burge and Todd, 1989),
- linked objects (First *et al.*, 1982),
- semantic network (Hertzberg *et al.*, 1987),
- production rules (Fisher, 1990).

Each of these will be discussed below, followed by an analysis of the differences and similarities of the approaches. The PLEXUS system (Jaspers, 1990) which has been developed at Delft University of Technology will be discussed in detail in Section 3.5.

The statistical approach. A statistical system relies on a very large number of cases to be available. However, the occurrence of brachial plexus injuries is relatively rare, and the number of different possible injury combinations is very large. Therefore it is not possible to develop a purely statistical model. For this reason, Burge and Todd (1989) have adopted a statistical approach in which the need for a large number of test cases is avoided by using a model based on the anatomy.

The system is meant to assist specialists in localizing peripheral nerve injuries. In order to construct a model of the nerve pathways and the muscles they supply, Burge and Todd (1989) needed to know by which pathways muscles and various areas of the skin are supplied, to determine the proportion of innervation which is received via each pathway and to determine the proportion of the torque which individual muscles contribute to each joint movement. The a priori probability of each combination of lesions occurring is also needed. The probabilities of a lesion occurring were determined subjectively. It was assumed that lesions occur independently. From this model, the conditional probability of a particular lesion given certain manifestations can be calculated.

Linked objects. The knowledge based approaches which follow, have all separated the knowledge representation from the inference algorithms which are used to localize a specific case. LOCALIZE (First *et al.*, 1982) is a system which uses a network of objects and links to represent the anatomical knowledge. The system is meant to assist physicians with localization of lesions in the peripheral nervous system.

The knowledge is represented as a network. Nerve segments make up the nodes of the network. A nerve segment is a portion of a nerve between two points where it branches. Each nerve segment has certain information which is related with it, for instance, spinal segment origins where the fibres enter and exit from the spinal cord, and a list of muscles which would be expected to be affected after complete transection of the nerve segment. The links represent the anatomic connections between the nerve segments. The systems contains a large number of nerve segments (2244) and links between the segments (9796).

The inference which is performed is based on the notion that clinically the most likely lesion is that with the fewest number of injured locations. Many different combinations of injured locations can explain a certain set of manifestations. For example, a number of distal (situated away from the centre of the body) lesions could give the same motoric deficit as one more proximal (situated towards the centre of the body) lesion. This idea has been implemented using a rule (sometimes called Occam's razor) which favours a single all-encompassing solution when possible. A convergence algorithm is used for this. The program traces through the network by using the links between the segments.

The algorithm starts distally, at the individual nerves supplying the muscles. When distal nerve segments join together to form a larger more proximal nerve segment and all the distal segments have been found to be affected, then the more proximal nerve segment will replace the distal segments as the injured location. The pattern of strict convergence does not hold for the point where the nerves form the brachial plexus. A special plexus algorithm is instantiated at that point. As nerve fibres proximally leave the plexus, they join to form spinal nerve roots, so the convergence algorithm applies for root lesions.

Semantic network. Hertzberg *et al.* (1987) describe a system which uses a semantic network approach for knowledge representation. The system is a prototype for testing the representational method, and is meant to assist in neurological diagnosis. The brainstem and brachial plexus were chosen as representative parts of the central and peripheral nervous system, and the prototype was developed for these two areas. The brachial plexus part of the system will be discussed below.

There are two trees of nodes, one containing the neuroanatomical knowledge and one containing the physical signs. Causal links connect the physical signs to the anatomy tree. The anatomy tree consists of a hierarchical structure of nodes containing the nerves and the muscles, with hierarchical links representing the anatomic connections between the nerves and muscles.

The algorithm used to find the injured locations is based on two different rules, the parsimony rule and the specificity rule. The parsimony rule is

interpreted as follows in the program. If a given set of signs is causally linked to a group of location nodes and the group of location nodes shared a common parent, then the parent node is the most likely location of the injury. The specificity rule is used as follows in the program. The relative value of each sign in localizing a lesion (specificity) depends on the anatomical extent of the structure causing the sign. The value of specificity for each sign is determined by dividing 100 points by the number of anatomic nodes each sign is linked to. This permits a more specific sign to make a greater contribution to a location. The algorithm finds the parent node whose children have the highest average numerical value. A threshold for reporting the results can be selected by the user.

Production rules combined with algorithms. A system which can determine the site of a lesion in a brachial, lumbar or sacral plexus injury is described by Fisher (1990). The system is called PLEXXUS (with double x). Its aim is to provide assistance in cases of complex injuries when experts are unavailable. The anatomical knowledge has been represented in the form of production rules and two additional algorithms are used.

The brachial plexus is divided into 89 nerve segments. When muscles are found to be weak on examination, the production rules infer the nerve segments which may be involved.

As was stated above, two algorithms are used in the system. The first algorithm, called the sharing algorithm, determines shared nerve segments among weak muscles. The second algorithm, called the proximal working algorithm, narrows the list of potential lesion sites, if possible, to one specific location. The algorithm confirms the hypotheses and tries to establish a more proximal site by querying the user for additional muscle weaknesses.

With multiple lesions, there is no perfect match in which at least one segment is shared by each weak muscle. If there is no single common pathway to explain the patient's findings, an arbitrary cut off of which nerve segments should be investigated and which should not, was set at 50%. The system continues to select those segments that are shared by at least 50% of the weak muscles.

3.4.2.1. Discussion

A brief overview of the systems which were discussed above is shown in Table 3.4. All these programs have in some way modelled the anatomy of the brachial plexus, and by reasoning about the structure and function of the nerves and

Table 3.4. Knowledge based systems containing brachial plexus knowledge.

reference	Burge and Todd (1989)	First *et al.* (1982)
program name		LOCALIZE
goal	to assist specialist in localizing peripheral nerve lesions	to assist physicians in localization of lesions in the peripheral nervous system
input	muscle power, sensibility, joint movement, Horner	clinical and electromyographic evidence of muscle weakness
knowledge representation	statistical approach; but using model based on anatomy, therefore relying on estimation of relatively few statistical parameters	- nerve segments (2244) with attributes, such as muscles innervated by segment - links between segments (9796)
inference	probability calculation	convergence algorithm (Occam's razor) to find injury with fewest number of loci plexus algorithm
output	probability of block or partial block in nerves	- certain data inconsistencies; deviations from expected values - wounded nerve segments
validation	-compared to 3 orthopaedic surgeons on 26 cases -number of errors compared using Wilcoxon signed rank test, shows that program performed significantly better at 5% level	sample patient cases of varying complexity

Hertzberg *et al.* (1987)	Fisher (1990)
	PLEXXUS
- to assist in neurological diagnosis - prototype system to test one method of representation	- prototype system for assisting neurosurgeons in determining lesion sites of brachial, lumbar, or sacral plexus injury
for brachial plexus part of system: motor and sensory manifestations	muscle weakness (present or absent)
for brachial plexus part of system: - semantic net of nodes (150) and links (70) - two trees of nodes; one for anatomy of nerves and muscles (anatomical links within hierarchy) other for manifestations - causal links between trees	- production rules for each muscle a list of nerve segments which could be involved when muscle is weak 89 brachial plexus segments
- parsimony rule: if signs causally linked to group of nodes and nodes shared a common parent, then parent most likely location - specificity rule: the relative value of physical sign depends on anatomical extent of structure causing sign; value of sign inversely dependent on number of nodes to which it is causally linked, numeric value given to signs. - parent nodes with subordinate nodes with highest average value found	- sharing algorithm: determines shared segments among muscles - proximal working algorithm: narrows list of potential lesion sites, if possible, to one specific location - with multiple lesions there is not a perfect match; program continues to select those segments shared by at least 50% of the weak muscles
localized lesions	location of lesion, also in graphic representation
hypothetical cases abstracted from literature	none mentioned

muscles, the site of the injury is determined. All the programs relate the function of the muscles to locations in the brachial plexus which may be injured. The way in which the structure and function are represented in the computer is different for all four programs and the way in which actual inference takes place is also different, although some basic principles can be found in all four. Most neurological localization systems incorporate the principle of parsimony, which implies that the most likely lesion is that with the fewest number of injured locations. This principle may be termed convergence algorithm, parsimony rule or proximal working algorithm. In the statistical system, the parsimony idea is not explicitly modelled. However, certain locations will have a higher prior probability than other locations, therefore the question of redundancy in the network is dealt with probabilistically. The paper by Burge and Todd (1989) is the only one to report a validation study of the program. The others merely state that the program did well on test cases.

3.4.3. RELATED PROGRAMS

There are various packages which are related to brachial plexus injuries and which are commercially available. Three of these will be mentioned below.

The first software package is called the Lesion Game™ (Guiteras, 1989). This is a learning tool that is designed as an adjunct to physical therapy curricula. The program shows a graphical representation of the brachial plexus and is completely mouse driven. It allows the user to view and study muscle innervations, which the program can automatically draw into the graphical representation.

In addition to the possibility of studying muscle innervations, the program can randomly select a lesion, from 44 different possible single site lesions, which the user has to attempt to find in as few guesses (manual muscle tests) as possible. As muscles are selected (muscle tested) using the mouse, the computer searches a table to find the appropriate strength. The program has a table consisting of the 44 single site brachial plexus injuries, 50 muscles and 2 sensations. When the user thinks there is enough information to determine the location of the lesion, the user clicks the appropriate location in the graphical representation, and the computer program indicates whether this is indeed the correct location.

The EVAL™ examination system from Greenleaf Medical Systems, Palo Alto, CA, can be used to evaluate impairment of the hand and upper extremity. It is a

computer based system that links measurement tools with software in order to be able to conduct tests, collect data, generate reports and analyse results.

The tests include, for instance, strength, range of motion, and sensation. The instruments which are coupled to the computer are a dynamometer for grip strength, an electronic pinchmeter, an electronic hand goniometer and an electronic upper extremity goniometer. Step-by-step prompts guide the user through selected tests or a complete examination.

Another software package is a specialised module for the Medical Electronic Desktop™. This module may be used for obstetrical brachial plexus injuries. The system keeps records on each patient in an electronic 'paper like' format with especially designed input and output forms. These forms include detailed evaluative checklists, treatment protocols, correspondence and reports.

3.5. Knowledge based system PLEXUS

The object of the knowledge based system PLEXUS (Jaspers *et al.*, 1989; Jaspers, 1990) is to assist neurologists, neurosurgeons, orthopaedic surgeons, rehabilitation physicians and traumatologists in the diagnosis and treatment planning of brachial plexus injuries. The system has been developed in cooperation with two Dutch brachial plexus experts, prof. dr. R.T.W.M. Thomeer of the Academic Hospital in Leiden and dr. A.C.J. Slooff of the 'De Wever Hospital' in Heerlen. The system is meant for physicians who are not specialised in the domain of brachial plexus injuries.

In order to request advice from PLEXUS, the physician enters patient data into the computer, the computer will then reason with the patient specific data, and will use the general knowledge concerning brachial plexus injuries which is stored in the system, to generate patient specific advice regarding:
- the locations which are injured,
- the severity of the injured locations,
- the preferred treatment.

The advice is shown to the physician on the computer screen. The most important aspects of the system will be discussed below. A detailed description of the architecture of the knowledge based system PLEXUS may be found in Jaspers (1990).

In contrast to the systems which were mentioned above, PLEXUS uses patient history information and results of radiological examinations, in addition to the

usual neurological and neurophysiological data which are incorporated in other systems as well.

Furthermore, PLEXUS not only localizes the lesion but also gives an indication of the severity of the injury, recommends additional diagnostic tests to be conducted, and suggests a treatment plan.

A hybrid representation has been used for the knowledge based system PLEXUS. Part of the knowledge has been represented in the form of production rules. The knowledge based system shell Delfi2+ (de Swaan Arons, 1991) has been used for this purpose. This knowledge based system shell facilitates forward and backward chaining of the rules, and also allows external programs to be activated. Various external programs which have been written in conventional programming languages (C, Pascal), are activated at certain points in the consultation.

The reasoning mechanism, knowledge bases and external programs have been implemented on a SUN® workstation. The knowledge representation, in the form of production rules and external programs, will be discussed below.

In order to improve the possibilities for acceptance of the system (van Daalen, 1988), some work has been carried out in the area of explanation generation (van Daalen and Jaspers, 1989), however, it is felt that the explanations which can be provided using the methodology which was described by van Daalen and Jaspers (1989) is more suitable for somewhat smaller systems. Therefore, the system which will be described, and referred to in the following chapters, is the implementation without possibilities for extensive explanation of the advice.

During the course of the project an additional implementation has been developed. This is a prototype system in which the brachial plexus knowledge has been represented in the form of objects and relations (ter Haar, 1989; van Heerebeek, 1991; Jaspers, 1990), using the knowledge based system shell Delfi3 (de Swaan Arons, 1991). However, this implementation has not been as extensively validated as the Delfi2+ version. Since the main issue concerns the validation of knowledge based systems, the Delfi2+ version will be described below.

3.5.1. THE KNOWLEDGE BASE

PLEXUS comprises two knowledge bases. The first knowledge base, PLEXAKT, contains the knowledge necessary for the localization of brachial plexus injuries and the second knowledge base, TREAT, contains knowledge concerning the severity of injuries and regarding treatment planning. The architecture of both knowledge bases will be discussed below.

3.5.1.1. The diagnostic knowledge base PLEXAKT

The aim of the diagnostic module is to determine the exact location of the structures within the brachial plexus which are injured. The solution strategy has been implemented according to the following general method. The data which are entered into the system by the physician are first abstracted into meaningful intermediate concepts. Following the data abstraction, the concepts are checked for possible inconsistencies. Significant inconsistencies will be reported to the physician.

Using the intermediate concepts, a rough localization of the injury is then performed using production rules. Based upon the rough localization, the exact injured locations are found by means of a hypothesize and test algorithm which hypothesizes possible injury combinations and tries to find the combination which best explains the motoric deficit in the arm.

As explained above, three different tasks may be distinguished:

- data abstraction,
- heuristic match,
- refinement.

The representation of these tasks will be explained in detail below.

Data abstraction. The data abstraction knowledge converts the data into intermediate concepts to be used for further reasoning, and then detects possible inconsistencies and incompleteness in the data which have been entered. The data abstraction knowledge has been represented in the form of production rules. An example of a data abstraction rule is shown in Figure 3.6.

```
DATA ABSTRACTION RULE

IF
[[ myelography.not_visible = "c5" ]
OR
 [ CT_scan.not_visible = "c5" ]
OR
 [ MRI.not_visible = "c5" ]]
AND
NOT [ c5.radiology_visible ]
THEN
CONCLUDE c5.radiology := "root_not_visible"   CF (1.000)
FI
ENDRULE
```

Figure 3.6. Data abstraction rule.

There is a number of production rules which test whether the data which is entered into the computer is consistent. There are, for instance, various tests which provide similar information. By checking whether the results of these examinations contain the same information, the consistency of the data can be checked.

When inconsistencies are detected, certain rules containing messages are present which indicate the inconsistencies which have been found. The system will show these messages on the computer screen and request the user to perform the tests again, but will also go on reasoning with the evidence it has, using the results of the tests which are most reliable. The production rule formalism is well suited to this kind of knowledge, since a certain action has to be taken when certain evidence is found. An example of a consistency checking rule can be seen in Figure 3.7.

```
CONSISTENCY CHECKING RULE

IF
[ Tinel.location = "supraclavicular" ]
AND
[ c5.Tinel_radiating ]
AND
NOT [ c5.lesion ]
AND
NOT [ c5.sensibility = "anaesthetic" ]
AND
NOT [ c5.sensibility = "hypaesthetic" ]
THEN
CONCLUDE patient.required_examination := "check_sign_of_Tinel"   CF (1.000)
CONCLUDE patient.required_examination := "sensibility"   CF (1.000)
EXECUTE plexusremark
WRITE ************************************************************
WRITE   The Tinel-Hoffman sign is radiating towards the c5-dermatome, indicating
WRITE   a lesion of spinal nerve c5. But there is no sign of motoric or sensory
WRITE   disability of this spinal nerve. Therefore I advise you to check again the
WRITE   sensibility of the c5-dermatome as well as towards which dermatome the
WRITE   Tinel-Hoffman sign is radiating.
WRITE ************************************************************
```

Figure 3.7. Consistency checking rule.

It is clear that inconsistencies can only be found when redundant information is present. The detection of inconsistencies allows the system to deal with uncertain information (measurement noise) to a certain extent, and it also makes the

system more robust for uncertain knowledge, such as individual variations, because it can detect whether this is present.

In addition, there are rules which detect that insufficient evidence is available to find the injured locations. Messages will be shown to the user when this is the case. When additional information would be required to perform a better localization, the system will provide a diagnosis based upon the information which is available, but will also indicate that it would be able to provide an improved localization if more information were present.

Heuristic match. The heuristic match task provides a rough localization of the injury. It contains empirical associations between the intermediate concepts and conclusions which may be drawn. Production rules have been used to represent the surface knowledge used for the heuristic match task.

Evidence from patient history, neurological, neurophysiological and radiological examinations are used to draw possible conclusions. An important drawback of rule based systems is that when a large collection of rules is used, the structure will usually not be transparent. To overcome this problem, all the evidence has been classified into five different categories. The categories of evidence are used in different kinds of production rules.

The categories of evidence are:
- Triggering facts: facts which immediately lead to a certain conclusion, regardless of any other facts which may be present.
- Necessary facts: facts which have to be present for a certain conclusion to be true.
- Exclusionary facts: facts which immediately lead to exclusion of a certain conclusion.
- Corresponding facts: facts which will lead to an increase in the certainty of a conclusion. However, the presence of such a fact alone will not lead to the conclusion being true.
- Irrelevant facts: some facts may not be relevant for a certain conclusion.

This classification of evidence makes the uncertainty in the suggestive strength of each piece of evidence for each hypothesis explicit, rather than quantifying it. The transparency of the system also improves by using the classification of evidence, since it shows the relation of each piece of evidence to each hypothesis (Jaspers, 1990). The various kinds of evidence are used in different kinds of production rules.

For transparency reasons the rules have been divided into different categories of rules. The strategy is to firstly try to confirm and disconfirm as

```
TRIGGERING RULE

IF
[ patient.extraforaminar_trauma ] {bruises in neck}
AND
[ patient.location_extraforaminar_trauma = "supraclavicular"]
THEN
CONCLUDE lesion.supraclavicular := TRUE   CF (1.000)
FI
ENDRULE
```

```
PRUNING RULE

IF
[Tinel.location = "supraclavicular"]
AND
[c5.Tinel_radiating]
THEN
CONCLUDE c5.exclude_avulsion := TRUE   CF (1.000)
FI
ENDRULE
```

```
EVALUATION RULE

IF
[[c5.radiology = "root_not_visible"]
OR
 [c5.proc_avulsion = TRUE]] {fracture of protruding part of spine}
AND
NOT [c5.exclude_avulsion]
THEN
CONCLUDE c5.avulsion := TRUE   CF (0.900)
FI
ENDRULE
```

```
CONFIRMATION RULE

IF
[ c5.avulsion = TRUE]
AND
[ c5.radiology = "meningocele"]  {pouch filled with fluid visible}
THEN
CONCLUDE c5.avulsion := TRUE   CF (0.500)
FI
ENDRULE
```

Figure 3.8. The heuristic match task.

many hypotheses as possible, and then using this knowledge to try to prove any additional hypotheses by weighing positive evidence against negative evidence.

The kinds of rules are the following:
- Triggering rules: rules which use triggering facts which can immediately confirm a hypothesis.
- Pruning rules: rules which use exclusionary facts which can immediately exclude a hypothesis.
- Evaluation rules: if no sufficient exclusionary evidence is available to rule out a hypothesis, and sufficient positive evidence is present, the hypothesis is postulated to be processed further.
- Confirmation rules: corresponding facts are used in order to become more certain about a hypothesis which has already been confirmed.

Examples of the different kinds of production rules are shown in Figure 3.8.

The heuristic match task provides a rough localization of the injury. After the heuristic match task, it may be certain for some locations whether they are or are not injured. For other locations, the empirical associations which are used, are not deep enough to be able to exactly determine whether they are injured. Therefore, the heuristic match task is used to quickly try to eliminate or confirm certain branches of the search tree, and a deep refinement task based on the structure and function of nerves is used to find the exact injured locations.

Refinement. The refinement task, for exact localization of the injury, is based on deep knowledge of the structure and function of the nerves of the brachial plexus. In the brachial plexus, 41 different possible injury locations have been distinguished. Usually, one brachial plexus lesion consists of more than one injured location. Therefore, in theory, there are 2^{41} different injury combinations.

It is not feasible to enumerate all possible combinations and to test which combination best explains the data. Thus, it is not possible to use a hypothesize and test approach for every combination which may occur. Therefore, during the heuristic match task, all evidence about certain hypotheses is gathered and as many hypotheses as possible are confirmed or excluded based on shallow knowledge and on knowledge which builds up the combinations from individual locations. In this way, the search tree consisting of all possible injury combinations is pruned.

Following this, the possible combinations which remain can be hypothesized and tested to see which combination best explains the motoric deficit in the patient's arm. This could be done by testing all possible combinations, which

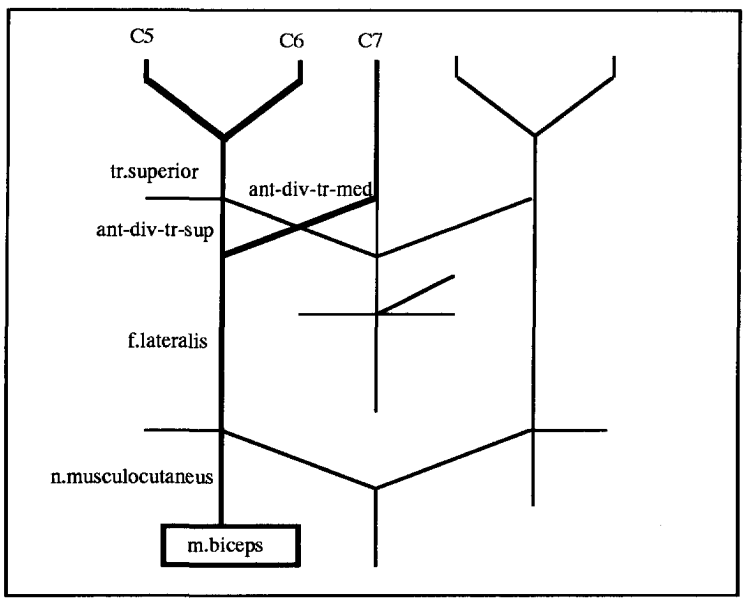


Figure 3.9. Graphical representation of the structure of the brachial plexus.

would still take a significant amount of time. However, in the literature (Rich, 1983) heuristic search algorithms have been described which search for the optimal solution without requiring every possible combination to be tested.

For PLEXUS the abstraction level of the deep representation is at the level of nerves, and not of the structures within the nerves, since these structures may show too much inter-personal variation to allow a robust representation which can be used for all traumatic brachial plexus injuries.

The structure and function of the nerves have been represented in the computer. The structure of the brachial plexus is the way in which the nerves are connected to each other and finally to the muscles of the shoulder, arm and hand. The function of the system is represented as the conduction of signals which are directed through the network towards the muscles.

If there is no injury, then the signals from the central nervous system can be conducted through the network of nerves and the muscles have full function. If there is an injury somewhere in the pathway between the spinal cord and a muscle, the signal will not be conducted completely and the muscle will only function partially. If all pathways to a muscle are blocked, then the muscle will not function at all.

Thus when a certain injury combination is hypothesized, from the structure of the pathways and the function of passing on the signal, a prediction can be made as to whether a muscle will fully function, partially function or not function at all. This can then be compared to the actual muscle strengths which were measured by the physician during the neurological examination.

Part of the structure of the nerves is shown graphically in Figure 3.9. The innervation of the biceps muscle has been highlighted. The structure of the highlighted part of the brachial plexus may be represented in the computer in the following way.

```
biceps = (((C5 + C6) * truncus-superior * anterior-division-truncus-superior + C7 *
anterior-division-truncus-medius) * fasciculus-lateralis * n-musculocutaneus)
```

The function of the pathways of nerves which lead to the m.biceps can also be represented. Assume that C5 and C6 have a more prominent part in the innervation of the m.biceps than the spinal nerve C7, and assume that an intact nerve has an innervation value of 1, and a defect nerve has an innervation value of 0, then the function can be represented in binary relations. An example of a binary relation for the m.biceps is the following.

```
innervation-biceps = (((2/5*C5 + 2/5*C6) * truncus-superior * anterior-division-
truncus-superior + 1/5*C7 * anterior-division-truncus-medius) * fasciculus-
lateralis * n-musculocutaneus)
```

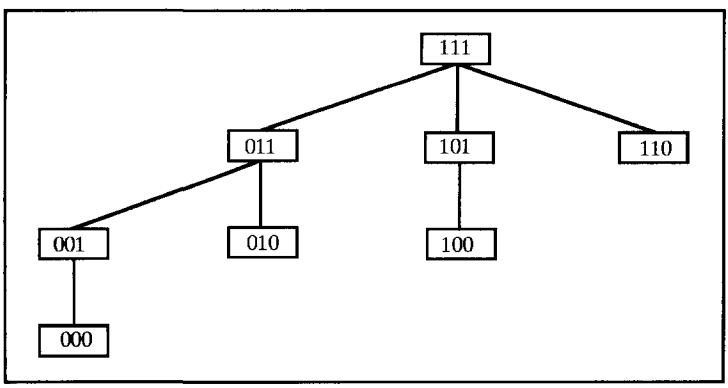


Figure 3.10. Possible injury combinations represented in a tree.

Depending on the parts of the plexus which show a defect, the innervation of the biceps can be calculated and can range from 0 to 1. All muscles can be represented in this way. When a certain combination of injured nerves is hypothesized, the innervation of the muscles can be predicted by these binary relations. The numerical value of the predicted innervation of each muscle is transformed into either intact, partially defect or defect. The predicted innervation of the muscles can then be compared to the real innervation which was found by the physician during the neurological examination.

However, as was stated above, it would still be very inefficient to hypothesize all possibilities, i.e. all combinations of injured locations, which remain after the heuristic matching task. Therefore, a heuristic tree-search algorithm is used to find the best possible combination of injured locations without hypothesizing every possible combination. The A* algorithm, adapted for trees instead of graphs (Rich, 1983) has been implemented in PLEXUS (de Lind van Wijngaarden and Furth, 1987; Meinders, 1989). This algorithm, called the Algorithm for Knowledgeable Trees (A^{kt}), will be explained below.

A^{kt} algorithm. The A^{kt} algorithm will be discussed using an example involving 3 possible injury locations and 3 muscles, and it can easily be extended to the 41 locations in the brachial plexus, which are represented in PLEXUS.

Suppose that there are 3 locations, and a location can either be injured (represented by a 0) or intact (represented by a 1), then there are eight different injury combinations. By representing all possible injury combinations in a tree, it becomes straightforward to see in which way the best solution can be found without testing all (8 solution possibilities in this example) solution combinations. All possible injury combinations are shown in Figure 3.10.

Assume that it is possible to test all combinations. This can be done in the following way. To begin, assume that none of the nerves are injured. This is shown at the top or first level of the tree in Figure 3.10. The muscle strengths in this case can be predicted by using the binary muscle function relations which were discussed above.

Now assume that there is one injured nerve. This gives the three combinations at the second level of the tree. All these combinations with one injured nerve may then be tested. Following this, assume that two nerves are injured. This gives the combinations at the third level of the tree. Finally, assume that there are three nerves which are injured, which gives the bottom or fourth level of the tree.

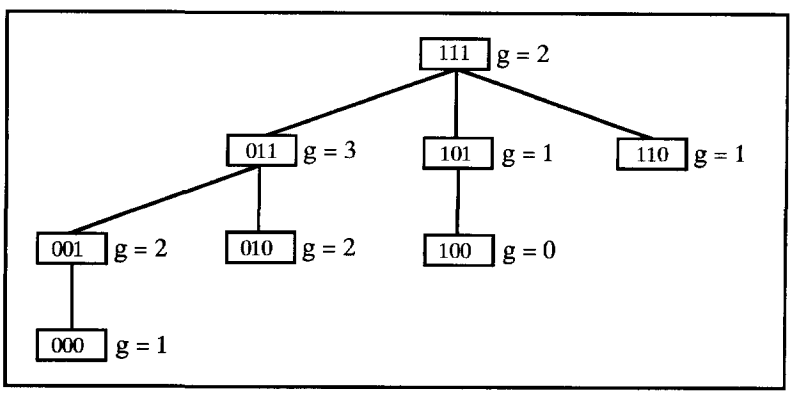


Figure 3.11. Values of g calculated for the complete tree.

Combinations in the tree are related to the combinations to which they are linked. For example, the combinations at a certain level of the tree are called the children of the combinations which are situated at the level above them, and the combinations at a higher level are called the parents of the combinations which are situated at the level below them. Starting at the top level of a tree in which all locations are assumed to be intact, the children can be generated by adding one more injured location at each next level. This procedure is called expanding the tree.

In order to determine which combination of injured nerves represents the best solution, some kind of closeness measure is necessary. For each combination of injured nerves, the difference between the actual (measured) muscle function and the predicted muscle function can be determined.

The muscle function is set at the numerical value of 1 for an intact muscle, 0.5 for a partially defect muscle and at 0 for a totally defect muscle. The actual muscle functions which are measured during the neurological examination can range from 0 which represents no function, to 5 which represents complete function. However, since it is difficult to objectively interpret values which relate to a partially functioning muscle (1 through 4), physicians may enter values ranging from 0 to 5 into the computer, but internally the computer converts this to the three classes which were mentioned above.

Let $m(i)_{act}$ be the actual muscle function which was determined during the neurological examination, and let $m(i)_{pred}$ be the muscle function of the ith muscle which is predicted by the binary muscle function relation. Then the total difference g between the predicted and actual muscle function is calculated over all of the 38 muscles, and is defined as follows:

$$g = \sum_{i=1}^{38} | m(i)_{act} - m(i)_{pred} |. \qquad [3.1]$$

In the above example, assume that there are only 3 muscles which correspond to the 3 locations, one muscle corresponds to exactly one location. Furthermore, let the first muscle be intact and the second and third muscles not be intact. The actual answer, i.e. combination of injured nerves, which should follow from this example is (100). This is the combination which should be found by the algorithm. In normal situations, with a larger number of nerves and muscles, it is obviously not possible to determine the answer straightaway.

For each of the combinations in the tree which is shown in Figure 3.10, the value of g can be calculated. The values of g are shown in Figure 3.11. It may be seen that the node containing (100) does indeed have the lowest value of g. Now it is

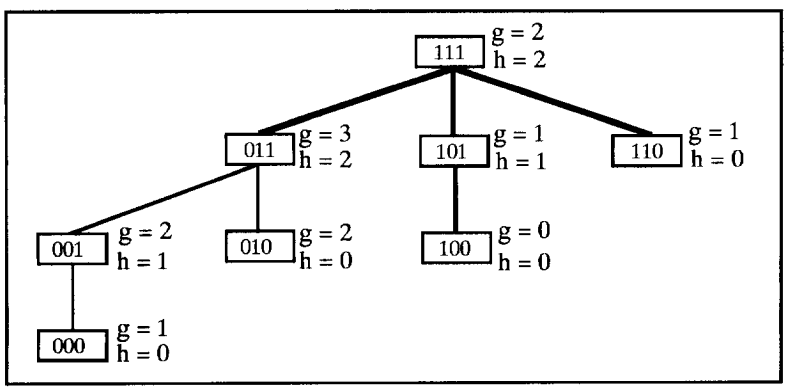


Figure 3.12. Heuristic function values in location tree.

possible to add another metric, called the heuristic function h, which will enable finding the optimal solution without traversing the complete tree.

The idea is that branches of this tree will be pruned and certain parent nodes will not have to be expanded, as it will be clear that there will be no need to progress further into certain branches.

The heuristic function of a node is a measure for the optimal solution which may be obtained by traversing downwards into a certain branch and expanding a node, if it is assumed that the locations which can still be varied in the children of this parent node are all correct (i.e. are equal to the muscle strengths which were measured). For instance, for the node (011) which can be seen in the tree, the second and third locations are varied in its children, and the first one always remains (0). Therefore, for the node (011), it can be seen that even if it is expanded, the first location will always be predicted incorrectly, but the second and third can still be changed. So optimally, it could correctly predict the second and the third node and the minimal g which can be obtained in that part of the decision tree is therefore equal to 1. In the tree (Figure 3.11) it can be verified that this is indeed the case.

Thus, the heuristic function is a measure for the improvement in g which can be brought about by expanding a certain tree and traversing that part of the tree. It is known that the optimal g which can be reached by expanding the node (011) is equal to 1. For the node (101), the g is already equal to 1, and it is known that by expanding the node, only the third location can be changed. Therefore only the final location can improve the present g, and the g of the node (101) can therefore maximally improve by 1 to become a minimum of 0 when the node is expanded. Since the minimum attainable g in the tree is 0, it is not necessary to expand the part of the tree containing the node (011) because there is another part of the tree which has a minimal g of 0, and that part is therefore more accurate.

The formal definition of the heuristic function h is shown below, where $m(i)_{opt}$ is the optimal estimation which could be obtained by expanding the tree down the present branch. If the location can still be changed, the value $m(i)_{opt}$ is equal to the muscle strength which was measured. The value $m(i)_{opt}$ is equal to predicted if the location cannot be changed.

$$h = \sum_{i=1}^{38} | m(i)_{opt} - m(i)_{pred} |. \qquad [3.2]$$

The value of the heuristic function for the nodes in the example can be seen in Figure 3.12.

The heuristic function overestimates the improvement in g or at best is a precise estimate, since for the locations which cannot be varied the values are calculated, and for those which can be varied it assumed that the values are equal to the actual values. An evaluation function f is now used to determine the minimum value of g which can be obtained by expanding the node and traversing down a certain branch. The evaluation function f is defined as follows:

$$f = g - h. \qquad [3.3]$$

At each level of the tree, only the node with the evaluation function which has the lowest value has to be expanded. Thus, in the search tree which is shown above, only 5 of the combinations (linked with the bold lines) have to be hypothesized and tested, and for the other 3 combinations this is not necessary.

For PLEXUS, the algorithm incorporating heuristic search consists of 41 different locations which may be injured. The locations which are known to be intact or defect after processing the production rules are set at their final value. This means that the search tree is pruned, since certain branches of the search tree do not have to be expanded, but the values of the locations are already known from the production rules. The locations which are unknown after processing the production rules, are set at intact when the algorithm is started, and the tree is expanded according to the values of the evaluation function.

There is one further principle which has been incorporated in the algorithm, this is called the principle of parsimony. This principle implies that the combination of the least number of injured locations which explains the symptoms, is most likely to be the correct answer.

The brachial plexus is a redundant network of nerves, this means that for a certain motoric deficit, there may be several explanations. For instance, an injury of two nerves more distally (lower down) in the plexus, may cause the same motoric deficit as one injury more proximal (higher up) in the plexus. This is, for instance, the case in a possible injury of both n.axillaris and n.radialis as opposed to an injury of the fasciculus posterior. This principle is used in most neurological localization programs (First *et al.*, 1982; Hertzberg *et al.*, 1987; Fisher, 1990).

In PLEXUS, the principle of parsimony is applied after the production rules have been processed, and is used only for the remaining essentially similar hypotheses (Jaspers 1990). The parsimony principle has been implemented by having the

more proximal locations on the left hand side of the list of possible injured nerves which is used by the algorithm. The algorithm traverses from left to right. It ends when it has found a combination of locations which explains the motoric deficit and will not go on to evaluate the other nodes.

In the above example (Figure 3.12), assume that the left hand item represents the fasciculus posterior, the second item represents the n.axillaris and the right hand item represents the n.radialis. Then the injury could be explained by the combination (011) and also by the combination (100). However, the combination (011) will be processed first, and will therefore be the answer.

Thus, the localizing strategy incorporated in PLEXUS consists of three phases:
- data abstraction takes place by consistency checking and transformation of the data into meaningful concepts,
- production rules are used for a rough localization of the injury and pruning of the search tree,
- the A^{kt} algorithm is applied for exact localization of the brachial plexus injury.

After localization of the injury, the treatment module is instantiated.

3.5.1.2. The knowledge base for providing treatment advice TREAT

In order to suggest a therapy, it is necessary to determine the severity of the injury, and to distinguish between injuries which show spontaneous recovery and those which will not recover spontaneously. When the severity of the injury has been investigated, the treatment plan can be determined.

The main aim is to differentiate those patients who should be treated conservatively only, from those who should be treated surgically, so that the patients who should be treated surgically may be referred to a specialist centre for surgery.

The treatment planning module is completely rule based. First, production rules are applied to assess the severity of the injury. These rules use information about the localization of the injury and additional data obtained from the patient history, radiological examination, and more recent physical and neurophysiological examinations, so it can be decided whether any improvement has taken place. For instance, an advancing Tinel's sign or reinnervation potentials on the EMG provide information about possible spontaneous recovery.

Three different groups of injuries may be distinguished:
- injuries which will recover spontaneously,
- injuries which will not recover spontaneously,
- injuries for which it is not yet known whether they will recover spontaneously.

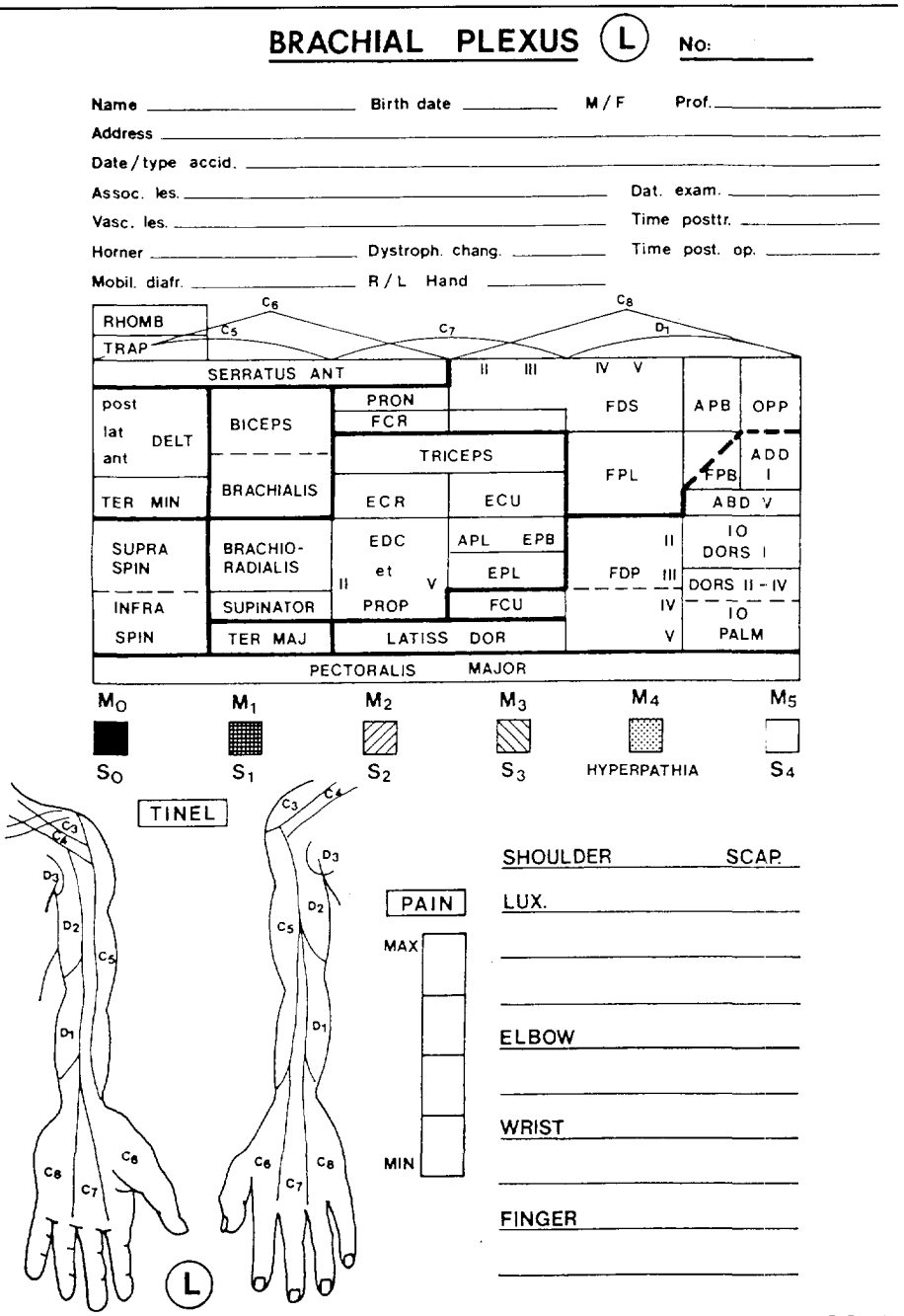


Figure 3.13. Brachial plexus data recording form (adapted from Merle d'Aubigné and Deburge, 1967).

After the severity has been determined, the treatment planning knowledge is applied. The knowledge about treatment planning of brachial plexus injuries is incomplete, and therefore expert heuristics are very important (Jaspers, 1990). Conservative treatment will be advised for injuries which will recover spontaneously. For nerve injuries which will not recover spontaneously, surgical treatment may be advised (although this will not always be the case).

The system distinguishes between three general surgical procedures. These are not specified in detail, for this is the task of the physician in the specialist centre who will perform the operation. When it is not yet possible to decide whether spontaneous recovery will take place, the system will advise further diagnostic testing and another consultation with the system to be performed after a certain period of time.

3.5.2. THE USER INTERFACE

The PLEXUS user interface is meant for the input of data and the output of advice. Users may enter all relevant data into the computer by means of the user interface. When data entry has been completed, a consultation with the knowledge based part of the system may be requested. The recommendations provided by PLEXUS are then shown on the computer screen.

The user interface has been designed according to the results of an investigation concerning the present practice of neurologists, and neurologists' requirements regarding computer advice and presentation of the advice (Grolman, 1989).

Presently, the user interface runs on an Apple Macintosh® computer, and has been implemented using the software package Hypercard™. Interaction with the system requires no previous typing and computing experience.

The user interface is based on a well-known scheme devised by Merle d'Aubigné and Deburge (1967). This scheme can be seen in Figure 3.13. The way in which this scheme has been represented on the computer screen can be seen in Figure 3.14. This scheme is the first page of the user interface, it shows a summary of the data which have been entered. The actual data entry is carried out on subsequent pages of the interface. An example of such a screen can be seen in Figure 3.15.

Most of the data entry is carried out using the mouse of the computer and clicking on the relevant answer possibility. The data entry pages have been divided into five different sections:

- patient history,

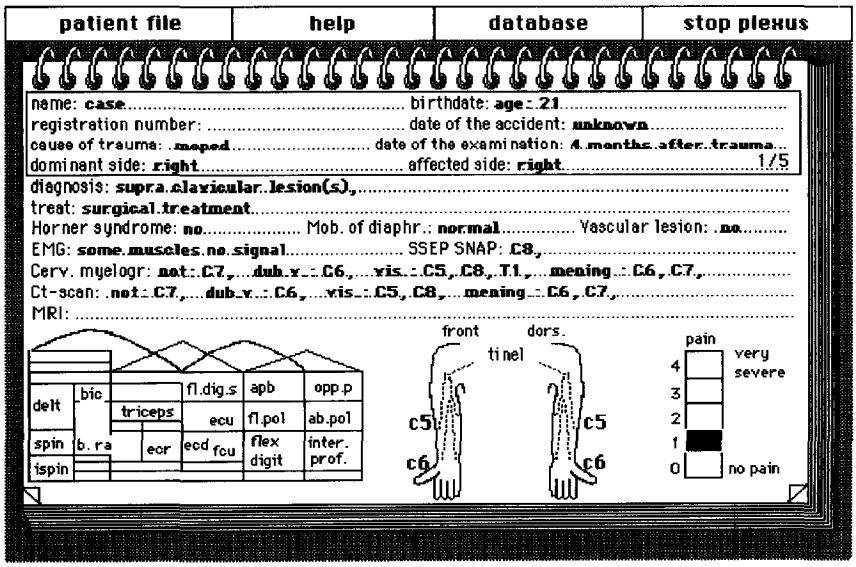


Figure 3.14. Summary information represented on the computer screen.

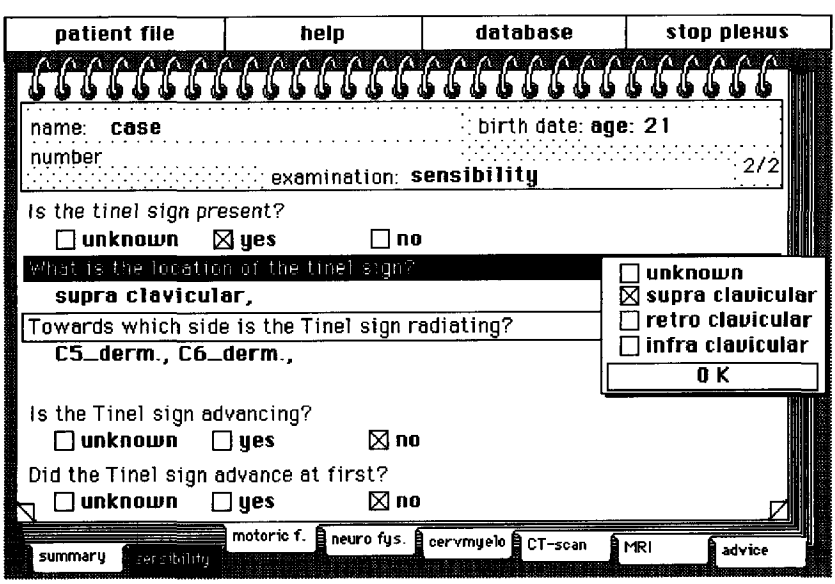


Figure 3.15. Data entry screen.

- neurological examinations,
- neurophysiological examinations,
- radiological examinations,
- advice.

It is possible to quickly skip to the next section by clicking on the relevant section name at the bottom of the screen. Each of these sections contains various screens on which data may be entered. One may proceed to the next page by clicking on the dog's ear in the bottom right hand corner of each screen.

When all relevant data have been entered into the computer, the physician can request advice from the knowledge based system. It is not necessary to answer all questions in order to perform a consultation. It is up to the physicians to gather the data which they think are relevant for a specific patient. A consultation can be requested by clicking on the consultation option in the advice section.

Upon a request for advice, the patient data which have been entered into the physician's Apple Macintosh computer are sent to the central SUN workstation at Delft University via a modem connection, since the user interface and knowledge bases run on different computers. The system will then reason with the data and the knowledge represented in the knowledge based system, and the advice is sent back to the physician's Apple Macintosh.

The diagnosis and the treatment plan are shown to the user in textual form. The injured locations are also shown in a graphical representation of the anatomy of the brachial plexus. An example of possible graphical output is shown in Figure 3.16.

For validation and development purposes, separating the interface from the knowledge based part of the system provides a number of advantages. Updating the knowledge can be done on the central workstation, and it is possible to keep track of the progress of evaluation studies centrally. For actual use, however, it would be advisable to implement the system on one computer, preferably the computer standard which is used in the hospitals.

3.6. Preliminary validation of PLEXUS

During the development of PLEXUS, the performance of the system was tested using about 100 test cases, consisting of retrospective patient data provided by the cooperating experts. The results of these cases were reviewed and the system was updated until it was felt that the system's diagnostic and treatment planning

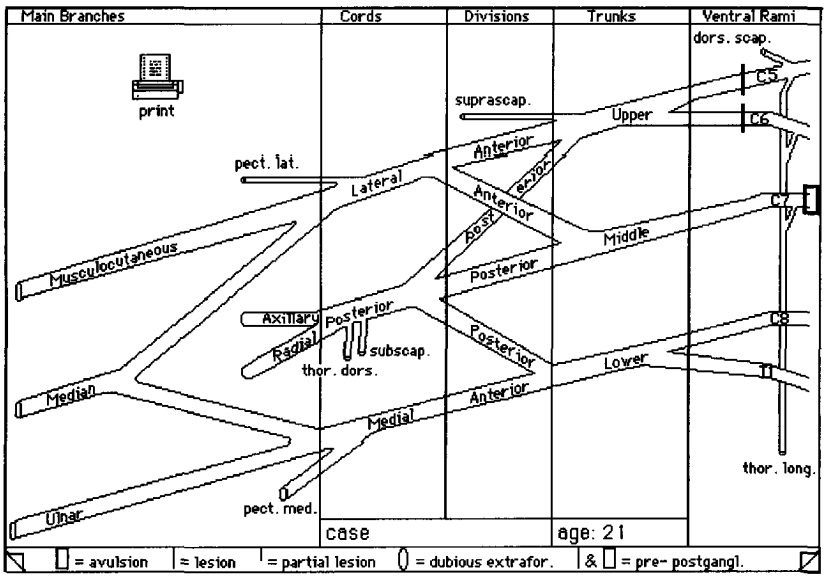


Figure 3.16. Injured locations shown in a schematic graphical representation of the brachial plexus.

performance on these training cases was at expert level. Upon which, preliminary evaluation studies of the systems output were performed.

3.6.1. PRELIMINARY PERFORMANCE EVALUATION

The preliminary performance evaluation studies which have been performed will be discussed along the lines which were indicated in Chapter 2, where a framework for evaluation was described. The framework can be seen in Figure 2.1. However, the present study was meant as an informal investigation of the problem solving performance of the system, and therefore it has a number of limitations which should be avoided when performing a formal evaluation.

3.6.1.1. Direct comparison of system output to cooperating expert's opinion

Goal. The aim of the system is to provide expert level advice to neurologists, neurosurgeons, rehabilitation physicians and traumatologists who are not familiar with the management of brachial plexus injuries. Therefore, the correctness of the recommendations provided by the knowledge based system has been evaluated. As a first test of system performance, system output was directly compared to one of the cooperating experts.

Selecting test input. The test cases consisted of 15 retrospective cases which originated from one of the cooperating experts.

Specifying who uses the system. The data were entered into the computer by the researchers.

Specifying a standard for performance. The diagnostic and treatment advice provided by PLEXUS was compared to the diagnosis and treatment which was determined by the expert who actually saw the patients.

Specifying physicians to test against. The expert who treated the patients was chosen as the standard, and no other physicians were involved in the evaluation.

Comparison. The diagnostic advice and the treatment advice was compared directly to the expert's opinion. The comparison was performed by the researchers. Since PLEXUS is a system which gives multiple answers which are non-exclusive, some kind of scoring scheme is necessary when comparing system output to the physician's opinion.

Table 3.5. Results of preliminary evaluation of the correctness of the advice (Jaspers, 1990).

quality of advice		poor	fair	good
PLEXUS LOCALIZATION				
supraclavicular	operative		1 (7%)	3 (20%)
	conservative			4 (27%)
infraclavicular	operative			2 (13%)
	conservative			2 (13%)
two-level	operative		2 (13%)	1 (7%)
total			3 (20%)	12 (80%)
PLEXUS TREATMENT				
supraclavicular	operative			4 (27%)
	conservative		2 (13%)	2 (13%)
infraclavicular	operative			2 (13%)
	conservative			2 (13%)
two-level	operative			3 (19%)
total			2 (13%)	13 (87%)

The following criteria were used:
- the system's advice is classified as 'good' if it corresponds to the standard,
- the system's advice is classified as 'fair' if it only slightly deviates from the standard,
- the system's advice is classified as 'poor' in all other cases.

Analysis of results. Percentages of poor, fair and good cases were calculated. The results of this evaluation are shown in Table 3.5. It may be seen that the system did not give any poor recommendations, and 80% of the diagnoses were judged to be good. Furthermore, 87% of the treatment plans were judged to be good.

Threats to the validity. There is a number of limitations to this study. These limitations are mainly due to the fact that it was an informal study which was meant to give some idea of the system's problem solving performance. In a formal evaluation these limitations should be avoided. Some of these limitations will be mentioned below:
- Circularity: The same expert who was involved in the development of the system was involved in the evaluation of the system.
- Subjective criteria: the criteria which were used to judge the results are subjective and depend on the person judging the results.
- Representativeness: The experts usually receive the more severely injured patients, therefore it is questionable whether the test cases are representative for the target population.
- Statistical conclusion validity: It is not possible to test hypotheses statistically with this number of test cases.

3.6.1.2. Blind evaluation of the system's problem solving capacity

A further preliminary performance evaluation entailed a double-blind evaluation involving both cooperating experts.

Goal. To gain a better insight into the level of expertise reached by the system in comparison to human experts.

Selecting test input. Both cooperating experts provided the data of the 10 latest patients whom they had operated on.

Specifying who uses the system. The test cases were entered by the experts themselves.

Table 3.6. Results of a preliminary blind evaluation of the advice (Jaspers, 1990).

		HUMAN EXPERT					
		--	-	0	+	++	total
PLEXUS	--						0
	-				1		1
	0		1	1			2
	+		1		1	4	6
	++			1	4	6	11
	total	0	2	2	6	10	20

crude agreement = 8/20 = 0.40
PLEXUS positive rate = 7/20 = 0.35
PLEXUS negative rate = 5/20 = 0.25

Specifying physicians to test against. The system was tested against the two cooperating experts, who each diagnosed the 10 cases which did not originate from them.

Specifying a standard of performance. The standard of performance is implicit in the judging experts. All test patients had been operated on by the judging experts. Therefore the judging expert had actually seen the exposed plexus of the patients. However, the part of the plexus which is exposed is determined by the pre-operative diagnosis. This means that the diagnosis may not be verified completely in all cases.

Comparison. The origin of the diagnoses was blinded, and the treating experts were asked to rate both the system's diagnosis and the expert's diagnosis on a five point scale.

Analysis of the results. The number of cases in which the system had the same score as the non-treating expert, the number of cases in which the system had a better score than the human expert and the number of cases in which the human expert scored higher were determined. The results can be seen in Table 3.6. It shows that in 7 out of 20 cases the knowledge based system's advice was judged to be better than that of the human expert, and in 5 out of 20 cases the expert's advice was judged to be better.

Threats to the validity. This study was a further preliminary evaluation. The limitations which were mentioned with regard to the previous informal evaluation still hold, although the judgement of the recommendations was left to the experts in the present study. However, some subjectivity may have been introduced due to the blinding of the diagnoses and treatment plans.

Evaluation of user interaction. During this evaluation, aspects of user interaction were addressed implicitly, since the physicians themselves consulted the knowledge based system. At that time, the user interface was a textual interface, and the interaction between the user and the knowledge based system consisted of the physician typing the answers to the questions which were posed by the knowledge based system. This user interface proved to be inadequate for physician use of the system. In order to solve this problem, the present graphical interface was developed (Grolman, 1989).

The graphical interface was informally evaluated by videotaping sessions during which the experts interacted with the knowledge based system. The user interface was updated on the basis of information which was obtained during the interaction sessions.

3.6.2. VERIFICATION AND DEVELOPMENT VALIDATION

In addition to the evaluation studies which were described above, a system should be verified and thoroughly validated. At the time when the knowledge based system PLEXUS was developed, most researchers were interested in developing working prototypes which simulated expert reasoning rather than building knowledge based systems which were meant to be put to into actual use. Most of the attention was devoted to representation languages and uncertainty calculation. Validation aspects only received attention once more serious applications were being developed.

3.6.2.1. Verification

In the past, verification was limited to checking the syntax of the program code. This also holds true for PLEXUS. Since then, more sophisticated verification tools have been developed, which can check rule based systems for completeness and consistency (Suwa *et al.*, 1982; Nguyen *et al.*, 1987). More recent systems can verify complete inference chains (Ginsberg, 1988; Preece and Shinghal, 1992). A number of different verification tools has been summarised by Voorhorst (1992). Verification tools seem to be a low cost method to check knowledge based systems, and therefore these should be applied to any serious application from the start of the development. However, the majority of verification tools which have been described in the literature are limited to rule based knowledge representation methods.

PLEXUS was developed using an expert system shell (de Swaan Arons, 1991). Therefore, it was assumed that the inference processes functioned correctly and did not need separate inspection. However, when this cannot be assumed, the complete system will have to be verified, including inference processes.

3.6.2.2. Development validation

In order to investigate whether the correct system has been developed, it is necessary to validate the knowledge based system. A knowledge based system should be tested on a range of test cases. However, knowledge based systems are often developed for domains for which large numbers of test cases are not available. This means that there is usually a lack of cases for testing the system. Furthermore, since solving the problem often requires expertise, it is usually not possible for the developer to directly draw up suitable test cases, and it is too time consuming for an expert when large numbers of test cases are needed.

In order to adequately validate a knowledge based system, some kind of method for test case generation will often be needed. One method of test case generation is to use an inverse knowledge based system, which upon being given a diagnosis generates the input data which correspond to the diagnosis.

This method has a number of disadvantages. It is necessary to involve independent developers in the development of the inverse system, otherwise the systems will be dependent and errors may not be found. The validation of the inverse system will also present a problem. When an inconsistency is found between the diagnosis entered into the case generator and the diagnosis provided by the knowledge based system, it may mean that an error is present in either of the systems. Therefore, empirical evaluation of the system will still be required.

Although this method has a number of limitations, validation on a large range of test cases, either using real or generated test cases, is necessary for all systems which are meant for actual use since it allows many more cases to be tested than in an empirical evaluation study. Shwe *et al.* (1990) describe a system which generates scripts of test data. This has been developed for testing the knowledge based system ONCOCIN.

A similar approach is being applied to the knowledge based system PLEXUS (Voorhorst, 1993). Upon entry of a diagnosis, a test case generator provides input data for the knowledge based system. The test case generator has two possibilities. Firstly, it can generate input data based upon the theory of the general anatomy of the brachial plexus and assuming that all examination results are according to what would be expected theoretically. Thus, using this possibility, the ideal patient is created. The knowledge based system PLEXUS should be able to correctly diagnose all ideal patients.
However, in practice:

- the anatomy of the brachial plexus may vary from person to person,
- the results of the examinations may not be what would be expected theoretically,
- the examination results may have been incorrectly interpreted by the physician or may not have been carried out at all.

Therefore, the test case generator also has the possibility of generating cases in which one or more of the above are present. These cases can be used to investigate the robustness of the system.

The test case generator uses diagnoses as its input. It reads these diagnoses from a file. This file consists of combinations of injured brachial plexus locations. Since in theory there are 2^{41} different injury combinations it is not possible to test all of these. Therefore, the diagnoses file is filled up by running a computer program which chooses different injury combinations based on certain criteria

which depend on the goal of the study. Possible choices could be to test extreme diagnoses or to test diagnoses which are clinically relevant. For instance, the program could determine all injury combinations existing of less than six injured locations, where the injury locations are situated close to each other. In this way choices can be made as to the diagnoses which are used for producing test cases.

3.7. Conclusions

The need for providing assistance to physicians in the domain of brachial plexus injuries was investigated from the patient management and from the physicians' point of view. A retrospective analysis of patient files showed that localization of brachial plexus injuries is extremely difficult and brachial plexus patients are often referred to the relevant specialist centre at a very late stage. These difficulties illustrate the need to assist physicians in the diagnosis and treatment planning of brachial plexus injuries. This need is recognised to a certain extent by neurologists themselves. About half of the respondents to a questionnaire indicated that they would use such a decision support system personally. Computer-assisted advice in the domain of brachial plexus injuries must fulfil certain requirements regarding quality of advice and regarding human-computer interaction in order to be acceptable for use in actual practice.

As a result, the knowledge based system PLEXUS was developed. In contrast to other programs for brachial plexus injuries, PLEXUS uses patient history information and radiological examinations results in addition to neurological and neurophysiological results, which are also used by other programs. Furthermore, besides assisting in localization, PLEXUS also provides treatment planning recommendations. PLEXUS is a hybrid knowledge based system consisting of a diagnostic and treatment planning module.

The diagnostic module uses production rules for rough localization of the injury and a heuristic tree search algorithm for exact localization. The treatment planning module is completely rule based. The system presently has a user friendly graphical user interface which requires no previous computing experience in order to enter data and receive recommendations from the system.

Knowledge based systems are often very large computer programs incorporating uncertain knowledge, which require thorough verification and validation. Verification and validation should be incorporated into the development lifecycle and require attention continually. When the knowledge based system PLEXUS was developed, sophisticated verification tools had not yet been developed. However, verification tools seem to be a low cost method to check knowledge based systems, and therefore these should be applied to any serious application

from the start of the development. However, the majority of verification tools which have been described in the literature are limited to rule based knowledge representation.

Due to a lack of relevant test cases, the validation of PLEXUS on a large range of test cases is being carried out by developing an inverse knowledge based system which provides input data upon entry of a diagnosis.

Methods for 'dynamic' validation of knowledge based systems, i.e. running the system with test cases, are an open research subject in knowledge based systems. This topic requires much more attention, since this kind of validation provides very worthwhile information. Validation on a large range of test input, which is either real or generated, is necessary for all systems which are meant for actual use since it usually allows many more cases to be tested than in an empirical evaluation study. However, as this method also has a number of limitations, empirical evaluation of the system is also necessary.

Various informal evaluation studies of the problem solving performance of the knowledge based system PLEXUS were performed. Although the number of cases involved was limited, the results of the evaluation studies were encouraging. Based upon the information obtained from the results of the informal evaluation studies which were carried out for PLEXUS, some of the knowledge in the system was updated, after which the system was frozen. The updated version of the system has been evaluated in a formal performance evaluation study involving four international brachial plexus experts, and the system has undergone a field evaluation in four different hospitals in the Netherlands. These evaluation studies will be discussed in Chapter 4 and Chapter 5.

4 Laboratory evaluation of the diagnostic and treatment planning performance of the medical knowledge based system PLEXUS

The problem solving performance of the knowledge based system PLEXUS has been evaluated in cooperation with four experts from different European countries. The evaluation setup allowed both direct and blind comparison of the system's recommendations to the diagnoses and therapies suggested by the four experts. Various methods of analysis were used to determine the level of performance which is achieved by the system. The results show that the accuracy of the recommendations provided by PLEXUS is comparable to those obtained from the experts. However, PLEXUS obtained a higher fraction of false positive answers. For a number of cases this is caused by the fact that PLEXUS tries to explain more of the dysfunction than the experts do. The intra- and inter-expert variability proved to be rather high in this study. These results are supported by the blind evaluation. During the blind evaluation, the experts were also asked to indicate which of the recommendations they thought originated from PLEXUS. The number of times the experts indicated that answers originated from the knowledge based system did not significantly deviate from the number of times this was expected to occur by chance. The relatively limited representativeness of the test cases and the fact that only domain experts cooperated in the evaluation are limitations of the present study.

4.1. Introduction

The diagnostic and treatment planning performance of the medical knowledge based system PLEXUS has been evaluated. The aim of the study was to investigate whether the system's problem solving performance is comparable to that of a number of international experts in the domain of brachial plexus injuries. The evaluation setup was a variant of the so-called Turing test, which can be used for evaluating medical knowledge based systems (Quaglini *et al.*, 1988; Yu *et al.*, 1979). The evaluation of PLEXUS consisted of three rounds:

- data collection,
- determining diagnoses and therapies, and direct comparison of opinions,
- blind subjective judgement of the opinions.

In the first round, four internationally known experts from different European countries, were asked to provide retrospective data of ten consecutive patients with a brachial plexus injury. This resulted in a total of forty test cases which were available for use in the evaluation.

Table 4.1. Summary of the laboratory evaluation setup. The evaluation consisted of three rounds. The output obtained after the second and third round was analysed to investigate system performance.

round	input	system	output	analysis
round 1		treating experts	patient data diagnoses treatment plans	
round 2	patient data	PLEXUS	diagnoses treatment plans	direct comparison: PLEXUS & treating experts
	patient data	non-treating experts	diagnoses treatment plans	direct comparison: non-treating experts & treating experts
round 3	diagnoses treatment plans	non-treating experts	judgements	study judgements

In the second round, the patient data which were obtained in the previous round, were entered into the computer and a consultation with the knowledgebased system PLEXUS was carried out. The diagnoses and treatment plans obtained from the knowledge based system were directly compared to those of the experts who sent in the data and actually treated the patients. Various methods of calculation were used to determine the performance of the system relative to the treating experts.

In addition, each expert was sent fifteen of the thirty cases which did not originate from that particular expert. The experts were asked to provide a diagnosis and treatment plan for these fifteen patients. The case notes only contained the relevant information needed to diagnose the patients and to determine a treatment plan, the original diagnoses and therapies were removed from the notes. At the end of this round, the diagnoses and treatment plans submitted by these non-treating experts were directly compared to those of the treating experts. The non-treating experts' results were also compared to the results obtained by the knowledge based system.

In the third round, each expert was sent the fifteen cases which did not originate from him, and which he did not diagnose in the second round, i.e. cases he had not seen before. This time, the diagnoses and treatment plans provided by the treating expert, the knowledge based system, and the non-treating experts who diagnosed the cases in the second round were attached to the case notes. The judging experts were asked to judge all the diagnoses and treatment plans on a five point scale. The evaluation was carried out blindly, therefore, care had been taken to make it impossible to distinguish the origin of the diagnoses and treatment plans.

A summary of the laboratory evaluation setup is shown in Table 4.1. Various methods of analysis were used to determine the performance of the knowledge based system compared to the treating experts and to the non-treating experts. In addition, the intra- and inter-expert variability were investigated in order to determine the level of agreement which exists in this domain.

The evaluation of PLEXUS will be described using the framework for evaluation design which was introduced in Chapter 2, and which is shown in Figure 2.1. The framework consists of a number of steps which has to be defined prior to performing an evaluation study. The choices that have been made for the evaluation of the knowledge based system PLEXUS will be discussed in Section 4.2. The results of the evaluation are described in Section 4.3.

The aspects which have been investigated include the following:

- direct comparison of diagnoses and treatment plans provided by the treating experts, non-treating experts and PLEXUS,

- intra- and inter-expert agreement in diagnosing and treatment planning,
- blind expert judgement of diagnoses and treatment plans provided by the treating experts, non-treating experts and PLEXUS,
- blind judgement of own diagnoses and treatment plans,
- analysis of whether it is possible for experts to distinguish the system's recommendations from those provided by humans,
- analysis of the differences between the results obtained in the direct comparison and in the blind judgement.

There are various sources of bias and confounding which may have influenced the results of the evaluation study. These possible threats to the validity of the study are discussed in Section 4.4. The most important findings which resulted from this evaluation study are summarised in Section 4.5. Finally, the lessons learned from the laboratory evaluation have resulted in a number of general recommendations for performance evaluation studies of medical knowledge based systems. The recommendations are also described in Section 4.5.

4.2. Design of the laboratory evaluation of PLEXUS

4.2.1. GOAL OF THE EVALUATION STUDY

The final aim of PLEXUS is to use the system in hospitals to assist physicians in the diagnosis and treatment planning of brachial plexus injuries. The system provides advice to the physicians, and this should be expert level advice. The goal of this evaluation study was to investigate whether the diagnostic and treatment planning performance of the knowledge based system PLEXUS is comparable to the performance of experts in the domain of brachial plexus injuries. Furthermore, it is usually impossible to study a whole range of potential users of the system in an evaluation. This was another reason to evaluate whether the level of performance of the knowledge based system is equivalent to the performance of a number of internationally recognised experts in a laboratory evaluation.

4.2.2. EVALUATION SETUP

4.2.2.1. Selection of test input

Representativeness. The test input in this evaluation study consisted of retrospective data of real brachial plexus patients. Four experts in the domain of brachial plexus injuries from four different European countries were asked to provide ten actual cases. This resulted in a total of forty test cases which were available for this study.

The objective was to obtain data which are representative for the actual situation in which the knowledge based system is to be used. Thus, the test patients should resemble the patients who are encountered by potential users of the system during daily practice. The potential users are neurologists and neurosurgeons who occasionally see patients with a brachial plexus injury.

Instead of asking potential users to provide the data, the four experts were asked to submit the data, as it would be easiest to obtain a relatively large number of test cases in this way. Furthermore, since brachial plexus injuries are their special interest, the patients would be well documented. However, the experts usually see the more severe cases, which means that the cases would not be representative for the target situation. Therefore, each expert was asked to provide the data of five patients who were treated surgically and five patients who were treated conservatively. It was assumed that the patients who were treated conservatively would be less severely injured than those who were treated surgically.

The treating experts (i.e. experts who actually treated the patients) were asked to submit the data of the first five new patients who visited them after the first of January 1987, and who were treated surgically. They were also asked to supply the data of the first five new patients who visited them after the first of January 1987, and who were treated conservatively.

Since the experts do not see many patients who are treated conservatively, it was necessary to go as far back in time as possible in order to obtain sufficient patients with milder injuries. However, prior to 1987 the treatment and diagnostic methods were not as advanced as they are at present. This meant that the data of patients who were treated before 1987 could not be used in the evaluation. Therefore, the earliest date for the first patient was set at the beginning of 1987.

If an expert did not have five patients who had been treated conservatively, he was asked to submit the patient data of additional surgically treated patients, in order to reach the total number of ten patients to be supplied by each expert.

The test input was restricted to traumatic brachial plexus injuries, since the diagnosis and treatment of other types of brachial plexus injuries, such as obstetrical or irradiation injuries, can be quite different. This, however, is not the same as limiting the input due to a certain diagnosis not being represented in the knowledge base, which has been done in various studies (see, for example, Miller *et al.*, 1982). In the latter situation, the actual diagnosis will have to be known before it can be determined whether the case is applicable. However, when the system is used in actual practice the users would apply the system because they want to be assisted in determining the actual diagnosis. Whereas,

physicians using PLEXUS in the target situation would know whether a patient had suffered a traumatic brachial plexus lesion.

The experts were asked to submit the data on special data entry forms containing various categories, allowing the expert to write down all the information which can be entered into the knowledge based system. Any other relevant comments could also be written down on the forms.

A number of problems arose with respect to the test input. Firstly, only one of the experts used the special data entry forms. Since the other experts took such a long time returning the forms, it was decided that they could also send in copies of their case notes, which were then transcribed onto the patient data entry forms by the researcher. This inevitably introduced subjectivity into the measurement, since the patient was not seen by the knowledge engineer, and some case notes were much more complete than others. For many of the cases, the information was either insufficient or not clear, so relevant additional information was requested from the experts.

Only two experts supplied data of conservatively treated patients. This meant that only seven of the cases which were submitted had been treated conservatively, instead of the twenty cases which were requested. The idea that those who were treated conservatively would also be patients with a milder injury did not prove to be true in all cases. For two of the patients, surgery had been indicated by the expert, but in one case the insurance company refused to pay, and in the other the patient refused the operation. Two further cases did not have a favourable prognosis, but in one case an operation was contraindicated due to the patient's cardiac state, and in the other case the injury had been sustained twenty years prior to the first visit and nothing further could be done. The way these four cases were dealt with will be explained in Section 4.2.2.2.

The fact that fewer cases with mild injuries were submitted than had been expected means that the requirement of representativeness of the patient data has not been satisfied.

This may also be deduced by looking at regularities which Narakas (1985) found in his series of patients. As Narakas is the expert in the domain, the patients treated by him will probably be more severely injured than the target population of patients. The regularities found by Narakas (1985) are presented in his law of the seven seventies:

- 70% of traumatic brachial plexus injuries are due to traffic accidents,
- 70% of the lesions in traffic accidents involve the use of a cycle of motorcycle,
- 70% of these patients have associated multiple injuries,
- 70% have a supraclavicular lesion,

- 70% of patients with supraclavicular lesions will have one or several roots of the plexus avulsed from the spinal cord,
- 70% of the patients with root avulsions will have the lower roots C7, C8, T1 or C8, T1 avulsed,
- 70% of the patients with lower root avulsions will experience persisting pain.

In the present study, 26 out of all 40 cases (65%) had at least one root avulsion. Furthermore, 26 of the 37 patients with a known cause of injury (70%) sustained the injury during a motorcycle accident. Whereas infraclavicular injury only was found in 7 of the 40 patients (18%), and these are usually the more mild injuries. Thus, the patients in the present study do not appear to have milder injuries than the patients seen by Narakas (1985).

A further problem with the test data concerns the fact that for a number of test cases the data which were sent in by the treating experts, were not as complete as would have been necessary for an optimal diagnosis, although all necessary information was probably available to the treating experts when they saw the patients. Therefore, in these cases the system probably had to perform the consultation with data of a lesser quality. The experts who did not treat the patients were also asked to provide a diagnosis and treatment plan for the patients, based upon the information which was available on paper.

Number of test cases. The number of test cases to be submitted was set at ten patients for each expert who would cooperate in the evaluation of the knowledge based system. This number was chosen for a very practical reason. It was thought that the experts would not be willing to take part if they were asked to provide more data. This relatively low number of test cases will limit the statistical conclusions which can be drawn from this study. The limitations caused by the sample size will be analysed below.

The aim of the evaluation is to investigate whether there is a difference between two proportions, for example, the proportion of correct answers given by the system and the proportion correct answers given by a physician. There is a null hypothesis (H_0) of no difference, and the aim is to see whether H_0 can be rejected in favour of the alternative hypothesis that there is a difference between the two proportions (H_1). The errors which can be made are to falsely reject the null-hypothesis, and a failure to reject the null hypothesis when there is in fact a difference. The former is often called a type I error and the latter is often called a type II error. This is illustrated in Table 4.2. The probabilities of these errors occurring are shown in Table 4.2, and they are indicated as α and as ß. The probability α is called the significance level, and 1- ß is called the power of a test.

Table 4.2. Different types of error which may occur when performing an experiment to test the null hypothesis of no difference between two samples. The errors which can be made are to falsely reject the null hypothesis, and a failure to reject the null hypothesis when there is in fact a difference.

	H_0 is true	H_1 is true
accept H_0	1-α	type II error ß
reject H_0	type I error α	1-ß

Cohen (1977) provides sample size tables for detecting a difference between proportions when using a normal curve test. A proportion is a special case of an arithmetic mean, one in which the measurement scale has only two possible values, zero for the absence of a characteristic and one for its presence. In order to use the sample size tables, the significance level (i.e. type I error; probability that the null hypothesis of no difference is falsely rejected) and power level (i.e. [1- type II error]; [1- probability that the null hypothesis is not rejected when there is an actual difference]) have to be specified in advance. This means that the maximum allowable probabilities of the errors occurring are chosen first.

The experimenter will want to conclude significance when there is a certain degree of departure from the null hypothesis. This degree of departure also has to be specified in advance, and is called the effect size. The smaller the effect size which has to be detected, the larger the sample size will have to be. The objective could be to conclude significance when there is an actual difference of 10% when the underlying actual proportions are 60% and 70%. However, the difference between the proportions (10%) cannot be used as the degree of departure (effect size), since the detectability of a difference in magnitude is not a simple function of the difference. The detectability also depends on the standard deviations which are unknown because the underlying actual proportions are not known. For instance, it is easier to detect a difference when the actual proportions are 95% and 85%, than when the actual proportions are 70% and 60%, since the standard deviations of proportions closer to the

values 0 and 1 will be smaller than the standard deviations of proportions situated nearer the middle. In order to solve this problem, a transformation, called the arcsine transformation, is applied to the probabilities. Let f represent the probability after the arcsine transformation has been applied. This is defined as shown below:

$$f = 2 \arcsin \sqrt{p.} \qquad [4.1]$$

After this transformation has been carried out for both proportions, the difference between the transformed proportions is used as the effect size. Since the curve of this transformation is steeper at the ends than in the middle, a difference between two proportions at one end will lead to a larger effect size than a difference between two proportions in the middle. This allows the detectability of a specified effect size to be the same, whatever the magnitude of the underlying proportions. The effect size h is defined as follows:

$$h = f_1 - f_2 . \qquad [4.2]$$

If an effect size of 0.2 is chosen, this will correspond to, for example, [p_1 = 5% and p_2 = 10%] or [p_1 = 60% and p_2 = 70%] or [p_1 80% and p_2 = 87%].

Four parameters of statistical inference have been described:
- power,
- significance,
- effect size,
- sample size.

Any one of these is a function of the other three. When power, significance and effect size have been chosen, then the corresponding sample size can be looked up in the sample size tables provided by Cohen (1977).

In an experiment investigating difference between two proportions, the significance level could, for example, be set at 0.05, a power level of 0.80 could be chosen and the effect size could be set at 0.2. The corresponding sample size can be looked up in the appropriate sample size table (Cohen, 1977). For this example, a sample size of 392 can be found in the sample size table. If the significance level is changed to 0.1 in this setup, a sample size of 309 can be found. If an effect size of 0.2 is chosen, significance is set at 0.1 and power is set at 0.25, then the sample size shown in the sample size table equals 47. Since 40 test cases were finally available for the evaluation of PLEXUS, the above estimations show the limitations of the design which has been chosen for this study.

4.2.2.2. Consultation

Specifying who uses the system. The test cases are entered into the knowledge based system by the knowledge engineer. In this study, the problem solving performance of the knowledge based part of the system is evaluated, rather than the problem solving performance of the complete man-machine system. Therefore, the aim is to exclude the user interface from the evaluation. This is achieved by attempting to perform optimal data entry, independent of the interface. User interaction will require separate evaluation, after which the complete system should be evaluated.

Specifying physicians to test against. The system has been tested in comparison with international experts in the domain of brachial plexus injuries. Six well known experts in the field were asked to take part in the evaluation. The number of experts was limited to six for practical reasons. However, two of the experts were not able to take part in the evaluation due to the amount of time involved. Since time limitations prevented requesting additional experts to take part in the evaluation, the final number of cooperating experts amounted to four. Four experts from four different European countries (Great Britain, Spain, Switzerland and France) finally took part in the study. These physicians are among the most prominent experts in the field of brachial plexus injuries.

In the literature there appears to be no real consensus concerning the methods used to treat patients with a brachial plexus injury. Due to this variability, more than four experts should ideally be involved in the study. However, practical limitations prevented extension of this investigation.

Specifying a standard of performance. In order to perform a direct comparison, a standard of performance is necessary. For a brachial plexus injury this would ideally be the actual diagnosis and the optimal treatment. However, these are not known. If a patient has undergone surgery, the brachial plexus will have been exposed, but the incision which is made during surgery will depend on the diagnosis which is made pre-operatively. Therefore, even a diagnosis which is established during surgery is not necessarily the same as the actual diagnosis.

Since there is no actual gold standard of performance in this domain, the clinical diagnosis established by the expert who actually saw the patient is used as the standard for patients who were treated conservatively, and the per-operative diagnosis is taken as the standard for patients who were treated surgically. The treating expert's treatment was chosen to be the standard in all cases.

As was discussed in Section 4.2.2.1 there were four patients whose injury was serious enough to be treated surgically, but who were treated conservatively for various reasons. The diagnostic standard for these patients is the same as for the conservatively treated patients, i.e. the clinical diagnosis. For the two cases who were not operated on due to insurance problems and refusal of the operation, the expert's operative plan was used as the standard for treatment. For the late case and the contra-indicated case, the conservative treatment proposed by the expert was taken as the standard. For the contra-indicated case, the fact that there was a contra-indication and the reasons for this were stated in the case notes.

The above standards were used for the direct comparison of both the system and the non-treating experts relative to the treating experts. No explicit standard was used for the double blind judgement of the diagnoses and treatment plans, which was carried out in the third round of the evaluation.

4.2.2.3. Comparison

Variables to be compared. During this evaluation, only system output has been measured. The output consists of a diagnosis and a treatment plan for every patient. No attempt has been made at evaluating the system's reasoning. One of the reasons for this is that it is not known whether suggested ways for measuring expert reasoning, such as think aloud protocols, actually resemble their reasoning. This was discussed in detail in Section 2.3.2.3.

Furthermore, time limitations on the part of the experts would prevent extensive measurements of intermediate conclusions and facts. It is of the utmost importance that the intermediate conclusions should be correct, since the final conclusions depend on the subconclusions. Detailed subsystem analysis should be carried out before performing an empirical evaluation study. Ideally, the correctness of intermediate conclusions should also be investigated in a laboratory evaluation, however this was not feasible for PLEXUS.

Judging the results. The results were judged in two different ways:
- direct comparison of diagnoses and treatment plans by the researcher,
- subjective blind evaluation of diagnoses and treatment plans by experts.

As was discussed above, a standard of performance was established against which the knowledge based system's recommendations could be directly compared, although the standard is not an actual gold standard.

The domain of brachial plexus injuries has an important feature which complicates the evaluation study and has an important influence on the remainder of this chapter. The brachial plexus is a network of nerves, and one brachial plexus injury usually consists of more than one injured nerve. The

combination of injured locations forms the diagnosis for a patient. This means that one diagnosis does not consist of just one answer, but consists of multiple answers (e.g. 'avulsion C5, avulsion C6, rupture C7') which are all true at the same time. This makes a comparison of different diagnoses very difficult.

The direct comparison was performed by the researcher. The answers were either classified as being correct if they were exactly the same, and classified as being incorrect if they were not exactly the same. The degree of difference was not taken into account, for this requires expert knowledge.

A further possibility for judging the results is to ask independent experts to judge the diagnoses and treatment plans given by the knowledge based system and by the experts. An expert will have enough domain knowledge to rate a complete case consisting of multiple answers and an expert will be able to take into account the degree of incorrectness of an answer, which was not possible during the direct comparison. Although involving expert judgement will introduce subjectivity into the measurement.

In order to rule out bias due to a pro- or anti-computer opinion on the part of the judging experts, the judges should not know whether the diagnosis originated from the computer or from another expert. The origin of the opinions has to be disguised. This is called a blind evaluation.

Thus, the direct comparison will give a better indication of where the differences between the experts and the system arise, and the blind comparison will give a better indication of the clinical applicability of the system. Both approaches for judging the output (i.e. direct comparison and blind expert evaluation) have been applied and have been compared to each other in this evaluation study.

The experts who supplied the test cases and the non-treating experts who diagnosed the cases in the second round, provided their diagnoses in the form of free text. In order to be able to directly compare diagnoses and to be able to blind the diagnoses, these were translated into a uniform terminology. The standard terminology was decided on by the researcher in cooperation with dr.ir. R.B.M. Jaspers who was the project leader when PLEXUS was designed (Jaspers, 1990). A number of examples of this translation will be given below.

- Some physicians distinguish three types of severity and some use a classification of five types. These were all converted to the three types.
- When physicians were not sure about the severity of an injury, they indicated more than one type of severity for one injured location. When it was obvious that they preferred on of these, the other was stated in brackets behind the main answer. An example of this can be seen in Figure 4.6.
- If a physician provided two possibilities for the severity of an injured location and both seemed to have an equal chance of being true according to the

physician, a more general term was used. For instance, if the physician did not know whether a nerve was ruptured or whether this involved a lesion in continuity, the term 'injury' was used.

4.2.3. ANALYSIS OF THE RESULTS

The diagnoses and treatment plans provided by PLEXUS have been analysed in various ways. In the direct comparison, the diagnoses provided by PLEXUS were firstly compared directly to the standard answers provided by the treating experts. The results have been calculated using a number of different ways of performance calculation which have been described in the literature. The different ways of performance calculation have also been compared, and a choice has been made as to the most appropriate methods to be used in the evaluation of PLEXUS.

The diagnoses and treatment plans provided by the non-treating experts have also been compared directly to the standard using the same methods of performance calculation. Using these outcomes, the results achieved by PLEXUS and by the non-treating experts were compared.

The expert judgement scores which were obtained during the blind evaluation have also been analysed. The third round of the evaluation resulted in judgements for the diagnoses and the treatment plans. Using these judgements, the performance of the knowledge based system can be calculated. Finally, the results obtained in the direct comparison and in the blind evaluation have been compared. The results of the analysis are described below.

4.3. Results

4.3.1. RESULTS OF THE DIRECT COMPARISON

4.3.1.1. Comparing the diagnostic recommendations provided by PLEXUS to the diagnoses provided by the treating experts

Firstly, the diagnoses obtained from PLEXUS after the second round of the evaluation are compared directly to the standard of performance, consisting of the diagnoses provided by the treating experts, i.e. the experts who actually treated the patients. After this, the diagnoses obtained from the non-treating experts will be compared to the standard of performance. This will be discussed in Section 4.3.1.2. The treatment planning performance of the knowledge based system and of the non-treating experts will be discussed in Sections 4.3.1.5 and 4.3.1.6.

For one of the 40 test cases, at first no diagnosis could be obtained from the computer due to the fact that there was a cycle in the production rules of the knowledge base. The cycle consisted of two production rules which repeatedly called on each other. When this cycle was stopped manually, a diagnosis was obtained and this diagnosis has been used in the analysis.

A cycle in the rules is an error which should have been identified prior to this evaluation. These errors can be found using software tools which are independent of the domain (i.e. verification tools). PLEXUS has not been verified as extensively as is possible at present. Hence this cycle was only discovered at a late stage of development, and this slight error in the knowledge base should be resolved.

For one case, no diagnosis was given by the treating expert. Therefore, a total of 39 diagnoses could be used in the comparison.

Various methods of calculating knowledge based system performance have been described by Indurkhya and Weiss (1989). Several of these can be applied to systems such as PLEXUS, which operate in a domain where one diagnosis consists of multiple answers (i.e. multiple injured nerves), rather than a diagnosis consisting of a single answer.

Five different models for calculating system performance were used. These are listed below:
- the case correctness model,
- the partial correctness model,
- the modified partial correctness model,
- positive negative correctness model,
- the diagnostic performance model.

These methods of performance calculation will be discussed in detail.

The case correctness model. The first model for calculating performance is the case correctness model (see, for example, Indurkhya and Weiss, 1989). System cases are only categorised as correct if they exactly match the standard. Let c_{csys} be the number of cases which is classified completely correctly by the system and let t_{cstan} be the total number of test cases. Accuracy can be defined as:

$$\text{accuracy} = \frac{c_{csys}}{t_{cstan}}. \qquad [4.3]$$

Only 5 out of a total of 39 cases were diagnosed completely correctly by the knowledge based system. Therefore, the accuracy of the system calculated by

case correctness is 13%. This result can be related to the experts from which these cases originate. The results show that for the cases which originated from one of the experts, PLEXUS diagnosed none of the cases correctly. For the cases originating from two other experts, one of each was diagnosed correctly by PLEXUS. Three out of the ten cases originating from the final expert were diagnosed correctly. This difference in scores is probably due to the fact this expert had included radiological results in all files and most patients were quite severely injured. Nine out of the ten patients originating from the latter expert had at least one avulsion, which, when the appropriate information is available, makes the diagnosis less difficult for the computer.

The partial correctness model. Another way of calculating the performance is to attempt some kind of closeness measure. For example, if a patient has 2 locations within the brachial plexus, which are actually injured, and the system is correct on 1 of the locations, then the system can be given a score of 1/2 on correctness. This has to be combined with some measure of how precise the system was. These measures are also given by Indurkhya and Weiss (1989).

This model can be used for systems which provide multiple answers. It is of the utmost importance to note the difference between cases (used in the previous model) and answers (used in this model), as one case usually consists of multiple answers. This means that the number of answers is always greater than (or equal to) the number of cases.

Let c_{asys} be the total number of answers correctly given by the system, let t_{astan} be the total number of answers in the standard, and let t_{asys} be the total number of answers given by the system. The accuracy and predictive value are defined as follows:

$$\text{accuracy} = \frac{c_{asys}}{t_{astan}}; \qquad [4.4]$$

$$\text{predictive value} = \frac{c_{asys}}{t_{asys}}. \qquad [4.5]$$

The values for these measures have been calculated for PLEXUS. This amounts to the following results:

$$\text{accuracy} = \frac{86}{155} = 0.55; \qquad \text{predictive value} = \frac{86}{190} = 0.45.$$

An example of the use of the partial correctness model is shown below.

Example 4.1.

A performance study is aimed at determining the accuracy of a system. The study incorporates three cases.

x case number	c_{asys} correct answers system	t_{astan} number of answers in standard
1	2	3
2	2	2
3	250	500

The accuracy, determined using Eq. [4.4], equals $\frac{2+2+250}{3+2+500} = 0.50$

As can be seen from the above, the number of answers in the third case bears heavily on the value of the accuracy.

To avoid the fact that the number of answers in a case influences the result, consider the following modified calculation:

$$\text{modified accuracy} = \frac{1}{3}\left(\frac{2}{3} + \frac{2}{2} + \frac{250}{500}\right) = 0.72$$

This way of calculating accuracy will be called the modified partial correctness model and the general model will be defined below.

The modified partial correctness model. If x represents a particular case, then the measures are defined as follows:

$$\text{accuracy} = \frac{1}{t_{cstan}} \sum_{x=1}^{t_{cstan}} \frac{c_{asys}(x)}{t_{astan}(x)} \; ; \qquad [4.6]$$

$$\text{predictive value} = \frac{1}{t_{cstan}} \sum_{x=1}^{t_{cstan}} \frac{c_{asys}(x)}{t_{asys}(x)} \; . \qquad [4.7]$$

The values of the above measures have been calculated for PLEXUS, which amounts to the following results:

accuracy = 0.52 ; predictive value = 0.45 .

This method of calculation is very suitable for systems such as PLEXUS which provide multiple answers to each case, since the number of answers per case does not influence the outcome.

The positive negative correctness model. This model can be applied in domains where only one answer is provided for each case. The answers given by the system are categorised into four different groups:

- When both system and standard agree that in a certain case a certain disease is present, then the answer is classified as being 'true positive'.
- When both system and standard agree that a certain disease is not present, then the answer is classified as being 'true negative'.
- If the system indicates a certain disease to be present, and it is absent according to the standard, then the answer is classified as being 'false positive'.
- If the system indicates a certain disease to be absent, and it is present according to the standard, then the answer is classified as being 'false negative'.

The categorisation is carried out with respect to a certain class, e.g. a certain disease.

For systems such as PLEXUS which give multiple answers for one case, it is not possible to use this categorisation. For PLEXUS, the following categorisation has been carried out for each case, where each answer in a case was classified.

- An answer was classified as 'true positive' if both system and treating expert had indicated this answer.
- An answer provided was classified as 'false positive' if the system had mentioned the answer but the treating expert had not mentioned this answer.
- An answer which had been mentioned by the treating expert, but which had not been mentioned by the system was classified as 'false negative'.

For PLEXUS the mean number of true positive, false positive and false negative answers per case have been calculated, and are shown below:

mean number of true positive answers per case = 2.21
mean number of false positive answers per case = 2.67
mean number of false negative answers per case = 1.77

Table 4.3. Comparison of various methods of accuracy calculation for PLEXUS relative to the experts who actually treated the patients. The left hand column shows the method of performance calculation which was used, and the right hand column shows the results achieved by PLEXUS calculated according to the different methods.

performance index	PLEXUS vs. treating experts
case correctness:	
accuracy	0.13
partial correctness:	
accuracy	0.55
predictive value	0.45
modified partial correctness:	
accuracy	0.52
predictive value	0.45
positive negative correctness:	
mean no. true positive answers	2.21
mean no. false positive answers	2.67
mean no. false negative answers	1.77

The system performance results which were calculated for PLEXUS using the case correctness model, the partial correctness methods, and the positive negative correctness method are summarised in table 4.3.

Diagnostic performance model (Indurkhya and Weiss, 1989). This is a measure which can be used to analyse the performance of a system for each class of answers. For PLEXUS a class of answers is the same as an injury location. A class could, for instance, be 'avulsion C5', and the performance of the system for this class may be determined.

In order to be able to calculate the performance for a class of answers, all the answers given by the standard are seen separately (i.e. not as a case), and these single answers are called p-cases. One p-case thus represents one answer given by the standard. The answers given by the standard are divided into single answers, but the answers given by the system are kept intact. The way in which this is done is illustrated below in Example 4.2.

Example 4.2.

A study involves two test cases. For these cases, the answers given by the standard and the answers given by the system are shown below.

case no.	standard	system
case 1	extraforaminal C5, extraforaminal C6	extraforaminal C6, fasciculus medialis
case 2	extraforaminal C5	fasciculus lateralis

The answers given by the standard are divided into three single answer p-cases. There are two different classes of answers which are represented in the p-cases, i.e. extraforaminal C5 and extraforaminal C6. The answers given by the system are left intact.

p-case no.	standard	system
p-case 1	extraforaminal C5	extraforaminal C6, fasciculus medialis
p-case 2	extraforaminal C6	extraforaminal C6, fasciculus medialis
p-case 3	extraforaminal C5	fasciculus lateralis

Table 4.4. PLEXUS accuracy calculated using the diagnostic performance model. The accuracy indicates how many times PLEXUS agrees with the standard on a certain injury, out of the total number of times a certain class of injuries is stated to be present by the standard.

class	accuracy
avulsion C4	1/1 = 1
avulsion C5	5/10 = 0.5
avulsion C6	10/20 = 0.5
avulsion C7	14/22 = 0.64
avulsion C8	14/19 = 0.74
avulsion T1	19/19 = 1

For each p-case, all answers given by the system can be compared to the p-case answer given by the standard. If the answers given by the system contain the p-case answer, then the p-case is classified as correct, otherwise it is classified as incorrect. In Example 4.2. it can be seen that the first and third p-case will be incorrect, and the second one will be classified as correct because the answers given by the system contain the p-case answer given by the standard.

A formal description of the principles illustrated in the example is presented below. Given that c_{psys} is the number of correct p-cases given by the system, and t_{pstan} is the number of p-cases in the standard. For a certain class of answers k, the accuracy for that class is defined as:

$$\text{accuracy}(k) = \frac{c_{psys}(k)}{t_{pstan}(k)}. \qquad [4.8]$$

The diagnostic performance model allows analysis of system performance for a class of answers. This model cannot be used to estimate the performance of the system over cases. If one tries to do so, the performance will be biased towards cases with more multiple conclusions (Indurkhya and Weiss, 1989). A case containing many answers would have a larger effect on the outcome than a case with fewer answers, as all answers are treated similarly.

The performance of PLEXUS on particular injury locations can be determined using the diagnostic performance model which was described above. Analysis of system performance on each specific injury location can help in finding the weak points of a system. This information can be used for updating the system.

The correctness rates for PLEXUS on the avulsion injuries are shown in Table 4.4. These have been calculated using Eq. [4.8] representing the p-case method which was explained above. It can be seen that the accuracy of the system is best on the lower avulsions. This can be explained by the fact that it is easier to diagnose lower root avulsions than it is to diagnose upper root avulsions. The system classifies all the avulsions of T1 correctly. However, avulsions of C6 are classified correctly in 10 out of 20 cases.

Analysing a system in this way can also give an idea of the frequency of occurrence of certain injuries in the test set, and of the representativeness of the test cases. In the present study consisting of the data of 39 patients, there are 31 different classes of injury locations which are mentioned by the treating experts (standard), of which 16 are only mentioned once. The number of avulsions (i.e. very severe injuries) which is said to be present by the treating experts in these 39 test patients is very large, it amounts to a total of 91.

Table 4.5. PLEXUS false positive rate calculated using the diagnostic performance model. The false positive rate indicates how many times PLEXUS falsely indicates a certain injury to be present, out of the total number of times this injury class is mentioned by the system.

class	false positive rate
avulsion C4	0/1
avulsion C5	2/7 = 0.29
avulsion C6	3/13 = 0.23
avulsion C7	0/14
avulsion C8	0/14
avulsion T1	0/19

In the performance measure used above, the answers given by the standard were divided into single answer cases. A further measure can be established by dividing the answers given by the system (as opposed to those given by the standard) into single answer cases. The single answer cases will be called q-cases. Now the answers given by the standard are left intact. Extending Example 4.2, the q-cases can be analysed as illustrated in Example 4.3.

Example 4.3.

The cases which were described in Example 4.2 are given. For these cases the answers given by the system and the answers given by the standard are shown below.

case no.	system	standard
case 1	extraforaminal C6, fasciculus medialis	extraforaminal C5, extraforaminal C6
case 2	fasciculus lateralis	extraforaminal C5

The answers given by the system can now be divided into single answer q-cases. There are three different classes of answers which are represented in the q-cases, i.e. extraforaminal C6, fasciculus medialis and fasciculus lateralis. The answers given by the standard are left intact.

q-case no.	system	standard
q-case 1	extraforaminal C6	extraforaminal C5, extraforaminal C6
q-case 2	fasciculus medialis	extraforaminal C5, extraforaminal C6
q-case 3	fasciculus lateralis	extraforaminal C5

It can be seen that in q-cases 2 and 3, the system gives an answer that does not appear in the standard. Since in these q-cases the system falsely indicates an injury to be present, this is called a *false positive answer*.

Given that f_{qsys} is the number of false positive q-cases given by the system, and t_{qsys} is the number of q-cases given by the system, then the false positive rate for a certain class of answers k, can be defined as follows:

Table 4.6. Performance of non-treating experts and knowledge based system relative to the experts who treated the patients. The different methods of performance calculation which can be seen in the left hand column were used to compare the diagnoses obtained from PLEXUS and from the non-treating experts to the diagnoses provided by the treating experts. The results of all non-treating experts combined is shown in the column denoted nt-experts. The individual non-treating experts are denoted as e1, e2, e3, e4. The treating experts are abbreviated as t-experts. The standard deviations are shown in the brackets behind the performance values.

perfor-mance index	PLEXUS vs. t-experts	nt-experts vs. t-experts	e1 vs. t-experts	e2 vs. t-experts	e3 vs. t-experts	e4 vs. t-experts
number of cases	39	47	13	7	13	14
true positive	2.21 (1.74)	2.47 (1.79)	2.77 (1.69)	1.86 (1.68)	2.31 (1.97)	2.64 (1.87)
false positive	2.67 (1.94)	1.53 (1.30)	1.77 (1.48)	1.29 (1.11)	1.77 (1.54)	1.21 (0.98)
false negative	1.77 (1.39)	1.51 (1.27)	1.54 (1.27)	1.57 (1.40)	1.69 (1.18)	1.29 (1.38)
case correct	0.13	0.17	0.15	0.14	0.15	0.21
partial correct: accuracy	0.55	0.62	0.64	0.54	0.58	0.67
predictive value	0.45	0.62	0.61	0.59	0.57	0.69
mod. part. correct: mod. accuracy	0.52 (0.35)	0.58 (0.36)	0.62 (0.34)	0.54 (0.40)	0.49 (0.39)	0.64 (0.36)
mod. pred. value	0.45 (0.36)	0.57 (0.38)	0.60 (0.36)	0.56 (0.44)	0.51 (0.42)	0.61 (0.35)

$$\text{false positive rate}(k) = \frac{f_{qsys}(k)}{t_{qsys}(k)}. \qquad [4.9]$$

For each possible class of injuries given by the system, the fraction of false positive answers can be determined. As explained above, this method cannot be used to calculate the false positive rate over cases.

Using Eq. [4.9] the false positive rate has been calculated for the answer classes which are given by PLEXUS. Some results of this method of calculation for determining the false positive rate of the classes obtained from the system are shown in Table 4.5.

In the present study, for the 39 test cases involved in the evaluation, there are 52 different classes of answers which are mentioned by the system, of which 20 are mentioned only once.

4.3.1.2. Comparing the diagnoses obtained from non-treating experts and from PLEXUS to the diagnoses established by the treating experts

Using the methods of calculation which were described above, the performance of the knowledge based system was determined relative to the treating experts. The methods of calculation which were used to determine system performance, have also been used to determine the performance of the non-treating experts relative to the standard, i.e. the treating experts. These results were obtained by direct comparison of the diagnoses by the researcher. The answers were classified as being correct if they were exactly the same, and classified as being incorrect if they differed. The degree of difference was not taken into account.

Table 4.6 shows the results of the direct comparison of the non-treating experts to the standard, using the metrics which were discussed above in Section 4.3.1.1. For reasons of anonymity, the non-treating experts are denoted as 'e1', 'e2', 'e3' and 'e4'. The total of all non-treating experts is shown in the column denoted 'nt-experts'. The term 't-experts' stands for the treating experts. In order to compare system results to the results obtained by the non-treating experts, the system results which were calculated earlier are shown in the second column of the table.

The number of cases which was diagnosed by each expert and which could be used in this comparison ranges from 7 to 14. One expert (e2) only diagnosed 7 cases in a manner which was detailed enough to allow a comparison to be made. The fact that only 7 out of 14 cases could be used, may influence the results obtained by this expert. If the cases for which a diagnosis is available are easier than the other 7 cases for which no clear diagnosis was obtained, then the scores achieved by this expert will be more positive. An informal analysis shows that

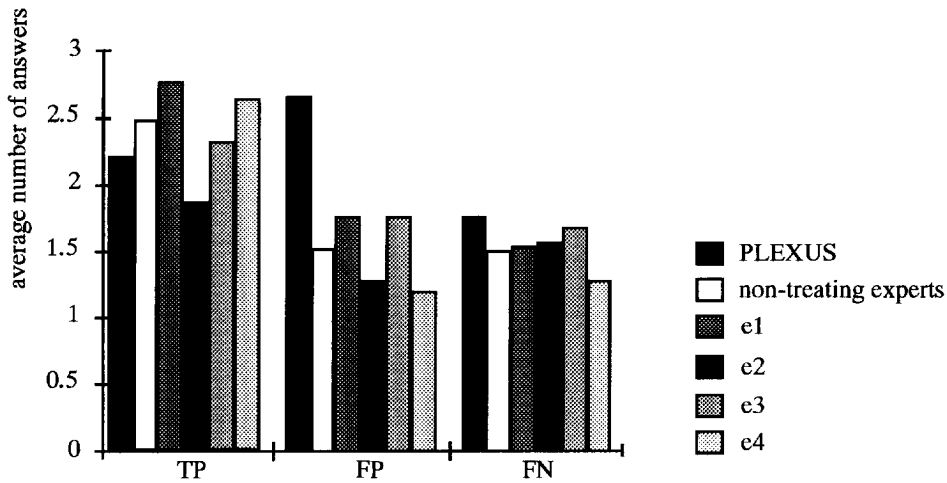


Figure 4.1. Average number of true positive (TP), false positive (FP) and false negative (FN) answers per case provided by PLEXUS and by the non-treating experts, relative to the experts who actually treated the patients. These values have been calculated for each non-treating expert individually, and for all non-treating experts combined.

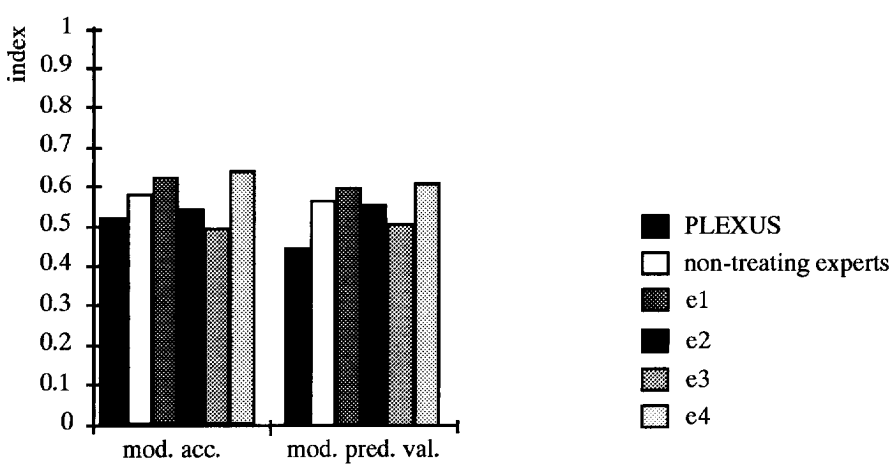


Figure 4.2. Modified accuracy and modified predictive values calculated for PLEXUS and for the non-treating experts relative to the treating experts. These values have been calculated for each non-treating expert individually, and for all non-treating experts combined.

the scores obtained by PLEXUS on the cases for which a diagnosis from e2 was available were somewhat higher than the scores obtained by PLEXUS on the other cases which were sent to e2 but for which no detailed diagnosis was available. These small numbers do not allow statistical analysis. It may be concluded, however, that by only taking certain cases into account, this expert is given the benefit of the doubt.

In Table 4.6, the numbers in the brackets represent the standard deviations. It can be seen that the standard deviations are substantial.

An estimation of the influence of the standard deviation on the results shown in the table will be demonstrated using an example. As can be seen in Table 4.6, the mean number of false negatives measured for PLEXUS amounted to 1.77. This is the sample mean. Assume that the tables for the t-distribution can be used. Since the sample size is 39, there are 38 degrees of freedom. The t-distribution for 38 degrees of freedom and a 95% confidence interval shows a value of t = 2.02. The confidence interval for the underlying actual value (rather than the measured value) of the number of false negatives can be calculated as follows:

$$1.77 - 2.02\frac{1.39}{\sqrt{39}} < \text{number of false negatives} < 1.77 + 2.02\frac{1.39}{\sqrt{39}}$$

$$= 1.32 < \text{number of false negatives} < 2.22$$

Since a normal distribution cannot be assumed in this case, the above is only an estimation. The 95% confidence intervals of the values shown in Table 4.6 have also been calculated using a nonparametric method called the bootstrap (Efron and Gong, 1983). For each sample, random draws with replacement were made from the sample until a new sample of the same size as the original sample was obtained, and the mean was calculated. This is repeated a number of times. In this study this was done 1000 times. In this way a bootstrap histogram of means is obtained. Using a method called the percentile method, the confidence interval can be determined. The 100α and $100(1-\alpha)$ percentiles are taken from the bootstrap histogram to obtain a $(1-2\alpha)$ confidence interval. The 95% confidence interval for the example above was calculated:

$$= 1.33 < \text{number of false negatives} < 2.23$$

All confidence intervals calculated in this way were approximately the same as those using the parametric standard deviation. Furthermore, since Efron and Gong (1983) state that the method of determining confidence intervals, though encouraging, is highly speculative, the parametric standard deviations will be given in this chapter.

The true positive, false positive, false negative values and the modified accuracy and modified predictive values obtained by PLEXUS were compared to the values obtained by the non-treating experts. The nonparametric Wilcoxon-Mann-Whitney test was used for hypothesis testing, since this avoids the t-test's assumption of a normal distribution being present. Significance was concluded at the 5% level (two-tailed $p < 0.05$). According to Siegel and Castellan (1988), Wilcoxon-Mann-Whitney is one of the most powerful of the nonparametric tests, and it is a very useful alternative to the parametric t-test when the researcher wishes to avoid the t-test's assumptions.

The mean number of true positive answers per case given by PLEXUS cannot be shown to differ significantly from the number of true positive answers given by the non-treating experts combined. This is also true for the average number of false negative answers. However, the number of false positive answers per case provided by the system is significantly higher (two-tailed $p < 0.01$) than the number of false positives given by the non-treating experts when combining all non-treating experts, i.e. third column of Table 4.6.

A similar trend can be seen in the lower half of the table for the modified predictive values, although the difference here was not found to be significant. The modified accuracy of the system cannot be shown to differ significantly from that of the non-treating experts. A graphical representation of the most important results is shown in Figure 4.1 and Figure 4.2.

In the third column of Table 4.6 it can be seen that the total number of cases diagnosed by the non-treating experts is larger than 39. This means that there are cases which have been diagnosed by more than one non-treating expert. Since fact that these patients are used twice could have influenced the results, the same calculations were carried out after removing these patients from the sample. However, the results which were described above were also obtained after performing the hypothesis tests on the reduced sample.

For a number of cases, a possible explanation for the low predictive value and large number of false positive answers suggested by PLEXUS, is that the system tries to explain all of the dysfunction which is present in the patient, thereby increasing the number of false positive answers when compared to the treating experts. However, since the standard is not an actual gold standard, the false positive answers do not necessarily have to be wrong. Furthermore, they may be seen as suggestions to the physician. The final round of the evaluation, where a number of experts has been asked to blindly judge the diagnoses and treatment plans provided by the treating experts, non-treating experts and by the knowledge based system, will show whether the false positives are indeed relevant answers

or whether they are not. The results of the final round of the evaluation will be discussed in Section 4.3.2.

4.3.1.3. Comparing the diagnoses suggested by PLEXUS directly to the diagnoses proposed by the non-treating experts

Table 4.7 shows the results of a direct comparison of the diagnoses proposed by PLEXUS and the diagnoses proposed by the non-treating experts. In contrast to the previous section, in which both PLEXUS and non-treating experts were compared to the treating experts, the calculations are carried out using the non-treating experts as if they are the standard. This has been done for each expert individually and for all non-treating experts combined. No significant difference (using Wilcoxon-Mann-Whitney) is found between the modified predictive value and modified accuracy of PLEXUS relative to the non-treating experts and of PLEXUS relative to the treating experts (second columns of Table 4.7 and Table 4.6).

Table 4.7. Performance of the knowledge based system compared to the non-treating experts. The different methods of performance calculation which can be seen in the left hand column were used to compare the diagnoses obtained from PLEXUS to the diagnoses submitted by the non-treating experts. The results of comparing PLEXUS to all non-treating experts combined is shown in the second column (nt-experts). The individual non-treating experts are denoted as e1, e2, e3, e4. The standard deviations are shown in the brackets behind the performance values.

performance index	PLEXUS vs. nt-experts	PLEXUS vs. e1	PLEXUS vs. e2	PLEXUS vs. e3	PLEXUS vs. e4
no. cases	48	13	7	14	14
true positive	2.40 (1.93)	2.38 (1.81)	1.57 (2.15)	2.57 (1.95)	2.64 (2.02)
false positive	2.25 (2.04)	2.46 (2.15)	2.57 (2.15)	2.07 (2.02)	2.07 (2.09)
false negative	1.58 (1.49)	2.15 (1.73)	1.57 (1.27)	1.43 (1.45)	1.21 (1.37)
accuracy	0.60	0.53	0.50	0.64	0.69
pred. value	0.52	0.49	0.38	0.55	0.56
mod. accuracy	0.55 (0.41)	0.49 (0.37)	0.40 (0.50)	0.61 (0.41)	0.63 (0.41)
mod. pred. val.	0.50 (0.39)	0.49 (0.39)	0.33 (0.41)	0.56 (0.40)	0.55 (0.40)

Table 4.8. Inter-expert variability. The different methods of performance calculation which can be seen in the left hand column were used to compare the diagnoses obtained from individual non-treating experts to those of their colleagues who also only saw the patients on paper. The individual non-treating experts are denoted as e1, e2, e3, e4. The standard deviations are shown in the brackets behind the performance values.

performance index	e1 vs. other nt-exp.	e2 vs. other nt-exp.	e3 vs. other nt-exp.	e4 vs. other nt-exp.
no. cases	10	4	10	8
true positive	2.80 (2.04)	3.00 (2.45)	3.10 (2.03)	2.88 (1.73)
false positive	1.20 (1.62)	0.50 (1.00)	0.50 (0.85)	1.00 (1.07)
false negative	0.80 (1.14)	0.50 (1.00)	0.70 (0.82)	1.25 (1.75)
accuracy	0.78	0.86	0.82	0.70
pred. value	0.70	0.86	0.86	0.74
mod. accuracy	0.72 (0.42)	0.75 (0.50)	0.71 (0.40)	0.73 (0.38)
mod. pred. val.	0.67 (0.42)	0.75 (0.50)	0.78 (0.42)	0.70 (0.34)

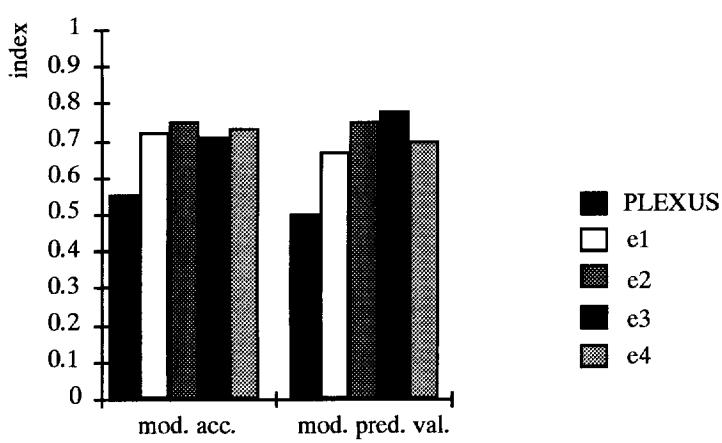


Figure 4.3. Comparing PLEXUS and the individual non-treating experts to the other non-treating experts. The figure shows the modified accuracy and predictive values obtained by comparing PLEXUS to all non-treating experts, and by comparing individual non-treating experts to the other non-treating experts.

4.3.1.4. Diagnostic inter- and intra-expert variability

Inter-expert variability. In order to compare each expert to all three other experts, each expert was asked to indicate a diagnosis for 3 groups of 4 patients, where each group of 4 was also sent to one other non-treating expert for analysis. This allowed the inter-expert variability to be analysed. It was evident that these numbers would be too small to analyse the variabilities properly. However, it would not have been feasible to have asked each of the experts to analyse more than the 17 patients which they were asked to diagnose in the second round.

Table 4.8 shows the inter-expert agreement for each expert compared to the other three experts. In theory 12 cases would have been available to calculate the inter-expert agreement for each expert. However, for a number of cases no adequate diagnosis was provided, which could be used for the comparison. This may again mean that only the easier cases may have been compared, and the pairs for which fewer cases could be compared are thus given the benefit of the doubt.

When comparing the results of PLEXUS using the non-treating experts as the standard (Table 4.7) to the inter-expert variabilities (Table 4.8), some differences can be seen. The measured modified accuracies and predictive values of the non-treating experts are higher than those achieved by PLEXUS. However, when using Wilcoxon-Mann-Whitney for hypothesis testing of the true positive values, false positive values, false negatives, modified accuracies and modified predictive values, the differences were only found to be significant for PLEXUS and one expert (e3) on the false positive value and on the modified predictive value. This is probably due to the limited amount of data. The modified accuracies and predictive values are also shown in Figure 4.3.

On the whole, the measured inter-expert results (Table 4.8) are better than the results which are obtained when comparing the non-treating experts to the treating experts (Table 4.6). However, when comparing true positive values, false positive values, false negatives, modified accuracies and modified predictive values the differences were only found to be significant for one of the experts (e3) for the number of false positives and the number of false negatives. The fact that no further differences were found may be due to the limited amount of data.

Possible differences between inter-expert results and results obtained when comparing the non-treating experts to the treating experts may have been caused by the fact that the non-treating experts did not actually see the patients. The non-treating experts received a patient file, and were asked to diagnose the patients from paper, which differs from the normal working situation. Furthermore, some information may not have been present in the patient files. In the latter situation, the system will also have been influenced by this.

Table 4.9. Comparing non-treating experts to treating experts using a more lenient scoring scheme. This scoring scheme allows answers, for which the experts could not distinguish between two possibilities, to be judged correct if either of the possibilities was actually true. The diagnoses of PLEXUS and the individual non-treating experts (e1, e2, e3, e4) were compared to those submitted by the treating experts (t-experts). The values in this table can be compared to the values in Table 4.6 where a more strict scoring scheme was used.

performance index	PLEXUS vs. t-experts	e1 vs. t-experts	e2 vs. t-experts	e3 vs. t-experts	e4 vs. t-experts
no. cases	39	13	7	13	14
accuracy	0.57	0.67	0.61	0.62	0.68
predictive value	0.49	0.69	0.77	0.68	0.72

A further influence which may explain the possible difference in performance of the non-treating experts relative to the treating experts and relative to the other non-treating experts has been investigated separately. The treating experts operated on most of the patients. Therefore, they would have been more certain of their answer than the experts who did not see the patients. This is illustrated in the following example. If a patient with an injury of C5 had been operated on by the treating expert, the expert would have been able to see whether the patient had an avulsion of C5 or whether the patient had a rupture of C5. The treating expert would therefore have given only one of these answers. Whereas a physician who had not seen the patient may have given both answers, i.e. 'rupture or avulsion C5'.

When the direct comparison is applied in the strictest sense, as was done in this evaluation, the answer given by the non-treating expert will be classified as incorrect. Since this will have created a negative influence on the results achieved by the non-treating experts, the influence of this effect was estimated by performing the direct comparison again but using less strict classification rules. This time, if an answer such 'rupture or avulsion C5' was given by the non-treating expert and the standard was either 'avulsion C5' or 'rupture C5', then the answer was classified as correct, rather than as incorrect.

By scoring the uncertain answer as being correct if either of the situations are actually true, the most positive result which can be achieved by the non-treating experts is determined. The changes this brings about in the accuracy values are shown in Table 4.9. It can be seen that although there is an increase in the performance values brought about by relaxing the strictness of the correctness classification, the complete gap between the values in Tables 4.6 and 4.8 (i.e. performance of non-treating experts relative to the treating experts and relative to the other non-treating experts) cannot be explained by this effect alone, which is illustrated in Figure 4.4 and Figure 4.5.

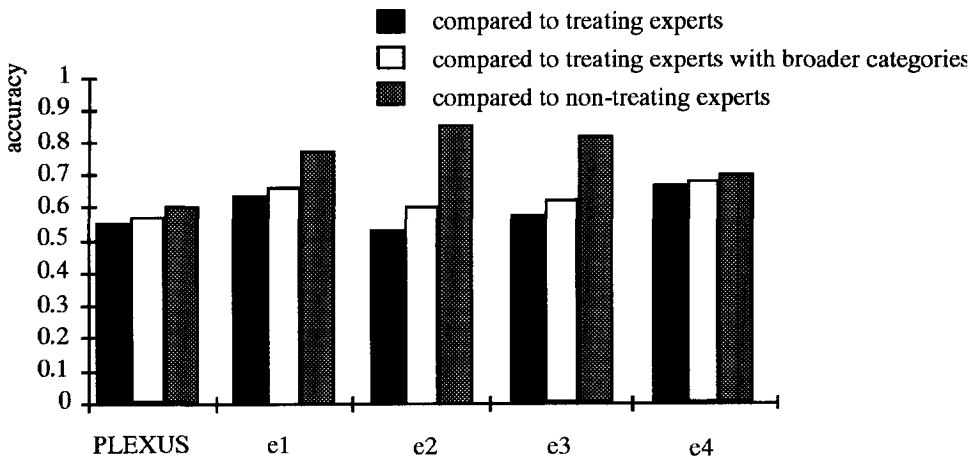


Figure 4.4. Change in accuracy when altering category definitions. The figure shows the accuracy when comparing the non-treating experts to the treating experts using the strict scoring scheme, using the more lenient scoring scheme, and when comparing the non-treating experts to the other non-treating experts.

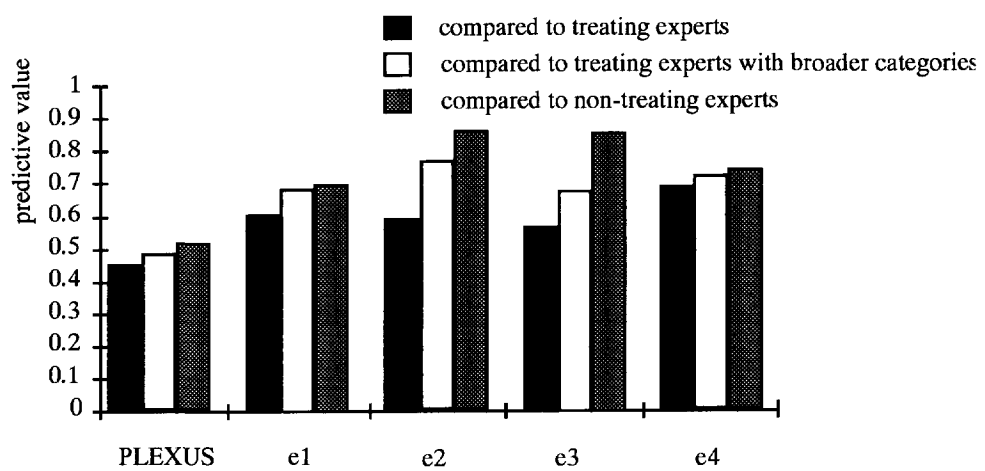


Figure 4.5. Change in predictive value when altering category definitions. The figure shows the predictive value when comparing the non-treating experts to the treating experts using the strict scoring scheme, using the more lenient scoring scheme, and when comparing the non-treating experts to the other non-treating experts.

Intra-expert variability. During the second round of the evaluation, the non-treating experts were also sent three of their own patients. The experts were not told they would be diagnosing some of their own patients. In order not to let the physicians know this, the patient names were removed and the files were randomly mixed with the other patients who had to be diagnosed.

By having the experts diagnose their own patients, the intra-expert variability could be studied. The results are shown in Table 4.10. Since the numbers involved are very small, care should be taken when interpreting the results. The four columns on the right hand side of Table 4.10 show that the intra-expert variability is strongly dependent on the level of difficulty of the injury, and this may vary greatly. The modified accuracies and modified predictive values over all non-treating experts are similar to the modified accuracies and modified predictive values which were obtained when comparing the non-treating experts to the treating experts (Table 4.6).

Table 4.10. Intra-expert variability. In the second round of the evaluation, the non-treating experts (e1, e2, e3, e4) were sent a number of their own patients. The non-treating experts did not know this. By comparing the diagnoses provided by the non-treating experts on their own patients to their own actual diagnoses, the intra-expert variability could be calculated. The treating experts are denoted by t-e1, t-e2, t-e3 and t-e4. In actual fact they are the same persons as e1, e2, e3 and e4. The total, calculated over all experts, is shown in the second column.

compared to	nt-experts vs. themselves as t-experts	e1 vs. t-e1	e2 vs. t-e2	e3 vs. t-e3	e4 vs. t-e4
no. cases	11	3	3	3	2
accuracy	0.66	0.55	0.83	0.22	1.00
pred. value	0.71	0.55	0.77	0.40	1.00
mod. accuracy	0.57 (0.43)	0.49 (0.71)	0.67 (0.96)	0.25 (0.38)	1.00
mod. pred. val.	0.63 (0.43)	0.49 (0.71)	0.67 (0.96)	0.50 (0.76)	1.00

Table 4.11. Categorisation of treatment plans provided by PLEXUS and by the treating experts. The treatments of the knowledge based system are shown vertically, and the treatments performed by the treating experts are shown horizontally. The diagonal shows the proportion of cases in which the system and the treating experts agree on the treatment. The values in the brackets are the probabilities of both system and expert stating the same treatment by chance alone.

STANDARD \ PLEXUS	graft	graft + trans	trans	neurol	decomp	wait	cons	total
graft	$\frac{6}{40}$ $(\frac{169}{1600})$	$\frac{7}{40}$						$\frac{13}{40}$
graft + trans	$\frac{1}{40}$	$\frac{3}{40}$ $(\frac{78}{1600})$	$\frac{2}{40}$					$\frac{6}{40}$
trans	$\frac{2}{40}$	$\frac{3}{40}$	$\frac{4}{40}$ $(\frac{54}{1600})$					$\frac{9}{40}$
neurol	$\frac{2}{40}$					$\frac{1}{40}$		$\frac{3}{40}$
decomp	$\frac{1}{40}$					$\frac{1}{40}$		$\frac{2}{40}$
wait	$\frac{1}{40}$					$\frac{1}{40}$ $(\frac{10}{1600})$		$\frac{2}{40}$
cons						$\frac{2}{40}$	$\frac{3}{40}$ $(\frac{15}{1600})$	$\frac{5}{40}$
total	$\frac{13}{40}$	$\frac{13}{40}$	$\frac{6}{40}$			$\frac{5}{40}$	$\frac{3}{40}$	$\frac{40}{40}$

The intra-expert agreement is similar to the results in Table 4.6. The measured results are lower than the inter-expert agreement shown in Table 4.8. However, when using Wilcoxon-Mann-Whitney to compare modified accuracy and modified predictive values no significant difference is found, which is probably due to the limited amount of data. Although no significant differences were found, possible differences may be due to the fact that the non-treating experts did not see and did not operate on the patients. If the non-treating experts were asked to diagnose the patients from paper again and the intra-expert agreement was then calculated, the intra-expert results would be expected to be similar to the inter-expert agreement.

4.3.1.5. Comparing PLEXUS treatment plans to those of the treating experts

In contrast to the diagnoses, the treatment plans do not consist of multiple answers. All treatment plans could be placed into one of seven categories. These categories can be seen in Table 4.11, and they correspond to 'nerve graft', 'nerve graft and nerve transfer', 'nerve transfer', 'neurolysis', 'decompression', 'conservative treatment for the time being' and 'conservative treatment only'. Since the aim of PLEXUS is to distinguish the patients who should be treated surgically from those who should only be treated conservatively, the treatment plans which are proposed by PLEXUS are stated at a relatively general level. To be able to compare PLEXUS to the experts, the treatments proposed by the experts have also been abstracted to this level of generality. However, the categories were not based solely upon the answers which can be given by PLEXUS. For instance, the category 'decompression' is never suggested by PLEXUS, however, it was mentioned by some experts, therefore this category is also used.

The Kappa coefficient (Cohen, 1960; Cohen, 1968) is a measure of agreement which can be used to analyse nominal scale agreement, and was therefore chosen for comparing treatment plans. This measure was also used in evaluation studies performed by Reggia (1983) and Kors *et al.* (1990). Kappa is a chance corrected measure of agreement, and was mentioned previously in Section 2.3.3. The way in which Kappa has been used for comparing the treatments provided by PLEXUS to those provided by the treating experts will be discussed in detail in this section.

Table 4.11 shows the categorisation of the treatment plans given by PLEXUS and by the treating experts. The total number of treatment plans obtained from each equals 40. The numbers in the cells are the fractions of treatment plans for which the treating experts have indicated the treatments shown in the first column and the system has suggested the treatments shown in the first row. The values recorded on the diagonal of the table are proportions of

cases for which both PLEXUS and the standard agree on the treatment plan. Sometimes this proportion of agreement, (called p_b) is used as a measure of agreement. However, a certain amount of agreement may be expected to occur by chance. Therefore, Cohen (1960) proposed a chance corrected measure for agreement. The numbers in brackets (Table 4.11) show the probability by chance alone of both PLEXUS and the standard giving the same answer. For instance, the probability by chance alone of both standard and PLEXUS choosing 'trans' equals the probability of the standard choosing 'trans' ([2+3+4]/40 = 9/40) multiplied by the probability of PLEXUS choosing 'trans' ([2+4]/40 = 6/40), which equals (54/1600). The proportion of agreement expected by chance is termed p_c.

Let p_o be the observed proportion of agreement and p_c be the proportion of agreement expected by chance. The Kappa coefficient of agreement K is then defined as follows.

$$K = \frac{p_o - p_c}{1 - p_c}. \qquad [4.10]$$

The upper limit of Kappa is 1, as p_c cannot equal 1 if there is more than one answer category and answers are placed in more than one category. It would seem more obvious that the probability of indicating a treatment purely by chance would be the a priori probability of this occurring. Using the Kappa methodology, the probability of equal opinions occurring by chance is taken from the sample itself, and therefore will depend on the sample. However, since the a priori probabilities are not known, the sample is the only information available.

In the literature there is some discussion concerning p_c (Siegel and Castellan, 1988; Gjørup, 1988). Gjørup (1988) states that Kappa is dependent on the prevalence of a diagnosis and suggests that Kappa values should be presented together with the original results in a contingency table.

The value of the Kappa coefficient of agreement between PLEXUS and the standard can be calculated using the values shown in Table 4.11 and equals 0.28 (p_o = 17/40 = 0.43 and p_c = 326/1600 = 0.20). For most purposes, values greater than 0.75 may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance (Fleiss, 1981).

The relatively low value found for PLEXUS has been influenced by the number of times PLEXUS indicates 'graft and transfer' when the treating expert indicates 'graft'. Using the results of performance of the non-treating experts relative to the treating experts, which will be described below, it will be possible to investigate whether the non-treating experts also suggest this procedure (i.e. 'graft and transfer') more often than the treating experts.

4.3.1.6. Comparing treatment plans obtained from non-treating experts and from PLEXUS to the treating experts

Table 4.12 shows the values of the Kappa coefficients of agreement and the proportions of agreement p_o when comparing the treatment plans of the non-treating experts to those recommended by the treating experts. Kappa has been calculated as was discussed above. The values in brackets are the standard deviations of Kappa which can be used to test whether the underlying value of Kappa is significantly different from a prescribed value other than zero. The formula for calculating the standard error (i.e. standard deviation divided by the square root of the sample size) is given in Appendix 1.

By studying the categorisation of the treatment plans it could be seen that there is a number of cases for which the non-treating experts suggest a 'graft and transfer' when the standard chooses 'graft' (10/53), which is similar to the pattern found for PLEXUS.

Table 4.12. Comparing the treatment plans obtained from PLEXUS and from the non-treating experts to the standard of performance. The Kappa coefficient of agreement and the proportion of agreement on treatment (p_o) have been calculated for PLEXUS relative to the treating experts (t-experts), and for the non-treating experts (nt-experts) relative to the treating experts. The standard deviation of Kappa is shown in the brackets.

agreement	PLEXUS vs. t-experts	nt-experts vs. t-experts	e1 vs. t-experts	e2 vs. t-experts	e3 vs. t-experts	e4 vs. t-expert
no. cases	40	53	14	13	13	13
Kappa	0.28 (0.61)	0.32 (0.52)	0.31 (0.51)	0.25 (0.49)	0.25 (0.43)	0.42 (0.54)
p_o	0.43	0.42	0.43	0.38	0.31	0.54

A further difference between the non-treating experts and the treating experts is that, although they are the same people, the non-treating experts more often advise the patient to wait for a period of time (13/53, as opposed to 2/53 for the treating experts). This may be due to the fact that the non-treating experts have to analyse the patients from paper and sometimes there may not be sufficient information in the patient files to establish a final treatment, whereas the treating experts do have this information but for instance did not include the results of certain examinations in the patient files. An exception is expert e4 who does not advise 'conservative treatment for the time being' as often as the other non-treating experts.

4.3.1.7. Directly comparing treatment plans obtained from PLEXUS to the treatments suggested by non-treating experts

Table 4.13 shows the results of the direct comparison between PLEXUS and the non-treating experts. Fleiss (1981) provides a method for comparing Kappa values. When comparing the Kappa value of PLEXUS versus non-treating experts and PLEXUS versus treating experts no significant difference could be found, probably due to the limited amount of data. Although no significant difference was found, the measured Kappa for PLEXUS versus the non-treating experts is higher than the agreement value obtained when comparing PLEXUS to the treating experts. On the basis of the results which were described in the previous section, this would have been expected, since both PLEXUS and the non-treating experts suggested the procedure 'graft and transfer' more often than the treating experts. This may be due to the fact that one will more readily suggest more complicated and uncertain operative procedures (such as nerve transfers) than would actually be undertaken.

Table 4.13. Comparing PLEXUS to non-treating experts. The Kappa coefficient of agreement and the proportion of agreement on treatment (p_o) have been calculated for PLEXUS relative to all non-treating experts combined (nt-experts), and to individual non-treating experts (e1, e2, e3, e4). The standard deviation of Kappa is shown in the brackets.

agreement	PLEXUS vs. nt-experts	PLEXUS vs. e1	PLEXUS vs. e2	PLEXUS vs. e3	PLEXUS vs. e4
no. cases	53	14	13	13	13
Kappa	0.45 (0.58)	0.48 (0.54)	0.39 (0.64)	0.43 (0.53)	0.42 (0.53)
p_o	0.57	0.57	0.54	0.54	0.62

4.3.1.8. Inter- and intra-expert agreement on treatment plans

The inter-expert agreement indicates whether the non-treating experts agree with the other non-treating experts on treatment planning. This was determined by asking each expert to indicate a treatment plan for 3 groups of 4 patients, where each group was also sent to one other non-treating expert for analysis. None of these experts has actually seen the patients. The results are shown in Table 4.14. When comparing the results obtained by e1 to those obtained by e4 no significant difference can be calculated although the measured Kappa value for e4 is lower. The low value obtained by e4 may have been caused by the fact that when the others indicated that they would wait for a period of time, this expert had already chosen a definite treatment. This was also found when comparing e4 to the treating experts in Section 4.3.1.6.

Table 4.14. Inter-expert agreement of non-treating experts with the other non-treating experts. The Kappa coefficient of agreement and the proportion of agreement on treatment (p_o) have been calculated for the individual non-treating experts (e1, e2, e3, e4) relative to all other non-treating experts (nt-experts). The standard deviation of Kappa is shown in the brackets.

agreement	e1 vs. nt-experts	e2 vs. nt-experts	e3 vs. nt-experts	e4 vs. nt-experts
no. cases	10	9	10	11
Kappa	0.52 (0.54)	0.43 (0.53)	0.49 (0.58)	0.17 (0.40)
p_o	0.60	0.56	0.60	0.27

The intra-expert agreement shows whether the experts agree with their own actual treatment. This was measured by asking each expert to indicate a treatment plan for 3 of their own patients. The experts did not know they would be analysing their own patients. The patient names were removed and the files were randomly mixed with the other patients who had to be analysed. For all experts added together the value of Kappa amounts to 0.27 (s.d. 0.47), and the proportion of agreement p_o equals 0.40.

These measured results are similar to the results obtained when comparing the non-treating experts to the treating experts (Table 4.12). However, there were only 10 patients which could be used for calculating the intra-expert agreement. The differences between the values of the inter- and intra-expert agreement, although not found to be significant, again indicate that there is a discrepancy

Scoring sheet for diagnosis

Diagnosis no. 1

rupture C5 rupture C6 (possibly avulsion) avulsion C7,C8,T1

++	+	o	-	--

Diagnosis no. 2

rupture C5 (possibly in continuity) rupture C6 avulsion C7,C8,T1

++	+	o	-	--

Diagnosis no. 3

rupture C5,C6 avulsion C7,C8,T1

++	+	o	-	--

ranking of the diagnoses:

rank 1	rank 2	rank 3	rank 4

computer diagnostic advice is number:

Figure 4.6. Sample scoring sheet.

between seeing the actual patient and forming an opinion from paper, and that there is an effect caused by missing information in the case notes. These issues were previously discussed in Section 4.3.1.4.

4.3.2. RESULTS OBTAINED IN THE BLIND EVALUATION

In the final (third) round of the evaluation each of the four experts was sent 20 patient files which had to be analysed. Each patient file contained a number of diagnoses and treatment plans which all had to be judged. The diagnoses and treatment plans originated from the treating expert (first round), and from the non-treating experts and PLEXUS (second round). The origin of the diagnoses and treatments were not shown to the judging expert, and the opinions were placed in random order. The experts were asked to answer a number of questions on separate scoring sheets which were attached to the case notes. A sample scoring sheet is shown in Figure 4.6. The scoring sheets consisted of three sections:

- Each diagnosis and each treatment plan had to be judged by the expert on a five point scale. The results of this part of the study are analysed in Section 4.3.2.1.
- The experts were asked to rank all the diagnoses and treatment plans which were shown on the scoring sheet (Figure 4.6). Since either three or four opinions were attached to each patient file, the scoring sheet contained the possibility to rank the maximum number of four answers. The results of the ranking procedure will be analysed in Section 4.3.2.2.
- The participants were asked to indicate which of the opinions they thought originated from the computer. The results of this part of the investigation are discussed in Section 4.3.2.4.

4.3.2.1. Ratings obtained by treating experts, non-treating experts and PLEXUS

In the third round, each file contained either three or four different diagnoses and treatment plans originating from the treating expert, the knowledge based system and one or two non-treating experts. The four judging experts were asked to judge all diagnoses and treatment plans which were attached to the patient files on a five point scale. The judgement categories were indicated with the symbols (++,+,o,-, --), and in the accompanying letter it was stated that this ranged from very good to very poor. No definitions as to the categories of the five point scale were given. It was suspected that, since the experts came from four different countries, if category definitions were given in the form of text, the interpretation in the different countries would have a greater unwanted effect on the outcome than if the categories were indicated with the symbols.

Table 4.15. Scoring results for diagnosis. The table shows the total score and average score per case received by PLEXUS, the treating experts (t-experts), non-treating experts combined (nt-experts), and individual non-treating experts (e1, e2, e3, e4). For each case the scores could range from -2 to +2. The standard deviations of the values are shown in brackets. The scores provided by only three of the four judges could be used.

scores	PLEXUS	t-experts	nt-experts	e1	e2	e3	e4
no. cases	37	37	37	8	11	9	9
tot. score	31	35	39.5	6	6.5	13	14
mean	0.84 (1.01)	0.95 (0.94)	1.07 (0.97)	0.75 (1.04)	0.59 (1.16)	1.44 (0.53)	1.56 (0.73)

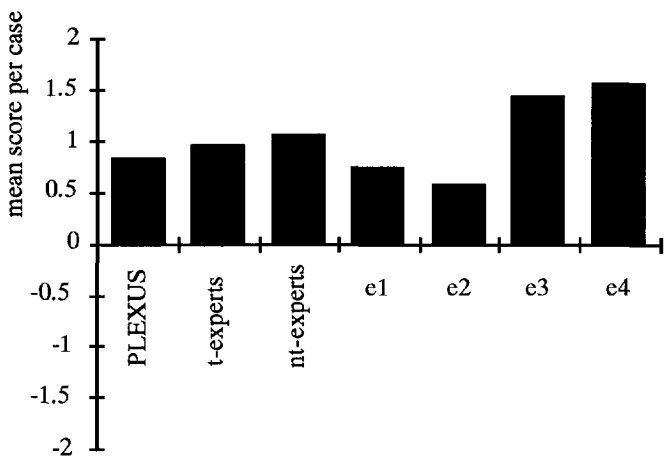


Figure 4.7. Mean score achieved for diagnosis. The mean score per case received by PLEXUS, the treating experts (t-experts), non-treating experts combined (nt-experts), and individual non-treating experts (e1, e2, e3, e4) is shown. For each case the score per case could range from -2 to +2.

Two of the four experts had some difficulty in rating the diagnoses on a five point scale. This had not been expected, since five point scales are very commonly used in investigations. One of the experts only produced complete scores for nine of the fourteen cases. For another expert, the five point scale was relaxed to a three point scale after the difficulties with the five point scale became clear, however this expert finally only indicated the diagnosis which was thought to be best in each case, rather than rating all diagnoses belonging to a certain patient. Two judges rated all diagnoses and treatment plans for fourteen patients as requested.

For calculating means and standard deviations, the ordinal five point scale has been converted to a scale ranging from -2 to +2, where -2 corresponds to (--) and +2 corresponds to (++). However, it is questionable whether the intervals of the scale (++,+,o,-,--) are equal, therefore the Wilcoxon-Mann-Whitney test has been used for hypothesis testing. This test is nonparametric and the magnitude of the intervals is not taken into account.

The total and mean scores received by the treating expert, non-treating experts and PLEXUS have been calculated. Table 4.15 shows the scores which are obtained by each of the participants of the evaluation for their diagnoses. The standard deviations are shown in the brackets. The mean scores are also shown in Figure 4.7. No significant difference was found when comparing the scores achieved by PLEXUS to the scores achieved by the treating experts and by the non-treating experts.

For Table 4.15 all raw scores were added and the mean was calculated. However, different judges may score in different ways. For example, some judges may be stricter than others. Therefore, all diagnostic scores were also normalised. This was done by subtracting the mean score given by the judge from the score given by that judge, and dividing this by the standard deviation of scores given by the judge. When hypothesis tests were performed on these normalised diagnostic scores, the results were the same as those described above.

Although the measured results obtained by PLEXUS, the treating experts and the non-treating experts do differ, this difference was not found to be significant. This may be due to a lack of test cases, which reduces the detectability of differences. A possible difference in means could to a certain extent probably be explained by the number of false positives given by the system.

In the direct comparison (Section 4.3.1) it was found that the system gave a larger number of false positive answers than the experts. It can now be determined whether the false positive answers influenced the score obtained by PLEXUS in the third round. The average number of false positive answers given by the system (when directly comparing system to the treating expert) over all the cases used in this round equals 2.4. If the average number of false positive answers is calculated only over those cases which received a score of 0 or less in

Table 4.16. Scoring results for treatment. The table shows total score and average score per case received by PLEXUS, the treating experts (t-experts), non-treating experts combined (nt-experts), and individual non-treating experts (e1, e2, e3, e4). For each case the scores could range from -2 to +2. The standard deviations of the values are shown in brackets. The scores provided by only two of the four judges could be used.

scores	PLEXUS	t-experts	nt-experts	e1	e2	e3	e4
no. cases	29	29	28	5	9	9	5
tot. score	40	33	21	0	13	0	8
mean	1.38 (0.78)	1.14 (0.69)	0.75 (1.35)	0 (1.22)	1.44 (0.53)	0 (1.73)	1.60 (0.55)

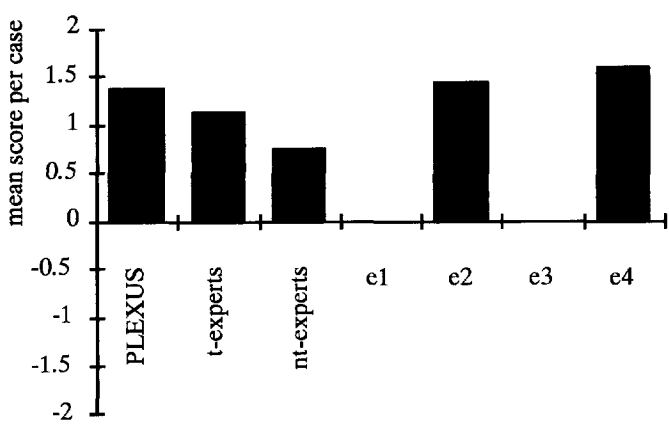


Figure 4.8. Mean score achieved for treatment. The mean score per case received by PLEXUS, the treating experts (t-experts), non-treating experts combined (nt-experts), and individual non-treating experts (e1, e2, e3, e4) is shown. For each case the score per case could range from -2 to +2.

this round, this amounts to 3.1. The average modified predictive value, which is also an indication of the false positives, drops from 0.5 on all cases which were judged, to 0.3 on the lower scoring cases. This difference was not found to be significant at the 5% level, however two-tailed $p < 0.1$. This difference is mainly due to the scoring of one of the experts. However, these results do indicate that the false positives given by the system should be reduced.

The scores which were given for the treatment plans are shown in Table 4.16 and the mean scores are also shown graphically in Figure 4.8.

4.3.2.2 Rankings received by treating experts, non-treating experts and PLEXUS

In the third round, the judges were also asked to rank the diagnoses and treatment plans attached to each patient file. Since either three or four opinions were attached to a file, the scoring sheet (Figure 4.6) contained the possibility to rank the maximum number of four answers. In the accompanying letter the judges were told that it was possible to use the same rank for more than one opinion. The results were analysed by counting the number of times an opinion was ranked highest of all opinions (or tied at that rank), the number of times an opinion was ranked lowest of all opinions (or tied at that rank), and the number of times an opinion was neither ranked highest nor lowest, in which case this will be termed 'middle'. Only three of the four judges provided complete rankings, therefore the ranking results are based on rankings provided by three judges. The results of this procedure are shown in Table 4.17 and Table 4.18. Using the Wilcoxon-Mann-Whitney test, no significant difference was found between the numbers of times the opinions originating from PLEXUS are ranked at a certain position, and the numbers of times the opinions belonging to the experts are ranked at a certain position.

Table 4.17. Rankings received by treating experts, non-treating experts and PLEXUS for diagnosis. The number of high, middle and low rankings were calculated for PLEXUS, the treating experts (t-experts) and the non-treating experts (nt-experts) and are presented as a percentage of the total number of cases. The absolute numbers of high, middle and low rankings are shown in brackets.

ranking diagnosis	PLEXUS	t-experts	nt-experts
no. cases	42	42	60
% high	40 (17)	45 (19)	45 (27)
% middle	24 (10)	10 (4)	18 (11)
% low	36 (15)	45 (19)	37 (22)

Table 4.18. Rankings received by treating experts, non-treating experts and PLEXUS for treatment planning. The number of high, middle and low rankings were calculated for PLEXUS, the treating experts (t-experts) and the non-treating experts (nt-experts) and are presented as a percentage of the total number of cases. The absolute numbers of high, middle and low rankings are shown in brackets.

ranking treatment	PLEXUS	t-experts	nt-experts
no. cases	38	38	54
% high	53 (20)	45 (17)	41 (22)
% middle	8 (3)	24 (9)	11 (6)
% low	39 (15)	31 (12)	48 (26)

Table 4.19. Interexpert agreement in judging the diagnosis results. The mean proportion of agreement in ranking per case is shown for each judging expert compared to the other judging experts. The standard deviation is shown in brackets.

	j-e1 vs. other judges	j-e3 vs. other judges	j-e4 vs. other judges
no. cases	8	8	8
mean proportion agreement per case	0.70 (0.31)	0.54 (0.36)	0.51 (0.33)

4.3.2.3. Inter- and intra-expert variability after the third round of the evaluation

In order to investigate the inter-expert variability in judging (i.e. whether the different judges rank the same cases in the same way), each expert was asked to judge 3 groups of 4 patients, and each group of four was also sent to one other judge for analysis. The agreement in ranking was investigated. Since adequate ranking results were received from three of the four judges, each judge could only be compared to the others on 8 different cases, rather than on 12 cases (3 groups of 4) as was originally intended. It was evident that these numbers would be too small to analyse the agreement properly. However, it would not have been feasible to have asked the judges to analyse more cases. The mean proportion of agreement in ranking (high, middle, low) per case was calculated. The mean proportion of agreement in judging the diagnoses varies from 0.7 for judge-e1 compared to the others, to 0.5 for judge-e4 compared to the others. However these values could only be calculated over 8 cases. The results are shown in Table 4.19. As was also established when investigating inter-expert agreement in diagnosis and treatment planning (Section 4.3.1.4 and Section 4.3.1.8), there is a substantial degree of inter-expert variability.

The intra-expert variability was also investigated. This was done in two ways:
- By having the experts judge a number of patients which originated from these same experts (i.e. their own patients which they sent in during the first round of the evaluation)
- By having the experts judge a number of the patients which these same experts diagnosed during the second round of the evaluation.

The experts did not know that they were judging their own patients and their own diagnoses and treatment plans.

The patient names were removed and the patients were randomly mixed with the other patients which had to be analysed. The experts' judgement regarding their own diagnoses and treatment plans may be determined. The results are a measure of the intra-expert agreement, and are shown in Table 4.20 and Table 4.21. These results have to be interpreted with great care, since they are only based on a limited amount of data. The numbers in the cells are the mean scores (on a scale from -2 to 2) given by the judges.

Ideally a physician will agree with his own diagnosis and score this highest. The results are similar to what was found on intra-expert variability after the second round (Section 4.3.1.4). The judges gave higher scores to their own diagnoses from paper (column 5 of Table 4.21) than to their own actual diagnoses when they saw the patients (column 4 of Table 4.20), although the difference is not found to be significant. A possible difference can be explained by the fact that the judging experts had to judge the diagnoses and treatment

Table 4.20. Experts scoring on the patients which also originate from these experts. In the third round, the judges were sent a number of cases which these judges originally submitted in the first round. For these cases, scores were given to PLEXUS, to non-treating experts and to the treating expert. In this situation the treating expert is in fact the same person as the judge. The mean score per case given to the various opinions are shown in the table, the standard deviations are given in brackets.

	no. cases	opinion of PLEXUS	own opinion (t-experts)	other opinions (nt-experts)
diagnosis	7	0.29 (1.11)	1.14 (0.69)	1.29 (0.76)
treatment	6	1.57 (0.54)	1.29 (0.76)	0.29 (1.50)

Table 4.21. Experts scoring on the patients they diagnosed in the second round. In the third round the judges were sent a number of cases which they had also diagnosed in the second round. For these cases, scores were given to PLEXUS, to the treating experts and to the non-treating expert. In this situation the non-treating expert is in fact the same person as the judge. The mean score per case given to the various opinions are shown in the table, the standard deviations are given in brackets.

	no. cases	opinion of PLEXUS	other opinion (t-expert)	own opinion (nt-expert)
diagnosis	9	1.06 (0.73)	1.00 (0.71)	1.67 (0.50)
treatment	6	1.50 (0.55)	1.17 (0.75)	1.50 (0.55)

plans from paper and did not actually see the patients. As was stated before, great care has to be taken when interpreting the results, as the number of patients is very small.

4.3.2.4. Identifying the recommendations originating from the computer

The final measure was a kind of Turing test. The experts were asked to identify which of the opinions they thought originated from the computer. All judges answered this question for all of the 14 patients they were sent. The results were analysed to see whether the number of times PLEXUS was correctly identified deviated significantly from what would be expected to occur by chance.

The results of this procedure can be seen in Table 4.22 (diagnosis) in Table 4.23 (treatment). The third column of these tables shows the number of times the judges correctly identified the opinion originating from PLEXUS. These numbers are not integers as might have been expected, since in some cases the judges indicated more than one answer. If, for instance, a judge gave two answers one of which actually originated from PLEXUS, a score of 1/2 was given to the number of times PLEXUS was correctly identified. If PLEXUS was correctly identified and no other answers were provided, a score of 1 was given.

The fourth column shows the number of times PLEXUS would have been identified correctly purely by chance. For some patients the experts could choose between 3 different diagnoses, which meant that these contributed an expected score of 1/3 (as in multiple choice questions). Other cases consisted of four diagnoses, which meant that these contributed an expected score of 1/4. By adding these scores for all patients seen by a particular judge, the values in the fourth column were obtained.

To investigate whether the mean number of times PLEXUS was correctly identified differs significantly from what would be expected by chance, the Wilcoxon matched-pairs signed-ranks test was used. When all judges were combined, no significant deviation from chance was found. On the diagnoses, judge j-e3 did however mention PLEXUS more often than would be expected by chance ($p < 0.05$). Judge j-e4 identified the treatments provided by PLEXUS less often than would be expected to occur by chance ($p < 0.01$).

On the whole, the results show that it is not possible to distinguish the answers given by PLEXUS from the answers given by the experts. It must be stated, however, that the researcher introduced a uniform terminology which may have had a positive influence on the results. In order to be able to perform a fair judgement a uniform terminology is necessary.

Table 4.22. Identification of computer diagnosis. The results are shown for all judging experts combined (j-experts) and for each judging expert individually (j-e1, j-e2, j-e3, j-e4). The third column shows the total number of times PLEXUS was correctly identified. The fourth column shows the total number of times this was expected to occur by chance.

judge	no. cases	no. times PLEXUS identified	no. times expected by chance
j-experts	56	21.8	16.9
j-e1	14	3.9	3.9
j-e2	14	4.0	4.5
j-e3	14	8.0	4.3
j-e4	14	5.8	4.3

Table 4.23. Identification of computer treatment plan. The results are shown for all judging experts combined (j-experts) and for each judging expert individually (j-e1, j-e2, j-e3, j-e4). The third column shows the total number of times PLEXUS was correctly identified. The fourth column shows the total number of times this was expected to occur by chance.

judge	no. cases	no. times PLEXUS identified	no. times expected by chance
j-experts	53	18.7	16.3
j-e1	14	5.2	4.0
j-e2	14	6.0	4.7
j-e3	14	7	4.4
j-e4	11	0.5	3.3

One interesting feature was present in the judgements provided by judge j-e2, who indicated a lack of confidence in the computer program when he was visited by the researcher, even though he had never actually seen the program. The expert thinks that the computer program is not good. However, when asked to identify the computer's answer, the expert always wrote down the number of the answer which, on the top section of the form (Figure 4.6), he had indicated as being the best answer. Thus, contradictory to his opinion about the computer, he indicated that the best answer originated from the computer. Therefore, he would have expected to have identified the computer less often than would be expected by chance. However, this does not occur, as can also be seen in the rows indicating 'j-e2' in Table 4.22 and 4.23.

4.3.3.COMPARISON OF THE RESULTS OBTAINED AFTER THE SECOND AND THE THIRD ROUND OF THE EVALUATION

From the results it becomes clear that even if the domain and system are such that non-exclusive multiple answers are given, the results after the second round (direct comparison) can be analysed and provide additional information, which would not have been available if the results had only been analysed after the final round of the evaluation.

It is very difficult to perform a direct comparison in domains where one case consists of multiple answers. All diagnoses and treatments were only available in the form of free text. Therefore, all cases had to be converted into a uniform terminology. After all cases had been converted into a uniform terminology, all answers were placed into categories, so that it could be determined whether the answers were correct or incorrect, i.e. whether answers provided by the system and by experts belonged to the same category or to different categories.

For a number of answers the choice of category presented some problems. This occurred when one answer consisted of more than one possibility. For instance, when an answer was given for which the expert also indicated another possibility. An expert could have stated that a patient had an 'avulsion C5 or possibly extraforaminal lesion C5'. In these cases, the expert's first choice was the 'avulsion C5'. Therefore, for purposes of comparison the answer was placed in category 'avulsion C5'.

These difficulties arise when performing a direct comparison of the results. When a blind judgement round is used, there is no need to use such definite categories, and the full answers were given to the judging experts, albeit in a uniform terminology.

Another problem in the direct comparison was that the comparison consisted of determining whether certain answers were exactly identical or not, although some answers will obviously be more correct than other answers. However, for people who are not expert in the domain this is difficult to decide on. Furthermore, if differentiation is made as to the correctness of certain answers, a scoring scheme will be necessary, which will introduce subjectivity into the measurement.

Although it is difficult to determine the correctness and incorrectness of the answers, some interesting results do arise from the direct comparison. It could be concluded from the direct comparison that the number of false positive answers provided by PLEXUS is relatively high.

Since the standard is not an actual gold standard, the false positive answers do not necessarily have to be wrong. The final round can show whether the false positives are relevant answers or whether they are not. It may be investigated if the system gets a lower overall score due to the false positives which it gives. In the evaluation of PLEXUS the results from the final round showed that the average number of false positives in the cases which received a low score was higher than the average number of false positives over all cases, although this was only apparent in the results obtained from one of the judges.

The blind evaluation also has a number of disadvantages. A blind evaluation is a subjective measure and it is difficult to define an appropriate scoring scale. The five point scale used in the evaluation of PLEXUS did not prove to be suitable for all experts. This would not have been expected, since five point scales are very commonly used in investigations.

In conclusion, it is worthwhile to investigate the multiple answers after the second round because additional information may be obtained from these calculations. It is recommended for systems with non-exclusive multiple answers to calculate the results after both the second and the third round of the evaluation. In this study, the other results, such as treatment results and intra- and inter-expert agreement essentially show similar patterns in both rounds.

4.4. Bias and confounding

A number of different sources of bias and confounding which may threaten the validity of laboratory evaluations has been mentioned in Section 2.3.4. These include pro- and anti-computer bias, coding, circularity and parochial bias (Chandrasekaran, 1983; Wyatt and Spiegelhalter, 1990). Pro- and anti-computer bias and circularity have been avoided in this study by performing a blind evaluation and by involving independent experts in the investigation. Possible threats to the validity of this evaluation study may come from, for instance, coding and parochial bias.

Parochial bias. This arises when the test cases are not representative for the complete population. PLEXUS has been designed using knowledge and cases which originate from domain experts, and the system has now been evaluated using cases which originate from other (independent) domain experts. The patients who are treated by domain experts are usually the more severe cases. This means that the test cases used in this study are not representative for the target population. Since PLEXUS is meant for neurologists and neurosurgeons with limited experience in the field of brachial plexus injuries who probably see less severe cases, this will limit the generality of the evaluation results. By looking at the test cases used in the clinical evaluation which will be discussed in Chapter 5, it may be concluded that the cases involved in the field evaluation were indeed less severely injured. However, in order to perform the laboratory evaluation a considerable number of cases were needed. This left no choice but to obtain the test cases from experts in the field of brachial plexus injuries.

Coding. A further influence on the results of this evaluation may have been caused by the various kinds of coding which were necessary. Firstly, for three of the experts the diagnosis had to be transcribed from the experts' own case notes onto a special data entry form. This procedure could introduce subjectivity.

There is also the coding which was mentioned by Chandrasekaran (1983). In order to perform a blinded evaluation, a uniform terminology has to be used. This may, however, also introduce subjectivity into the measurement, which may bias in favour of the system.

Another aspect of importance to the results of the study is the choice of categories. For the treatment, for instance, the experts often mention the precise surgical procedure in the case notes, but the system can only determine the general surgical procedures. Therefore, all operations have been abstracted to more general procedures. For the direct comparison it was also necessary to use categories for the diagnosis in order to compare the diagnoses. The results will depend on the choice of categories. When broader categories are chosen, more

answers will be correct. Therefore it is always necessary to compare the results to other physicians and not only to a standard.

The patient data which were given to the experts consisted mostly of the data which the expert system needs (with some additions where it was thought that these were important). This could imply that if the non-treating experts had seen the patients they may have produced recommendations which would have more closely resembled the treating experts' opinions. It can, for example, be seen that the inter-expert agreement is higher than the intra-expert agreement.

Ideally, the performance should be compared to the diagnoses of physicians who have seen the patient, but who have not yet operated on the patient, because in the actual situation when PLEXUS would be used the physicians who are assisted by the knowledge based system would use it in this way.

However, in this study the diagnoses of experts who did not see the patients and the diagnoses of physicians who have actually treated the patients are available, and the diagnoses of physicians who have seen but not yet treated the patients are not available. As was described above, the non-treating experts are probably influenced by the fact that they did not see the patients, thus in the ideal situation the physicians involved in the evaluation should see the patients.

4.5. Conclusions and recommendations

With regard to the evaluation of PLEXUS, conclusions may be drawn concerning the evaluation setup, the way in which the results have been analysed, and the results of the evaluation study. The conclusions drawn for this specific evaluation study will lead to general recommendations concerning laboratory evaluations of medical knowledge based systems.

Setup. The problem solving performance of PLEXUS was investigated in an evaluation involving four international experts in the domain of brachial plexus injuries. The evaluation consisted of three rounds. In the first round the experts were asked to provide test cases which could be used for evaluation. An analysis of the test cases showed that the patient data which were sent in for the evaluation were more severely injured than patients who would be encountered by potential users of the knowledge based system. Unfortunately, there was no possibility of obtaining test cases from the target situation.

The number of test cases available in this evaluation amounted to forty. The number of test cases involved in the evaluation was limited for practical reasons, as it was suspected that the experts would not cooperate if they were asked to supply more data. An analysis of the number of cases involved showed

that it would be difficult to statistically detect small differences between system and experts with this number of cases.

In the second round, each expert was asked for his diagnoses and treatment plans for a number of the test cases. The test cases were also entered into the knowledge based system and the computer's opinion was determined. After the second round, the system's advice could be directly compared to the opinions of the experts who provided the cases and to the experts who saw the cases in the second round. In the third round, each expert was sent a number of the cases, only this time the opinions of the treating experts, non-treating experts and PLEXUS were attached to the case notes. The origin of the opinions was not known to the experts and they were asked to blindly judge all the opinions.

This setup allowed both a direct comparison of the opinions after the second round of the evaluation, and a blind judgement of the opinions in the third round. The direct comparison in the second round allows a more in depth analysis of the opinions given by the knowledge based system and provides information about the areas in which the system performs well and the areas that require more attention. The third round allows a more subjective view of complete cases and gives information as to whether possible problem areas which were found in the second round are also perceived as limitations by the experts.

A difficulty in the second round was the choice of categories. All answers had to be placed in categories. Answers were classified as correct if they belonged to the same category otherwise they were classified as incorrect. The choice of categories will have had an influence on the evaluation results, as was demonstrated when the category limits were relaxed.

During the third round the experts were asked to blindly judge the opinions provided by other experts and by the knowledge based system. In order to do this, a five point scale was introduced. However, some experts had difficulties in using this five point scale for classifying the opinions. Furthermore, the experts were also asked to rank all the opinions. Finally, they had to identify the opinions they thought originated from the computer.

Analysis. The results of the direct comparison were calculated using a number of different metrics for performance calculation. The fact that in this domain one diagnosis consists of multiple answers largely determines the mathematical methods which may be used. The models for performance calculation which were used were:

- the case correctness model,
- the partial correctness model,
- the modified partial correctness model,
- positive negative correctness model,

• the diagnostic performance model.

The first method indicates whether a case is completely correct, or whether it is not. In situations with multiple answers it may be very difficult to obtain a completely correct case. This method is therefore less useful in these situations.

The partial correctness model, modified partial correctness model and the positive-negative correctness model all give an idea of whether the system can identify locations which are actually injured, and whether the system over- or under-estimates the extent of the injury. The modified partial correctness model seems to be a good model to use as a performance measure, since this gives equal weight to all test cases and the standard deviation can easily be calculated.

The diagnostic performance model is a useful model for calculating system performance for different classes of answers. It can be used to identify whether there are particular classes for which performance is lower than for other classes. This can indicate parts of the system which need to be updated.

For the direct comparison of the treatment plans, the Kappa coefficient of agreement was used. Since there is only a limited number of possible treatment categories, this coefficient allows calculation of a chance corrected agreement. However, there is some discussion as to the chance correction which is used. Therefore, the proportion of total agreement (which is the same as the case correctness discussed above) also has to be mentioned.

In the third round of the evaluation the experts had to blindly judge opinions on a five point scale. The results were analysed to determined whether there were significant differences in the scores received by the knowledge based system, the treating experts and the non-treating experts.

The judges were also asked to identify which of the opinions they thought originated from the computer. The results were analysed to investigate whether the number of times PLEXUS was correctly identified deviated significantly from what would be expected to occur by chance. This proved to be an interesting analysis.

Results. This can be divided into three different parts:
• results of the direct comparison,
• results of the blind judgement,
• inter- and intra-expert agreement.

Results obtained after the second round of the evaluation, i.e. the direct comparison, shows that the mean number of true positive answers per case given by PLEXUS cannot be shown to differ significantly from the number of true positive answers given by the non-treating experts. This is also true for the mean number of false negative answers. However, the number of false positive

answers per case provided by the system is significantly higher than the number of false positives given by the non-treating experts when combining all non-treating experts. A similar trend can be seen for the modified predictive values, although the difference here was not found to be significant The modified accuracy of the system cannot be shown to differ significantly from that of the non-treating experts.

For a number of cases, a possible explanation for the low predictive value and large number of false positive answers suggested by PLEXUS, is that the system tries to explain all of the dysfunction which is present in the patient, thereby increasing the number of false positive answers when compared to the treating experts.

In the blind evaluation, where the experts where asked to score the opinions on a five point scale, no significant difference was found between the diagnostic results obtained by PLEXUS, the treating experts and the non-treating experts. However, this may be explained by the fact that with this number of test cases a relatively small difference in actual means cannot be detected. It can be seen that the sample means are different. A possible difference in means could to a certain extent probably be explained by the number of false positives given by the system.

In the direct comparison it was found that the system produced a larger number of false positive answers than the experts. It was determined whether the false positive answers influenced the score obtained by PLEXUS. The average number of false positive answers given by the system (when directly comparing system to the treating expert) over all the cases in the third round is lower than the average number of false positive answers calculated only over those cases which received a score of 0 or less in this round. The average modified predictive value, which is also an indication of the number of false positives, calculated over all cases is different ($p < 0.1$) from the modified predictive value calculated only over the lower scoring cases. This difference is mainly due to the scoring of one of the experts. However, these results do indicate that the false positives given by the system should be reduced.

The areas in which these false positives mostly manifest themselves have been identified by looking at the diagnoses and treatment plans which were obtained after the second round. The knowledge based system PLEXUS should be updated accordingly.

In the third round, the experts were also asked to indicate which of the opinions originated from the computer. When combining all judges, no significant deviation was found from the number of times the computer would be identified by chance. The results show that it is not possible to distinguish the answers given by PLEXUS from the answers given by the experts. It must be stated,

however, that the researcher introduced a uniform terminology which may have had a positive influence on the results. In order to be able to perform a fair judgement a uniform terminology is necessary.

The intra- and inter-expert variability appear to be considerable in this domain. From the literature, it is known that there is no general consensus on treatment planning of brachial plexus injuries. However, the low agreement values found may also be partly explained by the fact that the direct comparison was performed in the strictest sense.

The modified accuracies and modified predictive values of the intra-expert agreement over all non-treating experts are similar to the modified accuracies and modified predictive values which were obtained when comparing the non-treating experts to the treating experts. The intra-expert agreement is lower than the inter-expert agreement. This supports the notion that there are differences due to the fact that the non-treating experts did not see and did not operate on the patients. If the non-treating experts were asked to diagnose the patients from paper again and the intra-expert agreement was calculated, the intra-expert results could be expected to be similar to the inter-expert agreement.

It may be concluded that the accuracy of the recommendations provided by PLEXUS is of expert level. The system does, however, produce a higher number of false positive answers. The system should be amended in order reduce the number of false positive answers. The evaluation method which has been used is largely satisfactory, although the number of test cases was limited. In some cases this limited number of test cases probably prevented significant conclusions from being drawn. A further limitation of this evaluation is the fact that no potential users were involved in the study.

General recommendations. For systems which give multiple exclusive answers, it is interesting to analyse the results both after the second and the third round of the evaluation. As these may provide complementary information. For PLEXUS, for example, the results obtained in the second round indicated that the number of false positive answers was relatively high, and the results from the third round showed to which extent this effect negatively influenced the system's performance. Asking judges to identify the computer's opinions was also an interesting aspect of this study and it did not place additional time demands on the experts.

This evaluation has taken almost two years to perform. From the literature, it becomes clear that these evaluations are always very time consuming. A possible explanation for this is that these experts are not potential users of the system and therefore often do not gain anything from cooperating in such an evaluation.

Therefore, as was also suggested in Chapter 2, running the system with actual or generated test cases prior to performing a formal laboratory evaluation, should be emphasised more, for this provides a large amount of additional information, and formal evaluations can usually not be performed in many iterations.

Although formal laboratory evaluation is necessary, it should only be carried out after extensive validation has been performed. Be sure that only seriously motivated people are involved in the evaluation, that they understand the time demands which will be placed on them in advance, and that they are prepared to seriously follow through the evaluation until the end. In order to reduce the time necessary to complete a formal laboratory evaluation, it is recommended to try to embed certain stages of the evaluation in workshops or regular conferences.

As was demonstrated, the laboratory evaluation of a knowledge based system which is designed according to the expert system paradigm is not easy. It is always fraught with difficulties such as a lack of test cases, lack of a gold standard and difficulty in judging answers. This is caused by the complexity of the problems which such systems aim to solve. As was previously stated in Chapter 2, and will be regarded from a different point of view in Chapters 5 and 6, it may be advisable to design somewhat less ambitious systems which, rather than aiming at solving the complete problem of diagnosis or treatment planning, are used in selected areas in which they cooperate with the user to solve the problem. This could possibly also reduce some of the problems which exist in knowledge based system validation.

5 Clinical evaluation of the medical knowledge based system PLEXUS

The medical knowledge based system PLEXUS was evaluated clinically in four different hospitals in The Netherlands. The performance of the human-machine system was studied, and the usability and acceptability of the system were addressed. Since the incidence rate of brachial plexus injuries is low, only qualitative results arose from the study. The results show that the performance of the knowledge based system in the hospitals is good, although a number of improvements is still necessary. The number of false positive answers given by the system is relatively high, as was also found in the previous chapter. Furthermore, in some cases the patient data were not as complete as was expected during the development of the system. This may cause the system to give an erroneous answer in cases where, due to a lack of data, it should not have suggested an answer at all. The usability of the user interface was investigated by means of videotaping actual interactive sessions during the field evaluation. This provided important information which may be used to update the user interface, so that the system satisfies a number of essential usability requirements. The acceptability of the system was studied by means of a brief questionnaire which was distributed among the cooperating physicians. The results are not conclusive, as a number of physicians indicated that they would use the system if it was generally available, whereas during the field evaluation the system was not used as readily as might have been expected.

5.1. Introduction

Only very few knowledge based systems have undergone a clinical evaluation. Some exceptions have been described by Adams *et al.* (1986), Bankowitz *et al.* (1989), Murray (1990), and Sutton (1989a). The limited number of reported evaluation studies may be partly explained by the fact that before a formal clinical evaluation can be performed, the system must have been verified (Nguyen *et al.*, 1987; Ginsberg, 1988; Preece and Shinghal, 1992) and validated (Shwe *et al.*, 1989). Furthermore, a formal laboratory evaluation must have demonstrated adequate performance, safety, potential usefulness and satisfactory human-machine interaction. In addition, if necessary the system can run in parallel with the normal situation in the background for a period of time prior to a formal clinical evaluation. Thus, a knowledge based system must have reached an advanced level of development before a clinical evaluation can be carried out. Clinical evaluation comprises many aspects, such as: acceptance, usability, safety, influence on patient care, legal and ethical aspects and cost-benefit analysis.

This chapter concerns the clinical evaluation of the knowledge based system PLEXUS. Two aspects were investigated:
- the performance of the human-machine system in the clinical environment,
- the usability and acceptance of the knowledge based system.

For the purpose of this evaluation, four computers were placed in four different hospitals in The Netherlands for the period of a year and a half. The physicians were asked to use PLEXUS for all traumatic brachial plexus patients who visited them during the evaluation period. The investigation of the performance of the human-machine system will be described in Section 5.2. This study was carried out according to the framework for evaluation design which was described in Chapter 2. The goal of this part of the evaluation was to investigate whether the knowledge based system PLEXUS does indeed have the capacity to assist physicians in the diagnosis and treatment planning of brachial plexus injuries. The evaluation setup is described in Section 5.2.1 and is discussed along the lines which are shown in Figure 2.1. Some problems and limiting factors which were encountered during the evaluation are mentioned in Section 5.2.4. The results of the performance study are described in Section 5.2.5.

The studies of the usability and acceptance of the knowledge based system will be described in Section 5.3. The usability was studied by means of videotapes of interactive sessions with PLEXUS. A brief questionnaire was distributed at the end of the evaluation period in order to study the acceptability of the system. The conclusions which may be drawn from the clinical evaluation of PLEXUS are summarised in Section 5.4.

5.2. Clinical performance evaluation of PLEXUS

5.2.1. GOAL OF THE PERFORMANCE EVALUATION STUDY

Most medical knowledge based systems are aimed at improving patient care by providing assistance in certain tasks. To investigate whether this objective is achieved, it is necessary to study whether there is a difference in final patient outcome between the unassisted and assisted situations. However, for PLEXUS the aim of the evaluation has been adapted, since final patient outcome in brachial plexus injuries is usually only known after several years and there is a large variability in treatment results. In this evaluation, the goal was to compare the diagnoses and treatment plans determined by physicians who use the knowledge based system to the diagnoses and treatment plans provided by physicians who do not use the knowledge based system.

5.2.2. EVALUATION SETUP

The evaluation setup will be discussed as it was planned prior to the investigation (van Daalen, 1992a). A number of practical limitations which became apparent during the study and which influenced the setup will be discussed afterwards.

The setup involved an evaluation of PLEXUS in five different hospitals in The Netherlands. For all traumatic brachial plexus injuries which were seen by the physicians during the evaluation period, the physicians were asked to enter the patient data into the computer. After entering the data, but before consulting the knowledge based part of PLEXUS, they were asked to enter their own opinion concerning the diagnosis and treatment plan into the computer. The physicians then performed a consultation with the system, after which they were asked to enter their own final opinion. At the end of the evaluation period, the opinions which were entered before and after consultation with PLEXUS should be judged in order to investigate the system's influence on task performance. Table 5.1 shows a schematic representation of the evaluation setup.

Table 5.1. Evaluation setup for the clinical evaluation of PLEXUS

For all traumatic prospective patients:
1) enter patient data into computer 2) enter own diagnosis and treatment plan 3) generate computer advice 4) enter own final opinion

The patient data are entered into the computer by means of the graphical interface which runs on an Apple Macintosh® at the hospitals. The knowledge based part of PLEXUS runs on a SUN® workstation at Delft University of Technology. In this way, it is possible to enter the data and receive advice locally at the hospitals, and to perform the reasoning and to keep track of the evaluation process centrally. The architecture of PLEXUS and the way in which the system works has been discussed extensively in Chapter 3.

5.2.2.1. Selection of test input

The test cases consisted of all prospective brachial plexus patients who were seen by the physicians over the evaluation period. The hospitals involved in the

evaluation were relatively large hospitals which were mainly situated in the western part of the country. This may mean that neither the patients nor the physicians are representative for the complete population. Furthermore, since the number of brachial plexus injuries which takes place each year is very small, the possibilities for quantitative analysis of the results are limited. Some of the limitations thus imposed on the investigation will be discussed in Section 5.2.4.

5.2.2.2. Consultation

Specifying who uses the system (human-machine system). A request was sent to six neurologists who worked in relatively large hospitals in The Netherlands, to ask whether they would be willing to participate in the evaluation of the knowledge based system. Of these neurologists, two referred to neurophysiologist-neurologists at the same hospital and one referred to a neurosurgeon who would be willing to cooperate. Since five computers were available, the first five who responded to the request finally participated in the evaluation. This meant that two neurologists, two neurophysiologist-neurologists and one neurosurgeon were involved in the evaluation. These physicians were provided with a computer which they could use for the duration of the evaluation. The physicians interacted with the knowledge based system themselves, as they are the potential users of the system. They entered the data, requested a consultation with the knowledge based part of the system and received the advice on the computer screen.

Specifying physicians to test against. Ideally a large group of physicians who use the system should be compared to another large group of physicians who do not use the system. A limited number of computers was available for the evaluation study. With these small numbers, differences between the physicians in the two groups would become too important. This meant that it was not possible to balance physicians with and physicians without a knowledge based system. Therefore, in this study the physicians acted as their own control. They first decided on a diagnosis and treatment plan without using the knowledge based part of PLEXUS and then entered their final opinion after consulting the knowledge based system. This setup may lead to a number of the effects which will be discussed in Section 5.2.4.

Specifying a standard of performance. As was mentioned in Section 4.2.2.2, a true standard of performance does not exist in the domain of brachial plexus injuries, because the actual diagnosis and optimal treatment plan are not known. In such situations, a panel of experts may be asked to establish standard answers for the test cases involved in the evaluation. However, even if an explicit diagnosis is present it is still very difficult to compare the standard and the

physician's diagnosis, since in the domain of brachial plexus injuries one diagnosis usually consists of multiple answers. Therefore, experts are asked to blindly judge the opinions provided by the unassisted physicians and by the human-machine system, and to motivate their judgements.

5.2.2.3. Comparison

In order to investigate whether the system can be of assistance in the diagnosis and treatment planning of brachial plexus injuries, the final patient outcome in the assisted situation should be compared to the final patient outcome in the unassisted situation. However, final patient outcome in brachial plexus injuries is usually only known after several years and there is a large variability in treatment results. Therefore, in this evaluation the comparison will concern the unassisted and assisted opinions regarding the injured locations, the indication of the severity of the injury, and the treatment plan for each patient. By asking the experts to judge the diagnoses, a blind judgement of the proposed diagnoses and suggested treatments is obtained.

5.2.3. ANALYSIS OF THE RESULTS

Various aspects of performance were analysed. Firstly, the differences between the assisted and unassisted diagnoses and treatment plans are investigated. The cases for which the unassisted opinion is judged to be superior to the assisted opinion are analysed in detail. This will also be done for the cases for which the system receives a less than optimal judgement. The limited number of test cases and the fact that few physicians are involved in the study severely restricts the conclusions which can be drawn from this evaluation and will limit generalisation of the performance evaluation results.

5.2.4. LIMITATIONS OF THE CLINICAL EVALUATION

Possible sources of bias and confounding in clinical evaluations were discussed in Section 2.4.4. A number of these sources of bias and confounding may influence the results of this evaluation (van Daalen, 1992b). A number of practical problems also arose during the study. These limitations affected the evaluation setup which was described above. Both the influence of the possible sources of bias and confounding, and the effects caused by the practical problems will be discussed below.

5.2.4.1. Bias and confounding

Carry-over effect. This is the possible effect on performance due to education of the user by the system (Wyatt and Spiegelhalter, 1990). This can occur in all evaluation studies which involve physicians as their own controls, since the unassisted situation may be positively influenced by the previous occasion when the knowledge based system was used. However, this effect may improve the unassisted situation, and this will mean that when a positive difference is found between the assisted and unassisted situation, the positive conclusion is still justified.

In the evaluation of PLEXUS, the number of test cases is very limited. Therefore, the influence of the carry-over effect in this evaluation is negligible.

Feedback effect. A decision-aid will often make it easier to monitor performance, and feedback to the physician may act as a stimulus to improvement (Wyatt and Spiegelhalter, 1990). At the time the system underwent the clinical evaluation only a preliminary laboratory evaluation had been carried out. Therefore, all cases were also shown to one of the domain experts, and feedback was given to the physicians in case this was thought to be necessary. It is not possible to determine the effect caused by providing feedback.

Checklist effect. The knowledge based system may encourage a more complete and structured data collection (Spiegelhalter, 1983). This means that in a comparison of physicians who use the system and physicians who do not use the system, the opinions of the physicians who use the system may be better due to the fact that they systematically collected the data, rather than due to the advice they obtained from the knowledge based system. In this evaluation, however, the physicians were asked for their own diagnosis after all patient data were entered. This means that in the control situation, the knowledge based system is used for data entry, but not for consultation. Both the unassisted and assisted situations benefit from systematic data collection.

It would have been interesting to investigate the influence of systematic data entry, since most of the physicians indicated that a system such as PLEXUS would force them to collect all the data systematically and this was regarded to be very positive.

Parochial bias/transferability. Most of the cooperating hospitals are situated in the western part of the country and are relatively large hospitals. Since the physicians at the larger hospitals probably more often see seriously injured patients, this would imply that the test patients may have been more severely injured than the target population. On the other hand, the fact that the number of motorcyclists in the north and east of the country may be larger than the number

of motorcyclists in the west, may also have influenced the representativeness of the test cases, since injuries caused by motorcycle accidents are usually the more severe injuries. As these effects have an opposite influence, the representativeness of the patients does not appear to differ from the target situation.

Trial size. The study only involved a limited number of hospitals and test cases. Therefore it is not possible to obtain quantitative results and to generalise the results.

5.2.4.2. Practical limitations

A number of practical problems arose during the evaluation. The most notable of these was an insufficient number of prospective test cases. Further problems were related to usability and acceptability which also had an effect on the results of the performance evaluation study. Usability and acceptability were studied in more detail, and will be discussed in Section 5.3.

Lack of data. The number of brachial plexus injuries which occurred over the evaluation period was lower than expected. After it became apparent that the evaluation setup that was used would not provide sufficient data, the evaluation period which was intended to last for a year was extended by half a year and a request was published in a newsletter which is regularly distributed among neurologists in The Netherlands. In this request any neurologist who had a brachial plexus patient under treatment and was interested in participating in the evaluation was asked to contact the researcher. Only two neurologists responded to this request. The researcher then visited the respective hospitals with the computer, and the physicians entered the brachial plexus data in the presence of the researcher. The patients almost all proved to be patients for whom the final treatment decision had already been made, i.e. retrospective patients. Since the number of patients would be even more limited if these retrospective patients would not be included in the evaluation, the retrospective data were used in the analyses. The total number of patients which were finally used in the analysis amounted to 19.

Limitations of the data. The completeness of the data showed a large variation. For a number of cases, the data entered into the system by the physician were not as complete as was expected. One especially notably lacking piece of information in the clinical evaluation patient files regards the muscle strength examination. Whereas all experts always perform complete muscle strength examinations, this is not always done by potential users. The PLEXUS system does, however, need these results in order to provide a reliable diagnosis. System

performance during the clinical evaluation was influenced by the fact that not all relevant information was available to the system.

Withdrawal. The cooperating physician at one of the hospitals had to withdraw from the evaluation because of problems in obtaining the relevant data. This only left four evaluation sites.

Judges. Two experts were involved in judging the opinions. One of the experts asked to judge the diagnoses and treatment plans at the end of the evaluation period had been involved in the development of the system. The other judge was a resident who specialises in brachial plexus injuries at the hospital of the other expert who was involved in the development of the system. This could possibly introduce a bias in favour of the system. On the other hand, experts who are not involved in the project may not analyse the answers as seriously and extensively as the two cooperating experts. Furthermore, independent experts would have to be found abroad which would certainly extend the response times.

No final opinion. In this evaluation setup, the physicians were first asked to enter the patient data into the computer and to enter their own diagnosis and treatment plan. They were then asked to perform a consultation with the knowledge based part of the system, and to enter their own final opinion after receiving the knowledge based system's advice.

For 17 of the cases, no final opinion was entered into the computer by the physicians after they had consulted the knowledge based system. This means that for most of the patients, the results of step 4 of Table 5.1 are not available. Therefore, it is not possible to compare the unassisted and assisted situations (i.e. comparing step 2 to step 4). Instead, the diagnoses and treatments provided by the physicians prior to performing a consultation (step 2) were compared to the diagnoses and treatments suggested by the knowledge based system (step 3).

Acceptability problems. It was noticed that even if the physicians did see a patient with a brachial plexus injury, they did not enter the data into the computer right away. By the time some of these patients were entered, the final treatment plan had been decided on and the patients were retrospective rather than prospective. When the evaluation period was finished, the acceptability of the system was addressed by means of a questionnaire. This will be discussed in detail in Section 5.3.2.

Usability problems. A number of assumptions were made during the development of the system which did not correspond to clinical practice. Two important problems only became apparent during the course of the evaluation, other usability aspects will be discussed in more detail in Section 5.3.1. The

knowledge based system always uses whole numbers of months (integers) in its calculations involving time information. This is related to the level of accuracy used internally in reasoning. Therefore, a possibility was only created for the physicians to enter the number of months in integers, although at first this was not explicitly mentioned to the physicians. However, when the physicians were for instance asked for the number of months since the accident, they would for example enter 2.5. The program cannot handle this and will terminate without indicating why. Another of these bugs concerned the fact that it is not allowed to use a space when typing the name of the patient. Such problems are very easy to resolve, however, they only come to light when people who are inexperienced in the use of the interface work with the system. If such incidents occur regularly it can hamper the acceptance of the system, whereas it is not a failure which is inherent in the system.

5.2.5. CLINICAL EVALUATION RESULTS

At the end of the evaluation period, the data of 19 patient cases had been entered into the system. The mean age of these patients is 37 years. However, 8 of the patients were younger than 25, and 6 of the patients were older than 50. None of the patients were between 40 and 50 years of age. The distribution of injury causes for these patients is shown in Figure 5.1. It can be seen that a substantial number of these injuries was caused by a fall (26%). The four oldest patients (aged 59, 69, 71 and 75) and only one of the younger patients (aged 19) sustained their plexus injury during a fall.

The percentage of injuries which was caused by traffic accidents was 58%. Narakas (1985) found that approximately 70% of his traumatic brachial plexus patients sustained their injury during traffic accidents.

The percentage of patients who have a supraclavicular (i.e. situated above the clavicle) injury according to the physicians in the hospitals equals 67%, and according to the knowledge based system supraclavicular injury can be found in 74% of the patients. This is approximately the same as for the patients seen by Narakas (1985). However, whereas Narakas found that 70% of his patients with a supraclavicular injury have at least one avulsion, in the present study both physicians and PLEXUS indicated that 50% of the patients with a supraclavicular injury had one or more avulsions.

In approximately 70% of the cases, the physicians proposed conservative treatment. PLEXUS proposed conservative treatment in 21% of the cases and suggested the physician to wait for a period of time to see whether recovery would occur in 32% of the cases. In the laboratory evaluation which was described in Chapter 4, PLEXUS advised conservative treatment in 8% of the

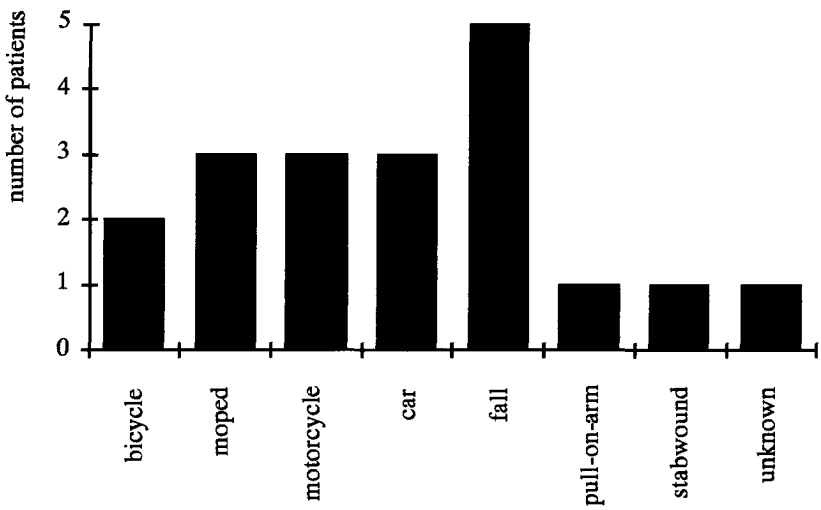


Figure 5.1. Distribution of injury causes in patients involved in the field evaluation of the knowledge based system.

cases and suggested the physician to wait for a period of time in 13% of the cases.

Although the number of patients is very low, these figures suggest that the patients in this field study are somewhat less severely injured than the patients seen by Narakas (1985), and the patients involved in the laboratory evaluation of PLEXUS.

Only 7 of the test cases were prospective cases. This means that the data were entered into the system at the time the system is meant to be used (i.e. when a treatment decision has to be made). The others were retrospective patients.

For one of the cases for which the system suggested that the physician should perform a further consultation after four months after the accident, the physician performed another consultation for this patient. Based upon the additional information which was available after four months, the physician entered his own opinion and computer advice was obtained. This second consultation is used as an additional case. For one of the patients, the physician did not enter his own diagnosis. Therefore, the total number of cases that could be used in the comparison of diagnoses and treatment plans still equals 19.

All case notes were printed. The opinions given by the knowledge based system and by the physician were attached anonymously and in random order at the bottom of each patient file. The two experts were asked to judge the opinions on a five point scale. This involved the same scale which was used in the laboratory evaluation (Chapter 4). For each of the opinions the expert had to choose one of the symbols (--,-,o,+,++) which best reflected the level of performance.

Ideally statistical analysis should be performed in order to determine whether assisted physicians perform differently to unassisted physicians. However, due to the small amount of data which became available in this evaluation, the data will be analysed qualitatively and all cases will be regarded separately

The results were analysed by investigating the number of times the knowledge based system received equal, better and worse judgements than the physician. When the knowledge based system received a worse judgement, the reasons for this were analysed in detail. Furthermore, the reasons why the knowledge based system produced any suboptimal suggestions, i.e. for which the system received any ratings lower than (++), were analysed.

On the whole, the system shows a good performance, although there is still some improvement possible. A number of general trends can be deduced from the analysis of the ratings obtained by PLEXUS. These general results will be discussed and will lead to suggestions for improvement of the system. First, the diagnoses produced by the system will be discussed, after which the

Table 5.2. Blind judgements of the diagnoses obtained in the clinical evaluation of PLEXUS. Two judges (jf-e1 and jf-e2) judged diagnoses established by PLEXUS and by physicians. The judgements provided by judge jf-e1 are shown in the columns, and the judgements provided by judge jf-e2 are shown in the rows. In the row and column marked 'total', the number of times the diagnoses were judged to be equal is shown, as well as the number of times the physicians recommended a superior diagnosis and the number of times PLEXUS recommended a superior diagnosis. The agreement between the judges is also shown.

judge	jf-e1				
		equal	physician	PLEXUS	total
jf-e2	equal	2	0	4	6
	physician	2	1	0	3
	PLEXUS	6	0	4	10
	total	10	1	8	19

Table 5.3. The number of times the diagnostic advice produced by PLEXUS was judged to be suboptimal (< ++). The judgements provided by judge jf-e1 are shown in the columns, and the judgements provided by judge jf-e2 are shown in the rows. The row and column marked 'total' show the number of times the diagnoses were judged to be optimal (++) and the times the diagnoses were judged to be suboptimal (< ++). The agreement between the judges is also shown.

judge	jf-e1			
		++	< ++	total
jf-e2	++	2	3	5
	< ++	5	9	14
	total	7	12	19

treatment plans will be analysed. A summary of the results for the diagnoses can be seen in Table 5.2 and Table 5.3, and a summary of the results for the treatments can be seen in Table 5.4. and Table 5.5. The two judges will be called jf-e1 and jf-e2. It can be seen that there is a number of cases on which the judges do not agree, the most important reasons for the judges disagreeing about certain cases will be discussed in Section 5.2.5.3. Possible implications this may have on the results of this evaluation will also be discussed.

5.2.5.1. Diagnosis

The most important results arising from the analysis of the diagnoses produced by PLEXUS in the clinical evaluation are the following:
- the information contained in a number of the patient files is insufficient,
- the system produces a number of false positive answers,
- a few of the answers given by the system are clinically unlikely,
- the system produces a number of false negative answers.

These conclusions lead to suggestions for improving the system. At the time the system was designed, it was expected that sufficient information would be present in the patient files, in order to be able to diagnose the patients. However, the system does provide facilities for informing the physicians when insufficient information or contradictory information is present, but it still establishes a diagnosis on the basis of the insufficient information. It was not expected that files would be so incomplete that it would not be possible to provide a correct diagnosis. Thus, the system should know the minimum amount of information which should be available in order to establish a diagnosis, and in cases where this is not available, the system should inform the physician of this and refrain from giving advice.

Since some of the files appeared to be incomplete, this could also lead to the notion that the system should be able to indicate to the physician which information is needed for performing a diagnosis, it should suggest examinations which should be carried out. The fact that the system should be able to assist in determining which examinations should be performed (early diagnosis) was also indicated by one of the physicians during the field evaluation. Presently, when the system suggests that additional information is needed, this is aimed at improving rather than at establishing a diagnosis.

As concluded in Chapter 4 on the laboratory evaluation, in some cases the system provides false positive answers. By updating the knowledge based system, it should be possible to avoid many of these false positive answers.

In a few cases, the diagnosis given by the system is regarded to be clinically unlikely, although all dysfunctioning is explained by the system. For

Table 5.4. Blind judgements of the treatments obtained in the clinical evaluation of PLEXUS. Two judges (jf-e1 and jf-e2) judged treatments suggested by PLEXUS and by physicians. The judgements provided by judge jf-e1 are shown in the columns, and the judgements provided by judge jf-e2 are shown in the rows. In the row and column marked 'total', the number of times the treatments were judged to be equal is shown, as well as the number of times the physicians recommended a superior treatment and the number of times PLEXUS recommended a superior treatment. The agreement between the judges is also shown.

judge	jf-e1				
		equal	physician	PLEXUS	total
jf-e2	equal	4	3	1	8
	physician	0	2	2	4
	PLEXUS	3	1	3	7
	total	7	6	6	19

Table 5.5. The number of times the treatment advice produced by PLEXUS was judged to be suboptimal (< ++). The judgements provided by judge jf-e1 are shown in the columns, and the judgements provided by judge jf-e2 are shown in the rows. The row and column marked 'total' show the number of times the treatments were judged to be optimal (++) and the times the treatments were judged to be suboptimal (< ++). The agreement between the judges is also shown.

judge	jf-e1			
		++	< ++	total
jf-e2	++	6	4	10
	< ++	4	5	9
	total	10	9	19

instance, a case where the system finds injuries at three different levels in the plexus whereas it is more likely that the injuries will be situated at the same level. For some of these situations, it should be possible incorporate additional heuristic knowledge in the knowledge base, that would allow the system to detect certain diagnoses which are clinically unlikely. The effort involved in updating the knowledge base will depend on the kinds of situations which are to be detected. However, there are probably certain situations which are easily recognised.

The system provides a number of false negative answers. In these cases, the system leaves out an injury location which should have been mentioned. Here, the false negative answers did not have any therapeutic consequences for the patients. The occurrence of the false negatives is probably due to the fact that sufficient muscles have to show a dysfunction in order for PLEXUS to conclude that the locations are injured. This may be due to the fact that the importance of certain muscles with regard to a certain injury location may be greater to the experts (sometimes one expert and sometimes both experts) than it is to the system.

5.2.5.2. Treatment

The most important results arising from the analysis of the treatments produced by PLEXUS in the clinical evaluation are the following:
- the information contained in a number of the patient files is insufficient,
- the system makes a few erroneous assumptions,
- the system does not indicate all the different treatment possibilities,
- in some cases the system makes the final decision when it should not yet have done this.

These conclusions also lead to some suggestions for improving the system. The measures which can be taken to avoid the system producing advice when insufficient information is present was discussed above. There are a few erroneous assumptions which were made in the knowledge base and which led to the incorrect treatment plans. The knowledge base should be updated to avoid this.

In some cases, the system suggests one treatment plan where there is some doubt as to the actual location of the injury and the severity of the injury. In these cases the system should present all therapeutic possibilities, which it does not do at present.

When it is not yet known whether an injury will recover, the system usually suggests that the physician should wait until four months after the accident before making a final decision. However, for some kinds of injuries this

should be 6 to 8 months. The system should therefore differentiate more between the different injuries.

5.2.5.3. Disagreement

It can be seen in Table 5.2 and Table 5.3 that for a number of cases, the judgement provided by judge jf-e1 is not the same as the judgement decided on by judge jf-e2. In most cases these differences can easily be explained. With regard to the extent of the disagreements, it can be seen in Table 5.2 that there were no cases for which one of the judges thought the system produced a superior diagnosis and the other judge thought that the physician suggested a superior diagnosis. The most important causes of differences in judgement of the diagnoses are the following:

- When there is not enough information in the patient file to establish a good diagnosis, one of the judges gives the score (o) to both physician and knowledge based system, whereas the other expert, although indicating that there is insufficient information, does distinguish between the opinions.
- In some cases neither of the opinions are completely correct and it is a matter of opinion how heavily certain errors or omissions are penalised.
- In a few cases one of the judges regards an answer to be false positive, whereas the other does not.

The last point is important for PLEXUS. The system produces a number of false positive answers. However, the judges do not agree about all the false positives given by the system. Therefore, further investigation will have to show in which cases the false positives are actually false and in which cases they may be seen as suggestions to the physicians.

The most important causes of differences in judgement of the treatments (Table 5.4 and Table 5.5) are the following:

- In a number of cases one of the judges differentiated more between the treatments, whereas the other indicated that it was most important to recognise that an operation should be performed.
- In one case there was a disagreement between the judges as to whether an operation should be performed or whether it should not.

These remarks indicate that in cases where there is no consensus, the knowledge based system should present the different points of view. Although the system does have the ability to do this in a limited number of situations, this capacity should be extended.

5.2.5.4. Conclusions

On the whole, it may be concluded that the system has a good performance, although some improvement is necessary. This is also shown in Tables 5.2 , 5.3, 5.4 and 5.5. Any reason why the system produced a suboptimal suggestion at all, in the opinion of either one or both of the experts, was analysed in detail in order to discover possible areas of improvement. The analysis of the suboptimal diagnoses and treatment plans could be generalised to a limited number of measures which should be taken in order to improve the performance of the knowledge based system. The most important recommendations are the following:

- The system should assist the physicians more in gathering and entering all data which are necessary to perform an adequate analysis of a brachial plexus injury. This includes suggesting possible examinations which should be performed for a particular patient.
- The system should be better at identifying certain diagnoses which are clinically unlikely, or parts of diagnoses which are clinically unlikely, and should inform the user of this. This requires incorporating additional heuristic knowledge in the knowledge base, which will allow detection of certain clinically unlikely diagnoses.
- In PLEXUS, the importance of certain muscles to certain injury locations is based on the literature. The importance of some of the muscles may have to be altered in the light of what is found clinically.
- Some of the treatment plans provided by PLEXUS should be accompanied by possible alternatives when the exact nature of the injury is not clear.
- In some cases PLEXUS advises the physician to wait for a period of time after which a final treatment decision should be made. This period of time should vary according to the nerves which are injured.

For a number of cases, there were differences between the ratings provided by the judges. In most cases these differences can easily be explained. Some of these differences lead to a number of recommendations for improving system performance:

- The system produces a number of false positive answers. However, the judges do not agree about all the false positives given by the system, therefore further investigation will have to show in which cases the false positives are actually false and in which cases they may be seen as suggestions to the physicians. The knowledge base should be updated accordingly.
- In cases where there is no consensus concerning the appropriate treatment, the system's capacity to present different points of view should be extended.

These measures should positively influence the performance of the system and reduce the number of suboptimal suggestions made by the system to a minimum.

5.3. Usability and acceptability of the knowledge based system

In addition to the performance evaluation of the human-machine system, the interaction between the human and the machine was addressed. Howard and Murray (1987) describe five main types of formal evaluation of user interfaces:
- expert-based, where expert knowledge and scientific principles are used,
- theory-based, where the mapping relationships between formal representations of the users and the device are examined with a view to identifying any mismatch,
- subject-based, studies involving four components: metric, task, user and system
- user-based, relates to personal evaluation by the user,
- market-based, relates to the final evaluation conducted by the market place.

Most evaluations concern subject based evaluations. Sutcliffe (1988) distinguishes three kinds of subject-based evaluation:
- diagnostic analysis, which aims to pin-point the poor design features in an interface design in an intuitive manner by examining recordings of dialogue sessions,
- monitoring one or more features of interface usage such as error rates, frequency of command use and duration of usage,
- experimental analysis designed to test empirically two different interface designs or two different features of a design (see, for example, Sassen, 1993).

Data collection for response times, error rates etc. may for example consist of on-line logs and observation. Data collection for user attitudes may consist of questionnaires, interviews, protocol analysis or factor analytic methods. The data analysis may consist of statistical or observational methods.

While the consequences of using an inappropriate technique may range from unnecessary expenditure of resources to the collection of data irrelevant to the evaluation questions posed, there is little advice available to aid an evaluator in the selection of an evaluation package (Howard and Murray, 1987).

The studies performed for PLEXUS addressed the usability and the acceptability of the knowledge based system. The usability and acceptability of a system can been described as follows:
- Usability concerns the extent to which an end-user is able to carry out the required tasks successfully, and without difficulty, using the computer application system (Ravden and Johnson, 1989).

- User acceptability is how willing users are to use a system in their own organisational context (Vainio-Larsson and Orring, 1990).

Vainio-Larsson and Orring (1990) also distinguish the functionality of a system which describes how well a system fits a set of particular task needs. However, they also state that in theory it may be possible to separate these three concepts, but that this separation is more difficult in practice. Ravden and Johnson (1989) include functionality in their definition of usability.

For PLEXUS, two investigations were performed, the first investigation was directed mainly at usability and the second one mainly at acceptability, although aspects of usability and acceptability will have an influence on the results of both studies. The investigation of the functionality will be included in the usability study.

5.3.1. USABILITY EVALUATION

The usability evaluation of the knowledge based system is directed towards investigating whether the user is capable of adequately using the system. Usability criteria must be specified in a way that makes them not only measurable but verifiable as well (Vainio-Larsson and Orring, 1990).

Ravden and Johnson (1989) describe an evaluation checklist for assessing usability of computer-based application systems. Each of the first nine sections of the checklist is based on a criterion which a well-designed user interface should aim to meet. The nine criteria are the following:

- Visual clarity: the information displayed on the screen should be clear, well-organised, unambiguous and easy to read.
- Consistency: the way the system looks and works should be consistent at all times.
- Compatibility: the way the system looks and works should be compatible with user conventions and expectations.
- Informative feedback: users should be given clear, informative feedback on where they are in the system, what actions they have taken, whether these actions have been successful and what actions should be taken next.
- Explicitness: the way the system works and is structured should be clear to the user.
- Appropriate functionality: the system should meet the needs and requirements of users when carrying out tasks.
- Flexibility and control: the interface should be sufficiently flexible in structure, in the way information is presented and in terms of what the user can do, to suit the needs and requirements of all users, and to allow them to feel in control of the system.

- Error prevention and correction: the system should be designed to minimize the possibility of user error, with inbuilt facilities for detecting and handling those which do occur; users should be able to check their inputs and correct errors or potential error situations before the input is processed.
- User guidance and support: informative, easy-to-use and relevant guidance and support should be provided, both on the computer and in hard copy form, to help the user understand and use the system.

Besides the issues mentioned above, the checklist allows the indication of system usability problems and has a section with general questions on system usability.

The usability investigation of PLEXUS entailed a study of the user interface of the knowledge based system. Some of the most important aspects of the interface will be discussed below.

PLEXUS user interface. The user interface facilitates the physician in entering data into the system, allows the physician to perform a consultation with the knowledge based part of the system and presents the advice produced by the knowledge based system on the screen. A representative example of a data entry screen can be seen in Figure 3.15. The interface was discussed in general in Section 3.5.2. A number of additional features which are necessary to understand the results of the user interface evaluation are discussed below.

Figure 3.15 shows a data entry screen. It can be seen that a folder metaphor has been used to facilitate the user in moving through the interface and to make the subsections of the interface explicit to the user. All data entry screens are made up of the same three sections:

- Heading: a heading can be seen at the top of each screen.
- Questions: data entry is performed in the middle part of each screen.
- Moving: the physician can reach all screens of the interface by using the buttons at the bottom of each screen.

In order to enter data into the system, the user answers the questions which are stated in the middle part of each screen. There are two kinds of questions: multiple choice questions and questions requiring textual input. The textual input is entered using the keyboard, and the multiple choice questions are answered using the mouse of the computer to choose the appropriate answer. Most of the questions are multiple choice questions, textual input has been avoided as much as possible. There are three kinds of multiple choice questions:

- questions showing all answer possibilities, and the most appropriate answer has to be chosen.
- questions which have so many answer possibilities that it is not possible to show all answer possibilities on the screen at all times. By clicking on the

question, the answer possibilities pop up on the screen and the user can then choose the appropriate answer.
- questions for which more than one answer can be true at the same time. By clicking on the question, the answer possibilities appear and the user can choose the appropriate answers. The user then has to acknowledge that all relevant answers have been chosen by clicking an 'OK' button.

A consultation with the knowledge based part of the system can be requested by clicking a button, upon which the computer automatically starts the knowledge based system. After this, the advice is provided on the screen in the form of text and in a graphical representation of the brachial plexus. An example of the graphical representation of an injury is shown in Figure 3.16.

Since the system is to be used for a relatively rare injury, it should be straightforward to use even for people who are inexperienced in working with the system. After a brief explanation about interacting with PLEXUS, the physicians should be able to use the system without consulting a manual, although a manual has been written for the use of the system (van Daalen, 1991).

5.3.1.1. PLEXUS usability evaluation

Goal. The final goal of this study was to investigate whether a user is capable of adequately entering patient data into the system and personally performing a consultation. Furthermore, possible improvements to the usability of the system should be identified. In this study, the aim was to obtain general information about the usability of the system and possible improvements to the usability, rather than performing a formal study.

Method. The usability was investigated by means of videotaping interactive sessions with the knowledge based system. Videotaping has also been used by Grolman (1989) to evaluate the prototype interface of PLEXUS. For the evaluation of PLEXUS, it was not feasible to ask the physicians to use the system and to fill out the complete usability checklist proposed by Ravden and Johnson (1989), due to the fact that they had volunteered to use the system and the response to a previous relatively extensive questionnaire had been disappointing because of the time involved in answering it. Therefore, the videotapes were observed, and the usability criteria stated in the questionnaire developed by Ravden and Johnson (1989), and which were mentioned above, were used as the basis for the analysis of the videotapes.

The present study was performed at five Dutch hospitals at the time when the knowledge based system was introduced in the hospitals, i.e. at the beginning of the performance evaluation period. For these physicians, this was the first

interactive session they themselves carried out with the system. The system had been demonstrated to these physicians on a previous occasion.

During the interactive session, the physician entered the data of one patient into the computer, performed a consultation with the system and received the computer's advice. Two of the physicians used actual patient data and the other three entered data of virtual patients.

When the researcher was present during any of the subsequent consultations with the knowledge based system, notes were made of any possible shortcomings and problems in the interaction. These notes were used in addition to the videotapes and were also incorporated in the results of the usability evaluation.

Analysis. Five videotapes were made which each lasted approximately one hour. The videotapes were played back and the time spent on each screen was recorded on paper. In addition, any important remarks made by the physicians and any observations which could be of interest to the usability of the interface were written down for each screen. These transcripts were then analysed using the nine usability criteria which were mentioned above.

5.3.1.2 Results of the usability evaluation

The most important findings resulting from the evaluation of the user interface are mentioned below. The information obtained during the evaluation will be discussed along the lines of the usability criteria stated by Ravden and Johnson (1988).

Visual clarity. The general visual clarity of the user interface is good. All screens have an informative title and overall the screens are uncluttered. There is a number of specific points concerning the visual clarity of the interface which requires improvement. This is mainly related to the organisation of information on the screen and to some of the answer formats which are not clear to the user.

- Organisation
  Where large amounts of data are displayed on the screen, they are clearly separated. However, the computer screen is very small. Some lists of muscles do not fit on the screen and scroll bars are used. Sometimes this scroll bar is not noticed by the physicians who think that the complete list is present on the screen. This may be solved by using graphical representations rather than lists of muscles. Most of the screens appear uncluttered, only those displaying the lists of muscles appear somewhat cluttered which should be solved.

Most of the information is very easy to see and to read. In a few cases various physicians missed a question which they had not noticed. By using a somewhat different layout for those screens, this can be solved.

- Answer format
  Most of the questions in the interface are multiple choice questions, in which case it is completely clear where and in what format the information should be entered. There are various questions which require typing, although this has been avoided as much as possible. For most of these questions, the answer format is not immediately clear to the user. This concerned questions informing about dates, number of months after the accident, name of the patient, age, percentages, and own diagnosis and treatment. This problem can easily be resolved by marking the entry field, for instance as ../../.. , and by performing a check on the answers which have been entered.

  For the questions concerning muscle strength, the place where the answers should be entered was not clearly indicated.
- Graphics
  A number of schematic and pictorial displays are drawn automatically by the computer. These are very clear and are regarded to be very useful by the physician. Although the legends should be more noticeable and one of these needs to be adapted.

Consistency. By using the same layout for all screens, predictability is maintained across the interface. A number of inconsistencies was found in the methods used for data entry. These have to be resolved, which will make it more straightforward to work with the system. There were also a few inconsistencies between this system and other programs, which should be altered in this program.

- Data entry
  In order to see the answer possibilities of some of the multiple choice questions, the physician has to click on the question and a pop-up menu appears containing all the possibilities. However, it is not always clear to the physicians that they have to click on the question itself for the pop-up menu to appear. Furthermore, in the case that multiple answers are possible for one question, the physician has to click on an 'OK' button, whereas when only one answer is possible it is not necessary to click on the 'OK' button and the pop-up menu disappears spontaneously. The ways in which the multiple choice questions are answered should be more consistent, because the reasons for these different kinds of questions are not clear to the user.

  The text questions could also benefit from more consistency. Sometimes the mouse has to be clicked first to be able to type in answers and sometimes

this is not necessary. Unlike some multiple choice questions which require the answer to be acknowledged, this is not necessary for the text questions.

- System messages
  When the computer is busy, sometimes a little watch is shown on the screen and sometimes a little rotating ball is shown. Only one of these indicators should be used.
- Menus
  At the top of the screen a number of menus can be pulled down. However, rather than having to click on the items, which has to be done in most other computer programs, the menus are pulled down by merely placing the mouse on the item. This should be made consistent with other software packages.
- Diagrams
  It was suspected that physicians would enter the strengths of all muscles for every patient. When this is not done, an inconsistency arises in the diagram showing muscle function, which is produced by the computer. In the muscle diagram, muscles with full strength are left white and unknown muscle strengths are also left white. This should be resolved. The shading of the muscle diagram is not exactly the same as the shading on the standard forms used in the hospitals. Furthermore, the paper muscle diagram in the hospitals also shows the terminal nerves (nerves at the end of the brachial plexus, which go down into the arm). This should be added to the computer's diagram.

Compatibility. On many of the screens, the way in which the system looks and works is compatible with user conventions and expectations. There is a number of areas in which the compatibility can easily be improved. Some incompatibilities regard differences which exist in the terminology used among different medical disciplines and others regard assumptions which were made by the system developers and which did not hold true.

- Terminology
  A few abbreviations are used in the interface. This should be avoided, since not all of these are completely clear to the physician.

  The terminology which is used in the interface is compatible with the terminology used by neurosurgeons. However, during the field evaluation, the system was also used by neurologists and neurophysiologist-neurologists. They indicated that some of the terms used in the interface did not conform to the way in which they used these terms. This was particularly relevant for the EMG examinations. However, the terminology used in different hospitals also differs. The terminology will at least have to be made explicit or may have to be altered. However, this will require further investigation. The order in which the EMG questions are asked is not compatible with the order in which the

examination is carried out. The order of the questions should be changed accordingly.
The difference in terminology is also present, to a lesser extent, for the muscle strength examinations. Not every hospital uses the scale proposed by the Medical Research Council (1986), however, everyone does know this scale. Therefore, the scale which is used has to be explicitly stated in the question. The scale used in the question about the level of pain is not known to all physicians, therefore clearer definitions of the answer possibilities will have to be given.

- Organisation
  For some of the screens the organisation and structure of the questions is not completely clear to the physicians. There are a few questions which do not belong to the subsection in which they have been placed, which should be altered. The order of a few of the questions and of some answer possibilities to certain questions also requires changing to a logical order.
- Conventions
  Some of the questions require the physician to fill in the number of months from the accident until a certain examination has been done. The computer expects an integer answer in these cases. However, this does not conform to user conventions. In addition, the wording of these questions is not always specific enough. Therefore, the interface should be changed accordingly. Furthermore, the computer does not allow the use of spaces in the name of the patient. This is not compatible with user conventions and should be altered.
- Missing
  For a number of questions there are additional answer possibilities which the user cannot choose and which are considered relevant by the physicians. These will have to be added.

Informative feedback. In general, the user interface is self-explanatory. There is, however, a number of questions for which the system does not adequately inform the user of the correct way to respond, or does not clearly indicate the actions it is performing.

- Domain
  The system's domain is not clearly delineated. For example, sometimes the physicians asked whether patients of a certain age or with a certain injury cause could be entered into the system. It was not clear to them whether the computer could handle these cases. This should be clearly indicated.
- Unavailable information
  For a number of questions, when the user had not carried out an examination or did not know the answer to a question, it was not clear what had to be entered. Sometimes this could be to click 'unknown' and sometime they have to leave a

blank, which is also inconsistent. It has to be made clear to the user what to do in case they do not know an answer and this has to be consistent throughout the interface.

- Defaults
  For some answers a default value is given by the computer. However, this is not always clear to the physician. Furthermore, the default values are not consistently used throughout the user interface. Care has to be taken with the default values, since sometimes it will not be known whether an examination has not been performed, or whether the physician has forgotten to answer the question.
- Editing
  There is one screen which shows a list of the patients which have been entered previously. Using this screen it is not immediately clear to the physician what has to be done to be able to see the patient data. Furthermore, it is not always clear to the physician that editing a patient is the same as entering new information.
- Saving
  At a number of stages during the interaction, the physician is asked if he wants to save the patient data. However, for the physicians it is not clear why they are suddenly asked this question. After asking whether the physician wants to save the data on the hard disk, the system always asks whether the physician wants to save the patient data on a diskette, this is also done when no diskette is present in the computer and should be avoided in those cases.
- Data entry
  When physicians are asked to enter their own diagnosis and treatment plan, they are required to enter this in the form of free text and this is not immediately clear to the physicians.
- Messages
  When the system performs a consultation, messages are shown on the screen which inform the user about what the system is doing. These messages are informative to the system designer, but probably not to users. Therefore, the content of these messages have to be updated to make them informative for the physician.

  Sometimes additional messages are required which inform the user that the system has finished a certain procedure, for instance colouring in the diagrams.

  When system errors occur, system messages are shown to the users. This should be prevented and messages should appear which inform the user of what can be done by the user at that moment.
- Consultation
  The telephone is used to perform a consultation with the knowledge based system. However, most physicians do not know when the phone is being used,

how long this takes and whether they can use the phone for conversations. This is not made clear to the physicians.

Explicitness. The user interface uses a folder metaphor. By clicking on the dog's ear at the bottom right hand corner it advances one page and by clicking on the left hand corner it goes back one page. The subsection of the interface the current screen belongs to is highlighted at the bottom of the screen. This metaphor works well, as physicians have no difficulty in moving through the interface, and it makes the user interface transparent to the user. However, a few alterations are necessary. These regard the distinction between various actions performed by the knowledge based system and changing of a number of terms.

- Distinction
  The fact that working with PLEXUS presently consists of entering data and then performing a consultation has not been made as clear as might have been, as performing a consultation is done on a similar kind of page as data entry. If this difference between data entry and consultation is to remain, a clearer distinction will have to be made between the two. This should also help in indicating to the physician that the diagnosis shown on the screen has been produced by the knowledge based system.
- Terminology
  The terms used in the menu which allows a new patient to be entered and old patients to be updated are computing terms and it is not made clear to the user what each of the options means.
- Organisation
  One of the questions informs about the examinations which have been performed for a patient. However, the interface does not proceed to only ask questions regarding the examinations which have been entered, all other examination results can also be entered. This question should be removed, and the internal reasons for the question being stated should be solved without involving the user.

  When entering a new patient or editing an old patient, the first page shown to the users is a page showing summary information. However, since for new patients there is no summary information as yet about the patient, the entries are left empty. When physicians see this page, some think that the information has to be entered on this page, whereas this is not the case. Therefore when entering a new patient, the summary screen should be skipped as a first page.

Appropriate functionality. The user interface is meant to assist physicians in entering data and performing a consultation. The interface also shows the advice provided by the system on the screen. The physicians indicated that improved assistance in entering the information would be useful. They also indicated that,

in addition to these tasks, they would be interested in some additional functionality with regard to the choice of the medical examinations to perform for a patient.

- Data entry

  The system could offer more assistance in entering the data into the computer. This can be done by adding some 'intelligence' to the data entry. For instance, if the physician does not enter certain information, the system could emphasize and explain the importance of this information with respect to the diagnosis. The system already has this capacity for various examinations, however, this should be extended. The system could also help the physician in choosing the most appropriate answer in a certain situation.

  The physicians also indicated that an important function of the knowledge based system is that it allows them to systematically record patient data. Presently, the data which can be recorded in the system mostly consist of the data needed by the knowledge based system in order to produce its advice. This means that for a number of tests only the results of the first and last tests can be recorded, and data of any intermediate tests are replaced by the results of the last test. The interface should be amended in order to allow the recording of more than two examinations.

  One physician suggested that the system should automatically read the values obtained from the EMG, rather than this having to be entered manually.

- Advice

  At present, the system provides diagnostic and treatment advice, and detects contradictions or a lack of information for a number of tests. However, some of the physicians indicated that they would appreciate it if they could first enter their clinical data into the system upon which the system should suggest the additional tests which have to be performed for the patient.

- Output

  The diagnosis is presented in text and in a graphical representation of the nerves of the brachial plexus. The graphical representation is very satisfactory, however, the textual representation of the diagnosis and treatment suggested by the system is taken directly from the knowledge based part of the system and will require some restructuring and rewording in order to make it informative to the user. At present, the system does not explain its advice. Some explanation facilities would be useful.

- Printing

  The system can print the patient file. However, it is not possible to only print parts of the patient file. At present it is only possible to print the file after a consultation with the knowledge based part of the system has been carried out. Printing a patient file takes quite a long time, and during printing nothing else can be done. A faster printer will decrease the printing time.

- Time
  The time needed to enter patient data into the system is quite substantial. Methods to reduce this time and to increase efficiency of the interface should be considered.

  After all the data have been entered, a consultation with the knowledge based system takes about 10 minutes, using the computers which were applied in the field evaluation. Some of the physicians see this as a negative point, for other physicians this is not a problem. Some parts could be speeded up. However, as computers are becoming a lot faster, this will partly solve itself.

Flexibility and control. The user interface is very flexible. The user can easily reach all parts of the interface and is in control of the actions which are performed. It is always possible to go back to the previous page. Shortcuts are available by choosing the appropriate label at the bottom of a page and the user can look through a series of screens in either direction. One of the physicians was surprised that the computer did not proceed to the next page automatically. However, this was done on purpose so that the physicians can volunteer as much information as they want, and then choose the next page. On the whole, this appears to work very well. There are a few situations in which the computer could take some more control in order to improve the efficiency of data entry.

- Computer control
  One item on which the computer may possibly take control and presently does not, regards the entering of muscle strengths. The strengths of 38 muscles have to be entered manually and the cursor does not go to the next muscle automatically. This is not very efficient.

  A further item on which the system leaves control to the physician is in the diagrams where the system specifically asks the physician if he wants the diagram coloured in, which is obvious, otherwise the physician would not have chosen this diagram. This will also prevent the physician from wondering whether he himself should colour in the diagram.

Error prevention and correction. Most of the inputs are multiple choice inputs, which helps to prevent errors from occurring. There are a few areas in which error prevention, detection and correction should be improved.

- Error detection
  For the text input, more facilities for checking the answers given by the user should be provided. For instance, muscle strength is measured on a scale of 0 to 5, and the computer could easily detect values outside this range.

- Error prevention
  When entering the EMG data, errors are easily made due to the fact that the muscles are listed closely together. This can easily be solved by choosing other ways of listing the muscles.
- Error correction
  The textual questions posed some problems when physicians wanted to delete their answers. The user first has to delete the answer manually. This means that the new information is sometimes added to the old information if this has not been cleared by the user. This problem does not exist for the multiple choice questions. For the multiple choice questions the user can undo an action by clicking the erroneous answer again.

User guidance and support. This consists of on-line guidance and off-line guidance.

- On-line help
  The on-line help facility is very limited at present and should be extended, so that the system clearly explains the possible actions which can be taken.
- Off-line help
  There is a hard copy manual which explains the complete interface in detail (van Daalen, 1991). However, this could benefit from a more detailed discussion of possible user and system errors, and care should be taken to maintain the manual up to date.

5.3.1.3. Conclusions of the usability evaluation

The user interface of PLEXUS is often regarded to be very user-friendly. On inspection, a number of areas of improvement can be identified. From the evaluation study it may be concluded that a number of aspects will require attention.

The method used to evaluate the user interface consisted of recording interaction sessions, transcribing the most relevant occurrences and comments onto paper and then analysing these transcripts taking into account the usability criteria discussed by Ravden and Johnson (1988). This has proved to be a very good method from which a lot of useful information has been obtained. This information can be used to update the user interface of PLEXUS so that the essential usability requirements are satisfied.

5.3.2. ACCEPTABILITY EVALUATION

During the clinical evaluation it became apparent that when the physicians saw a patient with a brachial plexus injury, the data were not readily entered into the system. This could indicate a lack of acceptability. The usability as described above is one aspect of acceptance, a user must be able to adequately use the system. However, there are many other aspects of importance to the acceptance of knowledge based systems, such as the need for the system, and the performance of the system. At the end of the evaluation period, a limited study of the acceptability of PLEXUS was conducted among the physicians involved in the clinical evaluation.

Since the system was not used as readily as it had been expected, and since it is often stated in the literature that hardly any knowledge based systems are used in practice, an extensive study into the acceptance of knowledge based systems in general has also been conducted. This investigation will be described in Chapter 6. The following sections concern the acceptability of PLEXUS.

5.3.2.1. Acceptability evaluation

Goal. The goal of the acceptability evaluation was to investigate whether the physicians would personally use the system if it was generally available.

Method. At the end of the performance evaluation period, a short questionnaire was distributed among the physicians who cooperated in the evaluation, in order to investigate their opinion regarding the acceptability of the knowledge based system. Rather than developing an extensive questionnaire, it was deliberately kept limited in order to ensure adequate response.

The questionnaire consisted of the following questions:
1) Do you think the system would be used if it was generally available?
2) Would you use the system it was generally available?
3) What are the positive aspects of the system?
4) What are the negative aspects of the system?
5) Do you have any further remarks or suggestions?
The physicians were asked to motivate their answers to the questions.

The first two questions address the acceptability of the system and the other questions are more related to the usability of the system, and the ways in which this could be improved.

Table 5.6. Answers to the question whether physicians think PLEXUS would be used if it was generally available.

Do you think the system would be used if it was generally available?
• probably only in centres where systematic examination is performed as a rule
• especially in centres which are more specialised in this kind of injury
• no; 'cost-benefit' analysis: learning to work with the system costs some time/experience, whereas plexus injuries occur infrequently
• I think so; the system is an aid in performing a diagnosis (especially neurological examinations)
• not in this form; 1) more surgical possibilities 2) more diagnostic possibilities

Table 5.7. Answers to the question whether the physicians themselves would use PLEXUS.

Would you use the system if it was generally available?
• yes
• yes, not only for the advice, but also because it forces good documentation and provides a good check on the way one works
• possibly; clinical analysis of the problem often provides the correct localisation of the lesion, the therapeutic decision arises too infrequently for a computer program
• probably
• not in this form; 1) more surgical possibilities 2) more diagnostic possibilities

Table 5.8. Positive aspects of the system according to the physicians.

What are the positive aspects of the system?
• it is logically consistent
• forces the physician to work systematically
• points at inconsistencies in testing or reasoning on the part of the physician
• provides good possibilities for analysis
• all diagnostic items in one system
• program is user-friendly and well-organised
• provides valuable advice
• specialist know-how and advice
• well-organised way to use patient data
• systematic organisation of Merle d'Aubigné

Analysis. The results were analysed qualitatively as the questionnaire was only sent to five physicians. However, this may still provide a general idea about the reasons for the lack of acceptance of the system.

5.3.2.2. Results of the acceptability evaluation

The answers to each of the questions will be discussed below.

Whether system would be used. The answers that were given by the physicians are shown in Table 5.6. On the whole, this question was answered in a relatively positive way. However, it is emphasised that learning to work with the system costs time and the number of brachial plexus injuries which occurs is very limited. There is one physician, more specialised in this domain than the potential users of the system, who indicates that the surgical possibilities the system presents should be elaborated. However, the system is meant for physicians who do not perform plexus operations themselves and who have to make the choice of either referring the patient to a neurosurgical centre for surgery or of treating the patient conservatively.

Whether the physicians would use the system. The answers to this question are shown in Table 5.7. Two of the physicians indicated that they would use the system themselves if it was available, and the others indicated that they may use the system. The final physician would like the system to be extended.

Positive aspects. The answers to this question are shown in Table 5.8. A number of positive aspects of the system was mentioned by the physicians. The most notable positive aspect is the fact that the system systematically organises the relevant brachial plexus patient data. Furthermore, the system is consistent and it was mentioned that the system provides valuable advice.

Negative aspects. The way in which the physicians answered this question is shown in Table 5.9. Although care has been taken to optimise the interaction between physicians and the system, it can be seen that it still takes time and effort to enter patient data into the system, especially for those who use the system irregularly. Most of the other aspects have already been discussed in Section 5.3.1. on usability.

Further suggestions. The answers to this question can be seen in Table 5.10. As was also mentioned in Section 5.3.1. on usability, besides giving diagnostic and treatment planning advice, the system could be directed more towards assisting

Table 5.9. Negative aspects of the system according to the physicians.

What are the negative aspects of the system?
• way of entering EMG data is unusual
• it is quite laborious
• need time to learn to work with the system
• present system provides few possibilities of following patient over time; only most recent data are available
• scheme lacks in differentiation for normal muscles, muscles which have not been investigated and affected muscles
• a practical problem is the fact that computer setup always has to be taken apart and stored in cupboard due to possible burglaries, which costs time
• entering data costs time
• communication with computer is laborious after receiving advice
• not easy to enter data
• not enough surgical possibilities

Table 5.10. Suggestions for improvement of the system put forward by physicians involved in the field evaluation.

Do you have any further remarks or suggestions?
• have various parts of the system looked at by clinical neurophysiologists and clinical neurologists, rather than only neurosurgeons
• connect database of literature to the system
• add list of addresses of neurosurgical centres which perform plexus surgery
• advice for early diagnostics, such as performing certain examinations, is missing from the system
• a more direct communication
• surgical possibilities should be extended

the physicians in deciding which examinations are appropriate for a particular patient and in systematically examining the patient.

At present, the system is divided into a data entry part and a consultation part. If any patient data are changed, a completely new consultation has to be performed. This has led to the observation that there is no direct communication with the system. By incorporating the suggestions stated above, and integrating consultation part in the rest of the system, this could possibly help to solve some of these problems.

5.3.2.3. Conclusions

On the whole, the physicians react positively to the acceptability questions. It is important to note that data entry costs time and effort. This was also identified during the course of the field evaluation, as most of the physicians did not enter the patient data right away and some preferred to use the system in the presence of the researcher. Furthermore, during the field evaluation, none of the physicians entered a different opinion after having consulted the knowledge based part of the system.

A number of useful suggestions was made with regard to the functionality of the system. The system should communicate more directly with the user and should also be directed towards assisting the physicians in determining the examinations which should be performed for a particular patient.

5.4. Conclusions and recommendations

The conclusions arising from this clinical evaluation will divided into the three areas which have been discussed: performance, usability and acceptability. After a separate discussion of these three aspects, general recommendations resulting from this evaluation will be mentioned.

Performance. The system's performance during the clinical evaluation was largely satisfactory. In this chapter, any suboptimal suggestion produced by the knowledge based system was analysed in detail. This has led to the identification of a number of areas of the system which should be improved. The most important recommendations are the following:

- the system should provide better assistance in data gathering and data entry,
- the system should have better possibilities for identifying clinically unlikely diagnoses,
- the importance of certain muscles to certain injury locations possibly requires updating,
- the number of false positive answers given by the system should be reduced,
- in some cases where the diagnosis is not certain, the system should provide alternative treatment plans,
- when the system advises the physician to wait for a period of time before making the final treatment decision, this period of time should vary according to the nerves which are injured.
- in cases where there is no consensus concerning the most appropriate treatment, the system's capacity to present different points of view should be extended.

These measures should positively influence the performance of the system and reduce the number of suboptimal suggestions made by the system to a minimum.

Some conclusions may also be drawn with regard to the evaluation method used during the clinical performance evaluation of PLEXUS. Ideally, a large group of physicians who use the system should be compared to another large group of physicians who do not use the system. The assisted and unassisted situations should be investigated using the actual patient outcome. However, there are probably few studies in which such a setup is possible. In the evaluation of PLEXUS, the resources and number of computers available did not allow the involvement of two groups of physicians, since with a small number of physicians in each group, the differences between the physicians would become too large. Furthermore, since final patient outcome in brachial plexus injuries is usually only known after several years and since there is a large variability in treatment results, the aim was to investigate assisted and unassisted diagnoses

and treatment plans. However, the actual diagnosis and optimal treatment of a patient are not known. Therefore, in this evaluation experts were asked to judge the results.

The setup included asking the physicians using the system for their opinion before and after consulting the knowledge based part of the system. However, the physicians entered a final opinion in only two of the cases. This caused the fact that, rather than studying the physicians in the unassisted and assisted situations, the unassisted opinions were compared to the knowledge based system's advice. A somewhat stronger emphasis on entering the final opinion may have improved this situation.

It may be concluded that although the number of test cases was small and practical limitations prevented an optimal evaluation setup, the clinical evaluation of PLEXUS proved to be worthwhile in providing information about the performance of the knowledge based system in the clinical environment, and a number of areas of improvement could be identified.

Usability. From the usability evaluation study it may be concluded that a number of aspects will require attention. The recommendations include the following:

- the consistency of the methods for data entry have to be improved,
- some of the items have to become more compatible with user conventions and with the information physicians need for diagnosing brachial plexus patients,
- the system should give somewhat more feedback to the users to make clear what kind of information is expected from them and to inform the physicians of the actions being performed by the system,
- the on-line help facilities should be improved,
- the system requires some additional functionality.

The results which were obtained with regard to the functionality of the system were especially interesting. It was identified that the system could offer more assistance in entering the data into the computer. This can be done by adding some 'intelligence' to the data entry. For instance, if the physician does not enter certain information, the system could emphasize and explain the importance of this information with respect to the diagnosis. The system already has this capacity for various examinations, however, this should be extended. The system could also help the physician in choosing the most appropriate answer in a certain situation. At present, PLEXUS gives diagnostic and treatment advice, and indicates contradictions or a lack of information for a number of tests. However, some of the physicians indicated that they would appreciate it if they could first enter their clinical data into the system, upon which the system should suggest the additional tests which have to be performed for the patient. The fact that the possibilities for assisting the physicians in data gathering and data entry should be improved, also arose from the performance study which was discussed above.

The method used to evaluate the user interface consisted of recording interaction sessions on video, transcribing the most relevant observations and comments onto paper and then analysing these transcripts using the usability checklist developed by Ravden and Johnson (1988). This has proved to be a very good method, from which useful information has been obtained. This information can be used to update PLEXUS so that essential usability requirements are satisfied.

Acceptability. The acceptability questionnaire was answered in a relatively positive way. One important aspect which was noted was that it takes some time and effort to enter the data of a patient into the knowledge based system. This problem was also identified during the course of the field evaluation, as the physicians did not enter the data into the system as readily as was expected. One further drawback was the fact that the number of brachial plexus injuries was smaller than expected. These aspects may be barriers to the acceptance of the system.

The physicians identified a number of possible areas of improvement. One physician indicated that a more direct communication with the system is required, whereas others stated that the system should also be directed at suggesting the examinations which should be performed for a particular patient and at helping the physicians in systematically documenting the patient.

It was noted during the field evaluation that, after consulting the knowledge based system, in only two cases a final opinion was entered. Thus, the relatively positive results obtained in the acceptability questionnaire also have to be seen in the light of the findings resulting from the course of the field evaluation. Therefore, the acceptability question still remains largely unanswered and will require further thorough investigation, because incorporation of some of the improvements will require considerable effort.

General recommendations. With respect to the data entry process, some physicians mentioned that an advantage of such a system is the fact that it helps in systematically organising and storing the patient data. In the near future quality control in medicine will become more important, just as it is in industry, where the ISO 9000 norm is used in many branches. It will be important for the physicians to systematically examine the patients and store the data accordingly, for they will have to be able to justify their actions when confronted with medical audits.

Assisting physicians in systematic data entry and advising them about the information which will be important, and which needs to be collected in different situations, should be emphasised more in PLEXUS. This conclusion arose from both the performance evaluation as well as from the investigation of usability and acceptability. The system could help in determining which examinations

should be performed for a particular patient and could also help physicians in entering the appropriate information in the computer. However, care should be taken with regard to the time spent per patient and the effort involved in entering patient data, for this seems to be an important barrier to the actual use of the system. The advantages obtained from using the system should outweigh the time and effort involved in using the system.

Systematically collecting data about brachial plexus patients could also serve another purpose. It became clear at conferences and in various hospitals that there is no real consensus on the treatment of brachial plexus injuries, neither among different disciplines such as neurosurgery, rehabilitation medicine and orthopaedic surgery, nor within a single discipline such as neurosurgery. It will only be possible to reach a consensus in this area by gathering sufficient information about results of various kinds of treatment.

6 Attitudes of physicians and process-operators towards knowledge based systems[1]

Knowledge based systems are rarely used in actual practice. A number of problems which may explain this lack of acceptance has been identified by various authors. Possible solutions to these problems lead to requirements which may have to be met by knowledge based systems. The importance of these requirements has been studied by means of a questionnaire which was distributed among physicians and process-operators. Results show that the introduction of a knowledge based system should not lead to a shift in responsibility from the human to the machine. Therefore, it is important for the user to understand how the system works. This requires a system design which helps the user to build up an adequate internal representation of the reasoning process.

6.1. Introduction

Knowledge based systems have not met the expectations which existed some fifteen years ago. Many knowledge based systems have been developed since then, and only very few are used in actual practice. Many authors have suggested reasons for the lack of acceptance of these systems (see, for example, Bramer, 1984; Kidd and Cooper, 1985). Some important problems which were identified, concern the fact that knowledge based systems often contain shallow heuristics, rather than containing knowledge based on a deep understanding of the problem domain. Furthermore, the explanation possibilities which are provided by these systems are inadequate.

More recent papers (Miller and Masarie, 1990; Woods and Roth, 1988) indicate that there is a problem associated with the role that most current knowledge based systems are programmed to adopt. The user collects the data and implements the actions for the machine, and the machine has the role of problem solver. The human-machine interface focuses on features to help the human to collect the data and to accept the machine's solution. Since the user's role in solving the problem is reduced to that of an interface between the machine and the environment and it seems like the user's thinking is replaced by the system, this kind of knowledge based system design has been termed the 'cognitive-tool-as-prosthesis' paradigm by Woods *et al.* (1990). Miller and Masarie (1990) refer to this style of diagnostic consultation as the 'Greek Oracle' model.

[1] Co-author of this chapter is J.M.A. Sassen.

Although many researchers describe reasons for the lack of acceptance of knowledge based systems, only very few of these (Teach and Shortliffe, 1981; Shortliffe, 1989; Roth *et al.*, 1987) have carried out practical investigations into this problem. Teach and Shortliffe (1981) addressed physicians' attitudes regarding computer-based clinical consultation systems by means of a questionnaire. An important conclusion with regard to the demands concerning the performance capabilities of such systems is that physicians will reject a system which dogmatically offers advice, even if it has impressive diagnostic accuracy and an ability to provide reliable treatment plans. They seem to prefer a system which can be used as a tool to assist them with patient management decisions in order to improve the quality of patient care.

These conclusions are supported by Shortliffe's (1989) observations of physicians' attitudes towards knowledge based systems. In this study, groups of physicians were first shown a medical knowledge based system in operation on a videotape, after which they spent some time discussing different computing issues in medicine. A frequently expressed concern was related to a fear of loss of control in decision making on the part of the physician. One of the participants stated that he never hoped to see a computer that would tell him exactly how to treat a patient.

The study performed by Roth *et al.* (1987) is of a different nature. It involved a study of the performance of technicians diagnosing faults in electro-mechanical equipment with the use of a knowledge based system. The technicians in the investigation varied in level of experience and in interactive style (active or passive). The faults they had to diagnose varied in level of difficulty. The knowledge based system was designed in the conventional way, i.e. according to the 'prosthesis' design.

Results of the study revealed that, contrary to the implicit assumptions in the design paradigm of the knowledge based system, technicians actively and substantially contributed to the diagnostic process. The more the human functioned as a passive data gatherer for the machine, the more joint system performance was degraded. Those who passively followed the directives of the machine expert dwelled on unproductive paths and reached dead-ends more often than participants who took a more active role. Active human participation led to more successful and rapid solutions. However, the machine expert not only failed to support an active human role, it actually retarded technicians from taking or carrying out an active role. Although acceptance was not explicitly addressed in this study, the same reasons which account for performance problems, can provide additional information about the lack of acceptance of knowledge based systems.

These three studies provide valuable information for the design and implementation of future systems. They have shown that users will reject a system that dogmatically offers advice, even if it has impressive accuracy. Furthermore, active human participation enhances task performance. Since Teach and Shortliffe's study in 1981, many new developments in knowledge based system and cognitive engineering technology have been reported (see, for example, Woods and Roth, 1988; Miller and Masarie, 1990; Steels, 1990; Struss, 1992). At present, users' opinions about these new concepts are not known, and experimental evaluations have not yet taken place. It is necessary to investigate these issues in order to be able to design and develop advisory systems which will be used in actual practice.

At the Man-Machine Systems Group of Delft University of Technology in The Netherlands, two areas of knowledge based assistance are being investigated. The first application is the assistance of operators of large industrial plants in fault detection and fault diagnosis, and the second is the assistance of physicians in diagnosis and treatment planning of nerve injuries in the neck. In order to develop systems which can adequately assist users in these two domains, an investigation of user opinions has been carried out among physicians and process-operators. The results of this investigation are described in this chapter.

6.2. Possible causes for the lack of acceptance of knowledge based systems

Possible causes for the lack of acceptance of knowledge based systems which have been mentioned in the literature will be discussed below. The discussion of these problems will lead to a number of requirements which may have to be met by knowledge based systems in order to provide better possibilities for actual use. At least four potential areas may affect the acceptance:
(1) human,
(2) machine,
(3) human-machine interaction,
(4) environment.

6.2.1. HUMAN

Loss of professional status. The psychological barriers to the acceptance of knowledge based systems may be high (Bramer, 1984). Systems which cause a user to fear that his job may be taken over by the system, or systems which

operate in an area in which a user feels his competence will be diminished will not be accepted. The level of skill which a human possesses is a major aspect of his status, both within and outside the working community. If the job is 'deskilled' by a knowledge based system, this is difficult for the individuals involved to come to terms with (Bainbridge, 1987).

Shortliffe (1989) reports a fear among physicians that the challenge of independent problem solving will be eliminated by knowledge based systems, whereas it is this challenge that attracted them to medical practice in the first place. The aspect of 'deskilling' is one of the reasons why knowledge based systems should be advisory systems, and it is often stated explicitly that overall responsibility remains with the user.

Furthermore, when someone does not want to use a knowledge based system, for instance because he fears that he will be replaced by it, the system will be expected to have an unrealistic high performance (Muir, 1987). Since the system will usually not attain this kind of performance, its advice will be rejected. In contrast, when someone wants to abrogate his responsibility, for instance because he feels incompetent or finds the job tedious, he will expect an unrealistically low performance of the machine (Muir, 1987) and will usually accept its advice. The latter is not an acceptance problem, but it may cause other undesirable situations to occur.

Lack of trust. The system must gain the trust of the users (Bell, 1985). They must be able to rely on it, not necessarily to perform perfectly, but to perform predictably in a manner which they understand. Shortliffe (1989) reported the pervasive assumption among physicians that computer-based decision aids will never be able to cope with certain 'distinctly human tasks'. Moreover, physicians are often sceptical about the opinion of human experts. They claim that it will always be possible to find another expert who says something a little different. By putting expert knowledge into a knowledge based system this distrust remains, and its advice may not be accepted as being useful.

Muir (1987) claims that one reason for distrust in a decision aid is the fact that the human's mere presence implies that the machine may be irresponsible or incompetent, and a distrusting human is required to monitor its output. If a machine is distrusted, the user will, if possible, carry out the tasks himself. This leaves little or no opportunity for him to reevaluate his distrust because the machine is not used. Therefore, the system cannot produce the behavioural evidence necessary to support a reevaluation. Furthermore, the human is left with little or no time for the reevaluation process since he is busy performing the task himself. An implication of this is that a human's trust in a machine,

once betrayed, may be difficult to recover. In contrast, if someone trusts a machine, the system will be allowed to perform its task, leaving available both the evidence and the time needed for the user to reevaluate its trustworthiness as necessary.

6.2.2. MACHINE

Functionality mismatch. In the early days of knowledge based systems, problems to be solved by a knowledge based system were usually chosen because they suited the technology, although there was no user need for assistance in that particular domain. However, demand will come only from perceived need on the part of the intended users (Shortliffe and Clancey, 1984).

Even when a particular problem is identified for which a need to solve it does exist, a functionality mismatch may arise. If intended users are not involved in the development of a knowledge based system, it may not be noted until system delivery that its functionality is not useful in practice. However, even if users are involved in the development, some difficulties will still remain, for it is much easier to criticise an actual system which does something different than expected, than it is to appraise a system design. Some intended users who join the design team can become enthusiasts and converts to the system design. If this happens, they may no longer be 'representative' users, but start to share the goals, values and assumptions of the other developers (Hart, 1991). In which case it is also very likely that the resulting system will not match user needs.

Technological limitations. Knowledge based systems which are developed to solve a recognised problem will still suffer from technological shortcomings (see, for example, Steels, 1990; Woods and Roth, 1988). Often, the knowledge contained in these systems consists of an enumeration of the situations which are to be recognised, the evidence which signals that these situations have arisen, and the responses that should follow. They consist of pre-planned routines that can anticipate to all situations which are foreseen to occur. However, significant events in natural systems almost always involve novel or unexpected features, which, by definition, cannot be contained in the knowledge base. Hence, such systems break down in these cases, and cannot deliver correct advice. They demonstrate brittleness in the face of unanticipated variability.

To handle this problem, some knowledge engineers choose a narrowly constrained domain, for instance a very narrow speciality in medical diagnosis, or single fault problems in troubleshooting domains. The domain is simplified to such an extent that complete coverage of the potential problems is ensured.

The advice which such systems deliver on challenging cases that occur in actual practice, and which violate the strict assumptions made by the knowledge engineer, can be unrealistic, a nuisance or erroneous (Woods and Roth, 1988). These same authors also note that very often, the range of problems which can be solved by the knowledge based system is almost isomorphic to the range of problems the target user is able to solve.

Two further problems which are often mentioned are the lack of common sense (see, for example, Bell, 1985; Buchanan, 1986) which is demonstrated by knowledge based systems, and the fact that they do not have knowledge of their own limitations. The problems due to lack of common sense may, for instance, manifest themselves in the interaction between the system and the user. This may lead to trust related acceptance problems. The absence of knowledge of its own limitations leads to the problem that a knowledge based system cannot distinguish when its advice is useful, from when it is not. It is up to users to judge the value of the advice, although they have only limited mechanisms available to perform this judgement (Woods and Roth, 1988). The judgement mechanisms which are available to the user will be discussed in more detail below.

6.2.3. HUMAN-MACHINE INTERACTION

Difficulties concerning data entry. Physicians recognise data entry as being a major barrier to the effective use of computers in clinical practice (Shortliffe, 1989). Whereas in other areas, such as the process industry, data entry is performed automatically via data-acquisition systems. When systems do require a certain amount of data to be entered, this may place time demands on the user. If the benefits of the system are outweighed by the time it takes to consult the system, the system may not be used. A related problem is the way in which dialogues between the user and the machine are conducted. The interaction may require typing on the part of the user. If the user is not an experienced typist, this may be very time consuming. Furthermore, some systems may be very difficult to learn to use, which may also lead to rejection.

Inability to provide adequate internal representation. When people refer to a human specialist, they generally pass on both authority and responsibility together (Woods, 1986). When using a machine advisor, it is often the user who is responsible. Therefore, he should have the possibility to take this responsibility. This implies that the user should be able to judge the advice on its merits. This requires the user to have an adequate internal representation of

the process. The interaction which takes place between user and machine should help the user in acquiring this internal representation.

One method of helping the user to build up the internal representation is to explain the advice and the reasoning which is carried out by the system to the user. However, present knowledge based systems are often not suited to providing this kind of explanation and justification of their reasoning and advice. Their capabilities for explanation are usually limited to showing a trace of production rules which were used to solve the problem. The production rules may not be at all clear to the users, since the rules consist of compiled knowledge, heuristics and programming-constructs. However, they have no other way of assessing the intended objectives of the knowledge based system. This makes it very difficult to judge and accept the final advice.

Problems related to the advice. Another problem deals with the type of advice which knowledge based systems supply. Coombs and Alty (1984) and Pollack *et al.* (1982) have studied naturally occurring situations in which people asked human experts for advice. These studies show that good advisory interactions involve cooperative problem solving. People actively participate in the definition and resolution of their own problems. The advisor does not merely respond to immediate requests, rather the advisor assists during problem formulation and plan generation.

The function of an advisor seems to be to broaden the user's horizon by raising and helping to answer questions like: What would happen if? Are there side-effects to this response? How do x and y interact? (Woods, 1986). One of the physicians involved in Shortliffe's (1989) study made a similar request: "I would find a computer useful in pointing out potential pitfalls - drug interactions that I didn't remember..." Most knowledge based systems are not capable of providing this type of advice. They do not allow actual cooperative problem solving, during which a negotiation takes place between user and system, and where the system possesses the knowledge which is needed to determine the user's intentions.

The output of a knowledge based system typically consists of some form of confidence or likelihood estimate over a set of possible diagnoses. A user is expected to act on the machine's solution, but this solution may be ambiguous. What is the machine's solution? The highest likelihood category? What if there are several high likelihood options or no high likelihood options? Should the likelihood be weighted by expected consequences? Choosing a solution to act on is further complicated by the non-standard procedures that are typically used to compute likelihood estimates. Likelihood is only one element of decision making under uncertainty and risk (Woods, 1986). Users confronted with such

a situation have no other means available than to solve the problem themselves, hence this may also hamper the acceptance of the knowledge based system.

6.2.4. ENVIRONMENT

There is a variety of additional problems which may influence the use of a knowledge based system. There are still ethical, legal and confidentiality issues to be resolved (Hart, 1991). Legal problems may arise when a user acts according to the system's recommendations, and the advice given by the system proves to be wrong. On the other hand, a user may neglect the system's advice that is actually correct. There are also various problems which are related to form rather than to content. If the system is not easily accessible, it may not be used. In addition, a system which only provides advice may not be sufficient, and the user may require the system to, for instance, also facilitate administrative tasks.

6.3. Possibilities to improve acceptance of knowledge based systems

Possible solutions to the acceptance problem of knowledge based systems have been encountered in the literature. These can be divided into the following five categories:
(1) alternative paradigms for knowledge based system design,
(2) improved knowledge models,
(3) minimising effect on current working practice,
(4) enhancing user friendliness,
(5) additional features.
These five categories of possible solutions will be discussed below. The same categories have been used in the study which was conducted in order to investigate user opinions. The results of this study will be described in Section 6.4, after the discussion of the possible solutions.

6.3.1. ALTERNATIVE PARADIGMS FOR KNOWLEDGE BASED SYSTEM DESIGN

Knowledge based system research has always been technology driven. Woods (1986) suggests that in order to design effective decision support systems, a problem driven, rather than a technology driven approach is required. In a problem driven approach, it is necessary to first investigate the factors which determine competence and incompetence in a domain. This may, for example, be done by performing a task analysis.

The results of the task analysis can then be used to develop tools which help people in improving their task performance. For instance, difficulty in diagnosis is rarely due to a global failure, but is more likely to be due to one or a limited number of steps which are difficult to overcome (Miller and Masarie, 1990). It should be possible to design a tool which helps the user to tackle these restraints, and hence to enable him to perform better. Note that most current knowledge based systems would assist the user by completely diagnosing the case, and presenting the diagnosis as an advice.

There is a variety of different system designs aimed at improving on the original systems which provide ready made advice. They range from variants of the traditional design, to a completely cooperative system which can be used as a tool. Miller (1984) and Langlotz and Shortliffe (1983) described critiquing systems. A critiquing system first asks how the user plans to solve the problem, and then critiques this plan. In its critique, the system discusses the advantages and disadvantages of the proposed approach, compared to other approaches that might be reasonable or preferred. Other possibilities are systems which process and combine input data in a way that allows users to make their own decision, or systems which predict the consequences of decisions which are taken by the user. Real cooperative problem solving systems have not yet been developed.

A knowledge based system which is developed according to the cooperative approach does not solve the entire problem by itself, but user and machine solve the problem together without diminishing the user's active problem solving role. It is like the machine is a subordinate of the human (Woods and Roth, 1988).

By allowing cooperative problem solving, the acceptance problems due to possible loss of professional status and lack of trust may be overcome. By carrying out repeated task analysis and involving the user in system design and development, the problems related to functionality mismatch should be reduced. Furthermore, since the system no longer solves the complete problem, the user's internal representation may improve and is kept updated. When the user directly participates in the consultative process by determining the steps to be taken, the burden of explanation is reduced (Miller and Masarie, 1990). The user knows that the assessment of the state that served as input to the machine's decision making, is sound, because he made the assessment himself. The user is now in a better position to follow the (relatively short) line of reasoning and evaluate the outcome of the machine (Woods and Roth, 1988).

6.3.2. IMPROVED KNOWLEDGE MODELS

In contrast to knowledge based systems, human experts can learn from previous experiences, have common sense, can explain their advice, do not suffer from brittleness, and know the limitations of their competence. Knowledge based systems of the current generation lack these kinds of knowledge, but according to Buchanan (1986) it is not to say that future systems will. It is often stated that systems which are based on a deep knowledge model instead of being based on a shallow knowledge model will function more like a human expert. Shallow knowledge consists of compiled knowledge and heuristics, whereas deep knowledge consists of underlying domain theories. Using deep knowledge may lead to systems which do not suffer from brittleness, which possess knowledge of their own limitations, and which have superior explanation possibilities (Steels, 1990). Furthermore, such systems could allow the user to establish a better internal representation of its functionality, and could possibly help to overcome a lack of trust in the system.

6.3.3. MINIMISING EFFECT ON CURRENT WORKING PRACTICE

One of Teach and Shortliffe's (1981) recommendations is that system designers should strive to minimise changes to current practices. This requirement can involve various aspects of current practice. Firstly, additional time commitment should be avoided. Furthermore, the system should ideally be available where and when the user customarily makes decisions, which may mean that the system should be portable. The minimal change requirement could also imply that the introduction of the system should not lead to a reduction of staff, and that the final responsibility remains with the user. This should help in overcoming some of the psychologically related problems and should facilitate a natural human-machine interaction.

6.3.4. ENHANCING USER FRIENDLINESS

Knowledge based systems should be easy to learn and be largely self-documenting (Teach and Shortliffe, 1981). If the system is (too) difficult to learn, the user may abandon the system before he even becomes aware of the functionality and other benefits which the system may have. Furthermore, as data entry may also be a barrier to system acceptance, alternative methods of interaction may be required, which should minimise the use of the keyboard and for which no previous computing or typing experience should be necessary.

6.3.5. ADDITIONAL FEATURES

Another aspect which has been proposed to improve the acceptance of knowledge based systems is to embed other useful support capabilities in the system, such as computerised data entry forms or automatic report generation (Langlotz and Shortliffe, 1983). This could also imply that the system should accompany its advice with references to the literature. A further possibility could be that the system also assists the user in systematic data entry. These additional features may improve the use of the system. However, a system with the same 'additional' functionality, but without a problem solving capability, could possibly have been equally useful, and much cheaper to develop.

6.4. User attitudes towards 'improved' knowledge based systems

To investigate which of the requirements that have been discussed above, are important to the potential user, an investigation of potential users' opinions was performed by means of a questionnaire.

Participants. Two different areas of knowledge based assistance are being investigated at the Man-Machine Systems Group of Delft University. Firstly, assistance to human operators of large industrial plants in fault detection and diagnosis, and secondly, assistance of physicians in the diagnosis and treatment planning of nerve injuries in the neck. Therefore, the questionnaire was distributed among process-operators and physicians. Since there may also be differences in attitude between novices and experienced persons, four groups of subjects were studied. The first group consisted of 66 medical students who had just started their clinical training at a university hospital, and whose mean age was 25.6 years (s.d. 2.7). The second group consisted of 66 experienced physicians with a mean age of 40.5 (s.d. 11.6) and a mean experience of 13.5 years (s.d. 11.3). The third group consisted of 168 students, who were being trained to become process-operators, at 4 different operator schools. The mean age of these students was 23.4 years (s.d. 6.7). Finally, there was a group of 63 experienced process-operators who worked at an oil refinery. The experienced operators had a mean age of 37.8 (s.d. 9.5), and a mean experience of 15.6 years (s.d 9.0).

Method. A questionnaire was developed to investigate user opinions regarding possible requirements for knowledge based systems. The questionnaire consisted of 41 different requirement statements. The statements corresponded to the requirements which were discussed in the previous section. However,

rather than being formulated in an abstract way, the statements consisted of specific implementations of the ideas. Fifteen of the statements were replicated from Teach and Shortliffe (1981) in order to allow a comparison between both studies. In accordance with Teach and Shortliffe, participants were asked to indicate their opinion on a five point scale, consisting the following categories: strongly agree, somewhat agree, not sure, somewhat disagree and strongly disagree. Both process-operators and physicians were asked the same questions, however, to make it as straightforward as possible for the subjects to answer the questions, the statements referred to respectively the industrial plant and to medicine. At the end of the questionnaire an open question was included with the aim of investigating the participants' general opinion concerning the use of knowledge based systems in their work.

Since the subjects were not familiar with knowledge based systems, it was not possible to distribute the questionnaire by mail among a large group. Instead, the participants were visited at process-operator schools, an oil refinery, and hospitals during one of their regular meetings. These meetings were usually attended by approximately 10 to 20 people. At the end of the meetings, the subject of knowledge based systems was introduced to the participants. This short introduction took about 10 minutes, and was aimed at explaining what knowledge based systems are and what kind of assistance they can offer. The issue of cooperation between the user and the machine was also explicitly addressed. After the explanation, each attendee was asked to complete the questionnaire. An advantage of this procedure in comparison to the distribution by mail among a larger group of subjects, is that the data do not consist solely of answers of interested or motivated people. A disadvantage is that the answers may have been influenced by the discussion which sometimes took place after the brief introduction.

Analysis. The opinions regarding the 41 different statements were transferred from the scale containing five categories to a scale ranging from -2 to +2, as was also done by Teach and Shortliffe (1981). For all four groups the mean and standard deviation of each of the requirement statements was calculated.

The differences between student operators and experienced operators, student physicians and experienced physicians, student physicians and student operators and the differences between experienced physicians and experienced operators were calculated for all requirements, by means of the Wilcoxon rank sum test (Mann-Whitney). Significance was calculated at the $p < 0.01$ level (two-tailed probability). The statistical analysis was carried out using the statistical software package SPSS for the Macintosh.

The final question was an open question in which the participants were asked to indicate their opinion regarding knowledge based systems. The

answers to the open question appeared to, quite naturally, fall into 4 different categories. The results of the open question provide a general idea of the participants' opinions concerning the use of knowledge based systems in their work.

First, the participants' opinions regarding the 41 different statements will be discussed. For this purpose, the statements have been subdivided into the five different topics which were mentioned in Section 6.3. Each topic will be discussed along the same general lines, consisting of a description of the statements which the participants strongly agreed with and somewhat agreed with, followed by the statements which the participants disagreed with. After this, the differences between this investigation and the study carried out by Teach and Shortliffe (1981) will be mentioned, and differences which were observed among the participating groups will be discussed.

Each section is accompanied by a table and a histogram. The tables show the statements which were used in the questionnaire. In the tables, the statements are classified according to the categories which were mentioned in Section 6.3, but in the questionnaire they were all placed in random order. In the actual questionnaire the statements were in Dutch and referred to the industrial plant and to medicine, however, for the purpose of this chapter, the statements have been transferred to a 'neutral' domain. The tables also show the means and standard deviations for each statement for all groups of participants. The histograms contain the mean values for each of the four groups along the vertical axis. The labels of the statements are shown along the horizontal axis of the histograms. The actual statements belonging to these labels are shown in the tables.

Label and question	st.ph m (sd)	exp.ph m (sd)	st.op m (sd)	exp.op m (sd)
COST1; KBS should improve the cost efficiency of tests and therapies	1.14 (0.94)	1.11 (1.04)	1.47 (0.75)	1.43 (0.73)
SIMPL2; KBS should process and combine data in such a way that it will be easier for user to make a diagnosis	1.15 (0.81)	1.30 (0.96)	1.40 (0.83)	1.52 (0.82)
CRIT3; KBS should indicate where and why possible differences occur between user's opinion and KBS's advice	1.39 (0.74)	1.56 (0.66)	0.98 (1.03)	1.02 (1.05)
OPINI4; KBS should take into account the user's opinion when giving an advice	0.89 (1.18)	0.36 (1.39)	1.14 (1.03)	0.95 (1.24)
PREDCT5; KBS should predict the consequences of a treatment plan suggested by the user	0.62 (1.09)	1.11 (1.04)	1.17 (0.97)	1.16 (0.97)
CONTR6; KBS should check whether the user's diagnosis is consistent with all data	0.45 (1.19)	0.95 (1.20)	1.01 (0.93)	1.19 (0.90)
SCARC7; KBS should be especially developed to deal with rare cases	0.00 (1.20)	-0.23 (1.38)	0.07 (1.40)	-0.03 (1.44)
TECH8; KBS should significantly reduce the amount of technical knowledge that a user must learn and remember	-1.12 (1.14)	-0.74 (1.42)	-0.98 (1.21)	-1.14 (1.29)
READY9; KBS should give a ready made diagnosis and treatment plan	-0.47 (1.22)	0.62 (1.17)	0.81 (1.01)	0.76 (1.15)

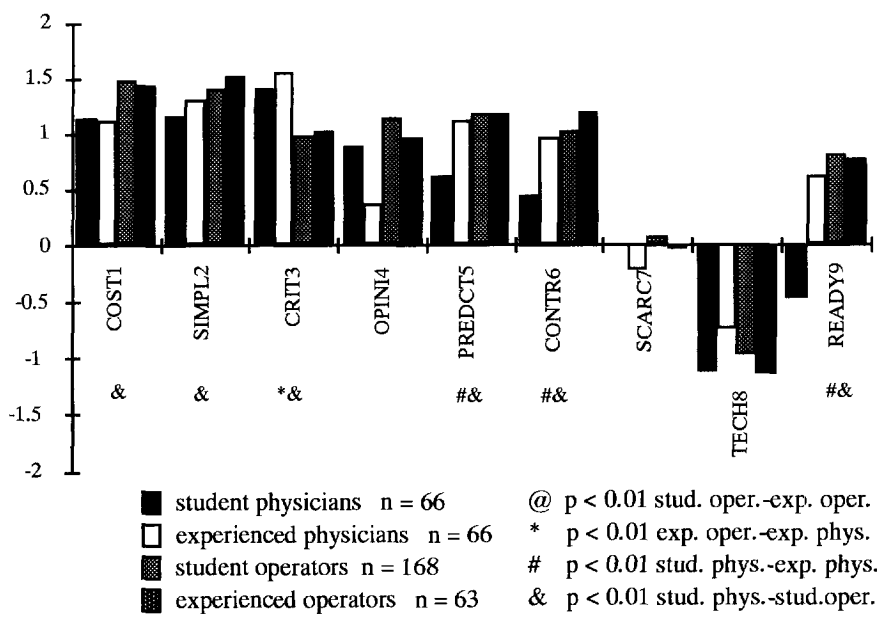


Figure 6.1. User attitudes towards alternative paradigms for system design.

6.4.1. USER ATTITUDES TOWARDS ALTERNATIVE PARADIGMS FOR KNOWLEDGE BASED SYSTEM DESIGN

The results show that the participants, both operators and physicians, agree with the requirement that a knowledge based system should improve cost efficiency of tests and therapies. This may be seen in Figure 6.1, where this statement corresponds to the first item, COST1. The operators emphasise cost efficiency slightly more than the physicians. The importance of cost efficiency was also established by Teach and Shortliffe (1981).

A knowledge based system's ability to process and combine data in such a way that it will be easier for the user to establish a diagnosis (SIMPL2) is also regarded to be an important requirement by all four groups. With respect to the ability to explain where and why there are differences between the user's opinion and the machine's advice (CRIT3), it can be seen that all four groups agree with the statement, although the physicians are more strongly in favour of this aspect than the operators. An explanation could be that operators have less time to make decisions and are used to computers in their job. They may therefore be less interested in this feature. The participants somewhat agree with the requirement that a knowledge based system should take into account the user's opinion when giving advice (OPINI4).

Also of importance are a system's ability to predict consequences of a treatment plan suggested by the user (PREDCT5) and the ability to check whether the user's diagnosis is consistent with all data (CONTR6), although the student physicians agree significantly less strongly with both requirements than the other groups. The participants are indifferent as to whether a knowledge based system should be especially developed to deal with rare cases (SCARC7).

The participants in this study somewhat disagreed with the fact that knowledge based systems should significantly reduce the amount of technical knowledge which a user must learn and remember (TECH8). This outcome could be related to the earlier reported fear of deskilling due to the use of a machine advisor (Section 6.2). Teach and Shortliffe's (1981) subjects were more indifferent regarding this statement.

There was a significant difference between student physicians and the other groups regarding the statement that a knowledge based system should give a ready made diagnosis and treatment plan (READY9). Student physicians somewhat disagree with this requirement, whereas the other participants somewhat agree with it. A possible explanation could be that student physicians may not yet be fully aware of the difficulties associated with diagnosis.

Label and question	st.ph m (sd)	exp.ph m (sd)	st.op m (sd)	exp.op m (sd)
OPUND1; The user should be able to understand how the KBS reached a diagnosis	1.67 (0.83)	1.74 (0.75)	1.76 (0.53)	1.87 (0.34)
XPL2; KBS should be able to explain their diagnostic and treatment decisions to the user	1.55 (0.53)	1.35 (0.81)	1.32 (0.80)	1.22 (0.89)
DIWRNG3; KBS should never make an incorrect diagnosis	1.03 (1.14)	0.76 (1.30)	1.33 (0.97)	1.43 (0.91)
TRWRON4; KBS should never make an error in treatment planning	1.18 (1.12)	0.77 (1.33)	1.18 (1.09)	1.19 (1.05)
AUTLEA5;KBS should automatically learn new information when interacting with experts/from previous experiences	1.02 (1.06)	1.11 (1.07)	1.40 (0.91)	1.54 (0.67)
DOMAIN6; KBS should contain knowledge about an entire specialism/subject, and not just a part of it	0.71 (1.31)	-0.05 (1.52)	1.48 (0.88)	1.54 (0.89)
UNDERS7; KBS should display an understanding of their own knowledge	0.89 (1.05)	0.73 (1.03)	0.92 (1.05)	1.03 (1.15)
COMSNS8; KBS should display common sense	0.61 (1.08)	0.21 (1.43)	0.46 (1.19)	0.02 (1.46)
SIMUL9; KBS should simulate users' thought processes	0.73 (1.12)	0.24 (1.23)	0.08 (1.25)	0.38 (1.24)

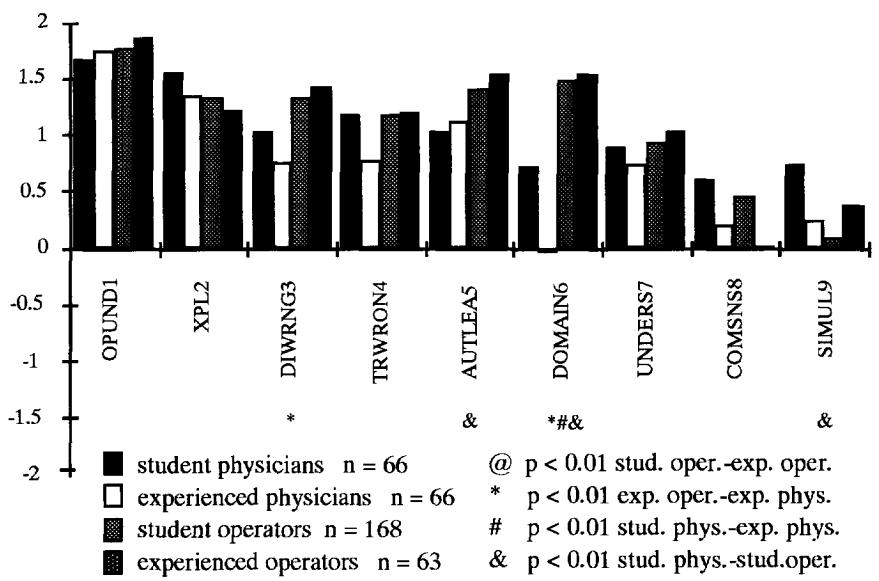


Figure 6.2. User attitudes towards improved knowledge models.

6.4.2. USER ATTITUDES TOWARDS IMPROVED KNOWLEDGE MODELS

For all participants, the most important requirement with respect to improved knowledge models is that a user should understand how the machine has reached a diagnosis. In Figure 6.2 the results of this statement correspond to item OPUND1. A possible method of achieving understanding of a diagnosis is by way of an explanation facility, and therefore it is not surprising that the subjects also attach importance to a knowledge based system's ability to explain the diagnosis and treatment planning (XPL2). In the study performed by Teach and Shortliffe (1981) explanation was also regarded as being very important for acceptance.

All subjects somewhat agreed with the statements that a knowledge based system should never make an incorrect diagnosis (DIWRNG3) and should never make an error in treatment planning (TRWRON4). The mean of the experienced physicians is lower than the mean of the other subjects; they probably recognise the fact that in medical domains this is not possible, and they are more realistic about this statement. A difference between this and Teach and Shortliffe's (1981) study is that their subjects did not think that a system has to display either perfect diagnostic accuracy or perfect treatment planning to be acceptable. A possible explanation for this difference is that the participants in the present study are more reluctant to use knowledge based systems and therefore expect a (unrealistically) high performance (see Section 6.2.1).

The participants somewhat agreed with the statement that knowledge based systems should automatically learn from previous experiences/when interacting with experts (AUTLEA5). The significant difference between student physicians and student operators on this item may be due to a disparity of the wording of the statement presented to the physicians (system should learn automatically when interacting with experts) and operators (system should learn from previous experiences). However, it also seems to be more obvious to develop an automatic learning system for the industrial domain, since such a knowledge based system can constantly monitor a process.

Operators and student physicians agree with the statement that a knowledge based system should contain knowledge about an entire subject (DOMAIN6). The operators feel more strongly about this than the student physicians, whereas the experienced physicians are indifferent with respect to this requirement.

The participants somewhat agree that a knowledge based system should display an understanding of its own knowledge (UNDERS7). Except for the

Label and question	st.ph m (sd)	exp.ph m (sd)	st.op m (sd)	exp.op m (sd)
RESPBL1; When using a KBS the user is responsible for the decisions to be made	1.92 (0.32)	1.94 (0.30)	1.56 (0.75)	1.71 (0.66)
SPEC2; KBS should not reduce the need for specialists/operators	1.38 (0.96)	0.61 (1.23)	1.54 (0.95)	1.43 (1.01)
PARA3; KBS should not reduce the need for paraprofessionals/supporting staff	0.98 (1.07)	-0.03 (1.20)	1.20 (1.08)	1.17 (1.09)
SECOP4; KBS should perform the task of a second opinion	1.26 (0.92)	0.97 (1.05)	1.01 (0.86)	1.29 (0.85)
NORMAL5; Data-entry should resemble current practice	1.03 (0.94)	1.06 (0.99)	1.18 (0.85)	0.92 (1.05)
ASK6; KBS should only give advice when explicitly asked for by the user	1.12 (1.13)	0.89 (1.29)	0.59 (1.39)	0.75 (1.34)
TASK7; KBS should not take over any (specialist) task from the user	1.05 (1.40)	0.24 (1.55)	0.55 (1.29)	0.67 (1.33)
PORT8; KBS should be portable and flexible so that the user can access them at any time and place	0.76 (1.08)	0.82 (1.11)	0.42 (1.30)	0.54 (1.32)
STAND9; KBS should become the standard for acceptable medical/operating practice	-0.92 (1.19)	-0.48 (1.27)	-0.32 (1.25)	-0.02 (1.21)
TIME10; Time necessary to solve a problem should not increase when using a KBS, even if diagnosis improves	0.44 (1.29)	0.20 (1.35)	0.90 (1.19)	1.37 (0.87)

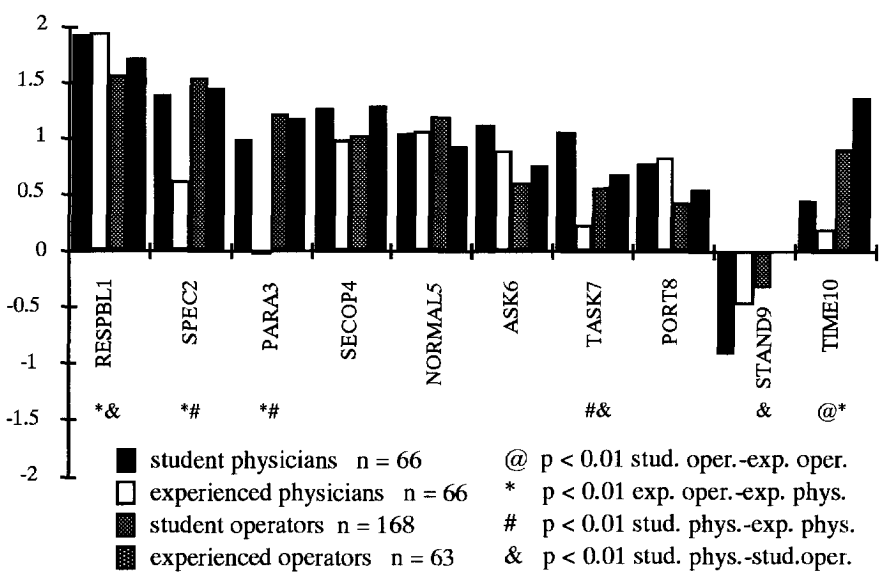


Figure 6.3. User attitudes towards minimising effect on working practice.

experienced operators, who are indifferent, the participants slightly agree with the statement that a knowledge based system should display common sense (COMSNS8). The results show that the student physicians are somewhat more in favour of the requirement that a knowledge based system should simulate users' thought processes (SIMUL9) than the other three groups.

6.4.3. USER ATTITUDES TOWARDS MINIMISING EFFECT ON CURRENT WORKING PRACTICE

The participants strongly agree with the statement that the user remains responsible when using a knowledge based system. This can be seen in Figure 6.3, where this statement corresponds to the item RESPBL1. The physicians agree somewhat more strongly with this requirement than the operators. The student physicians and the operators agree with the statements that knowledge based systems should not reduce the need for the user (SPEC2) and should not reduce the need for supporting staff (PARA3). Experienced physicians feel less strongly about these requirements. This may be caused by the pressure of work which they would prefer to be reduced.

Other statements with which the participants somewhat agree are the requirement that knowledge based systems should perform the task of a second opinion (SECOP4), that data-entry should resemble current practice (NORMAL5), and the fact that knowledge based systems should give advice only in those situations when explicitly asked for (ASK6). Operators somewhat agree with the statement that knowledge based systems should not take over any task from the user (TASK7). Experienced physicians are more indifferent towards this item and student physicians more strongly agree with it. All groups somewhat agree with the statement that knowledge based systems should be portable and flexible so that the user can access them in any time and place (PORT8).

A significant difference in opinion between student physicians and student operators was found with respect to the statement that knowledge based systems should become the standard for acceptable medical/operating practice (STAND9). Student physicians somewhat disagreed with this statement, whereas student operators were more indifferent. Teach and Shortliffe's (1981) data show that the physicians involved in their investigation also would not accept the use of a consultation system as a standard for acceptable medical practice.

There was a significant difference in opinion concerning the statement that the time necessary to solve a problem should not increase (TIME10), even when the diagnosis improves. Answers to this statement showed that physicians were indifferent, but student operators somewhat agreed and experienced

Label and question	st.ph m (sd)	exp.ph m (sd)	st.op m (sd)	exp.op m (sd)
EASY1; KBS should demand little effort from users to learn or use	1.50 (0.77)	1.58 (0.72)	1.47 (0.82)	1.63 (0.60)
HARDW2; KBS should run on the computer standard	0.76 (1.18)	1.08 (1.00)	0.87 (1.07)	0.86 (1.23)
GRAPH3; KBS should allow data-entry by means of menus and graphics using a mouse, rather than typing	0.50 (1.01)	0.65 (1.05)	0.75 (1.05)	0.56 (1.12)
VOICE4; KBS should respond to voice command and not require typing	-0.48 (1.22)	0.00 (1.18)	-0.74 (1.25)	-0.54 (1.37)
EXP5; The use of KBS should not require any knowledge of computers	0.82 (1.20)	0.80 (1.35)	0.10 (1.49)	0.75 (1.40)
NOINP6; KBS should not require any data-entry by the user	-0.80 (1.11)	-0.67 (1.23)	-0.11 (1.31)	0.22 (1.44)

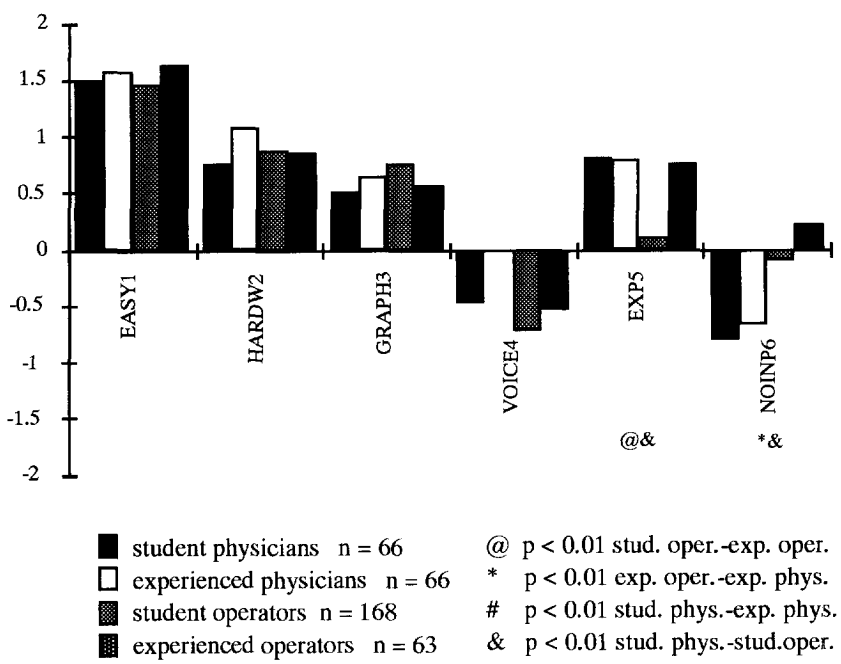


Figure 6.4. User attitudes towards enhancing user friendliness.

operators agreed more strongly. This may be caused by the fact that, in general, operators have to make decisions under greater time pressure than physicians.

6.4.4. USER ATTITUDES TOWARDS ENHANCING USER FRIENDLINESS

Answers provided to the statements related to enhancing user friendliness show that all groups of users agree with the fact that knowledge based systems should demand little effort from the user to learn or use. This can be seen in Figure 6.4, where this statement corresponds to item EASY1. Teach and Shortliffe's (1981) subjects were more indifferent towards this statement. Since so many advances have been made with respect to user friendliness of software during the last decade, the participants in the present study may be more critical. All groups somewhat agree with the fact that knowledge based systems should run on the computer standard (HARDW2), and that knowledge based systems should allow data entry by means of choices from a menu or graphics using a mouse, and not by means of typing (GRAPH3).

Student physicians and operators somewhat disagree with the statement that a knowledge based system should respond to voice command and should not require typing (VOICE4). It was noticed from remarks made by some of the participants, while answering the questionnaire, that they fear that the computer might mistake their remarks to a colleague or background noise as commands. The subjects who participated in the study performed by Teach and Shortliffe (1981), are more indifferent with respect to this requirement. In this study, the same is true for the experienced physicians, although the difference is not significant.

Differences in opinion between the student operators and the other groups were found with regard to the statement that the use of a knowledge based system should not demand any knowledge of computers (EXP5). The student operators were indifferent about this statement, whereas experienced operators and physicians somewhat agreed with it. This may be attributed to the fact that the younger student operators probably have more experience using a personal computer than physicians and the experienced operators. Another difference was found with respect to the issue that a knowledge based system should not require any data entry by the user (NOINP6). Operators were indifferent, whereas physicians somewhat disagreed with this statement. In a control room, almost all data is available in the computer system that controls the plant, whereas this is not always true for medical data. Operators may expect that all the important data is available to the knowledge based system, whereas physicians may expect that it is not possible to automatically acquire

Label and question	st.ph. m (sd)	exp.ph m (sd)	st.op. m (sd)	exp.op m (sd)
MPOSS1; KBS should give multiple diagnoses, where possible	1.53 (0.64)	1.53 (0.68)	1.06 (1.01)	1.19 (1.00)
NPUT2; KBS should also be able to assist the user with systematic data entry	1.33 (0.79)	1.48 (0.68)	1.19 (0.70)	1.33 (0.72)
VKNOW3; The user should be able to check the knowledge contained in a KBS in advance	1.18 (0.89)	0.94 (1.04)	1.11 (0.94)	1.13 (0.96)
ADAPT4; The user should be able to adapt the knowledge contained in a KBS	1.18 (0.99)	0.68 (1.49)	0.47 (1.33)	0.32 (1.52)
LITER5; KBS should accompany their advice with references to literature	1.45 (0.59)	1.59 (0.68)	0.30 (1.20)	-0.29 (1.36)
HIS6; KBS should be in contact with a Hospital Information System/Management Information System	0.97 (0.89)	0.79 (1.21)	0.53 (1.05)	-0.13 (1.14)
ADMIN7; In addition to providing a diagnosis and treatment planning, a KBS should perform administrative tasks	0.64 (1.21)	0.97 (1.23)	1.12 (1.12)	1.08 (1.07)

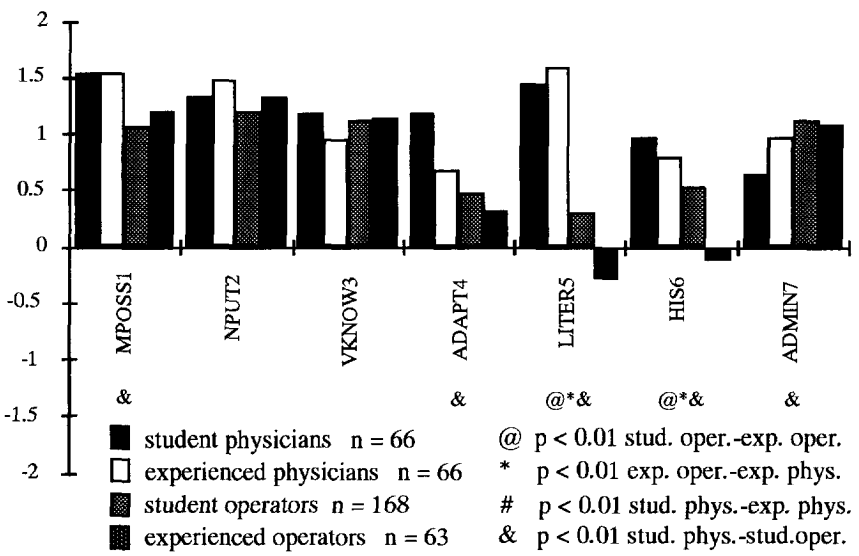


Figure 6.5. User attitudes towards additional features.

the necessary observations for diagnosing a certain disease. This is surprising since Shortliffe (1989) found that physicians recognise data entry as being a major barrier to the effective use of computers in clinical practice (see Section 6.2.3).

6.4.5. USER ATTITUDES TOWARDS ADDITIONAL FEATURES

Physicians strongly agree with the statement that a knowledge based system should give multiple diagnoses where possible, whereas operators somewhat agree with this statement. This can be seen in Figure 6.5, where this statement corresponds to item MPOSS1. All four groups somewhat agree with the requirement that a knowledge based system should also help the user with systematic data entry (NPUT2), and that the user should be able to check the knowledge contained in the system before actual use (VKNOW3).

Physicians somewhat agree with the statement that the user should be able to adapt the knowledge contained in the knowledge base (ADAPT4). Student physicians feel more strongly about this issue than the experienced physicians, whereas operators are more indifferent towards this statement. The disparity is possibly caused by the fact that in medicine new ways to treat patients are often introduced, whereas in the domain of process industry changes will only occur when new equipment has been installed. Furthermore, operators are used to machines which are placed and cannot be altered. Physicians somewhat agree that a machine advisor should accompany its advice with references to literature (LITER5). This item shows a significant difference between physicians and operators. In medicine the literature is often consulted in difficult cases, whereas in this respect no parallel can be drawn with process operating practice.

Concerning the fact whether a knowledge based system should be in contact with a management/hospital information system (HIS6), significant differences in opinion were found between experienced operators and the other groups, and also between student physicians and student operators. During discussions which were held after the introductory explanation, some experienced operators were concerned that all their actions would be stored in the management information system and could be misused by their superiors.

Finally, operators somewhat agreed with the statement that a knowledge based system should perform administrative tasks in addition to providing a diagnosis and treatment planning (ADMIN7), whereas especially the student physicians felt slightly less strongly towards this statement. Most routine administrative tasks in medicine will be performed by secretaries, so there may not such a clear advantage to a system that takes over this task. To operators,

the advantage may be more clear. One of their most time-consuming tasks is to log the values of the most important process-variables. In principle, this task can be automated, but they have to do it by hand to make sure that they periodically observe the state of the process they are supervising.

6.4.6. RESULTS OF THE OPEN QUESTION

At the end of the questionnaire an open question was added to investigate the participants' opinion regarding the use of knowledge based systems in their work. The answers given by the participants have been classified into four different categories: positive, maybe, negative and no answer. The general results may be seen in Figure 6.6. The percentage of respondents in each category is shown along the vertical axis of the diagram.

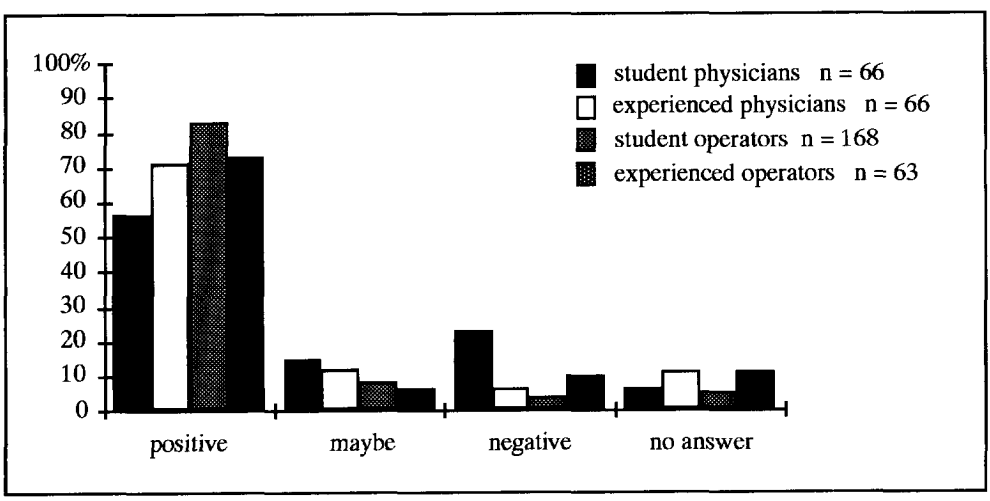


Figure 6.6. Results of the open question concerning general attitude towards knowledge based system.

It can be seen that the process-operators are on the whole slightly more positively inclined towards the use of knowledge based systems than the physicians. A difference may be seen when comparing the experienced and inexperienced physicians on the one hand, and the experienced and inexperienced operators on the other hand. A larger percentage of experienced physicians answered the question in a positive manner than student physicians, whereas a larger percentage of student operators answered the question in a

positive way than the experienced operators. The former seems to be in contrast with the idea that younger people are more used to computers and will therefore be more prepared to use knowledge based systems in their job. However, this question probably reflects their opinion on whether they think it is possible to develop such a system and may also concern their opinion regarding the need for assistance in the work they will be doing, as well as their attitude towards computer use.

A large percentage of participants (68%) that answered the open question in a positive way elaborated on the answer. They often mentioned that in their opinion the field seemed too complex to be able to develop a knowledge based system. They also mentioned certain requirements which would have to be met by these systems in order for them to accept knowledge based systems. Very often it was stated that the user should remain responsible, and the system should be used as an aid or a second opinion.

Another requirement which was often mentioned by the operators was that the system should not lead to a loss of jobs. In their opinion, loss of jobs could be prevented by creating advisory systems which cannot automatically interfere with the process. Other much less frequent remarks were that a knowledge based system should increase the quality of the outcome of the task, it should never make a mistake, and it should not take the initiative.

When the participants had a negative opinion, this was usually accompanied by a number of reasons. Some of the reasons which were mentioned, were the fact that it would be better to consult colleagues, a knowledge based system would be too time consuming, it would be too expensive, the user's knowledge would deteriorate, and there was some doubt as to whether such a system would work. Some participants answered the question very negatively. Participants stated that such a system is dangerous, encourages laziness, that users could become dependent on the system, users could become redundant, and additional automation leads to more disturbances in the industrial process. It was also mentioned that these systems are not necessary, that there may not be enough time to consult such a system, and it will overload the process-operator with even more information.

6.5. Conclusions

Although many knowledge based systems have been built, not very many are used in actual practice. In literature, several solutions to this acceptance problem are proposed. Since no evaluations of knowledge based systems based

on these new concepts have been carried out as yet, it is not clear which solutions will lead to knowledge based systems being used in actual practice. Therefore, user requirements regarding knowledge based systems have been investigated by means of a questionnaire. These requirements provide valuable information about the usefulness of new concepts in knowledge based system design. In this study, user requirements have been investigated among physicians and process-operators, both experienced and inexperienced.

The results show that it is very important to the participants that the introduction of a knowledge based system must not lead to changes in the current working practice. They are strongly in favour of the requirement that introduction of machine advisors should not lead to a shift of responsibility from the human to the machine. The system should be a subordinate of the human, therefore it is important for the participants to understand how it works, in order to judge the correctness of its advice. Another important requirement is that knowledge based systems should demand little effort from users to learn or use. This study also shows that the recommendations made by Teach and Shortliffe (1981) for knowledge based system design are still valid, although the importance of responsibility and the role of a machine advisor as a subordinate was not addressed in their study.

The participants consider their current knowledge to be necessary for judging the correctness of the advice given by the system, and therefore they somewhat disagree with the requirement that a knowledge based system should significantly reduce the amount of technical knowledge that a user must learn and remember. Moreover, their current knowledge is part of their professional status, which they do not want to lose.

A remarkable outcome of the study is that the different groups of subjects often shared the same opinion about the statements presented to them. Differences in opinion between operators and physicians are most often found with respect to those issues which concern a significant change in current working practice for one group, whereas due to the nature of the domain it will not affect the other group to the same extent.

Examples are the requirements that a knowledge based system should accompany its advice with references to the literature and that it should provide multiple diagnoses where possible. The physicians more strongly agree with these requirements than the operators. On the other hand, operators attach more importance to the requirement that a knowledge based system should improve cost efficiency and that a system should contain knowledge about a complete subject (specialism) and not just part of it. In our opinion, this

agreement among the participants indicates that other groups of users may have similar requirements regarding knowledge based systems.

The results of this investigation support the idea that in order to develop a knowledge based system which may have better possibilities for being used in actual practice, it is important to follow a problem driven approach. This implies that it is necessary to first establish the factors which determine competence and incompetence in the domain (Woods, 1986). The outcome of this procedure can subsequently be used to decide which tasks should be performed by the machine, in relation to the tasks which are performed by the user. The results show that a knowledge based system should be designed as a subordinate to the human, and the user must understand how it works. This requires a system design which facilitates the user in building up an adequate internal representation of the reasoning process of the knowledge based system.

These conclusions have some implications for knowledge based systems, which have been designed to tackle the complete problem of diagnosis, such as PLEXUS. The fact that these systems aim to solve the complete problem also influences aspects of validation. As was discussed in Chapter 4, validation of knowledge based systems is fraught with difficulties, such as lack of test cases and difficulty in judging the results. A cause of this is the complexity of the problems which these systems aim to solve. A less ambitious system design which incorporates cooperative problem solving may reduce some of the problems related to validation.

7 Conclusions

7.1. Validation

The validation of medical knowledge based system has been receiving more attention recently. It is recognised that knowledge based systems should be thoroughly validated before they can be used in actual practice. However, the number of actual validation studies which has been performed and described is relatively limited.

In the literature, the various kinds of validation are defined in different ways by different authors. In practice, three different types of validation can be distinguished. Each of these will briefly be summarised, and possible limitations and recommendations which have become clear from the evaluation of PLEXUS will be discussed. This will lead to general recommendations regarding the design and validation of medical knowledge based systems.

- Verification: is an activity which should ensure that product of one phase of the life cycle is consistent with itself and with the source from which it has been derived. As such, verification should be carried out at the end of each phase of development.

In practice, verification is usually performed on the knowledge based system itself, by carrying out checks on the knowledge base. These checks are more elaborate than the syntactic checks which are used in conventional programs. In the literature, methods have been described for checking rule bases for completeness and consistency (Nguyen *et. al.,* 1987). Ginsberg (1987) described a method which allows rule based systems to be analysed over complete inference chains.

Verification methods seem to be low cost methods to check knowledge bases. At the time PLEXUS was developed these methods were usually not applied. During the laboratory evaluation of PLEXUS, for instance, a cycle in the rules became clear. This could have been identified prior to the laboratory evaluation by performing more extensive verification. It is recommended to apply verification methods to any serious knowledge based system from the start

of the development. The majority of verification methods which have been found in the literature are unfortunately limited to rule based representation methods.

Verification should not only include the knowledge base of a system, but should be related to the complete development life cycle, and should, for instance, include the specifications and inference processes.

- Dynamic validation: is an activity which should ensure that the product at the end of each phase of development process complies with one of the formal requirements which it was intended to satisfy.

In practice, dynamic validation usually consists of investigating the behaviour of a knowledge based system by means of test cases. However, there are many areas in medicine, such as the domain of brachial plexus injuries, where insufficient actual test cases are available to thoroughly test the system. Furthermore, experts may have to be involved in establishing test cases and/or judging the outputs. This may severely restrict the number of cases which can be tested, due to the time involved. In order to overcome these problems, generated test cases (see, for example, Shwe *et al.*, 1989) may be used.

Test case generation has received little attention in the literature, but it deserves more research and should become more important in the future. Test case generation has also been applied in the validation of PLEXUS. Although care should be taken to develop the test case generator independently of the system, the method seems to provide useful information. It may be concluded that validation on a large range of test cases, either using real or generated test cases, is a necessary activity.

- Evaluation: is an activity which should ensure that the product at the end of each phase of development complies with one of the pseudo-formal requirements which it was intended to satisfy.

At the end of the development life cycle, after thorough verification, dynamic validation and evaluation have been carried out, two different empirical evaluation processes may be distinguished: laboratory evaluation and field evaluation.

- Laboratory evaluation: empirical evaluation of the knowledge based system in the laboratory environment.

Laboratory evaluation is the validation topic which has been discussed most often in the literature. Laboratory evaluation should provide additional evidence that a system is safe and potentially useful.

Most laboratory evaluation studies only involve the investigation of the advice produced by the knowledge based system. However, in addition to investigations of the system, the complete human-machine system should be studied in order to determine whether the system may be potentially useful.

- Field evaluation: empirical evaluation of the human-machine system in the target environment.

Field evaluations are carried out after a knowledge based system has been shown to be safe and potentially useful in a laboratory evaluation. Field evaluation encompasses investigation of a large number of aspects. The specific issues that are addressed depend on the nature of the system, on the domain and on the clinical role of the system. These issues include: the investigation of the impact of the system on physician actions, on patient care, and on health care processes, a cost benefit analysis, the examination of subjective reactions and the investigation of system use (Miller, 1986).

Only few field evaluations have been reported in the literature. It can be seen that many investigators are very positive after a laboratory evaluation. However, in many cases no further evaluations of the systems are reported. It will usually not possible to compare evaluation results of different systems, because domain characteristics such as domain size, system characteristics and evaluation designs have too much influence on the evaluation results. This is true for laboratory evaluations as well as for field evaluations.

Most of the work described in this thesis concerns the last two kinds of validation which were mentioned, laboratory evaluation and field evaluation. Amongst other aspects, the performance of PLEXUS was addressed during these evaluations. In this context the term performance is related to the quality of the accomplishments of the human-machine system, rather than implying technical performance measures.

The performance evaluation of PLEXUS was carried out according to a framework of performance evaluation design which was introduced in Chapter 2 and which is shown in Figure 2.1. The discussion of this framework led to an analysis of ways in which medical knowledge based systems can be evaluated in theory, taking into account some of the characteristics of medical knowledge based systems. These characteristics may raise a number of difficulties in the design of an evaluation. The difficulties which have been described in the literature include problems such as the specification of a standard, determining which variables should be measured, and various potential sources of bias and confounding.

In practice, some of the aspects of the framework were influenced by a number of further limitations, which had not been encountered in the literature and which became clear during the evaluation of PLEXUS. Since many of the problems will apply to medical knowledge based systems in general, these limitations will be discussed.

7.1.1. LIMITATIONS WHICH MAY ARISE DURING PERFORMANCE EVALUATIONS

The aspects of the framework which are influenced by limitations that became clear during practical evaluation studies will be regarded below.

Selecting test input
In many evaluation studies, the test input is required to be representative for the target situation. Therefore, it will be necessary to know the distribution of types of cases. However, very often exact knowledge of the distribution of cases will not be available. The best way to obtain representative cases is probably to use cases from the target environment. However, when retrospective test cases are required for an evaluation it may be difficult to obtain adequate test cases. In the target environment, the retrospective cases may not have been documented well enough to be used for evaluation. This means that one may have to resort to experts for supplying the cases, thereby diminishing the representativeness of the cases.

In some domains the number of test cases available may be restricted. If retrospective cases are used and physicians are asked to provide test cases, the time they are prepared to spend on retrieving the data will usually be limited. Furthermore, in some domains the number of prospective cases will also be limited, as was found during the evaluation of PLEXUS.

A limited number of test cases may mean that it will only be possible to draw statistically significant conclusions if there are very large differences between the measured values. This was true for the laboratory evaluation of PLEXUS. In the field evaluation, the number of test cases was too limited to perform quantitative analysis and only qualitative analysis could be carried out.

When using retrospective test cases, the physicians involved in the evaluation may have to form an opinion about the patient using only data which are available on paper. However, this does not resemble the normal situation. Since in the normal situation, the physicians would have seen the actual patient. This problem also arose during the laboratory evaluation of PLEXUS.

A further problem in the evaluation of PLEXUS was that not all relevant information was included in the patient files, which meant that neither the system nor the physician had all the relevant information.

It is necessary to provide all the physicians involved in the evaluation with the information which resembles the normal situation as closely as possible.

Specifying physicians to test against

It is necessary to decide on the level of experience of the physicians to test against, and to decide on the number of people who should be involved in the investigation. If the system is tested against experts, there is no certainty as to how good the experts are. If potential users are involved in the investigation, it is necessary to involve a range of potential users.

Since the agreement between physicians is not known in advance, it will be difficult to decide on the number of physicians who should be involved in the evaluation. A further problem which arises in practice is that the number of physicians involved will usually be restricted due to practical limitations.

Comparison

The final aim of a field evaluation may be to determine whether final patient outcome would improve due to the use of a knowledge based system. This requires final patient outcome to be measured. However, as could be seen in the field evaluation of PLEXUS, this may present some problems. In some domains it may take a long time before the final outcome becomes known. Furthermore, when the number of cases is limited, differences between patients will become an important factor in patient outcome. Therefore, in the evaluation of PLEXUS, the variables which were measured were the diagnoses and treatment plans.

A number of unforeseen situations may arise when judging the results of an evaluation study. The results may be judged by performing direct comparison of the outputs. The outputs may consist of, for instance, diagnoses and treatment plans. However, direct comparison may not be as easy as it seems.

In order to perform direct comparison it will be necessary to use a uniform terminology. However, in some domains the terminology used by physicians may not be uniform, and standardisation of the terminology will introduce subjectivity into the measurement.

A further problem arises in domains where one diagnosis consists of multiple answers. If direct comparison between various diagnoses is carried out it will be difficult to determine the agreement and disagreement between the diagnoses.

For any single answer it is often only possible to determine whether this answer is completely the same as another answer or whether it is different. It is

usuallly not possible to determine the extent of the agreement between answers, for this will require in depth domain knowledge and will probably introduce different problems, such as subjectivity.

An alternative to a direct comparison is a blind comparison by experts. For experts it will be possible to regard different degrees of correctness of an answer. However, blind expert evaluation of outputs will also present a number of difficulties. In a blind evaluation it is necessary to introduce a scoring scale for correctness of the outputs. It may be difficult to choose an appropriate scoring scale. Furthermore, the results will always be subjective.

Direct comparison and blind judgements can be seen as complementary to each other. A direct comparison may provide information about the way the answers are built up and may identify parts of the system which require improvement. The blind comparison will give more clinically relevant information about the performance of the system.

Analysis of the results
The choice of methods of analysis which can be used in a specific study depends on a number of factors, such as whether enough data is available for statistical analysis and whether multiple answers are possible for a case.

Furthermore, the statistical methods which may be used also depend on the distribution of the results. It is not always possible to use the parametric statistics, such as the t-test, since this requires a normal distribution to be present. For instance, in the laboratory evaluation of PLEXUS nonparametric statistics had to be used.

7.1.2. CONCLUSIONS REGARDING PERFORMANCE EVALUATION OF MEDICAL KNOWLEDGE BASED SYSTEMS

As demonstrated above, evaluations of medical knowledge based systems present many problems. Therefore, it is often difficult to obtain solid conclusions from these studies with respect to the goal of the evaluation study. However, the evaluation of PLEXUS shows that very worthwhile results may still be obtained from these studies. On the other hand, these studies are very time-consuming, and due to the complexity of the models implemented in these systems it is not possible to address every aspect of the system in these investigations.

From the analysis of the validation of medical knowledge based systems, it has become clear that most of the literature focuses on the system only rather than on

the complete human-machine system. The complete human-machine system is usually only addressed during the field evaluation. This is also true for the knowledge based system PLEXUS. However, for any knowledge based system it is necessary to study the complete human-machine system in the laboratory environment. Some of the advantages of this approach are the following:

- many of the problems and suggestions which would result from a field evaluation may be identified and corrected prior to the field evaluation,
- the acceptability of systems which have undergone human-machine evaluation in the laboratory may be better when these systems are finally tested in the field,
- in some domains, laboratory experiments of human-machine systems will allow testing of more scenarios than would become available during a field evaluation.

However, the human-machine investigations will also require careful design, and the problems which arise during other evaluation studies will also be present in these investigations.

7.2. Changing ideas about knowledge based decision support

As explained above, most researchers tend to focus on the system instead of on the human-machine system. This emphasis on the system is inherent in the design paradigm used in most present-day knowledge based systems, where the computer is designed to be a machine expert. The user collects the data and implements the actions for the machine, and the machine has the role of problem solver. As was mentioned in Chapter 6, this role has been questioned by a number of researchers (for example, Woods, 1986; Miller and Masarie, 1990; Rossi-Mori *et al.*, 1990). Roth *et al.* (1987) investigated the performance of users of a technical knowledge based system designed according to the traditional expert paradigm. Results of the study revealed that, contrary to the implicit assumptions in the design paradigm, technicians actively and substantially contributed to the diagnostic process. The more the human functioned as a passive data gatherer for the machine, the more joint performance was degraded. Active human participation led to more successful and rapid solutions. However, the machine expert not only failed to support an active human role, it actually retarded technicians from taking or carrying out an active role.

As opposed to designing knowledge based systems as machine experts, knowledge based systems may also designed according to the 'cognitive-tool-as-an-instrument' perspective (Woods *et al.*, 1990), where the system does not make or recommend solutions, but assists the user in the process of reaching a

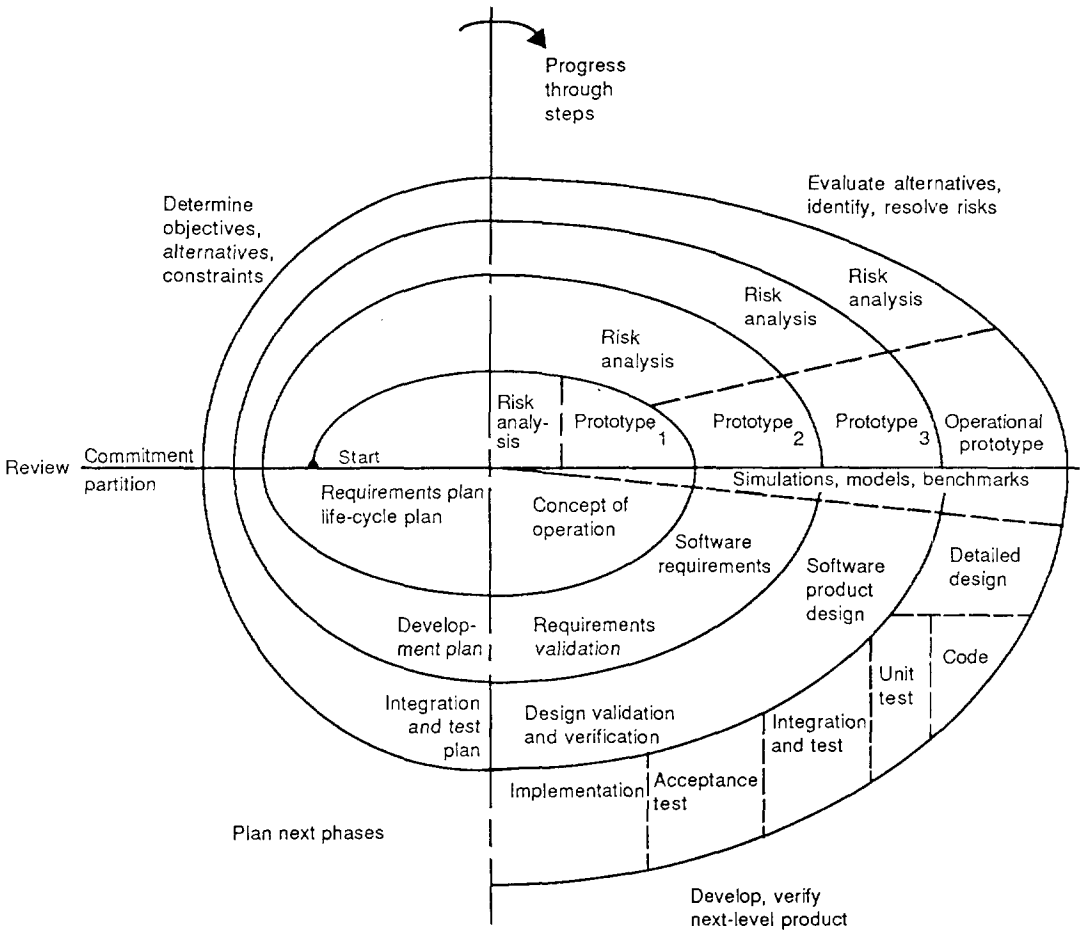


Figure 7.1. Spiral model of the software process (adapted from Boehm, 1988).

decision. This could, for instance, imply that the system presents the information in such a way that the user can determine a diagnosis.

In the field evaluation of PLEXUS, some physicians indicated that, besides providing diagnostic and treatment planning advice, the possibilities the system has for providing assistance in determining the examinations which should be performed for a certain patient should be extended. This would mean that the system would offer more assistance in the process of reaching a decision.

7.3. The design and validation of medical knowledge based systems

The fact that the role which medical knowledge based systems often are designed to play, may not be the most appropriate role, could indicate that the design, development and validation procedures which are commonly followed are incomplete. According to van der Spek (1991), one of the reasons that many systems do not answer expectations, is that during the design and development precisely defined methodologies are still little used at present. These methodologies involve more than methods for analysing, modelling and representing knowledge. Although, some models still tend to concentrate on these aspects.

The need to design quality into software from the beginning of the development life cycle is now generally accepted. The adoption of life cycle methodologies has helped to improve the quality of conventional software and promises to do the same for knowledge based systems (Fox, 1993).

By looking at a model of the software design and development process it will be shown that many aspects have to be taken into account in the design and development of a knowledge based system. In order to illustrate the points that will be made, the spiral model of software design (Boehm, 1988) will be taken as an example and will be explained below. However, the same could be said using different models.

The model that will be discussed is a model of software development and not of knowledge based system development. This means that the special techniques for, for instance, knowledge acquisition and formalisation are not taken into account. Van der Spek (1991) mentions several methodologies, such as SKE, which do take this into account.

However, the main aim of this section is to draw attention to the aspects of the development methodology which do not relate to knowledge analysis, knowledge acquisition etc., and the model that will be discussed has some

interesting aspects which will be described in detail below, including the evaluation of alternatives, and the validation at the end of each cycle.

The spiral model of the software process has been evolving for several years (Boehm, 1988). An adapted diagram of this model is shown in Figure 7.1. The spiral model consists of cycles that all address the same sequence of steps. The angular dimension represents the progress made in completing each cycle of the spiral. Each cycle begins with the identification of the objectives and the alternative means of achieving the objectives. This is shown in the upper left hand quadrant of Figure 7.1.

The alternatives are then evaluated relative to the objectives. Frequently this process will identify areas of uncertainty that are significant sources of project risk. The next step should then involve the formulation of a strategy for resolving the sources of risk. This may involve prototyping, simulation etc. Each level of software specification is followed by a validation step and the preparation of plans for the succeeding cycle. Each cycle is completed by a review involving the primary people or organisations concerned with the product.

Fox (1993) suggests an approach in which a safety life cycle is executed in parallel with the quality life cycle.

When planning to develop a system for providing assistance in performing a certain task, this requires the identification of objectives, the identification of alternatives, and a feasibility study. This should be taken very seriously and it is imperative to identify whether a real need for assistance does exist. The design of a knowledge based system should only be started if there is a commitment in the application domain. The development of knowledge based systems is too costly and time-consuming to undertake as a technology push. In the past, many knowledge based systems were developed because investigators wanted to apply the technology, rather than to solve a problem. Woods (1986) emphasises a problem driven approach, rather than a technology driven approach.

In order to know which parts of a task are to be performed by the human and which parts should be carried out by the machine, it is necessary to perform task analysis. It should be kept in mind that human and machine should cooperate. This is in contrast with the conventional paradigm for knowledge based system design which was discussed above. A different division of tasks may also lead to a change in emphasis of the validation methods which are appropriate.

Possible alternatives to a proposed solution should also be investigated. In some situations for which knowledge based systems have been developed, this would

not have been necessary because alternative means of assistance would have been just as adequate, if not more so.

Knowledge based system requirements should be explicitly specified. Otherwise, the requirements cannot be validated and the requirements cannot be used to validate products arising from succeeding cycles.

The interaction between the human and machine has not received enough attention in the knowledge engineering literature. The interaction has usually been a part of the system which has been added on, rather than being integrated into the design and development. It is necessary, however, to take into account the interaction between the user and the system from the start of the project. Furthermore, very often no potential users are involved in the development of such systems until a prototype system has been developed. In order to draw up an adequate requirements specification and to develop a system with the desired functionality, users should be involved during all stages of design and development.

7.3.1. CONCLUDING REMARKS CONCERNING THE DESIGN AND DEVELOPMENT OF MEDICAL KNOWLEDGE BASED SYSTEMS

From the above, it can be concluded that the design and development of knowledge based systems should proceed according to a methodology for system development. Besides addressing the analysis, modelling and representation of knowledge, it is of the utmost importance:

- to identify where and whether there is a real need for assistance,
- to investigate which parts of a task should be performed by the human and which by the machine,
- to identify alternatives,
- to develop an adequate requirements specification,
- to specifically address safety issues,
- to address human-machine interaction during all stages of the project.

This requires a multidisciplinary approach and user involvement from the beginning of the project.

Apart from the validation procedures which are described in the medical knowledge engineering literature, it is necessary:

- to evaluate alternatives,
- to validate requirements,
- to evaluate the human-machine system continually during the development.

Design and validation are closely related. Validation should be fully integrated in the design and development of a knowledge based system. By concentrating on the complete human-machine system and paying attention to all aspects of the design, development and validation life cycle, as opposed to emphasising the development of prototypes, more usable and acceptable systems should result.

7.4. PLEXUS

The analysis described above concerns general recommendations for the design and validation of medical knowledge based systems. This research was motivated by the knowledge based system PLEXUS and the objectives of validating the system and investigating the system's applicability in actual practice.

PLEXUS has been evaluated in a laboratory evaluation involving experts in the domain of brachial plexus injuries, and has been tested clinically in four hospitals in The Netherlands. The evaluation of PLEXUS has shown that:

- The system has a good performance, although certain areas have to be improved.
- The functionality of the system should be extended to include a more direct communication with the users, and the system's ability to assist the user in data gathering and data entry should be improved.
- It costs some time and effort to use the system.
- A number of physicians indicated that they would use the system if it was generally available.
- The number of brachial plexus injuries was lower than expected.

The answer to the question concerning the applicability of PLEXUS in actual practice remains largely unanswered, since a number of physicians indicated that they would use the system if it was generally available, whereas during the field evaluation the system was not used as readily as might have been expected. The acceptability of the system will require further investigation, since the incorporation of some of the suggestions resulting from the evaluation studies will require significant alterations to the system.

References

Adams, I.D. *et al.* (1986). Computer aided diagnosis of acute abdominal pain: a multicentre study. *British Medical Journal*, Vol. 293, pp. 800-804.

Adelman, L. (1991). Experiments, quasi-experiments, and case-studies: A review of empirical methods for evaluating decision support systems. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 21, No. 2, pp. 293-301.

Adlassnig, K.P. and Scheithauer,W. (1989). Performance evaluation of medical expert systems using ROC curves. *Computers and Biomedical Research*, Vol. 22, pp. 297-313.

Aikins, J.S.; Kunz, J.C.; Shortliffe, E.H. and Fallat, R.J. (1983). PUFF: An expert system for interpretation of pulmonary function data. *Computers and Biomedical Research*, Vol. 16, pp. 199-208.

Alnot, J.Y. and Narakas, A. (eds.) (1989). Les Paralysies du Plexus Brachial. Expansion Scientifique Française, Paris.

Alonso-Betanzos, A.; Devoe, L.D.; Castillo, R.A.; Moret-Bonillo, V.; Hernández-Sande, C. and Searle, N.S. (1989). FOETOS in clinical practice: A retrospective analysis of its performance. *Artificial Intelligence in Medicine*, Vol. 1, pp. 93-99.

Ayel, M. and Laurent, J-P. (eds.) (1991). Validation, Verification and Test of Knowledge-Based Systems. John Wiley & Sons, Chichester, England, ISBN 0-471-93018-0.

Bainbridge, L. (1987). Ironies of Automation. In: J. Rasmussen, K. Duncan and J. Leplat, (eds.), New Technology and Human Error, John Wiley and Sons, Chichester, pp. 271-283.

Bankowitz, R.A.; McNeil, M.A.; Challinor, S.M.; Parker, R.C.; Kapoor, W.N. and Miller, R.A. (1989). A computer-assisted medical diagnostic consultation service: Implementation and prospective evaluation of a prototype. *Annals of Internal Medicine*, Vol. 110, pp. 824-832.

Banks, G. and Weimer, B. (1984). Symbolic coordinate anatomy for neurology (SCAN). *Journal of Medical Systems*, Vol. 8, No. 3, pp. 157-162.

Bell, M.Z. (1985). Why expert systems fail. *Journal of the Operational Research Society*, Vol. 36, pp. 613- 619.

Bemmel, J.H. van (1988). Systems evaluation for the health of all. In: R. Hansen *et al.* (eds.), Lecture Notes in Medical Informatics, Springer Verlag, Vol. 35, pp. 27-34.

Bernelot Moens, H.J. and Korst, J.K. van der (1991). Measuring performance of a Bayesian decision support system for the diagnosis of rheumatic disorders. In: M. Stefanelli, A. Hasman, M. Fieschi, and J. Talmon (eds.), Proceedings of the 3rd European Conference on Artificial Intelligence in Medicine, Maastricht, pp. 150-159.

Bischoff, A. (1975). The intimate anatomy of peripheral nerves. In: J. Michon, E. Moberg (eds.), Traumatic Nerve Lesions of the Upper Limb. Churchill Livingstone, Edinburgh.

Boehm, B.W. (1988). A spiral model of software development and enhancement. *Computer*, May, pp. 61-72.

Bramer, M.A. (1984). A survey and critical review of expert systems research. In: D. Michie (ed.), Introductory Readings in Expert Systems, Gordon and Breach Science Publishers, pp. 3-29.

Buchanan, B.G. (1986). Expert systems: Working systems and the research literature. *Expert Systems*, Vol. 3, No. 1, pp. 32-51.

Burge, P. and Todd, B. (1990). Computer-aided localization of peripheral nerve lesions. Manuscript, 12 pp.

Catanzarite, V.A.; Greenburg, A.G. and Bremermann, H.J. (1981). Computer assisted diagnosis and computer consultation in neurology: Preliminary testing of diagnostic accuracy for the NEUROLOGIST system. *International Journal of Neuroscience*, Vol. 13, pp. 43-54.

Chandrasekaran, B. (1983). On evaluating AI systems for medical diagnosis. *The AI Magazine*, Summer, pp. 34-37.

Clancey, W.L.; Shortliffe, E.H. (1984). Readings in Medical Artificial Intelligence: The First Decade. Addison Wesley Publishing Co., Reading, Massachusetts.

Coene, L.N.J.E.M. (1985). Axillary Nerve Lesions and Associated Injuries. Ph.D. Thesis, Leiden University, The Netherlands, ISBN 90-9001139-0.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37-46.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, Vol. 70, No. 4, pp. 213-220.

Cohen, J. (1977). Statistical Power Analysis for the Behavioral Sciences. Academic Press, New York.

Coombs, M. and Alty, J. (1984). Expert systems: An alternative paradigm. *International Journal of Man-Machine Studies*, Vol. 20, pp. 21-43.

Daalen, C. van (1988). Factors Influencing Medical Expert System Acceptance. Report N-284, Laboratory for Measurement and Control, Delft University of Technology, 47 p., ISBN 90-370-0016-9.

Daalen, C. van and Jaspers, R.B.M. (1989). Explanation improvement to enhance acceptance of the PLEXUS system. In: J. Hunter *et al.* (eds.), Lecture Notes in Medical Informatics, Vol. 38, Proceedings of the Second European Conference on Artificial Intelligence in Medicine, Springer-Verlag, Berlin, pp. 286-295.

Daalen, C. van (1991). User guide for the PLEXUS user interface (in Dutch). Report N-375, Laboratory for Measurement and Control, Delft University of Technology.

Daalen, C. van (1992a). Field evaluation of medical knowledge based systems. The medical system PLEXUS. In: H.G. Stassen (ed.), Analysis, Design and Evaluation of Man-Machine Systems 1992, IFAC, Pergamon Press, New York, pp. 275-282.

Daalen, C. van (1992b). Clinical performance evaluation of a medical knowledge based system (in Dutch). In: H. de Swaan Arons, H. Koppelaar, E.J.H. Kerckhoffs (eds.), Proceedings of the Dutch Conference on Artificial Intelligence (NAIC), Delft University Press, pp. 311-319.

Daalen, C. van; Stassen, H.G.; Thomeer, R.T.W.M. and Slooff, A.C.J. (1993). Computer assisted diagnosis and treatment planning of brachial plexus injuries. *Clinical Neurology and Neurosurgery*, Vol. 95, pp. S50-S55.

Dombal, F.T. de and Horrocks, J.C. (1978). Use of receiver operating characterstic (ROC) curves to evaluate computer confidence threshold and clinical performance in the diagnosis of appendicitis. *Methods of Information in Medicine*, Vol. 17, No. 3, pp. 157-161.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, Vol. 37, No. 1, pp. 36-48.

Fieschi, M. (1990). Towards validation of expert systems as medical decision aids. *International Journal of Biomedical Computing*, Vol. 26, pp. 93-108.

First, M.B.; Weimer, B.J.; McLinden, S. and Miller, R.A. (1982). LOCALIZE: Computer-assisted localization of peripheral nervous system lesions. *Computers and Biomedical Research*, Vol. 15, pp. 525-543.

Fisher, W.S. (1990). Computer-aided intelligence: Application of an expert system to brachial plexus injuries. *Neurosurgery*, Vol. 27, No. 5, pp. 837-843.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions. John Wiley and Sons.

Fox, J. (1993). On the soundness and safety of expert systems. *Artificial Intelligence in Medicine*, Vol. 5, pp. 1-21.

François, P.; Cremilleux, B.; Robert, C. and Demongeot, J. (1992). MENINGE: A medical consulting system for child's meningitis. Study on a series of consecutive cases. *Artificial Intelligence in Medicine*, Vol. 4, pp. 281-292.

Gaschnig, J.; Klahr, P.; Pople, H.; Shortliffe, E. and Terry, A. (1983). Evaluation of expert systems: Issues and case studies. In: F. Hayes-Roth, D.A. Waterman, D.B. Lenat (eds.), Building Expert Systems, Addison-Wesley, pp. 241-280.

Ginsberg, A. (1988). Knowledge base reduction: A new approach to checking knowledge bases for inconsistency and redundancy. In: Proceedings of the 7th National Conference on AI (AAAI), Vol. 2, pp. 585-589.

Gjørup, T. (1988). The Kappa coefficient and the prevalence of a diagnosis. *Methods of Information in Medicine*, Vol. 27, pp. 184-186.

Gorry, G.A.; Silverman, H. and Pauker, S.G. (1978). Capturing clinical expertise. A computer program that considers clinical responses to digitalis. *The American Journal of Medicine*, Vol. 64, pp. 452-460.

Green, C.J.R. and Keyes, M.M. (1987). Verification and validation of expert systems. In: Proceedings of the Western Conference on Expert Systems, Anaheim, IEEE Computer Society, pp. 38-43.

Grolman, J.R.D. (1989). A User Interface for the Medical Expert System PLEXUS (in Dutch). M.Sc. Thesis, Faculty of Industrial Design, Delft University of Technology.

Guiteras, D.J. (1989). The Lesion Game: A special communication. *Physical Therapy*, Vol. 69, No. 10, pp. 858-862.

Gupta , U.G. (ed.) (1990). Validation and Verification of Knowledge-Based Systems. The IEEE Computer Society Press, Los Alamitos, CA, ISBN 0-8186-8995-1.

Haar, F. ter (1989). A New Implementation of PLEXUS Incorporating Structural Knowledge (in Dutch). Report N-315, Laboratory for Measurement and Control, Delft University of Technology.

Haberman, H.F. *et al.* (1985). DIAG: A computer-assisted dermatologic diagnostic system-clinical experience and insight. *Journal of the American Academy of Dermatology*, Vol. 12, No. 1, Part 1, pp. 132-143.

Hart, A. (1991). The role of decision-support systems in medicine. Conference Report. *Expert Systems*, Vol. 8, pp. 286-287.

Haymaker, W. and Woodhall, B. (1945). Peripheral Nerve Injuries. W.B. Saunders, Philadelphia.

Heerebeek, W.P.M. van (1991). LORETREAT: Representing the Knowledge About the Treatment of Brachial Plexus Injuries in an Object-Oriented Formalism (in Dutch). M.Sc. Thesis, Report A-549, Laboratory for Measurement and Control, Delft University of Technology.

Hertzberg, T.M.; Tremblay, G.F. and Lam, C.F. (1987). Computer-assisted localization of nervous system injuries. *Computers and Biomedical Research*, Vol. 20, pp. 489-496.

Hickam, D.H.; Shortliffe, E.H.; Bischoff, M.B.; Scott, A.C. and Jacobs, C.D. (1985). The treatment advice of a computer-based cancer chemotherapy protocol advisor. *Annals of Internal Medicine*, Vol. 103, pp. 928-936.

Hilden, J.; Habbema, J.D.F. and Bjerregaard, B. (1978). The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods of Information in Medicine*, Vol. 17, No. 4, pp. 227-237.

Howard, S. and Murray, D.M. (1987). An outline of techniques for evaluating the human-computer interface. In: P. Zunde and J.C. Agrawal (eds.), Empirical Foundations of Information and Software Science IV; Empirical Methods of Evaluation of Man-Machine Interfaces, Plenum Press, New York, pp. 177-185.

Indurkhya, N. and Weiss, S.M. (1989). Models for measuring performance of medical expert systems. *Artificial Intelligence in Medicine*, Vol. 1, pp. 61-70.

Jackson, P. (1986). Introduction to Expert Systems. Addison-Wesley, Wokingham.

Jaspers, R.B.M.; Louw, C.J.M. de; Lunteren, A. van and Stassen, H.G. (1982). Evaluation of the rehabilitation process of patients with a brachial plexus lesion (in Dutch). Proceedings of the Boerhaave Course on Traumatic Brachial Plexus Injuries, Leiden, pp. 65-89.

Jaspers, R.B.M. (1986). Diagnostics of Brachial Plexus Injuries (in Dutch). Report N-259, Laboratory for Measurement and Control, Delft University of Technology, ISBN 90-370-0006-1, 32 p.

Jaspers, R.B.M.; Daalen, C. van; Helm, F.C.T. van der (1989). Modelling the rehabilitation of brachial plexus injuries: The PLEXUS expert system. *Journal of Medical Engineering and Technology*, Vol. 13, pp. 114-118.

Jaspers, R.B.M. (1990). Medical Decision Support: An Approach in the Domain of Brachial Plexus Injuries. Ph.D. Thesis, Delft University of Technology, ISBN 90-370-0028-2.

Kent, D.L.; Shortliffe, E.H.; Carlson, R.W.; Bischoff, M.B. and Jacobs, C.D. (1985). Improvements in data collection through physician use of a computer-based chemotherapy treatment consultant. *Journal of Clinical Oncology*, Vol. 3, No. 10, pp. 1409-1417.

Kerr, A.T. (1918). The brachial plexus of nerves in man, the variations in its formation and branches. *American Journal of Anatomy*, Vol. 23, pp. 285-395.

Kidd, A.L. and Cooper, M.B. (1985). Man-Machine interface issues in the construction and use of an expert system. *International Journal of Man-Machine Studies*, Vol. 22, pp. 91-102.

Kingsland, L.C. (1985). The evaluation of medical expert systems: Experience with the AI/RHEUM knowledge-based consultant system in rheumatology. In: Proceedings 9th Annual Symposium on Computer Applications in Medical Care, pp. 292-295.

Kors, J.A.; Sittig A.C. and Bemmel, J.H. van (1990). The Delphi method to validate diagnostic knowledge in computerized ECG interpretation. *Methods of Information in Medicine*, Vol. 29, No. 1, pp. 44-50.

Langlotz, C.P. and Shortliffe E.H. (1983). Adapting a consultation system to critique user plans. *International Journal of Man-Machine Studies*, Vol. 19, pp. 479-496.

Laurent, J-P. (1992). Proposals for a valid terminology in KBS validation. In: B. Neumann (ed.), Proceedings of the 10th European Conference on Artificial Intelligence, Vienna, pp. 829-834.

Leffert, R.D. (1985). Brachial Plexus Injuries. Churchill Livingstone, New York.

Lind van Wijngaarden, D.G. de and Furth, V.H. (1987). Diagnosis of brachial plexus injuries. Delft University of Technology.

Lipscombe, B. (1989). Expert systems and computer-controlled decision making in medicine. *AI and Society*, Vol. 3, pp. 184-197.

Lucas, P.J.F. and Gaag, L.C. van der (1988). Principles of Expert Systems (in Dutch). Academic Service, Schoonhoven, ISBN 90-6233-264-1.

Lundsgaarde, H.P. (1987). Evaluating medical expert systems. *Soc. Sci. Med.*, Vol. 24, No. 10, pp. 805-819.

Lydiard, T.J. (1992). Overview of current practice and research initiatives for the verification and validation of KBS. *The Knowledge Engineering Review*, Vol. 7, No. 2, pp. 101-113.

McDermott, P.A. and Hale, R.L. (1982). Validation of a systems-actuarial computer process for multidimensional classification of child psychopathology. *Journal of Clinical Psychology*, Vol. 38, No. 3, pp. 477-486.

McDonald, C.J. *et al.* (1984). Reminders to physicians from an introspective computer medical record. A two-year randomized trial. *Annals of Internal Medicine*, Vol. 100, pp. 130-138.

Medical Research Council (1986). Aids to the Examination of the Peripheral Nervous System, Ballière Tindall, London.

Meinders, L.W. (1989). Application of a Tree-Search Algorithm in a Diagnostic Expert System for Brachial Plexus Injuries (in Dutch). M.Sc. Thesis, Report A-431, Laboratory for Measurement and Control, Delft University of Technology.

Merle d'Aubigné, R. and Deburge, A. (1967). Etiologie, évolution et pronostic des paralysies traumatiques du plexus brachial. *Revue de Chirurgie Orthopédique et Réparatrice de l'Appareil Moteur*, Vol. 53, No. 1, pp. 23-42.

Meseguer, P. (1992). Incremental verification of rule-based expert systems. In: B. Neumann (ed.), Proceedings of the 10th European Conference on Artificial Intelligence, Vienna, pp. 840-844.

Miller, P.L. (1984). Critiquing: A different approach to expert computer advice in medicine. In: Proceedings of the 8th Annual Symposium on Computer Applications in Medical Care, pp. 17-23.

Miller, P.L. (1986). The evaluation of artificial intelligence systems in medicine. *Computer Methods and Programs in Biomedicine,* Vol. 22, pp. 5-11.

Miller, P.L. and Sittig, D.F. (1990). The evaluation of clinical decision support systems: what is necessary versus what is interesting. *Medical Informatics,* Vol. 15, No. 3, pp. 185-190.

Miller, R.A.; Pople, H.E.; and Myers, J.D. (1982). INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. *The New England Journal of Medicine,*Vol. 307, No. 8, pp. 468-476.

Miller, R.A. and Masarie F.E. (1990). The demise of the "Greek Oracle" model for medical diagnostic systems. *Methods of Information in Medicine,* Vol. 29, pp. 1-2.

Minsky, M. (1975). A framework for representing knowledge. In: P.H. Winston (ed.), The Psychology of Computer Vision, McGraw-Hill, New York.

Mirkin, B.G. (1979). Expert judgement analysis. In: Group Choice, John Wiley, pp. 141-179.

Muir, B.M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies,* Vol. 27, pp. 527-539.

Murray, G.D.; Murray, L.S.; Barlow, P.; Teasdale, G.M. and Jennet, W.B. (1986). Assessing the performance and clinical impact of a computerized prognostic system in severe head injury. *Statistics in Medicine,* Vol. 5, pp. 403-410.

Murray, G.D. (1990). Assessing the clinical impact of a predictive system in severe head injury. *Medical Informatics,* Vol. 15, No. 3, pp. 269-273.

Narakas, A.O. (1985). The treatment of brachial plexus injuries. *International Orthopaedics,* Vol. 9, pp. 29-36.

Narakas, A.O. (1993). Personal Communication.

Nelson, S.J. *et al.* (1985). Evaluating RECONSIDER. A computer program for diagnostic prompting. *Journal of Medical Systems,* Vol. 9, Nos. 5/6, pp. 379-388.

Nguyen, T.A.; Perkins, W.A.; Laffey, T.J. and Pecora, D. (1987). Knowledge base verification. *AI Magazine,* Summer, pp. 69-75.

Niggeman, J. (1990). Analysis and representation of neuroanatomical knowledge. *Applied Artificial Intelligence,* Vol. 4, pp. 309-336.

Nykänen, P. ed. (1990). Issues in Evaluation of Computer-Based Support to Clinical Decision Making. SYDPOL-5 Working Group, Institute of Informatics, Oslo University Research Reports 127, ISBN 82-7368-031-2, 85 p.

Ohe, K. and Kaihara, S. (1988). Representation of neuroanatomical knowledge by PROLOG. In: O. Rienhoff *et al.* (eds.), Lecture Notes in Medical Informatics, Vol. 36, Springer Verlag, pp. 249-255.

O'Keefe, R.M.; Balci, O. and Smith, E.P. (1987). Validating expert system performance. *IEEE Expert,* Winter, pp. 81-90.

O'Leary, T.J.; Goul, M.; Moffit, K.E. and Radwan, A.E. (1990). Validating expert systems. *IEEE Expert,* Vol. 5, No. 3, pp. 51-58.

O'Moore, R. *et al.* (1990). Methodology for Evaluation of Knowledge Based Systems. KAVAS (A1021), Report E.M. 1.2., EEC AIM Office, 62 Rue de Treves, Brussels.

Pollack, M.E.; Hirschberg, J. and Webber, B. (1982). User participation in the reasoning processes of expert systems. In: Proceedings of the National Conference on Artificial Intelligence (AAAI), pp. 358-361.

Preece, A.D. and Shinghal, R. (1992). Verifying knowledge bases by anomaly detection: An experience report. In: B. Neumann (ed.), Proceedings of the 10th European Conference on Artificial Intelligence, Vienna, pp. 835-839.

Pryor, T.A.; Gardner, R.M.; Clayton, P.D. and Warner, H.R. (1983). The HELP system. *Journal of Medical Systems*, Vol. 7, No. 2, pp. 87-102.

Quaglini, S.; Stefanelli, M.; Barosi, G. and Berzuini, A. (1988). A performance evaluation of the expert system ANEMIA. *Computers and Biomedical Research*, Vol. 21, pp. 307-323.

Ravden, S.J. and Johnson, G.I. (1989). Evaluating Usability of Human-Computer Interfaces: A Practical Method. Ellis Horwood Limited, Chichester.

Reggia, J.A. (1978). A production rule system for neurological localization. In: Proceedings of the Second Annual Symposium on Computer Applications in Medical Care, Long Beach, CA, IEEE Computer Society, pp. 254-260.

Reggia, J.R. (1985). Evaluation of medical expert systems: Case study in performance assessment. In: Proceedings of the 9th Annual Symposium on Computer Applications in Medical Care, pp. 287-291.

Reggia, J.A.; Tuhrim, S.; Ahuja, S.B.; Pula, T.; Chu, B.; Dasigi, V. and Lubell, J. (1986). Plausible reasoning during neurological problem solving: The Maryland NEUREX project. In: R. Salamon, B. Blum and M. Jørgensen (eds.), Proceedings of MEDINFO86, Elsevier Science Publishers B.V., pp. 17-21.

Rich, E. (1983). Artificial Intelligence. McGraw-Hill, New York.

Rossi-Mori, A.; Pisanelli, D.M. and Ricci, F.L. (1990). Evaluation stages and design steps for knowledge-based systems in medicine. *Medical Informatics*, Vol. 15, No. 3, pp. 191-204.

Roth, E.M.; Bennett, K.B. and Woods, D.D. (1987). Human interaction with an "intelligent" machine. *International Journal of Man-Machine Studies*, Vol. 27, pp. 479-525.

Rothschild, M.A.; Swett, H.A.; Fisher, P.R.; Weltin, G.G. and Miller, P.L. (1990). Exploring subjective vs. objective issues in the validation of computer-based critiquing advice. *Computer Methods and Programs in Biomedicine*, Vol. 31, pp. 11-18.

Salamon, R.; Bernadet, M.; Samson, M.; Derouesne, C. and Gremy, F. (1976). Bayesian method applied to decision making in neurology - methodological considerations. *Methods of Information in Medicine*, Vol. 15, No. 3, pp. 174-179.

Sassen, J.M.A. (1993). Design issues of human operator support systems. Ph.D. Thesis, Delft University of Technology, ISBN 90-370-0090-8.

Seddon, H.J. (1943). Three types of nerve injury. *Brain*, Vol. 66, pp. 237-288.

Shapiro, A.R. (1977). The evaluation of clinical predictions. A method and initial application. *The New England Journal of Medicine*, Vol. 296, No. 26, pp. 1509-1514.

Shortliffe, E.H. (1976). Computer-based medical consultations: MYCIN. Elsevier, New York.

Shortliffe, E.H. and Clancey, W.J. (1984). Anticipating the second decade. In: W.J. Clancey and E.H. Shortliffe (eds.), Readings in Medical Artificial Intelligence. The First Decade, Addison-Wesley Publishing Company, pp. 463-472.

Shortliffe, E.H. (1989). Testing reality: The introduction of decision-support technologies for physicians. *Methods of Information in Medicine*, Vol. 28, pp. 1-5.

Shwe, M.A.; Tu, S.W. and Fagan, L.M. (1989). Validating the knowledge base of a therapy planning system. *Methods of Information in Medicine*, Vol. 28, pp. 36-50.

Siegel, S. and Castellan, N.J. (1988). Nonparametric Statistics for the Behavioral Sciences, 2nd. edition, McGraw-Hill.

Slooff, A.C.J. (1993). Obstetric brachial plexus lesions and their neurosurgical treatment. *Clinical Neurology and Neurosurgery*, Vol. 95, pp. S73-S77.

Soula, G.; Thirion, X.; San Marco, J.L.; Vialettes, B.; Guliana, J.M. and Navez, I. (1988). A multi-centered validation of the fuzzy expert system PROTIS. In: R. Hansen *et al.* (eds.), Lecture Notes in Medical Informatics, Springer Verlag, Vol. 35, pp. 647-651.

Spek, R. van der (1991). The design cycle. In: P. Braspenning (ed.), Knowledge Based Systems: Applications of Artificial Intelligence (in Dutch), Stichting Teleac, Utrecht, pp. 157-177.

Spiegelhalter, D.J. (1983). Evaluation of clinical decision-aids, with an application to a system for dyspepsia. *Statistics in Medicine*, Vol. 2, pp. 207-216.

Spitzer, R.L. and Endicott, J. (1969). DIAGNOII: Further developments in a computer program for psychiatric diagnosis. *American Journal of Psychiatry*, Vol. 125, No. 7 Supp., pp. 12-21.

Stachowitz, R.A. and Combs, J.B. (1987). Validation of expert systems. In: Proceedings of the 20th Annual Hawaii International Conference on System Sciences, pp. 686-695.

Stassen, H.G; Lunteren, A. van; Hoogendoorn, R.; Kolk, G.J. van der; Balk, P.; Morsink, G. and Schuurman, J.C. (1980). A computer model as an aid in the treatment of patients with injuries of the spinal cord. In: Proceedings of the ICCS, Cambridge, Massachusetts, IEEE, pp. 385-390.

Stassen, H.G. (1989). The rehabilitation of severely disabled persons. A man-machine system approach. In: W.B. Rouse (ed.), Advances in Man-Machine Systems Research, Vol. 5, JAI Press Inc., pp. 153-227.

Steels, L. (1990). Components of Expertise. *AI Magazine*, Summer, pp. 29-49.

Struss, P. (1992). Knowledge-based diagnosis. An important challenge and touchstone for AI. In: B. Neumann (ed.), Proceedings of the 10th European Conference on Artificial Intelligence, Vienna, pp. 863-874.

Sunderland, S. (1968). Nerves and Nerve Injuries. Churchill Livingstone, Edinburgh.

Sunderland, S. (1982). Pathophysiology of brachial plexus lesions. In: Proceedings of the Boerhaave Course on Traumatic Brachial Plexus Injuries, Leiden, pp. 13-26.

Sutcliffe, A.G. (1988). Human-Computer Interface Design. MacMillan Education Limited, Basingstoke.

Sutton, G.C. (1989a). How accurate is computer-aided diagnosis? *The Lancet*, October 14, pp. 905-908.

Sutton, G.C. (1989b). Computer-aided diagnosis: A review. *British Journal of Surgery*, Vol. 76, pp. 82-85.

Suwa, M.; Scott, A.C.; Shortliffe, E.H. (1982). An approach to verifying completeness and consistency in a rule-based expert system. *AI Magazine*, Vol. 3, No. 4, pp. 16-21.

Swaan Arons, H. de (1991). Delfi: Design, Development and Applicability of Expert System Shells. Ph.D. Thesis, Delft University of Technology, ISBN 90-6275-734-0.

Teach, R.L. and Shortliffe, E.H. (1981). An analysis of physicians' attitudes regarding computer-based clinical consultation systems. *Computers in Biomedical Research*, Vol. 14, pp. 542-558.

Thomeer, R.T.W.M. (1991). Recovery of brachial plexus injuries. *Clinical Neurology and Neurosurgery*, Vol. 93-1, pp. 3-11.

Vainio-Larsson, A. and Orring, R. (1990). Evaluating the usability of user interfaces: Research in practice. In: D. Diaper *et al.* (eds.), Human-Computer Interaction (INTERACT '90), Elsevier Science Publishers B.V., pp. 323-328.

Voorhorst, F.A. (1992). Verification and Validation of Rule-Based Expert Systems (in Dutch). Report S-628, Laboratory for Measurement and Control, Delft University of Technology.

Voorhorst, F.A. (1993). Validating PLEXUS Using Generated Test Cases (in Dutch). M.Sc. Thesis, Report A-628, Laboratory for Measurement and Control, Delft University of Technology.

Whitbeck, C. and Brooks, R. (1983). Criteria for evaluating a computer aid to clinical reasoning. *The Journal of Medicine and Philosophy*, Vol. 8, pp. 51-65.

White, K.S.; Lindsay, A.; Pryor, T.A.; Brown, W.F. and Walsh, K. (1984). Application of a computerized medical decision-making process to the problem of digoxin intoxication. *Journal of the American College of Cardiology*, Vol. 4, No. 3, pp. 571-576.

Wong, W.S.F.; Leung, K.S. and So, Y.T. (1990). The recent development and evaluation of a medical expert system (ABVAB). *International Journal of Biomedical Computing*, Vol. 25, pp. 223-229.

Woods, D.D. (1986). Paradigms for intelligent decision support. In: E. Hollnagel, G. Mancini and D.D. Woods (eds.), Intelligent Decision Support in Process Environments, NATO ASI Series, Springer-Verlag, Berlin, Vol. F21, pp. 153-173.

Woods, D.D. and Roth, E.M. (1988). Cognitive Systems Engineering. In: M. Helander, (ed.), Handbook of Human-Computer Interaction, Elsevier Science Publishers B.V. (North-Holland), pp. 3-43.

Woods, D.D.; Roth, E.M. and Bennett, K.B. (1990). Explorations in joint human-machine cognitive systems. In: S.P. Robertson, W. Zachary, J.B. Black (eds.), Cognition, Computing and Cooperation, Ablex Publishing Corporation, pp. 123-158.

Wyatt, J. (1987). The evaluation of clinical decision support systems: A discussion of the methodology used in the ACORN project. In: J. Fox, M. Fieschi, R. Engelbrecht (eds.), Proceedings of the European Conference on Artificial Intelligence in Medicine, Marseilles, pp. 15-24.

Wyatt, J. and Spiegelhalter, D. (1990). Evaluating medical decision-aids: What to test, and how? In: J. Talmon, J. Fox (eds.), System Engineering in Medicine, Springer Verlag, Heidelberg.

Xiang, Z.; Srihari, S.N.; Shapiro, S.C. and Chutkow, J.G. (1985). A modeling scheme for diagnosis. In: Proceedings of the Expert Systems in Government Symposium, IEEE, pp. 538-547.

Yu, V.L. *et al.* (1979). Antimicrobial selection by a computer: A blinded evaluation by infectious diseases experts. *Journal of the American Medical Association*, Vol. 242, No. 12, pp. 1279-1282.

Zagoria, R.J. and Reggia, J.A . (1983). Transferability of medical decision support systems based on Bayesian classification. *Medical Decision Making*, Vol. 3, No. 4, pp. 501-509.

Appendix 1 Analysis of the results

In this appendix, various methods of analysis of system performance are shown. One example will be used throughout. Suppose that there are 100 cases. Each case either has a particular illness, or the illness is absent. This means that the classification is mutually exclusive. There is a definite standard which gives the correct classification. The standard classification consists of 65 cases for which the illness is present and 35 cases for which the illness is absent. Let us suppose that the knowledge based system classifies 70 cases into present and 30 into absent. Most methods of analysis are based on a classification matrix which shows a paired comparison of the classificiations carried out by the standard and by the system. Such a classification matrix is shown below.

A.1. Error rate methods

1 Confusion Matrix

			system	
		present	absent	total
standard	present	60	5	65
	absent	10	25	35
	total	70	30	100

Let i_{sys} be the number of incorrect answers given by the system, when compared to the standard, and let t be the total number of answers. The error rate of the system is defined as follows:

$$\text{error rate} = \frac{i_{sys}}{t} . \qquad [A.1]$$

The error rate for this example is given below.

$$\text{error rate} = \frac{10+5}{100} = 15\%$$

2 Positive Negative Correctness Model

This model is an extension of the error rate model. Four different situations may be distinguished. These are always stated with respect to a class. With respect to the presence of the illness, if both standard and system state present, then the answer is called True Positive (TP). If the system states present and standard states absent, then the answer is categorized as False Positive (FP) with respect to the illness. If the system classifies the case into absent and the standard is present, then the answer is False Negative (FN), and if system and standard both say absent, then the answer is True Negative (TN) with respect to the illness.

		system				
		present		absent		total
standard	present	TP	60	FN	5	65
	absent	FP	10	TN	25	35
	total	70		30		100

Let c_{tp} be the number of true positive answers, i_{fp} the number of false positive answers, c_{tn} the number of true negative answers and i_{fn} the number of false negative answers. The accuracy and the sensitivity and specificity with respect to an illness are defined as follows (see, for example, Indurkhya and Weiss, 1989).

$$\text{accuracy} = \frac{c_{tp}+c_{tn}}{t} \qquad [A.2]$$

$$\text{sensitivity} = \frac{c_{tp}}{c_{tp}+i_{fn}} \qquad [A.3]$$

$$\text{specificity} = \frac{c_{tn}}{c_{tn}+i_{fp}} \qquad [A.4]$$

The accuracy, sensitivity and specificity for this example are given below.

Using [A.2] the accuracy $= \frac{60+25}{100} = 0.85$

Using [A.3] the sensitivity $= \frac{60}{60+5} = 0.92$

Using [A.4] the specificity $= \frac{25}{10+25} = 0.71$

3 ROC Curves

When certainty factors are used, the matrix which was stated in Example 2 can be extended to the matrix below. When introducing a certainty factor threshold of 0.5, this table would be the same as in Example 2.

		system	
		present	total
standard	present	30 CF 0.6 30 CF 0.8	65
	absent	5 CF 0.3 10 CF 0.7 25 CF 0.2	35
	total		100

The threshold, for certain answers to be true, can be modified. Suppose there are 3 possibilities: CF threshold is 0, CF threshold is 0.5, CF threshold is 0.8 If the certainty factor of illness present is higher than the threshold, then the case is classified as belonging to class present otherwise it is absent.

First, calculation of the sensitivities and specificities will be carried out using the positive negative correctness model which was discussed above, after which it will be possible to draw a curve, called an ROC curve.

threshold 0:

		system	
		present	absent
standard	present	65	
	absent	35	

Using [A.3] and [A.4] the sensitivity and specificity can be calculated.

$$\text{sensitivity} = \frac{65}{65+0} = 1; \qquad \text{specificity} = \frac{0}{35+0} = 0$$

threshold 0.5:

		system	
		present	absent
standard	present	60	5
	absent	10	25

(see Example 2) sensitivity = 0.92 specificity = 0.71

threshold 0.8:

		system	
		present	absent
standard	present	30	35
	absent	0	35

$$\text{sensitivity} = \frac{30}{30+35} = 0.46 \qquad \text{specificity} = \frac{35}{0+35} = 1$$

A curve of sensitivity against (1-specificity) may be drawn. This is known as a receiver operating characteristic (ROC) curve. The area under the curve is a measure for the performance of the system (see, for example, Indurkhya and Weiss, 1989) with respect to the illness.

A.2. Confidence Level Methods

4 Accuracy Coefficient and Distance Metrics

There are various methods which take into account the weights given by the system. Consider the matrix containing the certainty factors which was described in the previous example.

		system		
		present	absent	total
standard	present	30 CF 0.6	CF 0.4	65
		30 CF 0.8	CF 0.2	
		5 CF 0.3	CF 0.7	
	absent	10 CF 0.7	CF 0.3	35
		25 CF 0.2	CF 0.8	
	total			100

A) Zagoria and Reggia (1983) define an accuracy coefficient Q'

$$Q' = \frac{2}{n} \sum_{i=1}^{n} (p_i - 0.5) \qquad [A.5]$$

For this coefficient, p_i is the weight (between 0 and 1) assigned to the outcome which is actually true (the standard outcome) in the ith case. Q' varies between -1 and 1.

$$Q' = \frac{2}{100} (30(0.6-0.5) + 30(0.8-0.5) + 5(0.3-0.5) + 10(0.3-0.5) + 25(0.8-0.5)) = 0.33$$

B) Bernelot Moens and van der Korst (1991) define an accuracy coefficient Q"

$$Q'' = 1 - \frac{\sum |d_i - p_i|}{n} \qquad [A.6]$$

Where d_i is the standard and p_i is the probability given by the computer. This coefficient contains the average deviation from the probability of the standard under the assumption that probability of standard diagnosis at definite level is 1.0, possible level = 0.5, and not included = 0. Q" varies between 0 and 1.

$$Q'' = 1 - \frac{(30(1-0.6)+30(1-0.8)+5(1-0.3)+10|(0-0.7)|+25|(0-0.2)|)}{100} = 1 - \frac{33.5}{100} = 0.665$$

C) Distance metrics

Indurkhya and Weiss (1989) describe distance metrics using an n-dimensional space, however in this example only a 2-dimensional space is used.
If present is seen as representing the x-axis of a graph, and absent is the y-axis, with the certainty factors as the values of the variables.
standard answer in TP and FN case (1,0) (present,absent)
standard answer in FP and TN case (0,1) (present,absent)

The average squared distance from the standard can be determined as follows. Let $\underline{c}_{stan}$ be a vector with the standard answer and let $\underline{a}_{sys}$ be a vector of the answer given by the system.

$$\text{average squared distance} = \frac{\sum_{i=1}^{t} (\underline{c}_{stan} - \underline{a}_{sys})^2}{t} \qquad [A.7]$$

For the above example, the average squared distance can be calculated as follows.

30 cases in	TP	(0.6,0.4) sqrd. difference with standard	$(0.4,-0.4)^2 = 0.32$
30 cases in	TP	(0.8,0.2)	$(0.2,-0.2)^2 = 0.08$
5	FN	(0.3,0.7)	$(0.7,-0.7)^2 = 0.98$
10	FP	(0.7,0.3)	$(-0.7,0.7)^2 = 0.98$
25	TN	(0.2,0.8)	$(-0.2,0.2)^2 = 0.08$

$$\text{av.sq.dist. from stand.} = \frac{30*0.32+30*0.08+5*0.98+10*0.98+25*0.08}{100} = 0.287$$

A.3. Agreement

5 Measure of Agreement

The Kappa coefficient of agreement (Cohen, 1968) is often used to calculate inter- and intra- expert agreement. The confusion matrix of Example 1 is shown, with the difference that instead of using the number of cases n, the entries in this matrix consist of the number of cases in a category divided by the total number of cases $\left(p = \frac{n}{N}\right)$.

		system		
		present	absent	total
standard	present	0.6 (0.455)	0.05 (0.195)	0.65
	absent	0.1 (0.245)	0.25 (0.105)	0.35
	total	0.7	0.3	1

The probability of both standard and system saying present, by chance alone, equals 0.65*0.7 = 0.455

Let p_o be the observed proportion of agreement, and p_c be the proportion of agreement expected by chance. Then p_o and p_c can be calculated as follows:
$p_o = 0.6 + 0.25 = 0.85$; $p_c = 0.455 + 0.105 = 0.56$

Kappa is defined (Cohen, 1968) as $K = \frac{p_o - p_c}{1 - p_c}$. [A.8]

Using [A.8], for this example $K = \frac{0.85 - 0.56}{1 - 0.56} = 0.66$

For investigating whether the underlying value of K is significantly different from a prescribed value D *other than zero*, Fleiss (1981) gives an estimation of the standard error of K.

$$s.e.(K) = \frac{\sqrt{A+B-C}}{(1-p_c)\sqrt{n}} \qquad [A.9]$$

Where $A = \sum_{i=1}^{k} p_{ii}[1 - (p_{i.} + p_{.i})(1 - K)]^2$,

$B = (1 - K)^2 \sum\sum_{i \neq j} p_{ij} (p_{.i} + p_{j.})^2$,

$C = [K - p_c (1 - K)]^2$.

For applying the normal curve test, z is determined as follows:

$$z = \frac{|K - D|}{s.e.(K)} . \qquad [A.10]$$

For this example, A, B and C can be calculated as follows:

$A = 0.6[1 - (0.65+0.7)(1-0.66)]^2 + 0.25[1 - (0.35+0.3)(1-0.66)]^2 = 0.3273$

$B = (1 - 0.66)^2 [0.05(0.7+0.35)^2 + 0.1(0.3+0.65)^2] = 0.0168$

$C = [0.66 - 0.56(1-0.66)]^2 = 0.221$

$$s.e.(K) = \frac{\sqrt{0.3273+0.0168-0.221}}{(1-0.56)\sqrt{100}} = 0.08$$

if D = 0.50 then, using [A.10], $z = \frac{0.66-0.50}{0.08} = 2$ significant $p < 0.05$

The confidence interval for this Kappa can be determined using 95% confidence limits.

Confidence interval : 95% limits $0.50 \leq$ Kappa ≤ 0.82

Appendix 2 Summary of reported evaluation studies

Table 1. Summary of reported laboratory evaluations.

reference	Adlassnig(1989)	Aikins(1983)
system characteristics	CADIAG-2/PANCREAS 10 pancreatic diseases; diagnoses and ranked according to score of support	PUFF: interprets measurements from respiratory tests in lung function laboratory
object evaluation	to prepare the clinical application	formal evaluation of the BASIC-PUFF performance system
experimental setup	direct comparison of system output against gold standard; also with varying thresholds for system output	compare system diagnosis to 2 physicians by directly observing agreement
test input	47 retrospective patients from a university medical school; with 51 diagnoses of pancreatic diseases	144 cases
way of entry		automatic
test against		2 pulmonary physiologists; of which one cooperating expert
standard	histopathological or surgical in some cases reliable clinical diagnosis	not explicit
comparison	direct	direct rule: close agreement defined as differing by at most 1 degree of severity (mild,moderate,severe)
analysis	-ROC curves, varying internal threshold -sensitivity, specificity, accuracy	-percentage agreement on diagnoses; patients often have more than one disease
results	extended data set (incl. lab tests): 86.3% correct in top diagnosis 96.1% correct in first three concl: application of system at early stages of diagnostic process seems achievable (history, physical exam, basic lab tests)	-agreement 2 physiologists 92%; s.d. 1.63 -between expert and PUFF 96%; s.d. 3.83 -between independent physician and PUFF 89%; s.d. 4.69 concl: PUFF has shown that if task, domain, and researchers are carefully matched, then application of existing techniques can result in a system which successfully performs a moderately complicated task of medical diagnosis. -in 1983 PUFF routinely used in pulmonary function laboratory

reference	Alonso-Betanzos(1989)	Bernelot Moens(1991)
system characteristics	FOETOS: assist obstetrician diagnosing antepartum & intrapartum foetal well-being with 5 obstetrical tasks	Bayesian system for differential diagnosis of principal rheumatic disease categories. probabilities assigned to each of 15 possible categories
object evaluation	analyze errors for improvement	measure correctness using various methods and metrics
experimental setup	direct comparison of system and clinician on final and intermediate results of 4 of the tasks	compare system diagnosis to physicians' diagnosis using direct comparison with final diagnosis after follow-up
test input	20 retrospective patients under condition they had received the 3 tests and availability of labour records	570 consecutive cases newly referred to rheumatological outpatient clinic
way of entry		
test against	various obstetricians who had actually diagnosed and prognosed	rheumatologists after 1st patient encounter
standard	obstetricians	final diagnosis after follow-up; definite if made at that level by 2 (of 3) rheumatologists; possible if partial agreement or lower confidence
comparison	direct comparison of final and intermediate results	direct
analysis	-Kappa coefficient of agreement and percentage agreement -measure for differences; degree of how much more or less favourable system than clinician results	-performance by diagnosis -correctness by rank order -ROC curve, sensitivity, specificity -performance by case using scoring matrix
results	-agreement varied on different subtasks from 35% -85% -possible to indicate where improvements necessary concl: fundamental approach of the system is sound	-average probability given to the actual diagnosis physicians: def 0.60, abs 0.016 system : def 0.50, abs 0.044 -correct in top 3 phys. 65%; system 82% -sensitivity phys 64%, sys 65% -specificity phys 98%, sys 96% concl: system performance comes close to human experts. concl: need for careful selection and description of measures of performance

reference	Catanzarite(1981)	Fieschi(1990)
system characteristics	NEUROLOGIST: consultation system for clinical neurology	SPHINX: treatment of diabetes 17 or 8 therapeutic categories
object evaluation	preliminary accuracy testing	evaluate performance of system by situating it among practitioners of varying expertise; and investigate sensitivity of the system
experimental setup	test case selection, debugging and direct comparison of system output to diagnosis in literature	-6 practitioners and system make prescriptions for 100 cases -submit blindly to experts: prescriptions where disagreement between system and 1 of experts
test input	30 cases from literature (book) only if diagnosis included in kbase	100 random cases from 2 hospital departments and 1 private surgery
way of entry		
test against		case physician, 3 GP's, 2 experts
standard	in literature	
comparison	direct	-direct -reconsidered files rated blindly by experts: equivalent, acceptable, insufficient, unacceptable
analysis	-error rate -separation	-concordance of opinions Kappa/% -study of non-concording prescriptions -reproducibility -sensitivity
results	-23/30 cases top diagnosis correct -20/23 cases no other hypothesis within 20 points on a scale [-99 to 99] concl: expectation confirmed that the localizing diagnosis provides focused framework for disease identification; several weaknesses identified	-concordance of opinions (Kappa) e.g. 2 GP's : 17 cats: K= 0.50; 8 cats: K = 0.57 e.g. system and expert : 17 cats:K = 0.61; 8 cats: K = 0.71 -for 80 prescriptions reconsidered there is a significant difference between the experts -sensitivity study by changing limits around different interpretation zones

reference	Francois(1992)	Gorry(1978)
system characteristics	MENINGE: consulting system for child's meningitis; expert system with linear model subsystem	considers clinical responses to digitalis; determination and modification of dosage
object evaluation	determine performance level of system; also to decide if clinical evaluation can be carried out	to establish potential utility of programs such as this prototype
experimental setup	on consecutive series of patients system subdiagnosis, final diagnosis and treatment compared to reference and to medical team who saw patients	compare system output to actual dosage of drugs administered in cardiology service of hospital
test input	212 consecutive cases from French hospital, collected over 30 months	19 patients of cardiology service of hospital during 1 month in which clinical situation changed (various recommendations for most)
way of entry		
test against	-reference diagnosis and treatment -medical team during admission	clinicians of cardiology service
standard	microbiological data and evolution known subsequently	no standard
comparison	direct if one answer, topmost if more	direct
analysis	-matrices of classes with reference against system, for subdiagnosis, diagnosis and treatment -accuracy, sensitivity, specificity -system treatment compared to medical team (Wilcoxon)	-percentage agreement
results	-correct germ first in 92% of cases -treatment useful and efficient 94.8% -medical team significantly more ($p<0.01$ Wilcoxon) therapeutic errors than the system concl: these encouraging results allowed us to install MENINGE in medical care unit to evaluate clinical relevance	-22/38 recommendations same -2/38 system higher -14/38 system lower -4 patients showed digitalis toxicity, in each case clinicians failed to appreciate significance of the early signs of toxicity which program correctly interpreted. concl: trial demonstrates that use of such programs might be utilized to distribute knowledge about digitalis therapy where cardiac consultation may not readily available; further evaluation is clearly required

reference	Haberman(1985)	Hickam(1985)
system characteristics	DIAG: assists in the formulation of the differential diagnosis of skin diseases	ONCOCIN: chemotherapy protocol advisor
object evaluation	preliminary evaluation of system's diagnostic accuracy to better appreciate how closely the system emulates its expert counterpart	compare quality of lymphoma therapy recommendations with Stanford oncologists
experimental setup	direct comparison of system differential diagnosis to that of expert	-direct comparison of chemotherapy administered by clinic physician and system -subset analysed blindly by experts -interobserver reliability
test input	50 patient profiles	-415 visits for 39 patients to one centre -blind analysis too many cases to have experts judge; choose 25% of cases system and clinician agreed (n = 47) and 70% = 137 of unique disagreed
way of entry		
test against	experienced dermatologist	clinic physicians who actually treated the patients
standard	not explicit; experienced dermatologist (above) as standard	implicit
comparison	direct	-direct comparison -4 experienced oncologists rating ideal,acceptable,sub-optimal unacceptable; score 4,3,2,1
analysis	-mean no. of diseases -% diseases omitted and included in error	-categorical variables Chi-square or Fisher exact -scaled variables t-test, MANOVA -interobserver reliability weighted Kappa
results	-mean no. of diseases per differential: derm 3.7; DIAG 5.4 -% diseases omitted: derm standard; DIAG 5.9% -% diseases included in error: derm standard; DIAG 33.9% concl: although not ideal, the frequency of diagnostic omissions committed by the system was minimal. System updated upon this information	-189 of 415 agreement 197 of 415 unique disagreement -subset of 137 disagreed: physicians score 3.1(S.D. 0.09) computer score 3.06(S.D. 0.09) concl: system provides lymphoma protocol advice similar to treatment delivered in Stanford

reference	Kingsland(1985)	Kors(1990)
system characteristics	AI/RHEUM: consultant system in rheumatology 26 diseases in knowledge base -list differential diagnosis with definite, probable or possible	ECG interpretation -6 categories each of which: definite, probable, possible, definitely not
object evaluation	judging accuracy of the system	to determine and possibly improve performance of program
experimental setup	-direct comparison of the output of the system to opinion of consensus of experts -to test repeatability of gold standard blinded review of 48 cases	2 rounds: first cardiologists shown computer output and indicated and motivated own if different from computer; second indicate opinion and rate motivations when all outputs from 1st round given blindly. -directly compare outputs and motivations.
test input	-74 unselected consecutive cases from 2 periods in arthritis unit -59 extremely difficult cases from Japan, many multiple diagnoses	30 ECGs; stratified random samples; 30% normal; mixture of pathological -from large international data set
way of entry	researchers	
test against		5 Dutch cardiologists; unaware of each others identity
standard	consensus of expert rheumatologists	not explicit
comparison	-direct -correct when at top or tied at top and multiple correct when all in differential	direct
analysis	-error rates correct, incorrect, no answer -percentage agreement, repeatability	-Kappa for inter and intraobserver agreement -physicians' motivations studied
results	-of 74 cases, diagnoses of 63 available in kbase these 63 were correct -of 11 not available; on 5 appropriately refused to conclude -repeatability at least 2 of 3 agreed on 46 of 48 cases -Japanese cases: 54/59 correct concl: performance warrants testing in additional settings	-6 basic categories; cardiologists round 1 K = 0.68, rnd 2 K = 0.81 -2 class coding; cardiologists rnd 1 K = 0.83, rnd 2 K = 0.83 -6 categories; computer-cardiol. rnd 1 K = 0.56, rnd 2 K = 0.5 concl: computer classification is not yet at an expert level

reference	McDermott(1982)	Miller(1982)
system characteristics	MAC: systems-actuarial process for multi-dimensional classification of child psychopathology 6 diagnostic areas	INTERNIST-1: diagnosis within broad context of internal medicine
object evaluation	to assess the validity of MAC as discerned through its classification congruence with diagnoses provided by experts	preliminary evaluation to compare clinical acumen with that of human experts and to highlight its strengths and weaknesses
experimental setup	compare system diagnosis directly to two independent child psychologists and determine classification agreement	system output compared directly to diagnoses given by treating clinicians and case discussants on cases taken from a journal
test input	73 children and adolescents referred by families or schools to a children's outpatient clinic for psychological services	19 cases from Massachusetts General hospital described in a journal; only those whose major diagnoses represented in system
way of entry		
test against	-2 independent and experienced child psychologists	-treating clinicians -case discussants
standard	not explicit	pathologists or when clinical syndrome universally agreed to be present
comparison	direct	-if definitive diagnosis; it should be correct -if tentative diagnosis; topmost diagnosis should be correct
analysis	-agreement by Kappa -added check upon system's verity Light's G statistic to measure program's agreement with both experts conjointly held as standard	-numbers of correct , incorrect definitive and tentative diagnoses and failure to make correct diagnoses -tentative is unresolved differential diagnosis, correct if real is topmost
results	-Kappa for each diagnostic area; average experts and MAC 0.86 average between experts 0.765 concl: findings support the validity of MAC through its classification congruence with expert psychologists	-system, clinician and discussants 17/43, 23/43, 29/43 definitive correct -8/43, 5/43, 6/43 tentative correct concl: performance on the 19 cases appeared qualitatively similar to that of the hospital clinicians but inferior to that of the case discussants. The evaluation demonstrated that the present form of the program is not sufficiently reliable for clinical diagnosis

reference	Murray(1986)	Nelson(1985)
system characteristics	system based on databank containing 2500 patients, to assist in predicting outcome in early stages of head injury -for purpose of study collapsed to 3 outcome categories and their probabilities (add to total of 1)	RECONSIDER: internal medicine; diagnostic prompting aid; listing diseases that might be considered for inclusion in differential
object evaluation	to compare the computer predict-ions with those made by clinicians, for model may be telling us no more than what is obvious clinically	whether list contains correct diagnoses
experimental setup	-computer and physicians provide diagnosis and probabilities; compared to real outcome -asked number again after 9 mths	abstract of positive findings in admission notes entered into computer by 7 people; outcome compared directly to actual
test input	10 cases selected at random from databank; neurosurgeon prepared case history	100 consecutive 1st admissions to internal medicine service at university medical centre
way of entry		
test against	24 overseas neurosurgeons & 5 experts	7 people (other state) : computer scientist to experienced clinicians
standard	real outcome after 6 months	discharge diagnosis
comparison	direct	-direct comparison -whether discharge diagnosis included in list of 40
analysis	-probability triangles -variability	-ordinal position on each disease list; multiple diagnoses each diagnosis counted separately -performance as function of number of terms entered
results	-besides probability triangles for 3 cases, no numeric details given -considerable variability after 9 months concl: examples show that under these conditions the prognostic system performs as well if not better than experienced neurosurgeons	-the correct diagnosis was present in at least one version 98/105 of the time -histogram; if diagnosis present, more likely to be at top of list -average for single diagnosis 67% correct; 33 % misses average for 3 diagnoses 42% correct; 58% misses concl: results suggest that useful diagnostic prompting tool can be constructed along these lines

reference	Quaglini(1988)	Reggia(1985)
system characteristics	ANEMIA: diagnosis and management of anemic patients 65 disease entities	TIA expert system: assists with management of transient ischemic attacks
object evaluation	-establish if performance can be distinguished from expert hematologist -interexpert consensus	-determine how accurately system reproduces decision making of stroke specialists at the institution -compare treatment recommendations to what was actually done to gain perspective into usefulness
experimental setup	Turing test: experts and system given patient data provide diagnosis and reasoning; experts blindly judge outputs of others and of system; reasoning measured using questionnaire	-localization and classification of system directly compared to stroke specialist who reviewed cases -treatment recommendation of system compared to actual treatment physicians
test input	representative sample of 30 cases from 1 hospital selected by 2 independent hematologists	103 seen at 1 stroke centre; 78 random selected; included if referring physician scored TIA 25 all classified as having TIA
way of entry		
test against	6 experts; unaware of each others identity; different countries Europe	-diagnosis: stroke specialist -treatment: non stroke specialists
standard	chose not to have one	implicit for diagnosis
comparison	6 experts judge output & reasoning unacceptable, weakly acceptable, acceptable, ideal	direct
analysis	-measures of association; T_b, T_c,γ -tables of ratings given to experts and system -consensus percentage agreement	weighted Kappa
results	-26 of 30 diagnoses at least 3 satisfactory ratings for system -agreement in grading: mean percentage agreement < 50% concl: combining ratings for both diagnostic accuracy & reasoning the performance cannot be distinguished from that of an expert hematologist	-62 patients both stroke specialist and system specified localisation $K_W = 0.79$ -treatment actual - system $K_W = 0.31$ concl: subjective analysis of these differences suggests potential second opinion

reference	Rothschild(1990)	Soula(1988)
system characteristics	DXCON: critiquing system discussing radiologic workup of obstructive jaundice	PROTIS: therapeutic advice to non-specialist practitioners in field of non-insulin dependant diabetes
object evaluation	explore the implications of subjective issues in design and implementation of a validation of a critiquing system	evaluate kbase and analyse problems arising in practical use
experimental setup	-test critiques given to judges to give free comments -comments given to expert for expert opinion -evaluators categorize opinions and study implications for the system	-3 centres; each examines 50 own cases plus 100 provided by 2 others; after entering own diagnosis computer diagnosis is shown -compare 4 diagnoses (3 + system) directly to each other
test input	data abstracted from 10 real cases generate test critique at certain point in series of tests; points chosen at random	150 patients; from 3 centres
way of entry	data entered by knowledge engineer	teams of physicians
test against		3 teams of diabetes experts in different French hospitals
standard		not explicit
comparison		direct
analysis	-3 independent judges with extensive expertise; free comments -domain expert; opinion on comments -evaluators study comments	-according to rank of advice given -isolated, missing, identical & distinct propositions
results	-judges comments broken down into 4 categories: detail, accuracy, scope, wording; number of comments counted -decide whether experts finds judges' comments valid and whether system should be changed	-% of times 1st rank also appears as 1st rank for other centre e.g. C1-system 64%; C1-C2 55% -in 13% of cases practical exploitation raised serious problems -in 7 cases at least 1 expert considered advice given "dangerous" for patient concl: confrontation between system and the experts gives equal and usually better results than confrontation between experts

reference	Spitzer(1969)	Wong(1990)
system characteristics	DIAGNOII: computer program for psychiatric diagnosis, based on a logical decision tree model 46 diagnoses comprising output; one diagnosis for each subject	ABVAB: diagnosis of abnormal vaginal bleeding
object evaluation	to test whether system simulates clinical judgement	whether preference of diagnosis improves in accuracy when both history and physical exam data used; relative importance of both
experimental setup	compare system diagnosis directly to psychologists of different levels of experience who know and those do not know the subjects or patients	-compare diagnosis by system to real diagnosis and also -CF's related to physical exam multiplied by 3 factors (0,0.4,0.7); afterwards also for history rules -outputs compared
test input	-100 real cases provided by 9 psychiatrists in private practice and 13 2nd year psychiatry residents at institute; any subject on whom fairly complete information available -46 hypothetical cases; one case for each possible output -divided into 4 packets with same mix; each packet given to two other	44 cases (there are 5 different diagnoses in the sample)
way of entry		
test against	-those provided test cases -3 psychiatrists, 5 psychiatric residents end 2nd year; given only data which is used by computer	
standard	not explicit	there is one (but not known which)
comparison	direct	direct
analysis	-weighted Kappa	-percentage correct -for all of the weights; percentage correct in graph
results	-24 Kappa's for real cases vary from 0.29 to 0.79; mean between clinicians 0.45; mean between computer and clinician 0.45 -lower Kappa's when comparing with those who supplied case (mean 0.41), since they had access to all information concl: system is able to simulate the clinical diagnostic process to a high degree	-of 44, 70.4% in 1st preference -history alone: 66% in 1st -best results using both history and physical exam data concl: the testing results are quite satisfactory in spite of the limitations of time and the small domain; laboratory and imaging investigations will be incorporated and will hopefully achieve higher level of accuracy

reference	Yu(1979)	Zagoria(1983)
system characteristics	MYCIN: advice diagnosis and treatment of infectious diseases -part evaluated: choice of antimicrobials in the management of meningitis	Bayesian system for early bedside evaluation of stroke patients -for study 3 categories & probabilities
object evaluation	determine whether therapeutic regimens are as reliable as those that infectious disease specialist would recommend	investigating transferability; hypothesis that probabilities from large distant data base could form basis of accurate Bayesian system for stroke etiology
experimental setup	-10 prescribers (incl. system) prescribe therapy for 10 test cases -evaluators (experts) assess prescriptions without knowing identity prescribers or knowing that one is a computer program	3 Bayesian systems (2 low cost and 1 based on data from geographically distant population) & 3 physicians given same input investigate output directly
test input	10 selected by independent physician; retrospective from 1 hospital; diagnostically challenging; diverse	100 random patients admitted to hospital over three years
way of entry		
test against	5 faculty members; 1 senior postdoctoral fellow; 1 senior resident; 1 senior medical student; treating physician	-3 clinicians of different expertise -2 other Bayesian systems based on local probabilities
standard	implicit	discharge diagnosis
comparison	8 experts other than Stanford; score: equivalent, acceptable alternative, not acceptable	direct
analysis	-one way analysis of variance for overall difference system-phys. -Tukey studentized range test; individual differences	-Chi-square (McNemar) for difference in error rates (E.R) -accuracy coefficient (Q')
results	-65% of system output acceptable corresponding mean rating of 5 faculty members 55.5% -significant difference among prescribers -system never failed to cover a treatable pathogen concl: system compared favourably to experts, therefore believe that it will be valuable resource to those physicians with limited experience in domain, however further investigations in clinical environment are warranted	system tested:Q' =0.53; E.R. =0.17 e.g.other system:Q'=0.4;E.R.=0.24 e.g. physician :Q'=0.41; E.R. =0.24 concl: results provide significant support for argument that data collected at other institution can form basis for relatively accurate Bayesian system

reference	Adams(1986)	Bankowitz(1989)
system	Bayesian system for computer aided diagnosis of acute abdominal pain	QMR (quick medical reference) diagnostic program, which at one level provides a ranked list of diagnostic hypotheses in domain of internal medicine
object evaluation	test hypothesis that the program could be transferred to various types of hospital; that it could be used by doctors with no previous experience of computers and that clinical and financial benefit results	-to evaluate accuracy of computer-aided consultation service -study impact on diagnostic behaviour
experiment setup	in 8 hospitals physicians performance compared directly to standard before and after introduction of system. In 4 hospitals doctors into 4 groups; data collection forms, forms&computers, forms&feedback, forms&feedb&comp.	ward team and consultants asked differential before consultation with computer; consultants carry out consultation and provide results to ward team; ward team and consultants asked differential afterwards
test input	prospective; baseline: 1 year; 4075 cases test period: 2 years; 12662 cases total 16737 cases -for 7757 patients the doctor was encouraged to obtain immediate computer feedback	31 patients in 2 hospitals; diagnostically challenging with uncertain diagnosis; only included if suspected main diagnosis in kbase
way of entry	-computer used personally by doctor in 44.5% of possible cases; -by research assistant in 29.6% of possible cases	consultants in the 2 hospitals
test against	unaided physicians during baseline period; over 250 doctors in 8 centres	-ward team -consultants; either fellows or assistant professors proficient with system
standard	final (discharge) diagnosis each set of patient data independently checked by at least 2 other people	definite diagnosis confirmed by histologic etc. data; or 2 physicians were convinced; or followed up for 6 months; 20 of 31 had final diagnosis
comparison	direct comparison, according to criteria decided in advance by consensus view of the project leaders	direct comparison
analysis	-accuracy -diagnostic/management errors -cost savings	-percentage of correct diagnoses contained in differential; percentage of correct top diagnoses & 95% confidence interval; differences -percentages of diagnoses added to differential after consultation -rating of educational value and use of consultation on 3 point scale
results	-computer feedback obtained in 75.1% of possible cases -initial diagnostic accuracy rose from 45.6% to 65.3% ($p<0.001$) -bad management error rate fell from 0.9% to 0.2% ($p<0.001$) -savings estimated at 4258 bed nights per year concl: computer aided diagnosis is a useful system for improving diagnosis and encouraging better clinical practice	-QMR 85% (56%-97%:conf.interval 95%) consultants 80% (55%-94%) ward teams 60% (33%-81%) perc. correct diagnoses in list prior to consult -consultation influenced 26 of 31 postconsultation differentials of ward team -rated educationally helpful in 25 of 31 concl: system provided reasonable diagnostic suggestions not previously considered by the ward teams

reference	Fieschi(1990)	Kent(1985)
system	SPHINX: treatment of diabetes *not really performance measurement	ONCOCIN: consultation system for use in the management of patients enrolled in cancer chemotherapy protocols *not really performance measurement
object evaluation	judge to what extent such a system can provide help, be useful and be used by GPs in their everyday practice	impact of a computer-based data management system on the completeness of clinical trial data
experimental setup	GPs use system over period of time, GPs both judges and system users -using system connected via the French terminal Minitel	data completeness measured before and after introduction of the system; also measured after introduction when the system could not be used
test input		prospective test input, patients with Hodgkin's disease; pre-ONCOCIN: patients enrolled in chemotherapy protocol at Stanford during period; 20 patients; 66 visits post-ONCOCIN: 29 patients; 114 visits ONCOCIN used by physician in 56 visits; not used in 58
way of entry	38 French GPs used system during 6 months; GPs selected by picking names at random form a year-book of physicians in the region	11 oncology fellows at Stanford
test against		10 oncology fellows at Stanford; 5 of the fellows also conducted 25 of the post-ONCOCIN visits
standard		
comparison		direct comparison
analysis	-study traces to identify difficulties and study what is used most: therapeutic advice,dietetic advice, issuing diet sheets, general diabetes info -quantitative (number of calls, duration etc.) -questionnaires	-differences in proportion of items completed for each of the levels of the factors complexity of protocol and experience; standardized and combined by weighting inversely by the variance, resulted in standard normal variate, two-tailed P values calculated
results	-86% of physicians declare they have learnt something through the system -of the functions proposed, the therapeutic aid is the most requested and seems to be the most useful in everyday practice	-percentage expected physical findings 74% (pre) to 91% (post) $p < 0.05$ -toxicity history <1% to 45% $p < 0.01$ -x-ray 44% to 73% $p < 0.01$ -post-ONCOCIN when system not used: physician-dependent data recording likely to revert to old levels when system not used routinely concl: system can greatly enhance recovery of those data expected for chemotherapy protocol patients

reference	McDonald(1984)	Murray(1990)
system	reminder messages to physicians from an introspective computer medical record (Regenstrief Institute)	predict outcome in early stages of head injury
object evaluation	determine effect of reminder messages on physician behaviour	suggested benefits have been in terms of more appropriate use of resources
experimental setup	-control teams did not receive messages but computer executed logic & kept records -directly compare response rates of control and study groups -27 practice teams; practice teams as units of randomization; each team randomized to study or control. (some served in both groups)	quasi-experimental design with 4 centres; baseline of 12 months then 12 months use and then a withdrawal period, with staggered introduction to each centre
test input	12467 prospective patients during two year study	
way of entry	-research technicians & automatically study group: -61 residents, 11 faculty members and 4 nurse clinicians received reminders	
test against	control group: -54 residents, 11 faculty members and 4 nurse clinicians -faculty members and nurse clinicians served on both groups	neurosurgical units experimental unit 4 centres
standard		
comparison	-response rates -patient outcomes	
analysis	-residents: ANOVA -faculty members on both teams: individual was unit of analysis & paired t-test -nurse clinicians: individual was unit of analysis & paired t-test & Wilcoxon signed rank (small sample)	-accuracy of predications in the field -use of resources relationship between prognosis and intensity or 'aggression' of therapy
results	-study group residents responded to 49% of computer's indications; control 29% ($p<0.0001$) -preventive care was affected -response rates of residents & faculty members identical in control situation, even though faculty members received reminders during study concl: although computer reminder messages are potent activators of existing physicians intentions they have little influence on the acceptance of new practices	

reference	Sutton(1989a)	White(1984)
system	CAD-A and amended program DIAG: Bayesian programs for diagnosis of acute abdominal pain * compare with Adams	application of HELP system to alert physician to a condition that could be of concern in management of patient with digoxin
object evaluation	to maximise and quantify computer diagnostic accuracy	determine effect of system alerts on patient management
experimental setup	CAD-A , DIAG , and clinical performance at three hospitals compared to final diagnosis. -physicians enter own diagnosis before learning computer's diagnosis. Junior doctors received feedback in form of score sheets	after trial period of use of system; double-blind randomized study of 3 months; study frequency of action with and without alerts
test input	6962 cases prospective CAD-A : 6379 prospective DIAG : 583 both retrospective on 6712 of these cases	all patients receiving digoxin; total 396 patients; randomized 211 to alert group and 185 to nonalert group (i.e. alert reports were withheld)
way of entry	2 hospitals: 1st doctor to assess patient completed structured form at all 3 hospitals: cases keyed in by doctor	data are routinely stored; no additional entry required alert messages printed
test against	physicians at 3 Scottish hospitals; hospitals representing range of practice	physicians who did not receive alerts
standard	diagnosis ultimately assigned by the consultant in charge of the case	no standard
comparison	direct comparison of diagnoses	direct comparison of frequency of action
analysis	-accuracy & McNemar's test for paired binary data (null hypothesis that computers simply echoed clinician's diagnosis and so had the same accuracy) -discriminant matrices of physicians' diagnoses against definitive diagnosis and of computer's diagnosis against definitive diagnosis -sensitivity, specificity	-frequency of occurrence of each of the alerts in both groups evaluated with Chi-square -percentages of possible alert related actions -proportion method used in evaluating distribution of physician actions
results	CAD-A : accuracy 48%-59% DIAG : accuracy 56%-62% physicians : accuracy 65% -where computer use was optional the use fell away although only two programs were evaluated, they concl: figures suggest that computer systems based on Bayes' formula have no useful role in the diagnosis of acute abdominal pain	-increase of 22% ($P<0.003$) in actions -patients in alert group 2.84 ($p<0.002$) times more likely to have digoxin withheld on day of alert concl: clinical response indicated that system was successful in increasing awareness of conditions predisposing their patients to digoxin intoxication -objective measurement of actual patient benefit will require study of more extensive patient population

Summary

The research described in this thesis concerns the validation of medical knowledge based systems in general, and of the knowledge based system PLEXUS in particular. PLEXUS is a computer system which is designed to assist physicians in the diagnosis and treatment planning of nerve injuries in the area between the neck and the arm. In order to validate the knowledge based system PLEXUS, a review of the literature was first performed. This led to recommendations for the design of performance evaluations.

Two evaluation studies were carried out for PLEXUS. The first was a study of the system which was aimed at investigating the system's problem solving capacity. The second study was an investigation of the complete human-machine system in a number of hospitals. During this investigation, performance as well as usability and acceptability were addressed. In addition, a more extensive investigation into the attitudes of physicians towards knowledge based systems was performed. Besides information about the knowledge based system PLEXUS, general conclusions and recommendations regarding the design and validation of medical knowledge based systems resulted from these studies.

Due to the multidisciplinary nature of knowledge based system design and development, the literature on the validation of knowledge based systems is very diverse. Most of the validation literature involves evaluation of the performance of a knowledge based system. In this context the term performance is related to the quality of the human-machine system, rather than implying technical performance measures. The empirical evaluation studies which are described can be divided into laboratory evaluations and field evaluations. The aspects of importance in the design of a performance evaluation include the choice of a goal for evaluation, evaluation setup, analysis of the results and possible sources of bias and confounding. Many different evaluation setups and methods of analysing the results have been encountered.

Most investigators are very positive after a laboratory investigation, however, quite often no further evaluations of the systems, such as field evaluation, are reported. The discussion on the evaluation methods found in the literature has led to recommendations for performing medical knowledge based system evaluation studies. From the review it becomes clear that performance evaluation is only a limited part of the validation process, and knowledge based system validation should be a continual process which should proceed in parallel with the design and development of a system.

The recommendations which resulted from the survey of the literature were used to evaluate the performance of the medical knowledge based system PLEXUS. This system is designed to assist neurologists, neurosurgeons, orthopaedic surgeons, rehabilitation physicians and traumatologists in the diagnosis and treatment planning of brachial plexus injuries. The system is meant for physicians who are not specialised in the domain of brachial plexus injuries. The brachial plexus is a network of nerves which is situated in the area between the

neck and the arm, and innervates the muscles of the shoulder, arm and hand. In order to obtain advice from the system, the physician may enter patient history information and results of neurological, neurophysiological and radiological examinations into the system. The system then uses the patient data and the knowledge about brachial plexus injuries which is stored in the system, to advise the physician about the injured locations and the severity of the injury, and suggests a treatment plan. The system's graphical user interface is based on a familiar scheme, and does not require previous computing or typing experience. Preliminary evaluation studies of the system's problem solving performance produced promising results.

The problem solving performance of PLEXUS was evaluated in cooperation with four experts from different European countries. During this evaluation, the diagnoses and treatment plans proposed by the knowledge based system were compared to the opinions of the experts. The opinions were compared directly, as well as being compared by the experts who did not know whether the opinions originated from the computer or from another expert (blind evaluation). Various methods of analysis were used to determine the level of performance which is achieved by the system. The results show that the accuracy of the recommendations provided by PLEXUS is comparable to those obtained from the experts. However, PLEXUS provided a higher fraction of false positive answers. In a number of cases this is caused by the fact that PLEXUS tries to explain more of the dysfunction than the experts do. The intra- and inter-expert variability proved to be rather high in this study. These results are supported by the blind evaluation. In addition, during the blind evaluation the experts were asked to indicate which of the recommendations they thought originated from PLEXUS. The number of times the experts indicated that answers originated from the knowledge based system did not significantly deviate from the number of times this was expected to occur by chance. The relatively limited representativeness of the test cases and the fact that only domain experts cooperated in the evaluation are limitations of the investigation.

PLEXUS was also evaluated clinically in four different hospitals in The Netherlands. The performance of the human-machine system was studied, and the usability and acceptability of the system were addressed. Since the incidence rate of brachial plexus injuries is low, only qualitative results arose from the study. The results show that the performance of the knowledge based system in the hospitals is good, although a number of improvements is still necessary. The number of false positive answers given by the system is relatively high, as was also found in the laboratory investigation. Furthermore, in some cases the patient data that were entered into the system by the physicians were not as complete as had been expected during the development of the system. This may cause the

system to give an erroneous answer in cases where, due to a lack of data, it should not have suggested an answer at all.

The usability of the user interface was investigated by means of videotaping actual interactive sessions during the field evaluation. This provided important information which may be used to update the user interface, so that the system satisfies a number of essential usability requirements. The acceptability of the system was studied by means of a brief questionnaire which was distributed among the cooperating physicians. The results are not conclusive, as a number of physicians indicated that they would use the system if it was generally available, whereas during the field evaluation the system was not used as readily as might have been expected.

In addition, a more extensive investigation into the acceptability of knowledge based systems was performed. It is often stated in the literature that knowledge based systems are rarely used in actual practice. A number of problems which may explain this lack of acceptance has been identified by various authors. Possible solutions to these problems have led to the formulation of requirements which could be of importance to the acceptability of knowledge based systems. The opinion of potential users of knowledge based systems regarding these requirements has been studied by means of a questionnaire which was distributed among physicians and process-operators. Results show that the introduction of a knowledge based system should not lead to a shift in responsibility from the human to the machine. Therefore, it is important for the user to understand how the system works. This requires a system design which helps the user to build up an adequate internal representation of the reasoning process.

From the analysis of the validation of medical knowledge based systems, it has become clear that most of the literature concerning medical knowledge based systems focuses on the system only, rather than on the complete human-machine system. The cooperation between the human and the machine has received little attention. It is necessary to take into account the complete human-machine system from the start of the development of a knowledge based system. This implies that the human-machine system should also be addressed during laboratory evaluation studies, rather than concentrating on the performance of the knowledge based part of the system only. Validation should be fully integrated into the design and development of a knowledge based system.

Samenvatting

Dit onderzoek betreft de validatie van kennissystemen in het algemeen en de validatie van het kennissysteem PLEXUS in het bijzonder. PLEXUS is een computerprogramma dat bedoeld is om artsen te assisteren bij de diagnostiek en behandelplanning van zenuwletsels in het gebied tussen de nek en de arm. Teneinde het kennissysteem PLEXUS te kunnen valideren is allereerst een literatuuronderzoek uitgevoerd naar de validatie van medische kennissystemen. Dit heeft geleid tot aanbevelingen voor het ontwerpen van een evaluatie.

Vervolgens is een tweetal evaluaties uitgevoerd. Het eerste betrof een evaluatie die erop gericht was het nivo van de adviezen van PLEXUS te onderzoeken. De tweede betrof een evaluatie van het gehele mens-machine systeem in een aantal ziekenhuizen. Hierbij werd behalve naar de kwaliteit van de adviezen ook gekeken naar de bruikbaarheid en acceptatie van het systeem. Een uitgebreider onderzoek naar de acceptatie van kennissystemen is uitgevoerd door middel van een vragenlijst. Naast specifieke informatie over het kennissysteem PLEXUS, hebben deze onderzoeken geresulteerd in conclusies en aanbevelingen ten aanzien van het ontwerpen en valideren van medische kennissystemen in het algemeen.

Vanwege het multidisciplinaire karakter van kennissystemen loopt de literatuur op het gebied van de validatie van kennissystemen zeer uiteen. De terminologie is niet eenduidig gedefinieerd en de validatiemethoden die gehanteerd worden variëren van auteur tot auteur. De meeste artikelen betreffen de evaluatie van de kwaliteit van de adviezen van kennissystemen. De empirische evaluaties die beschreven zijn, kunnen worden onderverdeeld in laboratoriumevaluaties en veldevaluaties.

Belangrijke aspecten bij het ontwerp van een evaluatie betreffen onder meer, de keuze van een doel van het onderzoek, de evaluatiemethode, analyse van de resultaten en factoren die mogelijk van invloed kunnen zijn op de validiteit van het onderzoek. In de literatuur worden veel verschillende evaluatie- en analysemethoden beschreven.

De meeste onderzoekers zijn zeer positief na een laboratoriumevaluatie. Vaak is er echter geen vervolgonderzoek beschreven, zoals een veldevaluatie. Het literatuuronderzoek heeft geleid tot aanbevelingen voor het uitvoeren van evaluatiestudies van medische kennissystemen. Het blijkt dat de evaluatie van de kwaliteit van de adviezen slechts een deel is van het gehele validatieproces. De validatie van kennissystemen zou een steeds terugkerend proces moeten zijn, dat parallel verloopt met het ontwerp en de ontwikkeling van een systeem.

De aanbevelingen uit de literatuurstudie zijn gebruikt om PLEXUS te evalueren. Dit systeem is ontworpen om neurologen, neurochirurgen, orthopaedisch chirurgen, revalidatie-artsen en traumatologen te helpen bij de diagnostiek en behandelplanning van plexus brachialis letsels. Het systeem is bedoeld voor artsen die niet gespecialiseerd zijn op het gebied van plexus brachialis letsels. De plexus brachialis is een netwerk van zenuwen dat gelegen is in het gebied tussen

de nek en de arm. Om advies van het systeem te kunnen verkrijgen voert de arts anamnesegegevens en resultaten van neurologisch, neurofysiologisch en radiologisch onderzoek in de computer in. Het systeem gebruikt deze gegevens tezamen met de kennis over plexus letsels, die in het systeem is opgeslagen, om de arts te kunnen adviseren omtrent de gewonde locaties en de ernst van de verwonding, en om een behandelplan voor te kunnen stellen. Het werken met het systeem vereist geen computer- of type-ervaring. Voorlopige evaluatie van het systeem heeft goede resultaten opgeleverd.

De kwaliteit van de adviezen van PLEXUS is geëvalueerd in samenwerking met vier experts uit verschillende Europese landen. Tijdens deze evaluatie zijn de diagnoses en behandelplannen van het systeem vergeleken met de meningen van de experts. De meningen zijn direct vergeleken en tevens hebben de experts de meningen vergeleken zonder te weten of deze van de computer of van een andere expert afkomstig waren (blind onderzoek). Verschillende analysemethoden zijn gebruikt om de kwaliteit van het systeem te bepalen. Hieruit blijkt dat de gevoeligheid van het systeem vergelijkbaar is met die van de experts. Echter, PLEXUS geeft een groter aantal fout-positieve antwoorden. In een aantal gevallen wordt dit veroorzaakt doordat het systeem meer van de dysfunctie probeert te verklaren dan de experts. De inter- en intra-expert variabiliteit waren aanzienlijk in deze evaluatie. De resultaten van de directe vergelijking worden bevestigd door de resultaten van de blinde evaluatie. Voorts is tijdens de blinde evaluatie gevraagd of de experts voor iedere patiënt wilden aangeven welke van de adviezen volgens hen afkomstig was van PLEXUS. Het aantal maal dat de experts aangaven dat het advies afkomstig was van PLEXUS week niet significant af van het aantal maal dat dit door toeval verwacht kan worden. De enigszins beperkte representativiteit van de test-cases en het feit dat alleen experts bij deze evaluatie betrokken waren zijn beperkingen van deze studie.

PLEXUS is tevens klinisch geëvalueerd in vier verschillende Nederlandse ziekenhuizen. Naast de kwaliteit van het mens-machine systeem, zijn de bruikbaarheid en acceptatie van het systeem onderzocht. Daar het aantal plexus letsels dat jaarlijks plaats vindt gering is, bleek het slechts mogelijk om de resultaten kwalitatief te analyseren. Uit de resultaten bleek dat alhoewel de prestatie van het systeem goed is, een aantal verbeteringen nog noodzakelijk is. Zoals ook volgde uit het laboratorium-onderzoek, geeft het systeem relatief veel fout-positieve antwoorden. Voorts waren in een aantal gevallen de gegevens die door de artsen in de computer ingevoerd zijn niet zo compleet als was verwacht tijdens de ontwikkeling van het systeem. Dit kan ervoor zorgen dat het systeem een verkeerd antwoord geeft in gevallen waarin het systeem vanwege gebrek aan gegevens geen uitspraak had mogen doen.

De bruikbaarheid van de user-interface is onderzocht door middel van het maken van video-opnamen tijdens de veldevaluatie. Dit leverde informatie op die gebruikt kan worden om de user-interface te verbeteren, zodat deze voldoet aan een aantal essentiële bruikbaarheidseisen.

De acceptatie van PLEXUS is onderzocht door middel van een korte vragenlijst die beantwoord is door de artsen die bij de veldevaluatie betrokken waren. De resultaten zijn niet doorslaggevend, daar een aantal artsen aangaven dat zij het systeem zouden gebruiken als het algemeen beschikbaar was, terwijl tijdens de veldevaluatie het systeem niet werd gebruikt in de mate die was verwacht.

Voorts is een uitgebreider onderzoek uitgevoerd naar de acceptatie van kennissystemen. In de literatuur wordt vaak melding gemaakt van het feit dat slechts zeer weinig kennissystemen daadwerkelijk in de praktijk toegepast worden. Door een aantal auteurs zijn verschillende oorzaken aangegeven die dit gebrek aan acceptatie zouden kunnen verklaren. Mogelijke oplossingen voor deze problemen hebben geleid tot het formuleren van eisen die van belang zouden kunnen zijn voor de acceptatie van kennissystemen. De mening van potentiële gebruikers van kennissystemen ten aanzien van deze eisen is onderzocht door middel van een vragenlijst. Deze vragenlijst is ingevuld door artsen en proces-operators. Uit de resultaten blijkt dat de introductie van een kennissysteem niet mag leiden tot een verschuiving van de verantwoordelijkheid van de mens naar de machine. Hiertoe is het van belang dat de gebruiker begrijpt hoe het systeem werkt. Dit vereist een systeemontwerp dat de gebruiker in staat stelt een goede interne representatie van het redeneerproces op te bouwen.

Uit de analyse van de validatie van medische kennissystemen is het duidelijk geworden dat de meeste literatuur op dit gebied zich slechts richt op het systeem, in plaats van op het gehele mens-machine systeem. De samenwerking tussen mens en machine krijgt zeer weinig aandacht. Het is noodzakelijk om vanaf het begin van de ontwikkeling van een kennissysteem het complete mens-machine systeem in aanmerking te nemen. Dit houdt in dat naast het kennisgedeelte van het systeem, het gehele mens-machine systeem tijdens de laboratoriumevaluatie in beschouwing genomen moet worden. Validatie moet volledig geïntegreerd zijn in het ontwerp en de ontwikkeling van een kennissysteem.

Acknowledgements

During the past six years I have had the pleasure of working at the Laboratory for Measurement and Control. The topic of my research, my friends and colleagues at the Laboratory have made this into a very enjoyable period. I would like to thank everyone I worked, ate and drank (very important for civil servants) with for their direct and indirect contribution to this thesis.

I would especially like to thank Henk Stassen, who has guided my research in a most enthusiastic way. Eric Backer, my co-promotor, was always prepared to read my papers and give helpful directions.

Rob Jaspers has continued to be very interested in the work he started, and has always helped me even after he had left the university. When I graduated he was somewhat concerned about the fact that I, being a mechanical engineer, had never mended a bicycle tyre, but I can now safely go out into the real world.

I had the pleasure of writing a chapter of this thesis with Anne-Marie Sassen. I have great respect for the way in which she succeeded in finishing her own thesis just after the birth of Eline. Our discussions helped me to focus on the way in which medical knowledge based systems should be designed and validated.

A knowledge based system for brachial plexus injuries can never be designed and validated without the cooperation of physicians who know everything there is to know about the brachial plexus. Dr. A.C.J. Slooff, prof. dr. R.T.W.M. Thomeer and Martijn Malessy were always prepared to check the medical content of my papers, to analyse patient data and to help in the evaluation of PLEXUS. I would also like to thank Mrs. J. Slooff for her hospitality on my trips to Heerlen.

The laboratory evaluation of PLEXUS would not have been possible without the kind help of dr. A. Santos Palazzi, prof. dr. A.O. Narakas, dr. R. Birch and prof. dr. J-Y. Alnot, who spent a great deal of time gathering data and analysing cases.

Dr. C.W.G.M. Frenken, dr. P.J. de Jong, dr. V. van Kasteel, dr. W. Perquin, dr. J.F. Ploegmakers, dr. M. Prick, dr. R. Schellens and dr. T.W. van Weerden are kindly acknowledged for their cooperation in the clinical evaluation of PLEXUS.

The help of Aad Gutteling, Jaap van Dieten, Fulko van Westrenen and Bob Goedhart has also contributed to this thesis.

Over the years, many people have been involved in the PLEXUS project. I would like to thank Jurriaan Grolman, Rutger de Vries, Willem van

Heerebeek, Bram Buitendijk, Edwin Franse and Fred Voorhorst for their contribution to the development and validation of PLEXUS.

Fred ter Haar was also involved in the PLEXUS project and taught me all about the Tour de France. This allowed me to win the Tour de France pool in '92 and to beat the organiser, Frans van der Helm, my room-mate and friend who taught me how to head a football (yes, inside the building) and has still not succeeded in explaining his sense of humour to me.

Finally, I would like to thank Ernst, Frans, Marjolein and Zwaantje, my family and best friends, who have always encouraged and supported me.

Curriculum Vitae

Op 22 april 1963 ben ik in Delft geboren. In 1982 heb ik de middelbare school afgerond op St. Catherine's School in Guildford, Engeland. Na deze periode op een meisjesschool kwam ik terecht op een jongensschool, de Technische Universiteit Delft, om Werktuigbouwkunde te gaan studeren. Tijdens mijn studie ben ik met veel plezier Commissaris Onderwijs geweest bij de Studievereniging Gezelschap Leeghwater. Ook ben ik een aantal jaar lid geweest van de Faculteitsraad van Werktuigbouwkunde.

In augustus 1987 ben ik afgestudeerd bij de Vakgroep Meet- en Regeltechniek op een expertsysteem voor de behandelplanning van plexus brachialis letsels. Dit onderzoek heb ik voortgezet als Assistent in Opleiding. Het onderzoek heeft zich uiteindelijk toegespitst op de validatie van medische kennissystemen. Een artikel op het gebied van de klinische evaluatie van medische kennissystemen is genomineerd voor de NVKI-prijs voor de beste toegepast wetenschappelijke bijdrage op de Nederlandstalige AI Conferentie (NAIC) 1992. Tijdens mijn aanstelling als AIO ben ik gedurende twee jaar lid geweest van de Universiteitsraad en van de Universiteitsraadscommissie voor Onderwijs en Onderzoek.