INTERACTIVE CONSULTATION AND FORMALIZATION OF KNOWLEDGE
applied to ovarian pathology

(INTERAKTIEF RAADPLEGEN EN FORMALISEREN VAN KENNIS)

(toegepast op de ovarium pathologie)

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR
AAN DE ERASMUS UNIVERSITEIT ROTTERDAM
OP GEZAG VAN DE RECTOR MAGNIFICUS
PROF. DR. C.J. RIJNVOS
EN VOLGENS BESLUIT VAN HET COLLEGE VAN DEKANEN.
DE OPENBARE VERDEDIGING ZAL PLAATSVINDEN OP
DONDERDAG 28 SEPTEMBER 1989 OM 16.OO UUR

DOOR

ASTRID MARIA VAN GINNEKEN

geboren te 's Gravenhage

PROMOTIECOMMISSIE

PROMOTOR:	PROF. DR. IR. J.H. VAN BEMMEL
PROMOTOR:	PROF. DR. IR. A.W.M. SMEULDERS
OVERIGE LEDEN:	PROF. DR. IR. J.D.F. HABBEMA
	PROF. DR. H.C.S. WALLENBURG
CO-PROMOTOR:	PROF. DR. J.P.A. BAAK

Ad majorem Dei gloriam

Aan mijn vader en moeder

Dankwoord

Mijn dank begint bij Prof. Dr. Ir. J.H. van Bemmel, die mij door zijn inspirerende benadering op het pad van de medische informatica heeft gezet. Ik waardeer het dat hij altijd met enthousiasme en vertrouwen in mijn werk achter mij heeft gestaan. Het originele idee voor dit onderzoek was al veel eerder geboren bij Prof. Dr. J.P.A. Baak en Prof. Dr. Ir. A.W.M. Smeulders en zij waren het, die de aanzet ertoe al bijna een jaar eerder waren begonnen. Prof. Smeulders heeft mij gedurende het hele onderzoek intensief begeleid. Door zijn kritische, eerlijke en opbouwende houding heb ik geleerd om meer voor mijn inzichten op te komen, kritiek te verwerken en mijn gedachten meer gestruktureerd weer te geven. Op de niet altijd even makkelijke weg leerde ik veel over "wetenschap en onzekerheid". Prof. Baak ben ik dankbaar voor het nakijken van de vele diagnose beschrijvingen en dia's. Zijn inzet was van grote waarde voor het tot stand komen van de evaluatie. Hij gaf mij door zijn enthousiasme altijd het gevoel met "iets veelbelovends" bezig te zijn en ik dank hem voor zijn persoonlijk in mij gestelde vertrouwen.

Dr. Jansen heeft met zijn grote doorzettingsvermogen en toewijding gezorgd dat binnen het zeer strakke tijdschema al de teksten en de beeldplaat gereed kwamen. Ik wil hem ook graag bedanken voor zijn onmisbare inzet bij het struktureren van kennis in de ovariumpathologie. Voorts wil ik Ing. I. Brooijmans, hartelijk bedanken voor haar inzet bij het schrijven van de programmatuur. Ondanks de nodige tegenslagen bleef zij altijd bereid om aanpassingen en uitbreidingen te maken tot de programmatuur aan alle wensen voldeed. Drs. A.J.G. Fiege, A. Bos en J.Hos hebben veel belangrijk werk verzet met het invoeren van alle teksten, het zoeken in coupe archieven en het bijhouden van het project archief. Aan de fijne loyale samenwerking en vriendschappelijke verstandhoudingen binnen ons "team" heb ik fijne herrinneringen.

Drs. J. van der Lei heeft mij met raad en daad bijgestaan bij het ontwerpen en programmeren van de software voor het opslaan van geformaliseerde kennis. Dankzij zijn hulp kon het systeem binnen de beperkte tijd gerealiseerd worden. Hij scherpte mijn gedachten aan en wist met de nodige dosis humor alles op zijn tijd te relativeren.

Ing. A. Dekker, Ing. M. van Zijl, Drs. M. Worring en Drs. E van Mulligen hebben als stagiares waardevolle bijdragen geleverd aan de programmatuur. Voorts wil ik mijn dank uitspreken voor de medewerkers van het Laboratorium van het Pathologisch Instituut onder leiding van J. Koevoets. Zij hebben grote hoeveelheden coupes verzorgd van hoge kwaliteit. Ik waardeer de fotografen J. van Veldhuizen en H. Oskam, die altijd voor een zeer snelle ontwikkeling van onze grote hoeveelheden dia's zorgden. Drs. C. Dekkers heeft op uitstekende en vooral flexibele wijze de financien van het project beheerd. Graag wil ik L. de Langen bedanken voor de enthousiaste en kundige hulp, waaraan dit proefschrift haar layout te danken heeft.

Mijn dank gaat ook uit naar de leden van de voormalige Ovarium Tumoren Comissie en de pathologische afdelingen van de ziekenhuizen die materiaal beschikbaar stelden uit hun archief.

Bij deze wil ik ook mijn vrienden bedanken voor hun belangstelling en de aangename ontspannende avondjes "primaatbridge". Tenslotte wil ik grote waardering en dankbaarheid uitspreken voor mijn ouders, die met warm medeleven, bemoediging en stimulans de moeilijke momenten en de hoogtepunten met mij deelden.

The study was financed by the Praeventiefonds (grant 28-1207)

CONTENTS

CHAPTER 6

CHAPTER 1

General Introduction

1.1 INTRODUCTION

The contents of this thesis concerns diagnostic support in pathology as one of those domains in medicine where diagnosis is primarily based on visual observations and where temporal reasoning plays a secondary role. The majority of diagnostic problems in pathology involve the histologic typing and grading of tissue abnormalities. In pathology, the diagnosis is primarily based on macroscopic and microscopic observations, although a final diagnosis requires them to be interpreted in the context the patient's present symptoms and disease history. In brief, the aim of this study is to offer efficient and versatile diagnostic support with a potentially large scope. The diagnostic process and the role of reference knowledge therein are now outlined.

1.2 REFERENCE KNOWLEDGE

A schematic representation of the analytical diagnostic process in pathology is shown in Fig. 1 [1-4]. By training and through experience, a pathologist has learned to extract relevant features from the histologic image. A simplified conception of the image is then compared with typical conceptions of diagnoses, which the pathologist has in mind. When no satisfactory match is found the pathologist may use books, atlases and well-documented cases, or consult colleagues to acquire reference knowledge to refresh or supplement the personal knowledge.

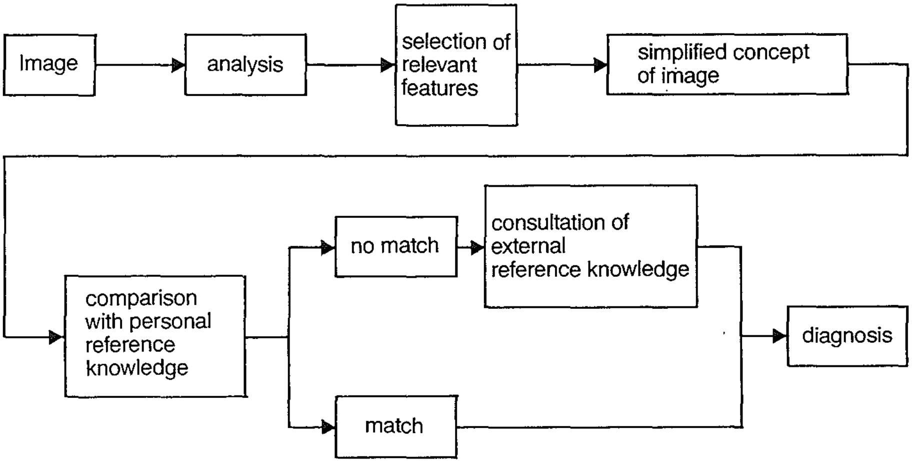


Image
analysis
selection of relevant features
simplified concept of image
comparison with personal reference knowledge
no match
consultation of external reference knowledge
match
diagnosis

Figure 1. A schematic representation of the analytical diagnostic process in pathology.

All sources of knowledge used in the diagnostic process constitute reference knowledge. We will call that part of the reference knowledge, which the pathologist accesses from the outside, external reference knowledge. Though books and atlases are the most widely used source of external reference knowledge, they are not the most suited medium for solving the "inverse problem of diagnosis", which requires information to be accessible via diagnostic features.

1.3 DIFFICULTIES IN DIAGNOSIS MAKING

Although many tissue abnormalities are diagnosed with high consensus among experts, part of these abnormalities are difficult to diagnose, reflected in a considerable amount of disagreement among pathologists. The scores for disagreement are studied in certain fields of tumor pathology and the percentages indicate the portion of the examined slides which were diagnosed without complete agreement. Examples in pathology are: the tumors of the ovary (40%), the liver (48%), the breast (14%-85%), and the lymphomas (10%-47%) [1,5-8]. These findings are not typical for pathology but occur in many fields of medicine [9]. In tumor pathology, both the type and grade of a the tumor potentially have consequences for therapeutic measures. A diagnosis covers a certain range of histological variations. As the demarcation of the diagnostic ranges varies in precision, part of the diagnoses present difficulties in typing. The grade of a tumor indicates the poorest level of differentiation of its composing cells. Here, the analogous changes in differentiation have to be made discrete.

Three important sources of error are recognized in diagnosis making: image interpretation, classification and verbal expression [1]. As to image interpretation, it is well known that artifacts and differences in the surroundings of a tissue component may result in a different interpretation of that component. As a result, features may be given too much weight or too little in the diagnostic considerations.

Classification errors basically are of the random or the systematic type. Random shifts are due to several variable influences: time of day, fatigue, recent experience, discussions with colleagues and so on. Random shifts are reflected by the phenomenon of restricted intra-observer agreement. For example, a pathologist, confronted with a highly malignant tumor may diagnose the next case with a bias to the benign side. The reverse may also occur. Systematic shifts concern more permanent differences between experts, which are rooted in differences in training and experience. Systematic shifts may also occur in the diagnosis of one pathologist as an expression of gaining new insights.

Finally, in their verbal expression, pathologists use different words to express the same intention. This can be distinguished in synonymy (different words, same meaning) and homonymy (same word, different meaning). Hence, diagnosis descriptions from books or colleagues may be interpreted differently by different pathologists [10].

Over the last decades much research has been done to make diagnoses more consistent. There are two distinct starting points for offering diagnostic support to achieve this goal:

- Increasing objectivity in the acquisition and interpretation of features
- Increasing the efficient use of external reference knowledge

1.4 TECHNIQUES USED IN QUANTITATIVE PATHOLOGY

When features become more objective their reproducibility usually increases and this is crucial for obtaining a better consistency in diagnosis. Many applications of diagnostic support involve extraction of objective features by means of morphometry, flowcytometry, and image analysis.

The term morphometry usually denotes simple and inexpensive methods for counting and measuring tissue and cell elements such as the differentiation of leukocytes, the number, form and area of nuclei, and the number of mitoses [11-14]. Examples of the significance of morphometric data in pathology are the positive correlation with malignancy of the nuclear area in pleural biopsies, the cellularity in estrogen receptor positive breast cancer, the number of mitoses in myometrial tumors and the volume percentage of epithelium in ovarian tumors [14,15]. Although for several types of tumors the application of morphometry can be a valuable contribution to the selection of a suitable therapy, the results of this technique are greatly dependent on the selection of the tissue section by the pathologist. In addition, it should be noted that morphometric methods are laborious when many tissue sections have to be examined to obtain reliable results.

The most well known application of flowcytometry is the assessment of the amount of DNA in the nuclei of cells. A flow cytometer can perform this measurement accurately and at high speed on large amounts of cells in suspension. In many tissues, such as lymphatic, ovarian, lung, bone and endometrial tissue, the amount of aneuploid DNA in the cells is an indicator of the degree of malignancy [16-20]. There are tumors where the presence of aneuploid DNA provides sufficient information about the malignancy of the tumor only when it is combined with histologic data [21-23]. Flowcytometers are still expensive and mostly used in research laboratories.

Image analysis usually involves a combination of image processing and pattern recognition. Image processing encompasses a variety of techniques, which can be applied to perform measurements on tissues and cells. In fact, many of the data, which can be acquired with morphometry could also be obtained quicker and more precisely with image processing techniques. Common steps in image processing are image digitization, image segmentation to isolate objects in the image, artefact rejection, and measurements describing the shape and density of the objects [24,25]. Image cytometry is an application of image processing to single cells and is especially suited for measurements on the geometry and architecture of cell components. When image cytometry is exclusively used to assess the amount of DNA in cells it is relatively time consuming when compared to flow cytometry [26]. Examples of applications on tissues are the epithelium stroma ratio in ovarian tumors [27] and the automated counting of mitoses [28]. At present, image processing is mainly used for research purposes.

Objectivity in the interpretation of measurements can be pursued by the application of pattern recognition techniques [29-31]. Many classification problems in pathology concern the differentiation of cells, which cannot easily be distinguished. Besides classification on the basis of features of known importance, statistical techniques can also be applied to detect features with discriminative power among a set of features of previously unknown relevance [32]. Pattern recognition techniques are often added to image processing applications to complete the task of identifying diagnostically relevant features [33]. Examples of applications in pathology can be found elsewhere [34-36]. The application of statistical pattern recognition for diagnostic support is still limited: extensive research on large sets of well-documented cases is necessary to provide, for a broad domain, the classifiers which can be used in diagnostic problem solving.

Although morphometry, cytometry and image analysis are found in an increasing number of applications in pathology, these techniques are very specialized and offer only part of the data necessary to arrive at a final diagnosis. In fact, these techniques serve for greater precision of diagnoses: a diagnosis based on routine techniques must reveal whether quantitative pathology is justified, i.e., whether well-defined quantitative criteria exist for that diagnosis.

1.5 CONSULTATION OF REFERENCE KNOWLEDGE

The availability of quantitative diagnostic support does not change the fact that diagnosis making is still primarily based on a variety of qualitative criteria for which the pathologist may want to consult reference knowledge. Specialized knowledge of colleagues as source of reference knowledge is only available for part of the time and a limited number of diagnostic problems. Therefore, it is important to promote efficient use of other sources of external reference knowledge such as books, atlases, journals and patient archives. These sources are characterized by the fact, that they consist of both text and pictures. Especially in pathology, pictures convey an important part of the knowledge as they have a permanent nature and can be studied repeatedly. Pictures may express features, which are difficult to describe in words. Specifically, pictures are an excellent medium to convey information concerning tissue architecture.

A variety of diagnostic support applications involve access to reference knowledge. Archives have been automated to permit efficient retrieval of previously diagnosed cases for the purpose of comparison or evaluation studies [37]. Large on-line network connections to medical libraries and storage of literature references and abstracts on optical discs facilitate the access to recent publications. Yet, most pathologists do not have direct access to these facilities and if they had, it would require too much searching effort to find the knowledge they need. Diagnostic support must be available at the pathologist's desk, adapted to the daily diagnostic needs.

Videodiscs are currently developed to offer large sources of pictorial reference knowledge. They are used in many fields of medicine such as normal histology [38], echocardiography [39], emergency care and student training programs [40]. Examples of videodiscs in pathology are the ones containing pictures of the ovary [41], the lung [42], the bone marrow [43] and the lymphatic system [44]. Videodiscs are an excellent medium for storing and accessing large sets of pathologic images. In order to take full advantage of videodiscs for diagnostic support their contents must be well-documented and integrated with reference knowledge concerning diagnostic and differential diagnostic criteria.

1.6 EXPERT SYSTEMS

At present, much attention is focused on the development of expert systems, which contain explicit knowledge with reasoning capabilities for the purpose of diagnostic decision support. Several kinds of knowledge representation and inference techniques have resulted in high-performance systems in pathology. Rule-based systems [45] have been developed for diagnosing diseases of the liver [46], for the classification of leucocytes [47], the diagnosis of bone marrow aspirates [43], and for distinguishing different types of leukemia on the basis of immune pathological techniques [48]. For small domains the rule-based approach is very straightforward, but explanation facilities have to be added as rule-bases are not an appropriate representation for inspection of the contents of the knowledge base. The main difficulties in the development of large rule-bases were the maintenance of consistency and the process of updating them without losing previous performance elsewhere [45,48].

A different approach is taken in the PATHFINDER system for diagnostic support in the field of lymphomas [44]. Here, knowledge is represented as lists of features per diagnosis with for each feature an indication of its specificity, sensitivity and importance. These values are used by the inference strategy to generate an ordered list of possible diagnoses. The advantage of the knowledge representation of PATHFINDER over a rule-based representation is its transparency: it is easy to see which diagnostic features are present for a certain diagnosis. However, PATHFINDER lacks the possibility to verbally express the context of features and tissue architecture. Although this information can be accessed by the pathologist via pictures, it cannot be propagated in the inference process.

In TEGUMENT [49], an expert system for skin diseases, the user passes through a decision tree, an approach which resembles the systematic classification of plants. A list of potential diagnoses and a definite diagnosis are made on a purely qualitative basis. As the other applications mentioned, TEGUMENT supports decision making but does not allow for consultation of knowledge.

Uncertainty plays an important part in diagnosis. Several expert systems incorporate uncertainty in their reasoning strategy [50,51]. Such systems use models for the expression of uncertainties in numeric parameters and their propagation in conclusions. As probabilistic data are scarce, experts are asked to specify the missing values for these parameters. The problems involved in reasoning with uncertainties are many [52,53]. Not all models allow for the explicit expression of different kinds of uncertainty. Furthermore, experts may have different conceptions of the parameters in which they express uncertainty. The values assigned by the experts are usually population specific. In a small domain the specification of uncertainty may be feasible, but in large domains the specified values become less reliable and the conditions for application of a model are often violated. Empirical testing must reveal whether or not the results of the reasoning process are valid. As the effect of changes in values remains obscure and unpredictable each update is trial and error.

The acquisition of knowledge for all of the systems mentioned above was a laborious process, which required the help of a knowledge engineer, who interrogated the expert and performed the conversion of the knowledge to a formal representation [54,55]. Passing knowledge onto a computer scientist completely unfamiliar with the knowledge domain may cause distortion or loss of information. Hence, knowledge engineering requires insight in computers as well as the domain of application. Such a combination of qualities is rare and to acquire it is a time-consuming process. For these reasons, attention is given to other methods of knowledge acquisition, which comes to expression in the development of systems, that extract knowledge from well-documented cases. Examples are IVY [3,56] for lung pathology and a system applied to breast cancer [57]. Although the acquisition of objective features and the access to reference knowledge have been mentioned separately for clarity, they are often integrated into expert systems reasoning with features obtained with quantitative pathology techniques [47,48,58]. Except for the PATHFINDER approach, which is currently in the process of being extended to a larger scope, expert system applications in pathology do not rise above an experimental stage and are not widely used because of their limited scope, limited transparency and the unfamiliarity of most physicians with the use of computers.

1.7 APPROACH IN THIS STUDY

The above considerations indicate that the most prominent drawbacks of existing applications for diagnostic support in pathology are one or more of the following:

- Limited applicability
- Limited scope
- Laborious knowledge acquisition
- Too much emphasis on decision support, based on findings only

Ideally, diagnostic support constitutes support in solving the most frequently occurring diagnostic questions: a search for possible diagnoses, differentiation among diagnoses and confirmation of diagnoses. To our knowledge no system exists which efficiently offers reference knowledge for consultation.

Furthermore, development of decision support on a large scale requires the availability of efficient tools for knowledge acquisition.

The research, reported in this thesis, is an attempt to answer the following two questions:

1. Is it feasible to increase the quality and efficiency of the diagnostic process in pathology by offering a large amount of reference pictures integrated with diagnosis descriptions and differential diagnostic information in the form of a computerized encyclopedia for routine use?
2. Is it possible to develop a tool, which allows for the storage and acquisition of formalized pathology knowledge directly from the expert?

The Chapters 2 and 3 deal with the first question. Chapter 2 describes the role of reference knowledge for the diagnosing pathologist and the form in which that knowledge usually is available. The impetus to the development of a computerized diagnostic encyclopedia are the current limitations in the access to reference knowledge and the present state of technology in the field of computers and storage media. The design of the encyclopedia, its contents and its user interface are described. Chapter 3 deals with a clinical evaluation study of the diagnostic encyclopedia as opposed to the use of books by two groups of pathologists.

Chapters 4 and 5 concentrate on the second question. To that end, Chapter 4, first addresses the problem of reasoning with uncertainties in decision support systems with focus on the field of pathology. Prior to developing a knowledge base and a reasoning system for diagnostic support, it is important to decide if and how uncertainty will be dealt with. For that purpose, criteria are formulated as basis for the comparison of five well-known strategies for expressing and combining uncertainties. In Chapter 5 attention is given to the fact that diagnostic support based on findings requires access to diagnostic features as separate entities. To eliminate the need for a knowledge engineer a system has been developed to acquire, directly from the expert, pathology knowledge as formalized diagnostic features.

REFERENCES

[1] Langley FA, Baak JPA and Oort J, Diagnosis Making: Error Sources, in: Baak JPA and Oort J (eds.), A manual of morphometry in diagnostic pathology, 1983, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, pp. 6-14.

[2] King LS, How does a pathologist make a diagnosis? Arch Pathol 84, 1967, pp. 331-333.

[3] Silbert JA and Hunter L, Expertise Through Experience: A Case-Based Approach to Expert System Design, Symposium on Computer Applications in Medical Care, March 1987.

[4] Connelly.DP and Johnson PE, The medical problem solving process, Hum Pathol 11, 1980, pp. 412-419.

[5] Baak JPA, Langley FA, Talerman, A and Delemarre JFM, Interpathologist and Intrapathologist Disagreement in Ovarian Tumor Grading and Typing, Analyt Quant Cytol 8, 1986, pp. 354-357.

[6] Bedossa P, Poynard T, Naveau S, Martin ED, Agostini H and Chaput JC, Observer variation in assessment of liver biopsies of alcoholic patients, Alcoholism 12(1), 1988, pp. 173-178.

[7] Dundas SA, Kay R, Beck S, Cotton DW, Coup AJ, Slater DN and Underwood JC, Can histopathologists reliably assess dysplasia in chronic inflammatory bowel disease?, J Clin Pathol 40(11), 1987, pp. 1282-1286.

[8] Cramer SF, Roth LM, Ulbright TM, Mazur MT, Nunez CA, Gersell DJ and Mills SE, Evaluation of the reproducibility of the World Health Organization classification of common ovarian cancers: with emphasis on methodology, Arch Pathol Lab Med 111(9), 1987, pp. 819-829.

[9] Koran LM, The reliability of clinical methods, data and judgments, N Engl J Med, 1975, pp. 642-646 and pp. 695-701.

[10] Toogood JH, What do we mean by "usually"?, Lancet, 1980, pp. 1094.

[11] Baak JPA, Basic Points in and Practical Aspects of the Application of Diagnostic Morphometry, Path Res Pract 179, 1984, pp. 193-199.

[12] Collan Y, Morphometry and stereology in pathology: An introduction, in: Collan et al. (eds.) Morphometry and Stereology in pathology, 1983, Kuopio University Press, Kuopio.

[13] Marchevsky AM, Gil J and Jeanty H, Computerized morphometry in pathology:current instrumentation and methods, Hum Pathol 18(4), 1987, pp. 320-331.

[14] Baak JPA, The Principles and Advances in Quantitative Pathology, Analyt Quant Cytol 9, 1987, pp. 89-95.

[15] Collan Y, Torkkeli T, Pesonen E, Jantunen E and Kosma VM, Application of Morphometry in Tumor Pathology, Analyt Quant Cytol Histol 9, 1987, pp. 79-88.

[16] Hedley DW, Friedlander ML and Taylor IW, Application of DNA Flow Cytometry to Paraffin-Embedded Archival Material for the Study of Aneuploidy and Its Clinical Significance, Cytometry 6, 1985, pp. 327-333.

[17] Friedlander ML, Hedley DW and Taylor IW, Clinical and biological significance of aneuploidy in human tumors, J Clin Pathol 37, 1984, pp. 961-974.

[18] Coon JS, Landay AL and Weinstein RS, Biology of Disease: Advances in Flow Cytometry for Diagnostic Pathology, Lab Invest 57, 1987, pp. 453-479.

[19] Heliö H, Karaharju E and Nordling S, Flow Cytometric Determination of DNA Content in Malignant and Benign Bone Tumours, Cytometry 6, 1985, pp. 165-171.

[20] Fu YS, Hall TL, Berek JS, Hacker NF and Reagan JW, Prognostic Significance of DNA Ploidy and Morphometric Analyses of Adenocarcinoma of the Uterine Cervix, Analyt Quant Cytol Histol 9, 1987, pp. 17-23.

[21] Sprenger E, Kowal S, Jütting U, Burger G and Rodenacker K, Malignancy diagnosis of breast cancers, in: Burger G, Ploem JS and Goerttler K (eds.), Clinical Cytometry and Histometry, 1987, Academic Press, London, San Diego, New York, Berkeley, Boston, Sydney, Tokyo and Toronto, pp. 399-401.

[22] Oud PS, Katzko MW, Pahlplatz MMM and Vooijs GP, Classification of normal and malignant endometrium using DNA and nuclear protein features, in: Burger G, Ploem JS and Goerttler K (eds.), Clinical Cytometry and Histometry, 1987, Academic Press, London, San Diego, New York, Berkeley, Boston, Sydney, Tokyo and Toronto, pp. 359-364.

[23] Baisch H, Otto U and Klöppel G, Malignancy Index Based on Flow Cytometry and Histology for Renal Cell Carcinomas and Its Correlation to Prognosis, Cytometry 7, 1986, pp.200-204.

[24] Bradbury S, Microscopical image analysis: problems and approaches, J Microsc 115, 1979, pp. 137-150.

[25] Bengtsson E, The Measuring of Cell Features, Analyt Quant Cytol Histol 9, 1987, pp. 212-217.

[26] Smeulders AWM and ten Kate TK, Accuracy of optical density measurement of cells. 1: Low resolution, Applied Optics, 26, 1987, pp. 3249-3257.

[27] Schipper NW, Smeulders AWM and Baak JPA, Quantification of Epithelial Volume by Image Processing Applied to Ovarian Tumors, Cytometry 8, 1987, pp. 345-352.

[28] Kaman EJ, Smeulders AWM, Verbeek PW, Young IT and Baak JPA, Image Processing for Mitoses in Sections of Breast Cancer: A feasibility Study, Cytometry 5, 1984, pp. 244-249.

[29] Fukanaga K, Introduction to statistical pattern recognition, 1972, Academic Press, New York.

[30] Duda RO and Heart PE, Pattern classification and scene analysis, 1973, John Wiley & Sons, New York.

[31] Sher PP, Mathematical and computer assisted procedures in clinical decision making, Hum Pathol 11, 1980, pp. 420-423.

[32] Bartels PH, Weber JE and Duckstein L, Machine Learning in Quantitative Histopathology, Analyt Quant Histol 10, 1988, pp. 299-306.

[33] Bibbo M, Bartels PH, Dytch HE, Puls JH and Wied GL, Rapid Cytophotometry and Its Application to Diagnostic Pathology, Appl Pathol 5, 1987, pp. 33-46.

[34] Burger G, Ploem JS and Goerttler K (eds.), Clinical Cytometry and Histometry, 1987, Academic Press, London, San Diego, New York, Berkeley, Boston, Sydney, Tokyo and Toronto.

[35] Proceedings of The First International Conference on Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer, February 1-3, 1987, Universal City, CA, USA.

[36] Proceedings of The Second International Conference of Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer, March 13-15, 1988, Chicago, Ill, USA.

[37] Stichting PALGA, Handleiding Koderen en Minithesaurus, 1988, Spinhex, Amsterdam.

[38] Medical Microscopic Morphology, Georg Thieme Verlag, Bereich Neue Medien, Stuttgart.

[39] Jaffe CC, Lynch PJ and Smeulders AWM, Hypermedia Techniques for Diagnostic Imaging Instruction: The Video-disc Echocardiography Encyclopedia, accepted for publication in Radiology, 1988.

[40] Videodisc Applications, status 1988, Telemedia Gmbh AG, Postfach 5555, D-4830 Gütersloh, Germany.

[41] Van Ginneken AM, Smeulders AWM and Jansen W, Design of a diagnostic encyclopedia using AIDA, Comp Meth Progr Biomed, 25, 1987, pp. 339-348.

[42] Silbert J.A. Producing a Videodisk for Pathology Slides, Informatics in Pathology, 1, 1986, pp. 38-44.

[43] Gyde OH, Computer assisted bone marrow aspirate reporting in: Proceedings of The Second International Conference on Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer, March 13-15, 1988, Chicago, Ill, USA.

[44] Horvitz EJ, Heckerman DE, Nathwani BN and Fagan LM, Diagnostic Strategies in the Hypothesis-Directed PATHFINDER system, in: Proceedings of The First Conference on Artificial Intelligence Applications, IEEE Computer Society, 1984, pp. 630-636.

[45] Hayes-Roth F, Rule-based systems, in: Communications of the ACM, 28, 1985, pp. 921-932.

[46] Chang E, McNeeley M and Gamble K, Strategies for choosing the next test in an expert system, in: Proceedings of the American Association of Medical Systems and Informatics Congress, 1984, Bethesda Md, American Association of Medical Systems and Informatics, 1984, pp. 198-202.

[47] Alvey PL and Greaves MF, Observations on the development of a high performance system for leukemia diagnosis, in: Bramer MA (ed.), Research and Development in Expert Systems III, 1986, Cambridge Unversity Press, pp. 99-110.

[48] Donovan RM, Nagel M and Goldstein E, An expert system-based leukocyte classification instrument, in: Proceedings of The First International Conference on Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer, February 1-3, 1987, Universal City, CA, USA.

[49] Potter B and Ronan SG, Computerized dermatopathologic diagnosis, J Am Acad Dermatol 17, 1987, pp. 119-131.

[50] Buchanan BG and Shortliffe EH, Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison-Wesley, 1984.

[51] Miller RA, Pople HE and Myers JD, INTERNIST-I, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine, N Engl J Med 307, 1982, pp. 468-476.

[52] van Ginneken AM, Smeulders AWM, An analysis of five strategies for reasoning in uncertainties and their suitability for pathology, in: Gelsema ES and Kanal LN (eds.), Pattern Recognition and Artificial Intelligence, 1988, Elsevier Science Publishers B.V. North Holland.

[53] Gammerman A and Creaney N, Modelling of Uncertainty in Expert Systems, Second International Expert Systems Conference - Oxford: Learned information, 1986, pp. 265-274.

[54] Gaines BR, An overview of knowledge-acquisition and transfer, Int J Man-Machine Studies 26, 1987, pp. 453-472.

[55] Musen MA and van der Lei J, Knowledge engineering for clinical programs: modeling the application area, Meth Inform Med 28, 1989, pp. 28-35.

[56] Hunter L and Silbert JA, Progress Report on IVY: A Learning System for Intelligent Information Retrieval in Pathology, Artificial Intelligence in Medicine Workshop, March 1987.

[57] Brunet M, Durrleman S, Ferber J, Ganascia JG, Hacene K, Hirt F, Jouniaux F and Meeus L, Conception d'un systeme expert d'aide a la decision en pathologie mammaire connecte a une base de donnees cliniques, Bull. Cancer 74, 1987, pp. 267-275.

[58] Paplanus SH, Graham AR and Bartels PH, Nuclear displacement in an expert system for colonic lesions, in: Proceedings of The Second International Conference of Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer, March 13-15, 1988, Chicago, Ill, USA.

CHAPTER 2

Design of the Diagnostic Encyclopedia Workstation (DEW)

Submitted for publication

Astrid M. van Ginneken, Arnold W.M. Smeulders, Wicher Jansen,
Jan P.A. Baak and Ingrid Brooymans

ABSTRACT

The Diagnostic Encyclopedia Workstation (DEW) contains reference knowledge to serve diagnostic decision making in the field of ovarian pathology. The database of DEW is accessed via a PC and a video disc device. Compared with the common source of reference knowledge, i.e. books, DEW has the following advantages: it can hold more verbal knowledge, pictures and case histories, and its information is accessible via several entries. Based on an analysis of the structure of this reference knowledge a relational database system was developed to hold the textual information. A videodisc is used for the pictorial part of the database. Software for data entry has been written in the language MUMPS with use of the relational database toolkit AIDA, which is particularly suited for manipulation with free texts. The software for consultation was developed as a separate entity, based on an analysis of the most efficient way for the user to access the information in the database. This part is written in C using MetaWindows, which allowed for the development of a graphical mouse-driven user interface. At present the encyclopedia has reached a stage whereby it can be clinically evaluated. The DEW contains 85 diagnoses in ovarian pathology, covering all frequent cases and many rarities. The diagnoses are illustrated by approximately 3000 pictures, divided among 158 cases.

2.1 INTRODUCTION

Currently, the personal computer (PC) with its ever increasing processing speed and storage capacity is used in many medical applications. They include storage and retrieval of information about the patient in primary care, information in specialized clinical departments and reference knowledge with respect to diseases [1,2]. In general, reference knowledge reflects the state-of-the-art in medicine, indicating what is known and should be known about a disease when making decisions.

This paper focuses on the design of an electronic encyclopedia, aimed to contain reference knowledge in the field of pathology. In pathology, decisions are primarily based on visual observations of well-preservable materials. Hence, decisions can be evaluated by repeating observations of the same materials. Because of the repeatability of observations, pictures have a lasting value as a carrier of knowledge. Therefore, reference knowledge in pathology is intrinsically pictorial for an important part.

We propose a system called Diagnostic Encyclopedia Workstation with acronym DEW, which contains verbal information as well as pictures. Since pathology is a wide-ranging domain, the DEW is, for the time being, restricted to the pathology of the ovary. This part of pathology is quite circumscribed and, compared to most organs, complex enough to be a good representative of pathology in general: the ovary contains several kinds of tissue and its hormonal status changes both monthly and in a lifetime.

This paper describes the considerations underlying the design of the DEW, the design itself and implementation of the DEW. Since the database is read-only the software for data entry and for consultation are developed as separate entities, using for either application the most suitable programming language and tools. Both parts of the application software are described.

2.2 CONSIDERATIONS UNDERLYING THE DESIGN OF THE DEW

An important part of the task of a pathologist is the visual classification and grading of histologic and cytologic slides in the context of the clinical data about the patient. However, the observations may be uncommon or lead to the consideration of several diagnoses, which are difficult to tell apart by their morphological similarity. Therefore, reference knowledge may be needed to confirm a diagnosis or to aid in the distinction of such diagnostically difficult cases. Reference knowledge consists of all information available to the pathologist in the diagnosis of a case. This reference knowledge usually differs slightly per pathologist: it includes the training received, the diagnostic experience gained in practice, books, atlases and the expertise of consulted colleagues. The knowledge in books and atlases is mostly descriptive, whereas the expertise of colleagues has a predominantly judgmental value.

Knowledge as laid down in books and atlases, and partly, the knowledge of experts can be stored and made accessible in a computer system. The reason for building such a system is the fact that there are several restrictions to the use of books as reference knowledge:

(1) An average textbook in pathology generally contains one picture for each diagnosis, only a minority of the diagnoses being illustrated with more than one picture. However, more pictures are usually desired to show alternative stains and laboratory techniques at several optical magnifications. In addition, more than one picture is needed to show the histologic variability within a certain diagnosis. Economical restrictions in publishing, especially where colour pictures are concerned, limits the number of illustrations for an individual diagnosis to just a few.

(2) Information with respect to morphologically similar alternative diagnoses and criteria to differentiate among them is usually limited in books. This is probably due to the practical fact that including extensive differential diagnostic information entails a considerable increase in the size of the book by the redundant character of that information.

(3) Pathology is such a wide-ranging domain that the information in a book has to be restricted to a part of this domain. As a consequence, a pathologist in a diagnostic session usually has to consult several books.

(4) The ordering of information in a book is one-dimensional, i.e. reading is scanning forward only. Contents and indexes provide the possibility to start anywhere in the book, but from thereon there is only the direction of reading. Combination of information at different places in a book can only be achieved by text repetitions in the book, the specification of internal references and/or the use of more than one copy of the book.

(5) The supply of information in a book cannot be adjusted to the need of the user. This need may change with the degree of experience of the user in a specific field, but also depends on whether the user wants to confirm a diagnosis or be informed about other potential diagnoses with criteria to differentiate among them.

2.2.1 System specifications

The restrictions mentioned can be overcome in a computer system with a large storage capacity and the possibility to structure the information such that it can be accessed along more than one direction without the need of redundancy. The contents of such a computer system should include all characteristics of books. In addition, it should contain:

(1) Many pictures in the form of illustrated case histories for each diagnosis to show the histologic variability of that diagnosis. The pictures must include all laboratory techniques and optical magnifications relevant to that diagnosis.

(2) For each diagnosis a list of morphologically similar diagnoses and criteria to differentiate among them.

(3) Information with respect to the clinical consequences of diagnoses.

Based on these considerations demands can be formulated for the computerization of pathology reference knowledge:

(1) A large amount of pictorial data must be stored, preferably in color. Storing a picture digitally would involve some 512 x 512 pixels in three colors each, yielding 0.75 mpixels per image. Obviously, at the present state-of-the-art, a thousand pictures or more cannot be contained on a hard disc of 20-30 mbytes, which is commonly used in a PC. Therefore, the pictorial information has to be stored on an external computer-addressable medium with large storage capacity and high retrieval speed.

(2) Apart from pictures, the database must contain verbal and numerical information. The verbal part of the information requires efficient storage of text of variable size.

(3) The database must allow for storage and manipulation of large amounts of data, comparable to the contents of many books.

(4) The database structure and application software must support flexible entry and update of data as well as retrieval of information via several entries with a self-explicable user interface.

(5) The system must run on a commonly used PC and accessories of moderate cost in order to compete with the constantly recurring costs of books, necessary to guarantee widespread use.

On the basis of these demands an IBM-AT personal computer was selected, equipped with 640Kb memory and a 20Mb Winchester drive. The computer is connected to a video long-disc player (VLP). A video disc can be mounted in the player and contains a maximum of 54000 images on either side, stored in pulse code modulation using one track per picture. Each image on the disc can be addressed by sending simple command codes via a standard RS-232 serial interface. The quality of the images on the disc are comparable with the average illustrations in regular books. Fig. 1 shows an overview of the DEW.

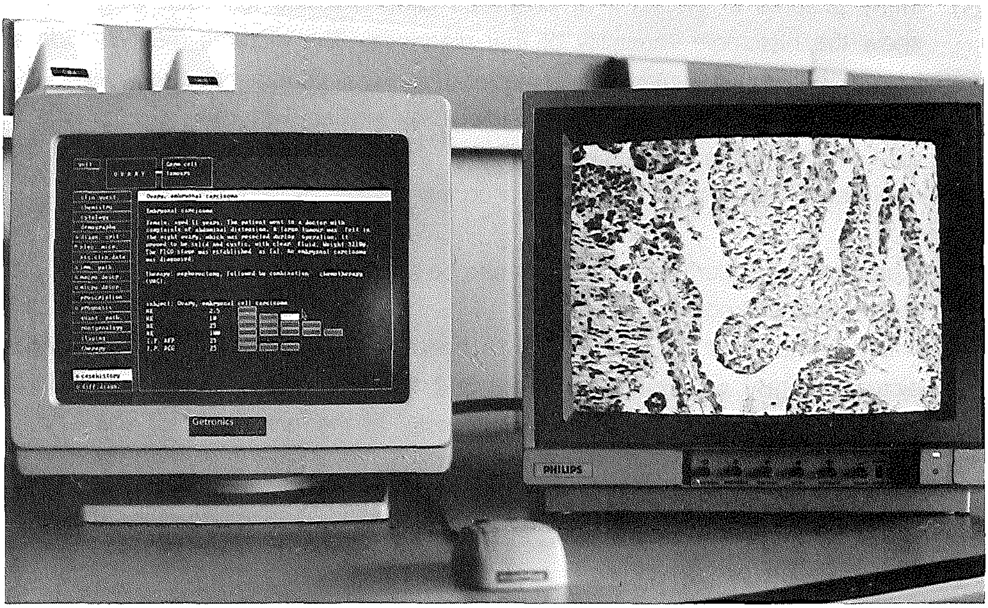


Figure 1. Overview of the DEW during a consultation session.

2.3 DESIGN OF DEW

2.3.1 The database

The choice of a database system and the design of the database, to be constructed, starts with an inventory of the structure of the knowledge it is intended to store and the demands for the retrieval of that information.

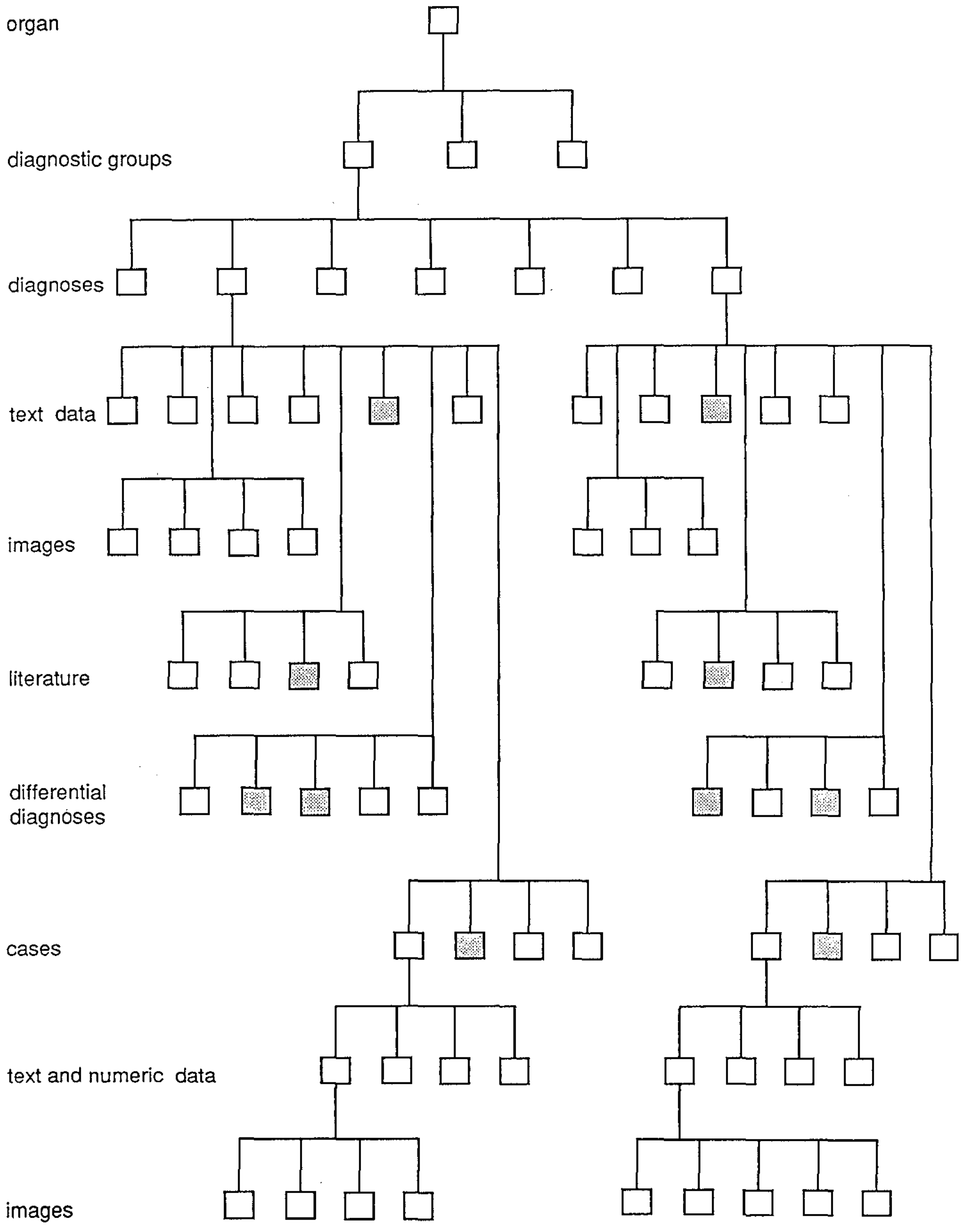


Figure 2. Structure of the information taken as the starting point for the construction of the database. Note, that some of the items appear at more than one place in the hierarchy: in this figure, these places are marked by a shaded box.

The structure of reference knowledge in pathology can be represented in a hierarchical way: diseases of the ovary are ordered by the WHO, which is reflected in a classification tree [3]. The top node of the tree is the ovary and the leaves of the tree are diagnoses. Between the top and the leaves of the tree there are one or more intermediate levels, containing subgroups of diagnoses with increasing refinement down the tree. In addition to a description, to each diagnosis there are several case histories, differential diagnoses and literature references. Each case, in turn, consists of a case history and a number of pictures. The pictures used to illustrate the described diagnosis are selected from the cases which belong to that diagnosis. When more than one diagnosis is assigned to a case, that case can be used to illustrate the corresponding diagnoses. In the same way, literature references can serve as source of information for several diagnoses. The structure of the reference knowledge can be represented by an extension of the classification tree: each diagnosis and diagnosis group has its own subtree with cases, literature references and differential diagnoses. However, this tree is not strictly hierarchical since part of the cases, differential diagnoses and literature references will have multiple connections with higher levels in the tree, i.e. diagnoses and diagnosis groups.

As to the demands for retrieval during consultation, the system must support separate retrieval of a diagnosis by its name, a case, an image, a differential diagnosis, and a literature reference. When entering and updating information more complex retrieval is needed. Deletion of information may require the removal of other information with which it is connected. For example, when an article is removed from the literature data all references to that article have also to be removed. In other words, for each node in the tree, it must be possible to retrieve the connecting nodes at higher levels.

When using a strictly hierarchical database redundancy of information is inevitable since multiple connections of a node with higher levels are not allowed in a hierarchical tree [Fig. 2]. The amount of redundancy would also be considerable as it involves the cases with multiple diagnoses, many literature references and all differential diagnostic information. The presence of redundancy implies inefficient use of memory and, more important, it is impractical in keeping the database internally consistent during entry and update of information. Redundancy can be avoided when using a relational database. In addition, separate retrieval of certain entities of information can be efficiently achieved by defining separate relations for each of them. By sharing the proper columns in the tables of the database connections can be represented in many directions.

Another point to consider in the selection of a suitable database system is the expected size of the database. According to the WHO there are eight major groups of diagnoses in the pathology of ovarian tumors [3]. Some of these major groups can be subdivided, making a total of 25 subgroups of diagnoses encompassing approximately 90 diagnoses, including rarities. In this number each tumor grading is counted as one diagnosis. Blaustein [5] recognizes four diagnosis groups in non-tumor pathology of the ovary, covering 62 diagnoses including rarely occurring diseases. When comparing pathology books some variation in these numbers occurs as some books use in some areas classifications other than the that of the WHO, or add various types of grading or further subclassification. A basic encyclopedia, excluding non-tumor rarities, will thus contain 90 tumor diagnoses and 35 non-tumor diagnoses, including a description of the normal ovary. The average amount of text is estimated at three pages per diagnosis and one page per case. With an average of three cases per diagnosis an encyclopedia of ovarian pathology would require storage of 950 pages of text. Starting from 2.5 kbytes per page, 950 pages consume approximately 2.5 mbytes of memory. For the purpose of flexibility the hardware and the selected database system should support storage and retrieval of larger amounts of data.

The relational database system AIDA meets all the database demands as specified in section 2.1 [4]. AIDA has a more than sufficient storage capacity and it efficiently supports the storage and manipulation of free texts. AIDA is a 4GS (fourth generation software) package, which is written in MUMPS as host language. Its availability as a toolkit make the AIDA-MUMPS combination very suitable for construction of the database and the software for the entry and update of information.

2.3.1.1 Implementation of the database

The present database contains six main tables for the storage of information:

(1) The relation DIAGNO holds the diagnosis description, divided among 17 different categories of information as shown in Table 1. Since a text may be shared by more than one diagnosis, a numeric field is present for each category to store a reference to another diagnosis.

(2) The relation CASUS contains information about patient cases. Its contents are also shown in Table 1.

(3) The relation CASIMA has only fields of fixed length. It holds all pictures belonging to a case, sorted by subject of photography, laboratory technique (both character fields) and magnification (numeric field).

(4) The relation DIFTXT consists of a free text field which holds differentiating criteria for each pair of morphologically related diagnoses. Since such a text sometimes applies to more than one pair of diagnoses, two numerical fields are added to store a reference.

(5) The relation IMAG contains for each image a numerical field for its address on the video disc and a free text field for a caption.

(6) The relation LITDAT holds all literature references. A character field of fixed length is used to store the name of the first author and a free text field stores the complete reference.

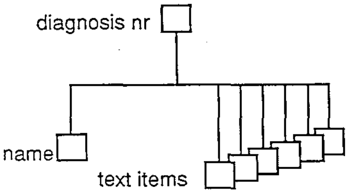
Relation DIAGNO		**Relation CASUS**	
Field name	**Type**	**Field name**	**Type**
Diagnosis number	integer	Casus number	integer
Diagnosis name	char	Code of case	char
Demography	text	Diagnosis of case	char
Clinical signs	text	Sex	char
Macroscopic description	text	Age	integer
Macroscopy recipes	text	Start of disease	char
Radiology	text	FIGO stage	char
Laboratory data	text	Macroscopic description	text
Staging	text	Microscopic description	text
Microscopy description	text	Mitoses/25 fields	integer
Electron microscopy	text	DNA index	integer
Immune pathology	text	Volume% epithelium	integer
Cytology	text	Therapy	char
Quantitative pathology	text	Follow-up	char
Quantitative pathology recipes	text	Case history	text
Diagnosis criteria	text		
Therapy	text		
Prognosis	text		
Clinical questions	text		

Table 1. Contents of the relations DIAGNO and CASUS. Text fields are free length fields, whereas character and numeric fields have a fixed length. Clinical questions are the questions, which a pathologist can expect from the clinician.

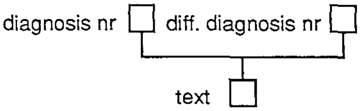
All six relations have at least one numerical field: holding a diagnosis-, case-, literature-, or image number for retrieval purposes.

Besides these main relations, which directly store information relevant to the user, the database has several relations for internal use. They represent one-to-many relations.

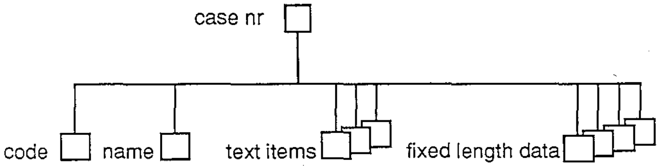
Relation DIAGNO



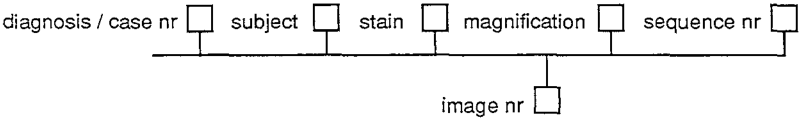
Relation DIFTXT



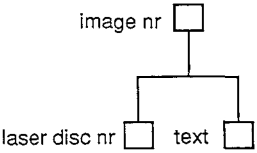
Relation CASUS



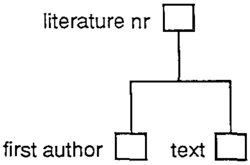
Relation CASIMA



Relation IMAG



Relation LITDAT



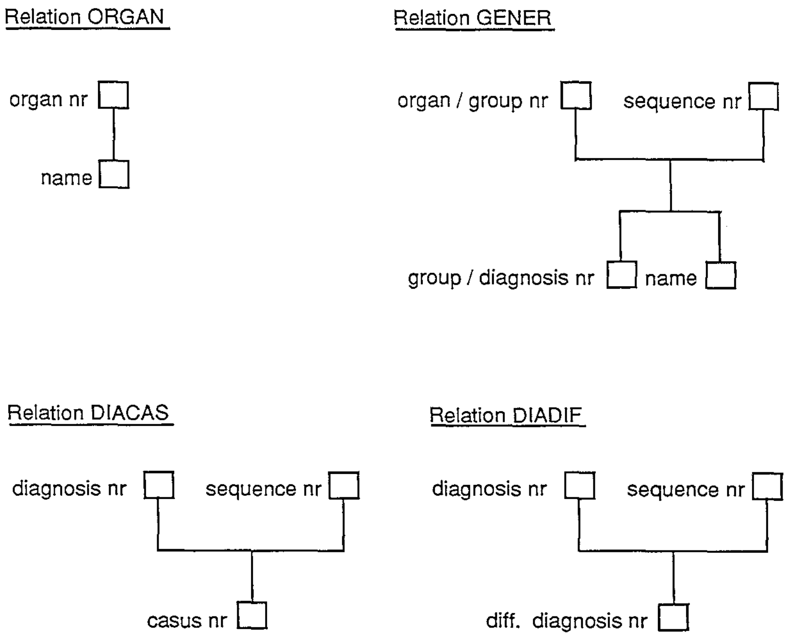


Figure 3. Structure of the database. Sequence numbers serve as second key in 'one-to-many' relations as shown in Fig. 2. Cases, literature references, differential diagnostic information, and images are defined as separate relations, because items of these categories can be referred to more than once.

For example, the relation DIACAS links cases to diagnoses: it stores at each diagnosis number the cases belonging to that diagnosis. Similar relations are present to link diagnostic groups with diagnoses and diagnoses with differential diagnoses.

In each one-to-many relation, sequence numbers offer the possibility to order the rows in the relation. This is especially desirable for differential diagnoses, which can then be ordered by decreasing similarity. The main structure of the database is shown in Fig. 3. In this Figure "text" refers to character fields of free length. The connections between the relations are clearly visible by the presence of common fields.

2.3.1.2 Contents of the database and the videodisc

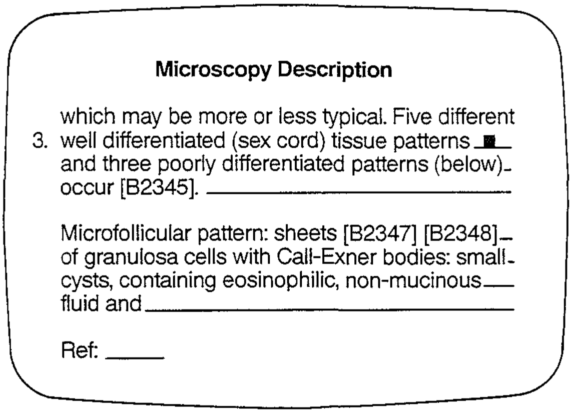
At present 85 tumor diagnoses with 158 cases have been entered in the database. These diagnoses belong to four major diagnosis groups of ovarian pathology: common epithelial tumors, sex cord stromal tumors, lipid cell tumors and germ cell tumors. The diagnosis descriptions are primarily based on textbooks and publications in journals [3,5-10]. Consultation of the descriptions is facilitated by the fact that they are uniformly organized: the items of information in each category are described in a fixed order, which is shared by all texts of that category. For each diagnosis a differential diagnosis (DD)-list has been made on the basis of a thorough search through the literature for morphologically similar diagnoses. In addition, tables are constructed with differentiating features for each pair of morphologically related diagnoses. For each diagnosis, pictures have been selected to illustrate its characteristics. The source of the pictures are cases, i.e. documented patient material.

As to the cases, some diagnoses are not illustrated by a case because of their rarity such as malignant serous adenofibroma and polyembryoma. Part of the cases serve to illustrate more than one diagnosis, implying a higher average of cases per diagnosis than 85/158. The illustrations belonging to the cases total approximately 3000. For the acquisition of macroscopic pictures an archive search was made, which was successful for only a minority of the diagnoses. The majority of the histologic slides have been especially cut from available paraffin material to obtain an optimal basis for photography. Standard histologic stains available for each case are Hematoxilin-Eosin and PAS. Dependent on the diagnosis involved and available material, additional stains are also available: PAS, PAS-diastase, astra-blue, Gomori, Giemsa, Grimelius and fat. The objective magnifications include 1.25x, 2.5x, 10x, 25x, 100x and 160x, with an additional magnification of 4x at the level of a color slide (24 x 36 mm^2). Some electron microscopic, immunomicroscopic and cytologic slides are included. The availability of techniques and magnifications varies per case.

2.3.1.3 Data entry

The software for data entry and update is written in MUMPS from which AIDA routines are called to perform database operations.

The user interface contains 30 different input screens. Numerical and fixed length character fields are combined in one screen insofar they belong to one relation. All free text fields have their own input screen. Screens can be bypassed, allowing for direct input of available information. Fig. 4 shows the input screen for the microscopy description.



Microscopy Description

```
   which may be more or less typical. Five different
3. well differentiated (sex cord) tissue patterns ■_
   and three poorly differentiated patterns (below)_
   occur [B2345]. ______________________________

   Microfollicular pattern: sheets [B2347] [B2348]_
   of granulosa cells with Call-Exner bodies: small_
   cysts, containing eosinophilic, non-mucinous___
   fluid and ___________________________________

   Ref: _____
```

Figure 4. Input screen for the microscopic description. The text is entered and edited per line. The line with the line number is the current line. Here, line 3 needs to be edited. When a text is shared by more than one diagnoses, it needs to be entered only once and can be referenced at the others by means of the reference field at the bottom. The image numbers will later appear as sense fields.

The entered data is checked by AIDA with respect to its type: text, character or numeric. More specific validation is performed by the application input program, which tests the validity of values entered for items such as age and DNA index. When names are changed or information is deleted, the input program takes care of database consistency.

2.3.2 Consultation

Once put together an encyclopedia system is read-only. As already pointed out, this enables separate development of the software for consultation.

In an encyclopedia system the presentation of data is of more than usual importance. The design of the user interface is based on an analysis of the items of information, which necessarily need to be specified by the user for the system to respond, and how this information can be most easily specified by the user.

A user who wishes to access a particular item of information about a diagnosis in a diagnostic encyclopedia must provide four different kinds of actions. The first concerns the specification of a diagnosis to the system. As organs, diagnostic groups and diagnoses cannot be displayed on one screen simultaneously, the specification of a diagnosis requires a sequence of interactions: the specification of an organ (at present only one), a major diagnostic group, possibly a number of subgroups and finally a diagnosis. At the level of a diagnosis the amount of available information is too large to be displayed on one screen, so the information is subdivided into a number of items, one of which is displayed by default, while the remaining items are optional. The second type of interaction concerns the specification of one of these items. Action type three allows the user to switch back to diagnosis selection again and the last type of interaction is the possibility to quit the session.

For specifying these choices to the system, there are two possibilities: active (user-entry by typing) or passive (selection from a list of presented possibilities). The obvious choice is passive entry since this is both convenient and it avoids the problem that nomenclature in pathology comprises many synonyms and different spellings. The presentation of possible choices has the additional advantage that the user is informed about what organs and diagnoses are available and how they are classified. All optional items should be on a fixed position of the screen with an indication as to whether or not they are available. This facilitates becoming acquainted with the system.

2.3.2.1 Implementation

In our opinion a graphical, mouse-driven interface, supporting windows is most suitable for construction of the user interface. A graphical interface, based on windows, is not yet available in combination with MUMPS. In addition, the interpreter language MUMPS is not as fast as a compiler-based language. Finally, MUMPS is not widely used, which is a disadvantage for software maintenance. Therefore, the consultation software is written in the "C" language and the graphics toolkit "MetaWindows".

Besides the graphical user interface, the software consists of two more parts. One of them is a conversion program, which creates a reformatted copy of the MUMPS database for read-only use. The conversion program takes each database relation as a sequential ASCII file, created by MUMPS, as input. As a result each relation of the MUMPS database is represented in the C-database as a sequential file of C-structures. The C-database is created only once. Hence, the conversion need not be carried out during runtime.

The other part is the retrieval program. For an information reference system, the overall response time is very important. The tolerated elapse time is set to the maximum considered to be still convenient in interaction, i.e. one second. The total time needed by the system to come up with a picture is the sum of the response time of the VLP and the retrieval time of the database. The response time of a Philips VLP 835 is specified in Fig. 5.

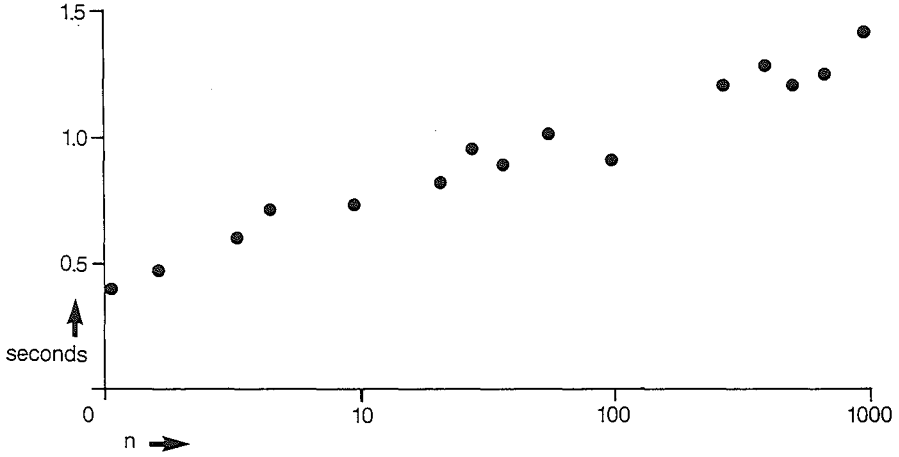


Figure 5. Response time of the video disc player: Philips VLP 835. The figure shows the time needed to come up with a picture n positions beyond the current position.

The times given are an average as the response is not always the same for the same jump. Remarkably enough, the response time does not increase linearly with the requested displacement (n). For n = 1000 the response time of the Philips exceeds 1 s. The maximum response time, when jumping from position 1 to position 54000, is 4.3 s. The Sony player is faster, but it is not possible to measure its response time accurately: an acknowledgement of the player indicates that a command is received, not that it has been successfully carried out. However, the majority of jumps will be within one case history and, to a lesser degree, within one diagnosis. Only when a different diagnosis is chosen, especially when this diagnosis belongs to a different diagnosis group, may a large jump occur. Taking this into consideration, it can be concluded that the access time of the player is acceptable. When the software system does not require much more than a few hundreds of milliseconds, the requested response time of one second is likely to be reached. The performance of the players as described above reflects the state of commercially available equipment in 1986. At present (1989), both Sony and Philips produce players with faster response times.

In the retrieval program, response time is shortened by using a binary search strategy and by keeping a part of the database in central memory. The contents of this part are based on the anticipation of the information, which the user is likely to ask for next. Therefore, the organs, diagnostic groups and diagnosis names reside in central memory permanently. As soon as a diagnosis is selected, all information about this diagnosis, its list of cases and its DD-list are also kept in central memory. Apart from the system itself this information requires approximately 15 kbytes. Keeping the differential diagnoses, which are relevant for the diagnosis under discussion, in memory would also require an average of 90 kbytes. These memory requirements can at present easily be met and do not form a limitation in the optimization of the response time.

The user interface is based on windows, which are used for display of information as well as selection of information with a mouse. The action to be taken upon a mouse click from the user depends on the selected window and the position of the mouse in that window. For example, in a diagnosis text window the user can ask for a picture or ask for the text to scroll up and in a DD-list the user can select one of the alternative diagnoses. The determination of the next action is triggered by a mouse click. Using the coordinates of the mouse this determination is performed in two steps. The first step involves a routine which returns the selected window as the current window. In the second step a routine is activated, which returns the selection of the user within the current window. For this purpose, the system continually logs the coordinates of the windows on display and for each of them the coordinates of each possible selection within it. As soon as the user selection has been determined, the user interface activates the retrieval program and displays the requested information on the text screen or the video monitor. The interface as it appears to the user is described in the following section.

2.3.2.2 Sample session of consultation

After booting the system, the user interface presents the major diagnosis groups of ovarian tumors. All names are displayed in boxes, which is visible in Fig. 6. The selection of a diagnosis group with the mouse elicits the display of a subsequent screen showing the subdivision of the selected group. This is repeated until a diagnosis is selected. Then, information on that diagnosis is automatically displayed on the screen. A few remarks need be made with respect to the specification of diagnoses. Each time a subgroup is selected, it is added to a row at the top of the screen, representing the choices made so far. In this way the user can always see the path, that leads to the current position in the classification tree.

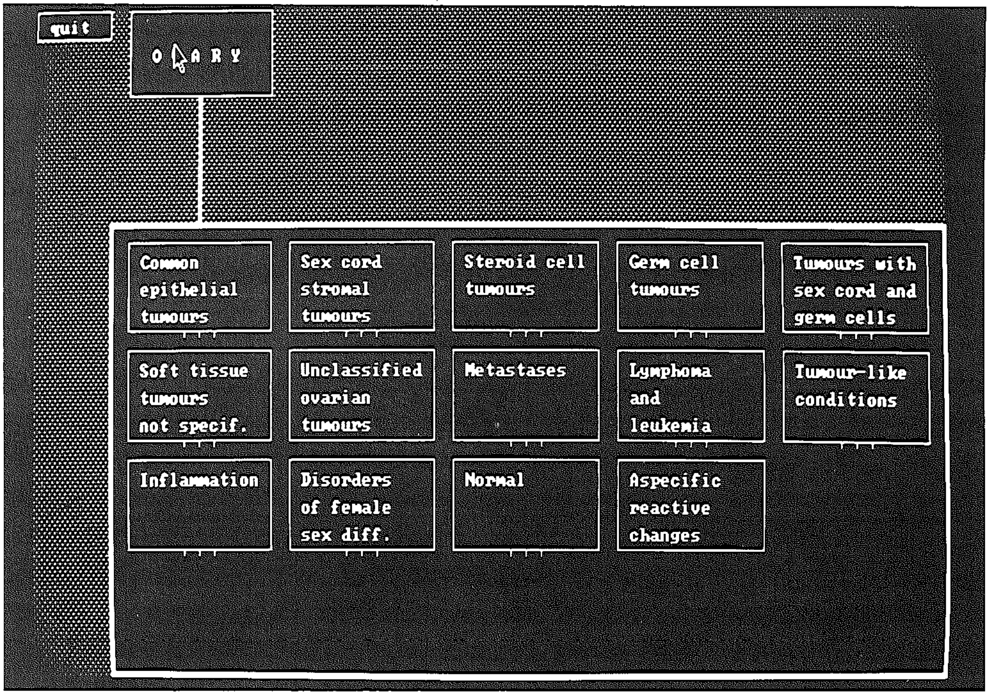
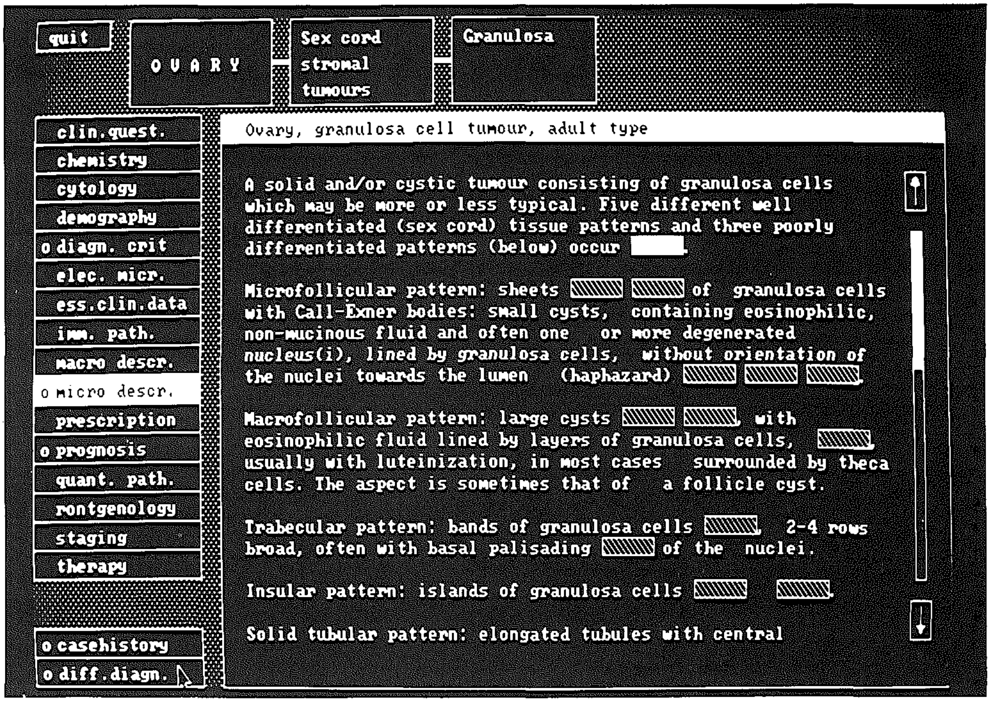

quit
OVARY
Common epithelial tumours
Sex cord stromal tumours
Steroid cell tumours
Germ cell tumours
Tumours with sex cord and germ cells
Soft tissue tumours not specif.
Unclassified ovarian tumours
Metastases
Lymphoma and leukemia
Tumour-like conditions
Inflammation
Disorders of female sex diff.
Normal
Aspecific reactive changes

Figure 6. First screen of the runtime version: the ovary and its major diagnosis groups of tumor pathology. Note, the small vertical lines at the bottom of each box, indicating the presence of a subdivision of the group. The cursor has the shape of an arrow, i.e. a click will result in the display of a screen with the subdivision of "sex cord stromal tumors".

In addition, all diagnosis groups are recognizable as a group by the presence of three small vertical lines at the bottom of their box. These lines indicate the presence of a further subdivision. Finally, information is not only available on diagnoses, but also on groups. This possibility helps the user to make the next selection, but requires that the user specifies to the system whether he or she wants information with respect to the selected group or the next screen with the subdivision of that group. As a consequence, two choices are possible at each box, representing a diagnosis group. When the cursor is in the left 2/3 of the box, it has its usual "arrow" shape and a click with the mouse will then result in the display of the subdivision of the selected group on the subsequent screen. In the remaining right section of the box the cursor changes to an "I", representing that a click with the mouse will now result in the display of information on the group as a whole.



Ovary, granulosa cell tumour, adult type

A solid and/or cystic tumour consisting of granulosa cells which may be more or less typical. Five different well differentiated (sex cord) tissue patterns and three poorly differentiated patterns (below) occur □.

Microfollicular pattern: sheets ▨ ▨ of granulosa cells with Call-Exner bodies: small cysts, containing eosinophilic, non-mucinous fluid and often one or more degenerated *nucleus(i)*, *lined by granulosa cells*, without orientation of the nuclei towards the lumen (haphazard) ▨ ▨ ▨.

Macrofollicular pattern: large cysts ▨ ▨, with eosinophilic fluid lined by layers of granulosa cells, ▨, usually with luteinization, in most cases surrounded by theca cells. The aspect is sometimes that of a follicle cyst.

Trabecular pattern: bands of granulosa cells ▨, 2-4 rows broad, often with basal palisading ▨ of the nuclei.

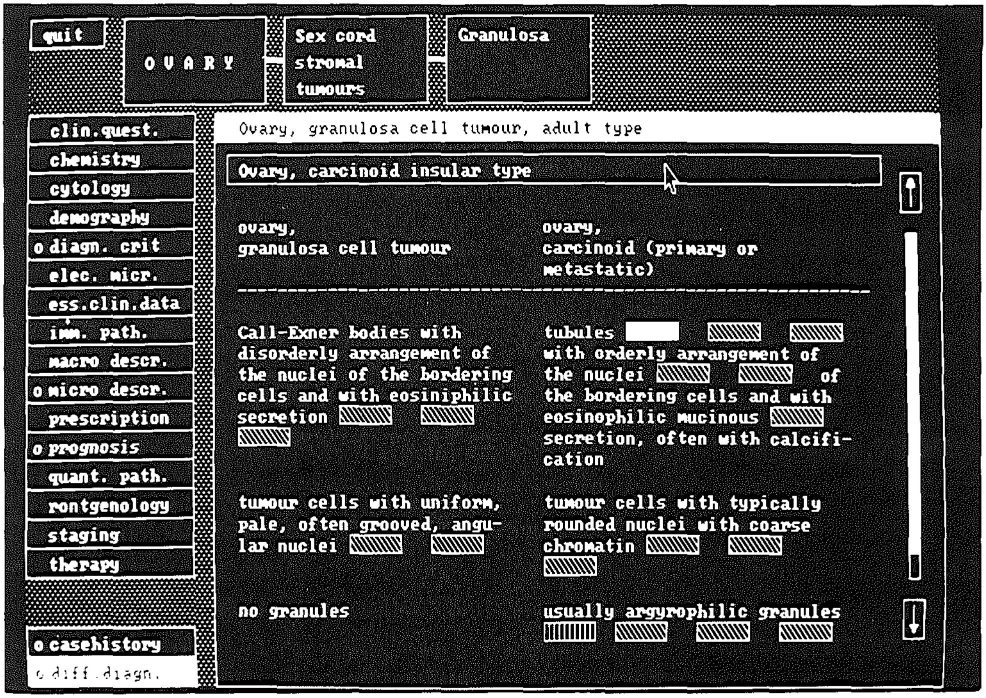
Insular pattern: islands of granulosa cells ▨ ▨.

Solid tubular pattern: elongated tubules with central

Figure 7. Screen layout at diagnosis level. The microscopic description is displayed by default. Note the sensefields, the scrollbar at the right, the optional items of information at the left and the path through the classification tree at the top.

As soon as a diagnosis has been selected, the system enters the diagnosis information mode. As to the screenlayout (Fig. 7), information about the selected diagnosis is displayed in a window at the right-hand side of the screen with the diagnosis name in the header. Initially, the microscopy description of a diagnosis is displayed by default. Optional information items are shown at the left of the screen. These items can be selected with the mouse, which causes the corresponding text to be displayed in the right window. A dot in front of the item name indicates that information on that item is available. As to the diagnosis text, a vertical scrollbar at the right indicates the portion of text, which is displayed. The two arrow boxes at the top and bottom of the scrollbar are used to scroll the text up and down. In diagnosis texts, small boxes may be visible with three different styles of shading. These boxes act as sense fields, which allow the user to retrieve pictures, glossary information or literature references by clicking on them. Pictorial information "behind" the sense fields appears on the video monitor, whereas glossary information and literature references are displayed in a toggle window on the text monitor. Each sense field applies to the remark in the text, which directly precedes the box.

In addition to information on the selected diagnosis, the optional items include information with respect to cases and differential diagnoses. When selecting "cases" the user is offered a list of available cases. Selection of a case from the list elicits the display of a screen with a case history at the top and a list of pictures below, sorted by subject of photography, stain and magnification. The pictures are accessible via sense fields.



Ovary, granulosa cell tumour, adult type

Ovary, carcinoid insular type

ovary, granulosa cell tumour	ovary, carcinoid (primary or metastatic)
Call-Exner bodies with disorderly arrangement of the nuclei of the bordering cells and with eosiniphilic secretion ▨ ▨ ▨	tubules □ ▨ ▨ with orderly arrangement of the nuclei ▨ ▨ of the bordering cells and with eosinophilic mucinous ▨ secretion, often with calcification
tumour cells with uniform, pale, often grooved, angular nuclei ▨ ▨	tumour cells with typically rounded nuclei with coarse chromatin ▨ ▨ ▨
no granules	usually argyrophilic granules ▥ ▨ ▨ ▨

Figure 8. Differential diagnosis between an adult granulosa cell tumor (current diagnosis) and a carcinoid (alternative) diagnosis. Carcinoid becomes the current diagnosis when the user clicks on its name (arrow).

The differential diagnoses are also presented as a list. When selecting one of them the user is offered a table, which shows both the common and different characteristics of the current diagnosis and the selected alternative (Fig. 8). When it happens that the alternative diagnosis is considered to be more likely than the current diagnosis, the user can click with the mouse on the name of the alternative diagnosis, which then becomes the current diagnosis. In this way a direct switch can be made to a diagnosis with a morphologically similar picture.

At all times, the top of the screen shows the path that takes the user to the current diagnosis. When the user selects one of the boxes from this path, the system switches back to diagnosis selection mode. As a consequence, it is not necessary to start the new selection at the top node "Ovary": any group in the path can be selected. At the upper left corner of the screen is a small field, which is used to quit the session with the system.

2.4 CONCLUSION

It can be concluded that the DEW offers the diagnostic support as formulated in the system specifications. As to its contents, the DEW offers all characteristics of books and in addition: documented case histories, lists with differential diagnoses and criteria to differentiate among them. The pictures per case include several relevant stains and magnifications. Where possible, macroscopic and electron microscopic pictures are included. The illustrations are available for the cases as well as the diagnosis descriptions.

Information with respect to diagnoses, items of information within a diagnosis, cases, differential diagnoses, pictures and literature references can be separately retrieved. The information is accessible by means of a mouse-driven interface.

The stage of development of the DEW permits clinical evaluation of the diagnostic support offered by the DEW versus books.

ACKNOWLEDGEMENTS

The help of S. Chadha-Ajwani M.D. Ph.D., A. Schaberg M.D. Ph.D., J.F.M. Delemarre M.D. Ph.D., C. Kooijman M.D. Ph.D., G. Wielinga M.D. Ph.D., Ch. Albus-Lutter M.D. Ph.D., M.J. Becker-Bloemkolk M.D. Ph.D., A.G.J.M. Hanselaar M.D., F.B. Lammes M.D. Ph.D., H.F. Heins M.D. Ph.D., G.J.M. Ras-Zeylmans M.D., E.J. Aartsen M.D., E. Engelsman M.D. Ph.D., M. v.d. Burg M.D. and J.H. Meerwaldt M.D. Ph.D. of the OTC

(Ovarian Tumors Committee) in providing clinical cases of ovarian pathology is gratefully acknowledged. B. Franken, P. van der Voort and R. van der Kooi are acknowledged for their software assistance and advice. In addition, the authors wish to thank A. Dekker and M. van Zijl for their help in the development of the software for consultation.

REFERENCES

[1] Boon WM, Westerhof HP, Duisterhout JS and Cromme PVM, The role of AIDA in a primary care information system, Comp. Methods Programs Biomed 25, 1987, pp. 287-296.

[2] Pressman NJ, Archival systems in diagnostic quantitative pathology, in: Proc. Int. Assoc. Pathol. Meeting, Vienna, 1986.

[3] Serov SF, Scully RE and Sobin LH, Histological typing of ovarian tumors, in: International Histological Classification of Tumours 9, World Health Organization, Geneva, 1973.

[4] Duisterhout JS, Franken B and Witte FSC, Structure and software tools of AIDA, Comp. Methods Programs Biomed. 25, 1987, pp. 259-274.

[5] Kurman RJ (ed.), Blaustein's Pathology of the Female Genital Tract, 3rd edn., Springer-Verlag, New York, Berlin, Heidelberg, London, Paris and Tokyo, 1987.

[6] Roth LM, Czernobilsky B, Tumors and Tumorlike Conditions of the Ovary, Churchill Livingstone, New York, Edinburgh, London and Melbourne, 1985.

[7] Scully RE, Tumors of the Ovary and Maldeveloped Gonads, Armed Forces Institute of Pathology, Maryland, 1979.

[8] Dallenbach-Hellweg G, Ovarialtumoren, Springer-Verlag, Berlin, Heidelberg and New York, 1982.

[9] Rosai J, Ackerman's Surgical Pathology, The C.V. Mosby Company, St. Louis, Toronto and London, 1981.

[10] Disaia PJ, Morrow CP, Townsend DE, Synopsis of Gynecologic Oncology, John Wiley & Sons, New York, Chichester, Brisbane and Toronto, 1975.

CHAPTER 3

Evaluation of a Diagnostic Encyclopedia Workstation for Ovarian Pathology

Submitted for publication

Astrid M. van Ginneken, Jan P.A. Baak, Wicher Jansen and
Arnold W.M. Smeulders

ABSTRACT

The Diagnostic Encyclopedia Workstation (DEW) is a computerized atlas of ovarian pathology. It provides completely integrated pictorial and textual information as reference knowledge to aid pathologists in their process of diagnosis making. The textual part of the reference knowledge comprises information per diagnosis such as a description of the macroscopic and microscopic images, clinical signs and prognosis. In addition, the system offers lists of differential diagnoses and criteria to differentiate among them.

The present study is meant to establish to what extent the system influences the efficiency of the diagnostic process and the agreement in the final diagnoses. Therefore, two groups of six pathologists each, covering the whole spectrum of experience in ovarian pathology, participated in the evaluation of the DEW. The quality of the resulting diagnoses was statistically analyzed with the Wilcoxon Ranksum Test with respect to five different viewpoints: classification, morphology, clinical consequences, duration of diagnostic process and consensus among the participants themselves. The results are discussed and it is concluded that books and the DEW are equivalent except for classification and morphology, which gave better results with books. The evaluation experiment was, however, very rigid and negatively biased for the DEW system. Subjectively, the majority of the participants preferred the computerized atlas over books as reference knowledge.

3.1 INTRODUCTION

The visual classification and grading of histologic or cytologic slides in the context of the clinical data about the patient is an important part of the clinical task of a pathologist [1-3]. However, some diagnostic classes may be uncommon or are easily confused with other diagnoses because of their morphological similarity. Therefore, a pathologist may need reference knowledge to confirm a diagnosis or to find criteria to aid in the distinction of such diagnostically difficult cases. Reference knowledge comprises the consultation of experts, documented cases and, especially, books.

The consultation of books can be very laborious for several reasons. First, books cover a limited number of diagnoses with a few (mostly black and white) pictures at each diagnosis in order to keep their size and price reasonable. Therefore, the pathologist may need more than one book to obtain a satisfactory amount of information. Second, the differentiation between morphologically similar diagnoses may be difficult, since differential diagnosis lists and uniquely defined criteria for differentiating each possible diagnosis are scarce. Third, books permit only a one-dimensional ordering of its contents: the direction of reading. Direct access to a particular item of information is possible to a limited extent and requires a search through the index, which has a fixed arrangement. Comparison of information at different places in a book can only be achieved by text repetitions or the specification of internal references. Fourth, the field of pathology is so wide that information in books has to be restricted to parts of pathology. Usually, pathologists consult several books to obtain sufficient information for the solution of a diagnostic problem.

The Diagnostic Encyclopedia Workstation (DEW) [4,5] is a computerized text and image atlas, which has been developed to offer the pathologist easy and flexible access to reference knowledge as laid down in books, extended with a large amount of color pictures, differential diagnosis lists and criteria. In addition, other subjects such as prognosis and clinical signs of diagnoses are provided. The DEW can be operated from the pathologist's desk. At present, the DEW covers 85 diagnoses of ovary pathology, including all common and many rare cases (see next section), which we considered sufficient to carry out an experiment to test the system's present performance.

The following sections of the paper include a short description of the DEW, the set-up of the evaluation experiment, a discussion of the results, and a conclusion.

3.2 DESCRIPTION OF THE DEW

The choice of the hardware, the design of the database and the user-interface, together with the considerations underlying them, have been extensively discussed elsewhere [5]. Here, only the most relevant aspects of the system are summarized.

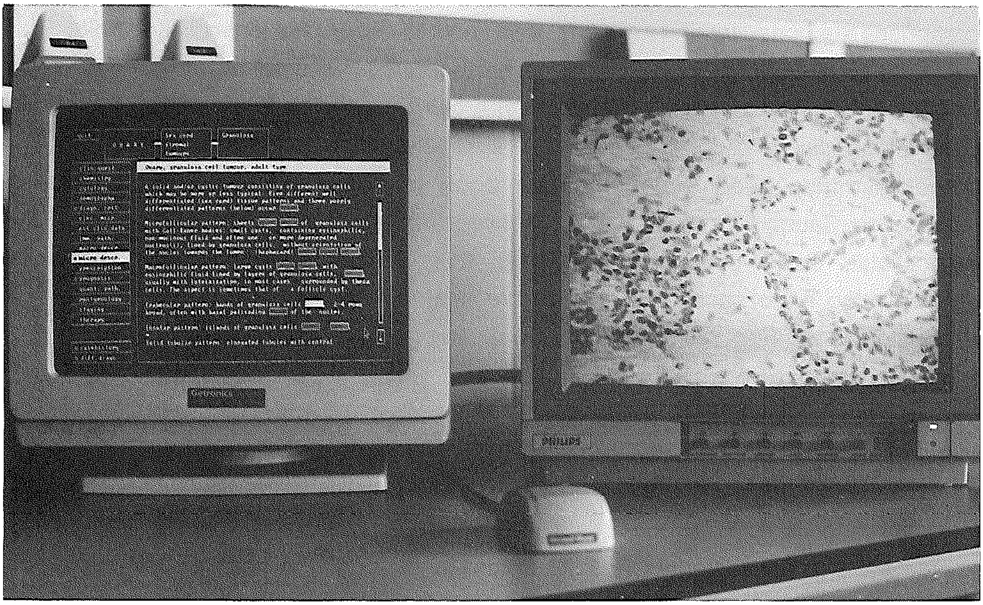


Figure 1. An overview of both monitors during a session with the DEW. On display are a microscopy description and a picture, illustrating a feature of the diagnosis.

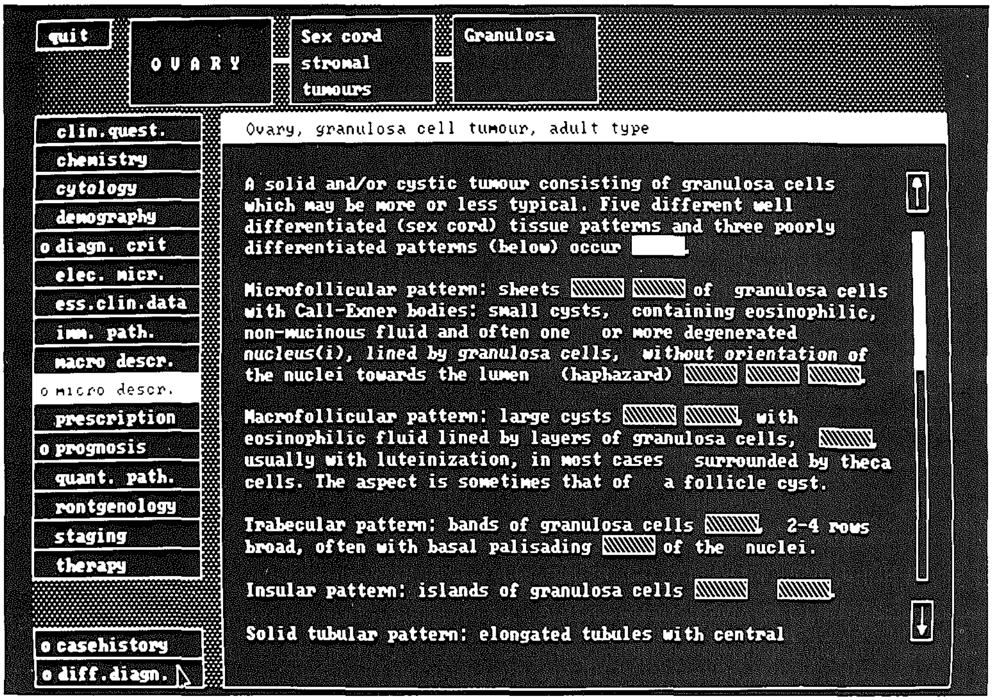
The DEW runs on an IBM-AT or compatible machine with 20 Mb hard disk, 640 Kb internal memory, a Hercules monochrome graphics card and an RS-232-C serial interface. Via the serial port the computer controls a videodisc player [1]. The monochrome monitor displays textual information, whereas a PAL video monitor displays the color pictures from the videodisc.

At present, the database holds information about 85 ovarian tumors. The tumor classification of the WHO is the basis for the ordering of the diagnoses in the

[1] Currently, the command codes of the Sony LDP 1500 P and the Philips VLP 835 players are supported.

database [6]. The 85 diagnoses cover almost completely the four main diagnostic groups of ovarian tumors: the common epithelial tumors, the sex cord stromal tumors, the germ cell tumors and the steroid cell tumors (see Appendix A). The pictures illustrating these diagnoses total approximately 3000, divided among 158 cases.

The user interface of the DEW is mouse-driven [7]. The first few screens serve as the table of contents as in a book and are used to specify the diagnosis to be retrieved. Each screen represents a level of choice, analogous to the chapters, sections and subsections in a book.



Ovary, granulosa cell tumour, adult type

A solid and/or cystic tumour consisting of granulosa cells which may be more or less typical. Five different well differentiated (sex cord) tissue patterns and three poorly differentiated patterns (below) occur □.

Microfollicular pattern: sheets ▨ ▨ of granulosa cells with Call-Exner bodies: small cysts, containing eosinophilic, non-mucinous fluid and often one or more degenerated *nucleus(i), lined by granulosa cells,* without orientation of the nuclei towards the lumen (haphazard) ▨ ▨ ▨.

Macrofollicular pattern: large cysts ▨ ▨, with eosinophilic fluid lined by layers of granulosa cells, ▨, usually with luteinization, in most cases surrounded by theca cells. The aspect is sometimes that of a follicle cyst.

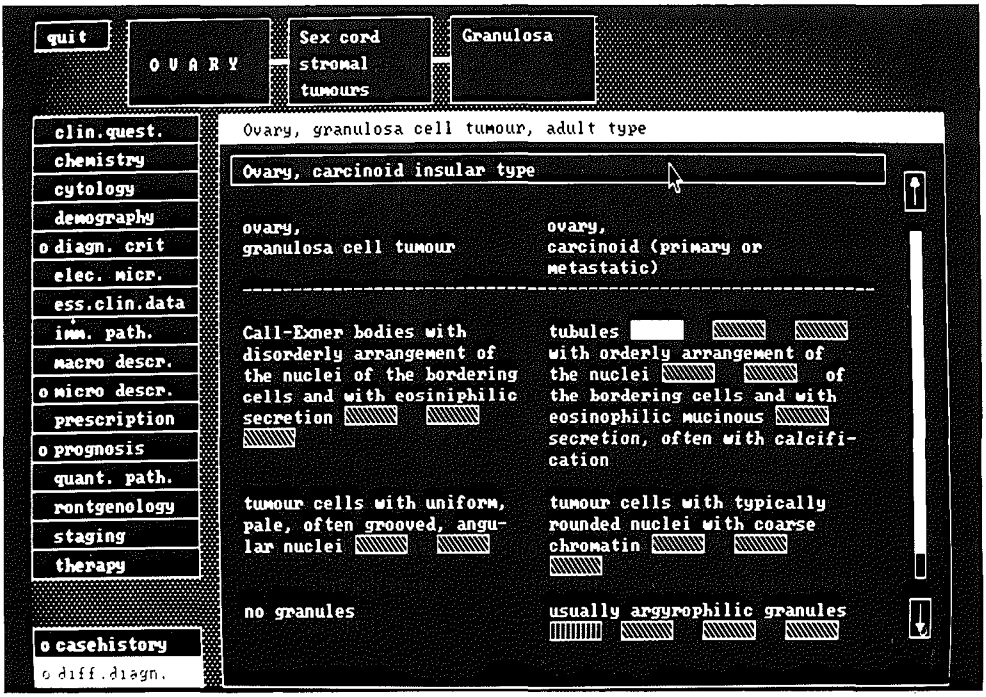
Trabecular pattern: bands of granulosa cells ▨, 2-4 rows broad, often with basal palisading ▨ of the nuclei.

Insular pattern: islands of granulosa cells ▨ ▨.

Solid tubular pattern: elongated tubules with central

Figure 2. Close-up of the screenlayout at diagnosis level. The microscopy description is displayed by default. Note the sensefields, the scrollbar at the right, the optional items of information at the left and the path through the classification tree at the top.

Once a diagnosis has been selected, a window with the microscopy description of that diagnosis is displayed on the screen and at the same time an overview of the histologic image is visible on the video monitor (Figs. 1,2). Small squares in the text are "sense fields", which result in the display of a picture when selected with the mouse. In this way, the user can call for illustrations of characteristics of a diagnosis, which are described in the text preceding the sense field. At the top of the screen the choices are visible, which lead to the selection of the current diagnosis. To the left of the text window is a list of other categories of information about the selected diagnosis, such as macroscopic description, immunopathology, electron microscopy, clinical data and prognosis. Since these categories are not always available or relevant, it is indicated on the screen, which categories are available for the current diagnosis. When one of the available categories of information is selected with the mouse, the text window changes accordingly. The category "diagnostic criteria" is always available. It contains a summary of all findings that have to be present in order to have sufficient proof for the selected diagnosis.



Ovary, granulosa cell tumour, adult type

Ovary, carcinoid insular type

ovary, granulosa cell tumour	ovary, carcinoid (primary or metastatic)
Call-Exner bodies with disorderly arrangement of the nuclei of the bordering cells and with eosiniphilic secretion ▨ ▨ ▨	tubules □ ▨ ▨ with orderly arrangement of the nuclei ▨ ▨ of the bordering cells and with eosinophilic mucinous ▨ secretion, often with calcification
tumour cells with uniform, pale, often grooved, angular nuclei ▨ ▨	tumour cells with typically rounded nuclei with coarse chromatin ▨ ▨ ▨
no granules	usually argyrophilic granules ▥ ▨ ▨ ▨

Figure 3. A table is shown, which lists the common and differentiating features of two morphologically similar diagnoses. Here, "carcinoid" is the selected diagnosis from the DD-list of the "adult granulosa cell tumor". It is possible to switch directly to "carcinoid" by selection of its name (arrow) with the mouse.

To support the differentiation between morphologically similar diagnoses, the user can call a list of these diagnoses at the lower left corner of the screen. A choice of one of the diagnoses on the list enables the pathologist to quickly compare the current diagnosis with the selected alternative [Fig. 3]. When the pathologist wants to switch from the current diagnosis to the alternative diagnosis, he or she can do so by touching the name of the alternative diagnosis with the mouse.

A session with the system is terminated by selecting the "quit" field at the upper left corner of the screen.

3.3 MATERIALS AND METHODS

3.3.1 Set-up of the evaluation experiment

To test the performance of the DEW versus books as sources of reference knowledge, 12 pathologists were divided into two equivalent groups, such that each group covered the whole spectrum of expertise in ovarian pathology. Each group was composed of one pathologist in early training, one pathologist in an advanced stage of training, two general pathologists and two experts in ovarian pathology. One expert, not taking part in the evaluation itself, selected the diagnostic test material. He selected two sets, A and B, such that 13 different diagnoses were represented by a different case in both sets.

Diagnoses in the test set	Order Session 1 Test set A	 Session 2 Test set B
insular carcinoid	1	5
Brenner tumor borderline	2	9
homologous mixed Mullerian tumor	3	10
mucinous cystadenocarcinoma well differentiated	4	2
cystic mature teratoma with malignant transformation	5	1
dysgerminoma	6	11
serous cystadenoma borderline	7	12
Sertoli cell tumor	8	3
endometrioid adenocarcinoma well differentiated	9	4
endodermal sinus tumor	10	7
mucinous cystadenoma borderline	11	8
immature teratoma	12	6
struma ovarii	13	13

Table 1. The diagnoses, which are represented by one case in each set. The numbers indicate the order in which the cases are presented in the first and the second session respectively.

Since evaluation of the quality of the diagnoses, made by the participants, requires a gold standard for the "correct" diagnosis, the cases were selected from the archive of the OTC (the Dutch national Ovarian Tumor Committee). However, part of the cases of the OTC archive were of great diagnostic difficulty and may have been diagnosed without complete consensus among the members of the OTC. Such diagnoses may be part of the test cases since the reports mentioned only the final diagnosis made by the OTC. The gold standard diagnoses of the cases included in the experiment are listed in Table 1.

	Session 1 **Test set A**	**Session 2** **Test set B**
Group 1	books	DEW
Group 2	DEW	books

Table 2. The order in which the participants use the books and the system. Note, that both groups start with slide set A.

The first group of pathologists (group 1) started with books as reference knowledge on the histologic slides of set A and used the DEW to diagnose set B. Group 2 started with set A as well, but used the DEW prior to the books. Table 2 shows the experiment schematically. Session 1 was always followed by session 2. The sessions were separated by more than one week. In both sessions the pathologists were offered a list, containing the names of the diagnoses, covered by the system (see Appendix A). Participants were allowed to use the list to find out which path should be taken in the system menu hierarchy to arrive at the diagnosis of their choice. To promote the comparability of the diagnostic results, the participants were asked to refine their diagnoses as much as possible, i.e. to choose only diagnosis names from the list. In the session with books three standard works on ovarian pathology were available [8-10].

A session with the DEW always started with a demonstration by the first author of approximately 15 minutes, followed by some time for the candidate to become familiar with the system. This required an average of 4-5 minutes. No time limits were imposed on the participants for the completion of the thirteen cases of each session. The cases were offered in a fixed order. During both sessions the first author (AMvG) observed the participants while making notes of the following:

- time mark when the participant started with a case
- time marks of every action of the candidate
    - looking through the microscope
    - looking at the list of diagnoses
    - consultation of the DEW or:
    - the book and the chapter, which was used
- time mark when the final diagnosis was made

The observer also recorded whether a diagnosis was made in doubt or not. During consultation the DEW system created a log file, containing all selections made by the user.

3.3.2 Viewpoints for evaluation

The experiment permits the evaluation of the following question:

- Does the diagnostic support of the DEW differ from the one provided by books, either qualitatively (agreement) or temporally (duration)?

It must be noted that the results of the second session can only be included in the evaluation of this question when both groups undergo an equal progress in their awareness of ovarian pathology in the first session. A learning effect, in the sense that it leads to improvement of diagnosis making, in itself is desirable and for that reason it is interesting to learn whether there is a difference in the learning effect between the use of books and the use of the system, but the set-up of the experiment does not allow for the differentiation between an unequal learning effect and the possibility that test set B is easier to diagnose than test set A. However, this differentiation is not relevant for the statistical evaluation of the books versus the DEW. When the analysis yields an equal learning effect in both groups after the first session the comparison of system versus books is not affected: the relative difference between the books and the system remains the same in both sessions. To evaluate the learning effect, achieved in the first session, the Wilcoxon Ranksum test first provides an answer to the following questions:

- Do both groups undergo an equal learning effect in the first session?
- Is the learning effect of the first session significant?

There are several viewpoints to evaluate the degree of diagnostic concordance with a gold standard:

- How well do the participants classify the cases of the test set?
- How strong is the morphological similarity between the diagnoses of the participants and the gold standard?
- What are, as compared to the gold standard, the clinical consequences of diagnoses, which differ from the gold standard?

The reason for this separation in different viewpoints is the fact that each of them represents a different ordering of diagnoses in groups of close resemblance: each viewpoint corresponds to a different interpretation of resemblance. The ordering in the WHO classification of ovarian tumors is a hybrid mixture in the sense that the division in major diagnostic groups reflects the origin of the tumors, whereas the minor divisions are based on morphological features. In consequence, some diagnoses show more morphological similarity with tumors in some other diagnosis group than with tumors in their own group. For example, an insular carcinoid has more in common, morphologically, with an adult granulosa cell tumor than with a mature cystic teratoma. Nonetheless, carcinoids and teratomas both belong to the group of germ cell tumors. Morphological similarity in turn, does not necessarily inform about the linical consequences of misdiagnosis. Two diagnoses may have many features in common and yet the treatment of patients with these tumors can differ considerably. The reverse may also occur.

In addition to these three viewpoints for evaluation (concordance, morphology, consequences), we have also analyzed the degree of consensus among the participants.

Finally, we tested the efficiency of the books versus the system, based on the time spent on each case.

3.3.3 Statistical analysis: scoring

For statistical evaluation of the diagnostic results a score is assigned to each diagnosis, given by the participants, to express its difference from the gold standard. A separate scoring is used for each viewpoint of evaluation. The scores expressing the degree of difference with the gold standard fulfill the requirements, which are posed to a metric. They can be rightfully said to express the distance between the diagnosis of the participant and the gold standard.

The "**classification score**" is based on the distance between the various levels in the classification tree of ovarian tumors. "Ovary" is the first level, "Common epithelial tumors" or "Germ cell tumors" are examples of the second level and diagnoses, such as "Serous adenocarcinoma" and "Dysgerminoma", are the leaves of the tree. The classification score is computed as follows. When the diagnosis of the participant is equal to the gold standard, the score is zero. In all other cases the score is equal to the difference between the level of the diagnosis of the participant and the smallest diagnosis group, which it has in common with the gold standard. For example, when a participant diagnosed a serous adenofibroma and the gold standard is a borderline endometrioid tumor, the smallest diagnosis group, which they have in common is common epithelial tumors (see Fig. 4).

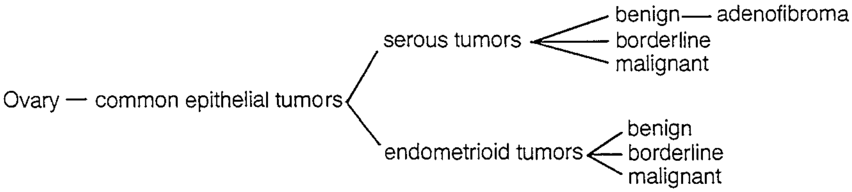


Figure 4. When a "serous adenofibroma" is diagnosed and the gold standard is a "borderline endometrioid tumor", then the smallest diagnosis group to which these diagnoses both belong is "common epithelial tumors". "Common epithelial tumors is two levels higher than "serous adenofibroma", yielding a classification distance of 2.

Starting upwards in the tree from a serous adenofibroma, it is one step to the group serous tumors and another step to the group common epithelial tumors. Consequently, the score assigned to the diagnosis of the participant is 2.

The "**morphology score**" to express the degree of morphological similarity is based on the consensus among gynecopathologists with respect to what diagnoses may give differential diagnostic problems with the gold standard. For this purpose eight experts in ovarian pathology made, for each of the 13 diagnoses of the experiment, a list of diagnoses, which they considered to pose diagnostic problems for a general pathologist. To our surprise the lists varied considerably among the experts (see Table 3 and Appendix B). Based on these lists the morphology score was determined as follows. The score was 0 when the diagnosis of the participant and the gold standard were the same. The score was 1 when all eight pathologists had included the diagnosis of the participant in their list, 2 when seven pathologists had mentioned that diagnosis, up to 8 when none of the pathologists considered the diagnosis of the participant morphologically confusable with the gold standard.

	DD's arranged by number of pathologists mentioning that DD.								Total of different DD's
Diagnosis	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
insular carcinoid	19	6	3	2	0	1	0	0	31
Brenner tumor borderline	12	2	0	0	0	2	1	0	17
homol. mixed Mullerian tumor	12	4	2	2	1	1	0	1	23
mucinous adenocarcinoma well differentiated	13	4	2	0	1	2	0	1	23
cystic mature teratoma with malignant tranformation	15	9	1	1	1	0	1	0	28
dysgerminoma	16	4	0	1	0	3	0	0	24
serous cystadenoma borderline	10	4	6	1	0	1	0	1	23
Sertoli cell tumor	20	8	6	0	1	2	0	0	37
endometrioid adenocarcinoma well differentiated	21	6	2	4	1	1	1	0	36
endodermal sinus tumor	12	6	1	2	0	1	1	0	23
mucinous cystadenoma borderline	17	7	3	4	1	0	0	2	34
immature teratoma	6	2	1	2	1	0	1	0	13
struma ovarii	11	6	2	0	0	1	1	0	21

Table 3 : The number of differential diagnoses, given by 8 experts for each diagnosis of the test set, related to the number pathologists agreeing with these differential diagnoses. The first column represents minimal consensus: this number of differential diagnoses is proposed by a single pathologist. The last column represents complete consensus: this number of diagnoses is suggested by all 8 pathologists.

The "**consequence score**" for the clinical consequences of misdiagnosis does not differentiate between the risk of overtreatment and the risk of undertreatment. For all diagnoses made by the participants, one expert assigned a score of 1 for slight, 2 for moderately severe, and 3 for severe differences in clinical consequences when compared to the gold standard.

The "**consensus score**" reflects, for each of both groups, the number of participants who made the same diagnosis. Consequently, a score of 1 reflects the minimal consensus, whereas a score of 6 represents complete consensus.

The "**time score**", used to compare the duration of diagnosis making, was equal to the number of seconds, which elapsed between starting with a case and making the final diagnosis.

The Wilcoxon Ranksum test [11] was used for the statistical analysis. It is important to realize that the Wilcoxon Ranksum test is used to detect the presence or absence of a significant difference between two small sets of data. It does not inform about the degree of difference between the two sets. At all times, when using the Wilcoxon Ranksum test the null hypothesis was the absence of a significant difference in diagnostic results between the system and the books. As level of significance we used 5%. In other words, the assumption that the system and the books are equivalent is rejected when the results of the experiment have a probability smaller than 5%. When the null hypothesis is rejected, the average of the test results of each group reveals whether books or system must be favored.

3.4 RESULTS

3.4.1 Scores

The average scores of both groups in relation to the use of the DEW and the books, together with the results of the statistical analysis, are shown in Table 4. Note that good results are reflected by low average scores in the upper three rows of Table 4, high average scores in the fourth row, and low averages in the last row of Table 4.

Subject of	average G1		average G2		significant
evaluation	books	DEW	DEW	books	in favor of:
classification	11.7	10.7	13.3	6.3	books
morphology	25.5	30.7	32.3	17.8	books
consequences	16.7	14.3	17.0	10.8	- -
consensus	43.0	36.7	43.3	49.7	- -
time in sec	5567	5323	4681	4060	- -

Table 4. The average score of all pathologists for each viewpoint of evaluation.

A significant learning effect was found both for the viewpoints of classification and clinical consequences. From the viewpoints of both classification and morphology the books were found to give better results than the DEW system. The data show that the diagnoses made in the system sessions less often belonged to the same diagnostic group as the gold standard than those made in the book sessions. The books were not significantly superior to the DEW with respect to consensus and clinical consequences, although the results of the second session show a clear tendency in favor of the books. This may be explained by differences in learning effect in favor of the DEW. It is important to realize that the statistical analysis of system versus books with respect to clinical consequences is less sensitive than the ones for classification and morphology, which is due to the fact, that there are fewer different therapies available than there are diagnosis names.

		Books		DEW	
		with	without	with	without
Session 1	E	16	24	17	17
	D	22	16	34	10
Session 2	E	26	23	14	25
	D	18	11	22	17

E = equal to gold standard
D = different from gold standard

Table 5. The number of correct classifications and misclassifications, related to the use of books as well as the DEW.

During the sessions, the books or the DEW were not always consulted. Table 5 shows the number of correct classifications and misclassifications in relation to the use of books and the DEW. Note that the majority of cases which were diagnosed without the use of reference knowledge were classified correctly as opposed to the cases which were classified with the use of books or the DEW.

Even when the books or the system were used, the participants did not always consult the "correct" (gold standard) diagnosis. With respect to the use of the DEW, the log files revealed that only for 18 of the 56 misclassifications with the system the "correct" diagnosis was consulted. It was not considered feasible to collect the same information from the sessions with the books since this would require to let the participants mention every diagnosis they consulted.

	certain	uncertain
books	82	5
system	87	13

Table 6 : The number of diagnoses, which were diagnosed with and without doubt, related to the use of books as well as the DEW. The remaining 125 diagnoses were made without the use of reference knowledge.

Finally, Table 6 shows the ratio between the number of "certain" and "uncertain" diagnoses in relation to the use of books or the DEW.

3.4.2 Comments by the pathologists

Apart from the results of the statistical analysis, we made notes of the participants' comments with respect to the use of the system. In this way we gained more insight in the weaknesses and strengths of the system. Apart from the use of the books or the DEW, all participants mentioned that the sessions differed considerably from the normal diagnostic routine. Since it was not possible to ask additional slides of a case or to put off a diagnosis till the next day, they sometimes felt themselves forced to make diagnoses for which they would not have taken responsibility in real practice. In the following, the positively valued properties of the DEW are discussed first, followed by their suggestions for improvement.

The first strength is the easy access to the information: only a few mouse clicks are necessary to consult diagnosis information and pictures. None of the participants experienced difficulties in working with the system. This is satisfying since none of them was experienced in using computers and all of them received no more than 20 minutes to become acquainted with the DEW.

The second strength concerns the availability of a large amount of color pictures. Books seldom contain more than two pictures per diagnosis and the majority of them are in black and white. The average photographic quality of the pictures was considered comparable to books with the exception of overviews with low contrast, even taking into account that they are displayed on a TV-screen.

The third property which was appreciated by the participants, concerned the availability of the pictures at the cases. Besides sense fields in the text to illustrate particular characteristics, pictures were also available as sense fields in a list sorted by subject, stain and magnification.

The fourth and last property to be mentioned is the availability of DD-lists and DD-criteria. In books, information about morphologically similar diagnoses is often scarce. Criteria to differentiate among them are even scarcer and are often found by looking at the alternatives themselves. The system has tables with DD-criteria, for each diagnosis in the DD-list to distinguish it from the selected alternative diagnosis.

As to suggestions for improvement, some of the participants mentioned the need for criteria to differentiate among diagnosis groups. Especially when a case is unfamiliar, such criteria would help them to find the appropriate path through the menu hierarchy. The design supports this option, but the criteria have not yet been entered due to shortage of time, available for development of the DEW.

Most participants preferred to see overviews prior to detailed pictures, since an overview might be sufficient to reject a diagnosis. They expressed a need for more overviews, especially available in the beginning of the diagnosis description. The relative scarcity of overviews is due to the fact that magnifications of 2.5x or smaller with low contrast require a higher resolution display than a normal video signal can offer. It must be noted, however, that the NTSC video signal produces better quality display than the PAL signal because of the higher frequency of 60 hertz of NTSC. More magnifications of 10x could be added and even magnifications of 2.5x of moderate contrast might be useful for a first impression.

Another remark concerned the topics of photography. Several participants, especially the more experienced ones, mentioned that part of the pictures were nonspecific for a diagnosis. They referred mainly to pictures of mitoses, atypia and stratification. These pictures correctly illustrate phenomena of the selected diagnosis, but they were not found to be helpful in the decision process. It is important to realize that increasing experience leads to easier interpretation of verbal descriptions and an increasing preference for highly specific illustrations.

Finally, the participants, almost unanimously, expressed the wish to have access to the pictures sorted by subject, stain and magnification also for diagnoses, in addition to the sense fields. It happened several times that the participants "tried" many sense fields to find an overview or a picture that might show an image comparable to what they had under the microscope. At times they gave up the effort long before all sense fields were tried. The availability of sorted pictures will allow access which is more adjusted to the needs of the user.

3.5 DISCUSSION

Although the clinical consequences of a diagnosis are more important than a correct classification, it is not sufficiently satisfactory that the DEW and the books yield equivalent results with respect to clinical consequences. Classification is the basis for therapy selection and, therefore, it is important to strive at an optimal classification of diseases.

3.5.1 Role of the DEW for the classification of ovarian tumors

It is important to realize that systems like the DEW can never solve the problem of consensus and in this they do not differ from books. An important explanation for the absence of complete consensus are the differences in education and experience among experts. As a result, experts differ in their judgments, putting different weight on certain characteristics within the context of other features. In addition, experts may differ in their precision to screen a slide. As a consequence, the contents of a the DEW as well as books will always be subject of discussion and even a complete agreement with respect to the contents of the DEW would not lead to complete diagnostic agreement among its users.

We observed several times that participants, using the same book as reference knowledge diagnosed a case differently. For example, two participants both considered an insular carcinoid and an adult granulosa cell tumor. They used the same book as reference knowledge. Finally, they made different decisions. Since they used the same reference knowledge, their attention was drawn to all relevant diagnostic criteria. Apparently, they put different emphasis on the morphological phenomena in the histological slide.

Apart from efficient access, the intended usefulness of the DEW in the classification of ovarian tumors lies in the fact, that it may enhance the user's awareness of:

- All criteria, relevant to confirm a diagnosis.
- The histologic variability of diagnoses.
- Potential diagnoses together with the criteria to differentiate among them.
- Differences in clinical consequences between diagnoses under consideration.

3.5.2 Causes of misdiagnosis

Prior to discussing potential causes of misclassification, it is important to realize that the experiment was negatively biased for the DEW, the main reason being the availability of only one H-E slide in the majority of the test cases. This posed difficulties in diagnosis making which, in practice, would have been easily solved with additional stains. In addition, the large set of 3000 pictures, as available in the DEW, could not be taken to its full advantage. The cases, for which several stains could be obtained, were already used for the video disc so we had to accept a selection of the remaining suboptimal cases for the experiment. However, the insights gained with respect to the functioning of the DEW are also applicable to situations where more stains are available. In the following, we concentrate on those potential causes of misdiagnosis, which can be explained on the basis of the experiment and which give insight into possible improvements of the DEW.

First, there is the problem of consensus [12,13]. Some diagnoses, which we have classified as a misdiagnosis on the basis of the gold standard, may in fact be judged correctly by one or more experts. It is interesting to mention, that the Sertoli cell tumor in set A, received eight different diagnoses. Apparently, the slide showed features fitting with several other diagnoses. Another example of a consensus problem are the two cases of a borderline Brenner tumor in the test set. This case was diagnosed as a benign Brenner tumor by 10 of the 12 participants regardless of the use of books or the DEW. Although consensus among the participants was very high, this case is responsible for 10 misclassifications with respect to the gold standard!

Second, a possible negative effect on the diagnostic result concerns the pictures. As the participants mentioned, part of the pictures did fit their diagnosis, but did not characterize it. Such pictures serve the purpose of completeness with respect to all possible histological manifestations of a diagnosis, including those shared with other diagnoses. There are also some rare diagnoses for which we found no cases at all and which we illustrated with pictures from other diagnoses. Then, it is the combination of pictures which characterizes the histologic image. However, the use of pictures from other diagnoses carries the risk of misinterpretation when they show more phenomena than the one(s) they were meant to illustrate. In general, when non-specific pictures dominate, insufficient scanning of the available pictures may cause the user to reject the diagnosis too soon.

Third, we observed that in 38 out of 56 misdiagnoses made with the aid of the system, the correct diagnosis was not consulted. A possible cause for not consulting the correct diagnosis may be found in the contents of the DD-lists per diagnosis. The example in Appendix B and Table 3 show that experts vary in their view on morphological similarity among diagnoses. Note, that for nine diagnoses of the test set the intersection of the DD-lists of the eight consulted pathologists is empty! In the same way, the DD-lists in the DEW differ from those made by the experts. As a consequence, it may very well happen that the user consults a diagnosis, which is considered to be morphologically similar to the correct diagnosis by part of the experts, but not by the system. Then, the DD-lists of that diagnosis will not help the user to find the correct diagnosis. A different situation, where the DD-lists of the system are not useful, occurs when the user consults a diagnosis far from the correct diagnosis. Here, the primary problem is not the contents of the DD-lists, but the fact that unfamiliarity with the case or a wrong interpretation of the observations causes the user to consider the wrong diagnoses. Then, it depends on the user whether the insight emerges that a different entry into the system is necessary.

In general, the search effort of the user is crucial for the diagnostic result: one user may search until a diagnosis is found, which fits moderately with the observations, whereas some other user may search for the perfect fit. Therefore, it is important to realize that long DD-lists and large numbers of pictures, which require a lot of scanning effort from the user, may have a negative influence on the diagnostic result.

3.5.3 Suggestions for improvements

Improvements of the DEW should include efficient support in finding the correct set of diagnoses to consider. As the participants mentioned, criteria to differentiate among diagnosis groups would facilitate consultation of the system for unfamiliar cases. The absence of these criteria is not a shortcoming of the system's design, but a consequence of the fact that this information is not yet completely entered.

As to the DD-lists, it is laborious to "try" all possibilities. The availability of a few overviews for each diagnosis on a DD-list offers the possibility to scan the list prior to making a selection.

In the diagnosis texts, the pictures are only accessible via sense fields in the text. Users, who want to see overviews or specific stains, probably the most experienced ones, may have to scan large portions of the diagnosis text before they find what they are looking for. It would facilitate a more directed scanning of the pictures when they were also available sorted by subject, laboratory technique and magnification (as the slides of the cases). When pictures are grouped together in categories, overviews are easy to find and the presence of non-specific pictures, for the purpose of completeness, may be found less inconvenient by experienced users. In general, the number of overviews should be increased.

For the selection of the cases it is worth to consider sampling from routine archives of experts. These archives probably contain many cases, which are very specific for a diagnosis and, therefore, do not give rise to consensus problems.

Naturally, the problem of consensus and wrong entries also occur when consulting books. In addition, it is important to realize that a considerable amount of diagnoses, including misdiagnoses have been made without the use of reference knowledge.

3.6 CONCLUSIONS

Based on an experiment with 12 pathologists, statistical analysis of the diagnostic support offered by either the DEW or books permits the following conclusions:
Books offered superior support to the DEW from the viewpoints of:

- classification.
- morphological similarity of diagnoses with the gold standard

Books and DEW differed, though not significantly, in favor of the books with respect to:

- the clinical consequences of misdiagnosis
- mutual consensus among the participants
- duration of the diagnostic process

However, it should be kept in mind that the evaluation experiment was tightly controlled and negatively biased for the DEW system: the large set of pictures of the DEW could not be used to full advantage.

Although, practically, a correct therapy is more important than a correct classification, the latter is necessary to select the therapy. In order to set goals for improvement of the system's support in classification, we analyzed its strengths and weaknesses. In the evaluation, it turned out that strong properties are:

- easy, mouse-driven access to the information
- the presence of many color pictures
- the availability of DD-lists and DD-criteria

The most prominent aspect to be improved is support in finding the way to the correct set of diagnoses for consideration. Useful extensions in the design include the availability of histologic overviews in DD-lists and the possibility to access the diagnosis pictures sorted by subject, stain and magnification in addition to the sense fields in the text. As to the contents, improvements would be the presence of criteria to differentiate among diagnosis groups, and expansion of the DD-lists based on information from experts.

Leaving the diagnostic responsibility with the user, the DEW is intended to make the diagnostic process less dependent on personal factors such as the user's preexistent knowledge and diagnostic approach. So far, the experiment has proven that the design of the DEW is successful in supporting efficient access to diagnosis information and differential diagnostic criteria for consultation. Provided that the design of the DEW is improved as indicated and its contents extended to other parts of pathology, it has the potential of becoming a welcome addition to daily diagnostic practice.

ACKNOWLEDGEMENTS

This work was supported by grant PF 28-1207 of the Praeventiefonds. We wish to thank J.G. van den Tweel M.D. Ph.D., C.E.M. Blomjous M.D. Ph.D., R.F.M. Schapers M.D. Ph.D., A.H. van Hattum M.D. Ph.D., H.C.M. van der Schoot M.D. Ph.D., J.R.J. Elbers M.D., J.C. van der Linden M.D. Ph.D., P. van der Valk M.D., H.V. Stel M.D. and H.E. van Ipenburg M.D., who participated in the evaluation experiment and H. Fox M.D. Ph.D., J.F.M. Delemarre M.D. Ph.D. and C. Kooijman M.D. Ph.D., who also helped with defining differential diagnoses. F. Nogales M.D. Ph.D., F.A. Langley M.D. Ph.D., A. Talerman M.D., R.E. Scully M.D. Ph.D. and A. Schaberg M.D. Ph.D. are also acknowledged for their valuable help in defining differential diagnoses. We acknowledge the members of the OTC (Dutch Ovarian Tumors Comittee), who have not yet been mentioned, S. Chadha-Ajwani M.D. Ph.D., F.B. Lammes M.D. Ph.D., A.G.J.M. Hanselaar M.D., Ch. Albus-Lutter M.D. Ph.D., G. Wielinga M.D. Ph.D., M.J. Becker-Bloemkolk M.D. Ph.D., E.J. Aartsen M.D., H.F. Heins M.D. Ph.D., G.J.M. Ras-Zeylmans M.D., E. Engelsman M.D. Ph.D., M. v.d. Burg M.D. and J.H. Meerwaldt M.D. Ph.D.

REFERENCES

[1] Langley FA, Baak JPA, Oort J, Diagnosis making: Error sources, in: Baak JPA, Oort J (eds.) A Manual of Morphometry in Diagnostic Pathology, Springer Verlag, New York, 1983, pp. 6-14.

[2] Giard RWM, Inflammatoire ziekten van het colon, Ph.D. thesis, Chapter 2, Rijksuniversiteit, Leiden 1986.

[3] Connelly DP, Johnson PE, The medical problem solving process, Human Pathol. 11 pp. 412-419, 1980.

[4] Jansen W, Baak JPA, Van Ginneken AM, Smeulders AWM, Diagnostic encyclopaedia workstation: an interactive system for diagnostic support, in: Imaging and visual documentation in medicine (Wamsteker K, et al., eds.), Elsevier Science Pub., Amsterdam, 1987, pp. 777-780.

[5] Van Ginneken AM, Smeulders AWM, Jansen W, Design of a diagnostic encyclopaedia using AIDA. In: Comp. Meth. Prog. Biomed. 25 pp. 339-348, 1987.

[6] Serov SF, Scully RE and Sobin LH, Histological typing of ovarian tumors, in: International Histological Classification of Tumours 9, World Health Organization, Geneva, 1973.

[7] Design of the Diagnostic Encyclopedia Workstation (DEW), submitted for publication in: Computers in Biology and Medicine, Chapter 2 this thesis.

[8] Blaustein A, Pathology of the Female Genital Tract, 3rd edn. Springer Verlag, Berlin, 1987.

[9] Scully RE, Tumors of the Ovary and Maldeveloped Gonads, AFIP, Washington, D.C. 1979.

[10] Fox H, Langley FA, Tumours of the ovary, Heineman, London, 1976.

[11] Hills M, Armitage P, The two-period cross-over clinical trial, Br. J. clin. Pharmac. 8 pp. 7-20, 1979.

[12] Baak JPA, Langley FA, Talerman A, Delemarre JFM, Interpathologist and Intrapathologist Disagreement in Ovarian Tumor Grading and Typing, Analyt. Quant. Cytol. 8 pp. 354-357, 1986.

[13] Baak JPA and Oort J, The Case for Morphometry in Diagnostic Pathology in: Baak JPA and Oort J, A manual of morphometry in diagnostic pathology, 1983, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, pp. 2-5.

Appendix A : This list represents the diagnoses and diagnosis groups, covered by the system. Subdivisions are represented by indentations in the list.

- OVARY
  - common epithelial tumors
    - serous tumors
      - benign serous tumors
        - serous (papillary) cystadenoma
        - serous (cyst)adenofibroma
        - serous surface papilloma
      - serous tumors of borderline malignant
        - serous papillary cystadenoma borderline
        - serous surface papilloma of borderline malignancy
        - serous (cyst)adenofibroma of borderline malignancy
      - malignant serous tumors
        - serous papillary (cyst)adenocarcinoma
          - serous (cyst)adenocarcinoma well differentiated
          - serous (cyst)adenocarcinoma moderately differentiated
          - serous (cyst)adenocarcinoma poorly differentiated
        - serous surface papillary carcinoma
        - malignant serous (cyst)adenofibroma
    - mucinous tumors
      - benign mucinous tumors
        - mucinous cystadenoma
        - mucinous (cyst)adenofibroma
      - mucinous tumors of borderline malignancy
        - mucinous cystadenoma of borderline malignancy
        - mucinous (cyst)adenofibroma of borderline malignancy
      - malignant mucinous tumors
        - mucinous (cyst)adenocarcinoma
          - mucinous (cyst)adenocarcinoma well differentiated
          - mucinous (cyst)adenocarcinoma moderately differentiated
          - mucinous (cyst)adenocarcinoma poorly differentiated
        - malignant mucinous (cyst)adenofibroma
    - endometrioid tumors
      - endometriosis
      - benign endometrioid tumors
        - endometrioid (cyst)adenoma
        - endometrioid (cyst)adenofibroma
      - endometrioid tumors of borderline malignancy
        - endometrioid (cyst)adenoma of borderline malignancy
        - endometrioid (cyst)adenofibroma of borderline malignancy
      - malignant endometrioid tumors
        - endometrioid adenocarcinoma
          - endometrioid adenocarcinoma well differentiated
          - endometrioid adenocarcinoma moderately differentiated
          - endometrioid adenocarcinoma poorly differentiated
        - endometrioid adenoacanthoma
        - malignant endometrioid cystadenofibroma
        - endometrioid stromal sarcoma
          - endometroid stromal sarcoma, good prognosis
          - endometrioid stromal sarcoma, poor prognosis

      - homologous mixed Mullerian tumors
      - heterologous mixed Mullerian tumors
      - adenosarcoma
  - clear cell tumors
    - clear cell adenofibroma
    - clear cell tumor of borderline malignancy
    - clear cell adenofibroma/ carcinoma
    - clear cell carcinoma
  - Brenner tumors
    - benign Brenner tumor
    - Brenner tumor of borderline malignancy
    - malignant Brenner tumor
    - transitional cell carcinoma
  - adenomatoid tumor
  - mixed epithelial tumors
    - benign mixed epithelial tumor
    - mixed epithelial tumor of borderline malignancy
    - malignant mixed epithelial tumor
  - undifferentiated carcinoma
    - undifferentiated carcinoma, large cell type
    - undifferentiated carcinoma, small cell type
  - unclassified epithelial tumors
- sex cord stromal tumors
  - granulosa cell tumors
    - granulosa cell tumor, juvenile type
    - granulosa cell tumor, adult type
    - granulosa cell tumor, poorly differentiated type
  - tumors in the thecoma fibroma group
    - thecoma
    - luteinized thecoma
    - stromal luteoma
    - fibroma
    - unclassifiable thecofibromatous tumor
    - sclerosing stromal tumor
  - Sertoli-Leydig cell tumors
    - Sertoli-Leydig cell tumors, well differentiated
      - Sertoli cell tumor
      - Sertoli cell tumor with lipid storage
      - Sertoli-Leydig cell tumor
      - Leydig cell tumor
    - Sertoli-Leydig cell tumor of intermediate differentiated
    - Sertoli-Leydig cell tumor, poorly differentiated
    - Sertoli-Leydig cell tumor with heterologous elements
  - gynandroblastoma
  - unclassified sex cord stromal tumors
    - sex cord tumor with annular tubules
      - SCTAT without Peutz-Jegher
      - SCTAT with Peutz-Jegher
- steroid cell tumors
  - stromal luteoma
  - luteinized thecoma
  - Leydig cell tumor
  - lipid cell tumor

- germ cell tumors
  - dysgerminoma
  - endodermal sinus tumor
  - embryonal carcinoma
  - polyembroma
  - choriocarcinoma
  - teratomas
    - immature teratoma
    - solid mature teratoma
    - benign cystic mature teratoma
    - cystic mature teratoma with malignant transformation
    - struma ovarii
    - malignant struma ovarii
    - carcinoid
      - carcinoid insular type
      - carcinoid trabecular type
      - goblet cell carcinoid
    - struma ovarii and carcinoid
    - other monodermal teratomas
  - mixed germ cell tumors

Appendix B: For each diagnosis of the test set, the columns represent the differential diagnoses as specified by each of 8 gynaecopathologists.

Testdiagnosis: borderline mucinous cystadenoma

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenoma			*			*			*
serous cystadenofibroma			*			*			
serous surface papilloma			*						
serous cystadenoma borderline			*		*				
serous surface papilloma borderline			*						
serous cystadenofibroma borderline			*			*			
serous cystadenoca well diff.			*			*			
serous surface papillary carcinoma			*						
serous adenofibroma malignant			*						
mucinous cystadenoma	*	*	*	*	*	*	*	*	
mucinous cystadenofibroma			*			*	*	*	
mucinous cystadenofibroma borderline						*			
mucinous adenoca. well diff.	*	*	*	*	*	*	*	*	
mucinous adenoca. mod. diff.	*		*	*			*		
mucinous adenoca. poorly diff.				*			*		
mucinous adenofibroma malignant				*					
endometrioid cystadenoma						*			
endometrioid cystadenofibroma						*			
endometrioid cystadenofibroma borderline		*		*	*	*	*		
endometrioid adenoca. well diff.	*	*	*				*		
endometrioid adenoca. mod. diff.							*		
clear cell adenofibroma			*						
clear cell tumor borderline	*		*			*			*
clear cell adenofibroma/carcinoma	*								*
clear cell carcinoma						*			*
mixed epithelial tumor benign			*	*					
mixed epithelial tumor borderline			*	*	*				
mixed epithelial tumor malignant				*					
Sertoli-leydig cell tumor with heterologous elements				*	*	*	*		
sex cord tumor with annular tubules					*	*	*		*
cystic mature teratoma with malignant transformation				*					
struma ovarii						*			
carcinoid insular type						*			
other monodermal teratomas			*						

Testdiagnosis: insular carcinoid

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenofibroma						*			
serous cystadenofibroma borderline						*			
serous cystadenoca. well diff.						*			
serous cystadenoca. mod. diff.						*			
serous cystadenoca. poorly diff.						*			
mucinous cystadenofibroma						*			
mucinous cystadenoca. well diff.						*			
mucinous cystadenoca. mod. diff.						*			
mucinous cystadenoca. poorly diff.						*			
endometrioid adenoca. well diff.							*		*
endometrioid adenoca. mod. diff.							*		*
clear cell adenofibroma			*						
clear cell carcinoma			*						
Brenner tumor benign			*		*	*	*		*
Brenner tumor borderline				*		*			*
Brenner tumor malignant				*		*			*
undifferentiated carcinoma								*	*
granulosa cell tumor juvenile type									*
granulosa cell tumor adult type		*	*	*	*	*	*		*
granulosa cell tumor poorly diff.		*	*	*					*
Sertoli cell tumor	*	*		*		*			*
Sertoli cell tumor with lipid storage	*					*			
Sertoli-Leydig cell tumor well diff.				*		*			
Sertoli-Leydig cell tumor mod. diff.		*				*			
Sertoli-Leydig cell tumor with heterologous elements						*	*		
sex cord tumor with annular tubules		*							
dysgerminoma				*					
cystic mature teratoma with malignant transformation			*						*
struma ovarii malignant							*		
struma ovarii and carcinoid	*	*	*						*
carcinoid trabecular type						*	*	*	*
carcinoid goblet cell								*	*

Testdiagnosis: endometrioid adenocarcinoma, well differentiated

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenofibroma						*			
serous cystadenofibroma borderline						*			
serous cystadenoca. well diff.		*	*			*			*
serous cystadenoca. mod. diff.			*						
serous adenofibroma malignant						*			
mucinous cystadenofibroma						*			
mucinous cystadenoma borderline						*			
mucinous cystadenofibroma borderline						*			
mucinous cystadenoca. well diff.		*	*	*	*	*			*
mucinous cystadenoca. mod. diff.	*		*	*			*		*
mucinous cystadenoca. poorly diff.			*						*
mucinous adenofibroma malignant	*								
endometriosis						*			*
endometrioid cystadenoma		*	*	*				*	
endometrioid cystadenofibroma	*	*		*		*			
endometrioid cystadenofibroma borderline	*		*	*	*	*	*	*	*
endometrioid cystadenoca. mod. diff.	*		*	*	*		*	*	*
endometrioid cystadenoca. poorly diff.				*					*
endometrioid adenoacanthoma			*						
endometrioid cystadenofibroma malignant			*	*					
homologous mixed Mullerian tumor			*						
heterologous mixed Mullerian tumor			*						
clear cell adenofibroma	*								
clear cell tumor borderline	*			*					
clear cell adenofibroma/carcinoma			*						
clear cell carcinoma		*		*					
Brenner tumor borderline					*				
mixed epithelial tumor borderline				*	*				
mixed epithelial tumor malignant			*		*				
Sertoli cell tumor			*	*	*	*			*
Sertoli cell tumor with lipid storage						*			*
Sertoli-Leydig cell tumor					*	*	*		*
Leydig cell tumor						*			*
endodermal sinus tumor			*		*				
immature teratoma			*						
cystic mature teratoma with malignant transformation			*						
carcinoid									*

Testdiagnosis: Endodermal sinus tumor

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenoca. mod. diff.			*						
mucinous cystadenoca. well diff.				*					
mucinous cystadenoca. mod. diff.			*	*					
mucinous cystadenoca. poorly diff.				*	*				
endometrioid adenoca. well diff.			*	*	*				
endometrioid adenoca. mod. diff.				*	*				
endometrioid adenoca. poorly diff.				*					
heterologous mixed Mullerian tumor			*						
clear cell tumor borderline				*			*		
clear cell adenofibroma/carcinoma				*	*				*
clear cell carcinoma	*		*	*	*	*	*		*
adenomatoid tumor			*						
undifferentiated carcinoma				*					
undifferentiated ca. large cell					*				
undifferentiated ca. small cell							*		
granulosa cell tumor juvenile type			*						
Leydig cell tumor			*						
dysgerminoma		*		*	*	*			
embryonal carcinoma	*	*	*	*	*	*	*		*
polyembryoma		*		*					
choriocarcinoma				*					
immature teratoma		*							
mixed germ cell tumors				*	*		*	*	*

Testdiagnosis: immature teratoma

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
homologous mixed Mullerian tumor			*	*		*			*
heterologous mixed Mullerian tumor	*		*	*	*	*			*
adenosarcoma					*				
undifferentiated ca. small cell				*					
Sertoli-Leydig cell tumor with heterologous elements					*				
dysgerminoma					*				
endodermal sinus tumor				*					
embryonal carcinoma			*	*					*
choriocarcinoma				*					
solid mature teratoma	*	*	*	*	*	*	*		*
cystic mature teratoma	*	*	*				*		*
cystic mature teratoma with malignant transformation		*		*		*	*		*
mixed germ cell tumor				*				*	*

Testdiagnosis: struma ovarii

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenoma				*					
mucinous cystadenoma			*	*					
mucinous cystadenofibroma			*						
endometrioid cystadenoma			*						
endometrioid adenocarcinoma						*			
clear cell adenofibroma			*		*				
clear cell carcinoma			*		*	*			
Brenner tumor benign				*					
undifferentiated ca. small cell								*	
granulosa cell tumor adult type			*	*				*	
Sertoli-leydig cell tumor well diff.				*		*			
Sertoli-Leydig cell tumor mod. diff.					*	*			
SCTAT without Peutz-Jegher			*						
SCTAT with Peutz-Jegher			*						
endodermal sinus tumor			*						
immature teratoma			*						*
solid mature teratoma	*			*					
cystic mature teratoma	*			*					*
cystic mature teratoma with malignant transformation									*
struma ovarii malignant	*	*	*	*		*	*	*	*
carcinoid insular type								*	
struma ovarii and carcinoid	*	*	*	*	*		*		
mixed germ cell tumor									*

Testdiagnosis: Homologous mixed Mullerian tumor

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenoca. poorly diff.	*	*			*				*
serous adenofibroma malignant			*		*				
mucinous cystadenoca. poorly diff.			*						*
endometrioid adenofibroma			*						
endometrioid adenoca. well diff.							*		
endometrioid adenoca. mod. diff.							*		
endometrioid adenoca. poorly diff.		*	*		*		*		*
endometrioid adenoacanthoma			*						
endometrioid adenofibroma malignant			*		*				
endometrioid stromal sarcoma		*	*	*	*		*	*	
heterologous mixed Mullerian tumor		*	*	*		*		*	*
Adenosarcoma		*	*	*	*	*	*	*	* *
Brenner tumor malignant									*
transitional cell carcinoma	*								
mixed epithelial tumor malignant							*		
undifferentiated ca. large cell							*		
undifferentiated ca. small cell	*		*						
granulosa cell tumor poorly diff.	*								
thecoma			*						
Sertoli-Leydig cell tumor poorly diff.			*	*	*				
Sertoli-Leydig cell tumor with heterologous elements				*					
immature teratoma		*		*		*	*		*
solid mature teratoma				*					
cystic mature teratoma with malignant transformation		*	*						

Testdiagnosis: Brenner tumor borderline

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenoca. mod. diff.			*						
mucinous cystadenoca. poorly diff.			*						
endometrioid adenoca. mod. diff.			*						
endometrioid adenoca. poorly diff.			*						
endometrioid adenoacanthoma					*	*			
endometrioid adenofibroma malignant					*				
homologous mixed Mullerian tumor			*						
heterologous mixed Mullerian tumor			*						
clear cell adenofibroma/carcinoma							*		
Brenner tumor benign	*	*	*	*		*		*	*
Brenner tumor malignant		*	*	*	*	*	*	*	*
transitional cell carcinoma	*	*	*	*	*		*		
mixed epithelial tumor malignant			*						
undifferentiated ca. large cell			*						
granulosa cell tumor adult type			*		*				
cystic mature teratoma				*					
cystic mature teratoma with malignant transformation			*						
carcinoid insular type									*

Testdiagnosis: mucinous cystadenocarcinoma well differentiated

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenoca. well diff.		*		*					*
serous cystadenoca. mod. diff.				*					
serous adenofibroma malignant			*						
mucinous cystadenoma	*	*		*		*	*		
mucinous cystadenofibroma				*		*			
mucinous cystadenoma borderline	*	*	*	*	*	*	*	*	*
mucinous cystadenofibroma borderline				*		*			*
mucinous cystadenoca. mod. diff.	*		*	*	*	*	*		*
mucinous cystadenoca. poorly diff.				*					
endometrioid cystadenoma			*						
endometrioid cystadenoma borderline			*						
endometrioid cystadenoca. well diff.		*	*	*	*	*	*		*
endometrioid cystadenoca. mod. diff.			*	*			*		
homologous mixed Mullerian tumor			*						
clear cell tumor borderline									*
clear cell adenofibroma/carcinoma			*						*
clear cell carcinoma									*
mixed epithelial tumor borderline				*					
mixed epithelial tumor malignant				*					
granulosa cell tumor juvenile type			*						
Sertoli-Leydig cell tumor with heterologous elements		*		*	*				
sex cord tumor with annular tubules									*
endodermal sinus tumor			*						
immature teratoma				*					
cystic mature teratoma with malignant transformation				*					
other monodermal teratomas			*				*		

Testdiagnosis: cystic mature teratoma with mal. transformation

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
mucinous cystadenoma borderline				*					
mucinous cystadenofibroma borderline				*					
mucinous cystadenoca. well diff.				*					
mucinous cystadenoca. mod. diff.				*					
mucinous cystadenoca. poorly diff.			*	*					
endometrioid adenoca. well diff.				*			*		
endometrioid adenoca. mod. diff.				*			*		
endometrioid adenoca. poorly diff.				*			*		
endometrioid adenoacanthoma				*					
endometrioid adenofibroma malignant				*					
endometrioid stromal sarcoma				*					
homologous mixed Mullerian tumor			*	*		*			
heterologous mixed Mullerian tumor			*	*					
adenosarcoma					*				
clear cell carcinoma			*						
Brenner tumor borderline			*	*					
Brenner tumor malignant			*	*					
transitional cell carcinoma			*	*					
mixed epithelial tumor borderline				*					
mixed epithelial tumor malignant				*					
undifferentiated ca. large cell				*					*
undifferentiated ca. small cell			*	*					
Sertoli-Leydig cell tumor with heterologous elements				*					
immature teratoma	*	*	*	*	*	*	*		*
solid mature teratoma		*	*	*		*		*	*
cystic mature teratoma		*				*	*	*	*
struma ovarii									*
struma ovarii malignant									*
carcinoid insular type			*						*
carcinoid trabecular type			*						*
mixed germ cell tumors									*

Testdiagnosis: dysgerminoma

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenoca. poorly diff.						*			
mucinous cystadenoca. poorly diff.						*			
Endometrioid adenoca. poorly diff.			*			*			
Clear cell carcinoma	*		*	*	*	*	*		*
transitional cell carcinoma			*						
undifferentiated ca. large cell			*	*	*			*	
undifferentiated ca. small cell				*					*
unclassified epithelial tumor								*	
granulosa cell tumor juvenile type	*								
granulosa cell tumor adult type							*		
granulosa cell tumor poorly diff.							*		
Sertoli cell tumor with lipid storage					*				
Leydig cell tumor				*					
stromal luteoma				*					
luteinized thecoma				*					
lipid cell tumor			*	*					
endodermal sinus tumor	*	*	*	*	*		*		
embryonal carcinoma	*	*		*	*		*	*	*
choriocarcinoma				*					
immature teratoma			*						
cystic mature teratoma with malignant transformation			*						
carcinoid insular type		*		*					
carcinoid trabecular type				*					
mixed germ cell tumor		*			*				*

Testdiagnosis: serous cystadenoma borderline

	Pathologists								
Diagnosis	1	2	3	4	5	6	7	8	DEW
serous cystadenoma	*	*	*	*	*	*	*		*
serous cystadenofibroma			*	*		*			*
serous surface papilloma			*	*		*			*
serous surface papilloma borderline			*						
serous cystadenofibroma borderline									*
serous cystadenoca. well diff.	*	*	*	*	*	*	*	*	*
serous cystadenoca. mod. diff.				*					
serous cystadenoca. poorly diff.				*					
serous surface papillary carcinoma			*	*		*			
serous cystadenofibroma malignant			*	*		*			
mucinous cystadenoma						*			
mucinous cystadenofibroma						*			
mucinous cystadenoma borderline			*	*	*	*			
mucinous cystadenofibroma borderline			*	*		*			
endometrioid cystadenoma	*								
endometrioid cystadenofibroma						*			
endometrioid cystadenofibroma borderline				*	*				
endometrioid adenoca. well diff.	*		*						
clear cell tumor borderline		*	*	*					*
clear cell adenofibroma/carcinoma									*
clear cell carcinoma		*			*				*
mixed epithelial tumor benign							*		
mixed epithelial tumor borderline				*	*				
solid mature teratoma				*					
cystic mature teratoma				*					

Testdiagnosis: Sertoli cell tumor

Diagnosis	Pathologists 1	2	3	4	5	6	7	8	DEW
serous cystadenoca. well diff.									*
serous cystadenoca. mod. diff.									*
serous cystadenoca. poorly diff.									*
mucinous cystadenofibroma			*						
mucinous cystadenoma borderline						*			
mucinous cystadenofibroma borderline						*			
mucinous cystadenoca. well diff.				*		*			
mucinous cystadenoca. mod. diff.				*					
endometrioid cystadenoma						*			
endometrioid cystadenofibroma						*			
endometrioid cystadenofibroma borderline						*			
endometrioid adenoca. well diff.	*	*		*	*	*	*		*
endometrioid adenoca. mod. diff.		*		*	*				*
endometrioid adenoca. poorly diff.		*		*					*
homologous mixed Mullerian tumor				*					
heterologous mixed Mullerian tumor				*					
clear cell adenofibroma			*			*			
clear cell tumor bord.			*			*			
clear cell adenofibroma/carcinoma						*			
clear cell carcinoma						*			
adenomatoid tumor			*						
undifferentiated carcinoma								*	
granulosa cell tumor juvenile type			*			*	*		
granulosa cell tumor adult type		*	*	*		*	*	*	*
granulosa cell tumor poorly diff.				*		*	*		
Sertoli cell tumor with lipid storage			*	*			*		*
Sertoli-Leydig cell tumor well diff.	*		*	*	*		*		*
Leydig cell tumor								*	
Sertoli-Leydig cell tumor mod. diff.			*	*					*
Sertoli-Leydig cell tumor poorly diff.				*					
Sertoli-Leydig cell tumor with heterologous elements			*	*					*
unclassified sex cord stromal tumor				*		*			
sex cord tumor with annular tubules	*	*	*						
stromal luteoma			*						
luteinized thecoma	*								
dysgerminoma						*			
endodermal sinus tumor			*						
struma ovarii			*						
carcinoid insular type	*					*	*		*
carcinoid trabecular type				*	*				*

CHAPTER 4

Reasoning in uncertainties: an analysis of five strategies and their suitability in pathology.

Accepted for publication
Analytical and Quantitative Cytology and Histology

Astrid M. van Ginneken and Arnold W.M. Smeulders

ABSTRACT

In reasoning systems, uncertainty plays a crucial part, especially for those fields where judgements are essential, as in pathology. Uncertainty has several aspects, such as prevalence of diseases, occurrence of findings and the sensitivity and predictive value of findings.

For the functioning of a reasoning system two aspects are crucial: first, the internal representation of the uncertainty and second, the way in which the uncertainty is propagated in the reasoning process when combining formal statements.

Five well known reasoning strategies are compared: probability theory, MYCIN's certainty factor model, fuzzy set theory, the theory of Dempster-Shafer and Pathfinder's scoring mechanism.

The comparison addresses, among others, the following questions:

- Under what conditions will the model function? In particular, what information is to be specified a priori to the system?
- Can the different aspects of uncertainty be dealt with as separate entities?
- How are unknown uncertainties dealt with?
- How is evidence in favor of a hypothesis combined with evidence against it?
- How does the model treat the simultanuous occurrence of more than one disorder, that is, how does the model support reasoning with compound hypotheses?

It is preliminary concluded, that only in Pathfinder and probability theory, the different aspects of uncertainty are expressed as separate entities. Hence, the other models do not accurately represent uncertain knowledge. Also, theoretically attractive models such as Bayes, MYCIN and the theory of Dempster-Shafer can only function properly under the tight condition of mutual exclusiveness of hypotheses, not always suited for broader parts of pathology. They may, however, be suited for smaller parts with a limited number of defined diseases and a limited number of features. All models but Bayes lack a predictable performance as there is no or only a partial underlying theory to guarantee minimization of the overall error.

4.1 INTRODUCTION

In pathology as in many other medical fields, the increase of the body of medical knowledge makes it increasingly difficult for a general pathologist to master the whole field of the discipline at the present level of available knowledge. The general scope of our project is to make available in an efficient way reference knowledge, now less accessible in books and journals in the form of a computerized system.

Such a reference system consists of two major parts: a knowledge base and a reasoning system (inference engine). The knowledge base holds a collection of interpreted facts, i.e. facts with a meaning in the context of a specific subfield. The reasoning system provides directives as to which facts of knowledge to combine and how to combine them to arrive at an answer.

In this process uncertainties play a crucial part in several ways. First, there is uncertainty with respect to the presence or absence of a finding either alone or in combination or, in general, the reliability thereof. For example, nucleoli may be considered large by one expert and moderately enlarged by another expert. Second, there is uncertainty as how to interpret facts, i.e. uncertainty in the knowledge. Certainty about the meaning of thyroid tissue strongly depends on its localization. If normal appearing thyroid follicles are found in the thyroid, pathologists will usually believe, that this section is obtained from a normal thyroid, and certainty of the diagnosis benignancy is close to 1. However, if the same morphological appearance is found in the lung or in other places, certainty about the diagnosis "normal ectopic thyroid tissue" is less than 1. Third, there is the problem of consensus about the knowledge: the same criterion may be assigned different meanings by different experts.

Though the reliability of histopathological diagnoses is well accepted in general, the certainty of a concluded diagnosis is not always equal to one, due to uncertainty with respect to one or more of the aspects mentioned. Often, the surgical pathologist will be aware of the uncertainty in a particular case and will seek additional information to confirm or exclude a certain diagnosis. Each new observation may add to the certainty of a particular conclusion, provided the new observations are not too strongly correlated with the previous findings.

A reasoning system can only deal with observational uncertainty if such uncertainty is explicitly provided with the data on which it operates. The problem of consensus about the knowledge is outside the scope of a reasoning system. Here, the term uncertainty refers to uncertainty in the knowledge, i.e. the interpretation of findings.

This paper will focus on techniques dealing with uncertainty and the combination thereof in the reasoning part of an expert system. When knowledge is combined, in what way are uncertainties combined (the combinatorics) to reach the conclusion and in what way do they affect the conclusion?

Strategies for the combinatorics of uncertainties have been studied in a number of papers. We will discuss the following five most well-known combination techniques to gain insight in their suitability for pathology:

- Bayes theorem (probability reasoning)
- Certainty factor model of MYCIN (rule-based reasoning)
- Theory of Dempster-Shafer (reasoning with hypothesis sets)
- Fuzzy logic (reasoning with hierarchical hypothesis groups)
- Pathfinder (heuristic, hypothesis directed reasoning)

Therefore, an analysis of each technique is given first, including the conditions for application, the way uncertainties are expressed, the combinatorics and the required data for application. Then, in section 3, the following comparisons are addressed:

- To what extent is the expression of uncertainties in these models similar?
- How are unknown uncertainties dealt with?
- How is evidence in favor of a hypothesis combined with evidence against that hypothesis?
- Is it possible to deal with uncertainties assigned to hypothesis sets? I.e. is it possible to zoom in from a large hypothesis set to a smaller one as more information becomes available?
- What part do uncertainties play in the selection of questions
- What are the conditions for application?

4.2 CHARACTERISTICS OF THE REASONING STRATEGIES

4.2.1 Bayes theorem

Given some piece of evidence e, and a set of disease hypotheses, Bayes' rule [1] will render the most likely hypothesis,

given the evidence. This calculation requires the availability of several data, represented by several symbols. $P(D_i)$ is the a priori probability of disease D_i. These probabilities must be known all and sum to 1. The cumulative probability of disease hypotheses, for which the a priori values are not known, is equally divided among them. $P(e|D_i)$ is the conditional probability of evidence e, given diagnosis D_i and $P(e)$ is the a priori probability of evidence e. $P(e)$ is equal to $\Sigma_i P(e|D_i) * P(D_i)$. Application of Bayes' rule gives $P(D_i|e)$, which is the conditional probability of diagnosis D_i given evidence e and is given by:

$$P(D_i|e) = \frac{P(e|D_i) * P(D_i)}{\Sigma_i \; P(e|D_j) * P(D_j)}$$

The most important assumption underlying the use of Bayes' rule is that all disease hypotheses must be mutually exclusive and exhaustive. This implies that exactly one of the diseases of the hypothesis is assumed to be present in the patient. If this condition is not satisfied, the calculated probability values will deviate from the correct value.

This is explained as follows: Let D_1 and D_2 be two disease hypotheses, whose intersection is not empty, reflected by the fact that $P(D_1) + P(D_2) > 1$. The correct way to deal with this situation is to define a third hypothesis $D_1 \cap D_2$. This new hypothesis functions as a new disease, namely the "combination of D_1 and D_2", which is then treated in the same way as other diseases. We now have three mutually exclusive hypotheses:

$$D_1' = D_1 \backslash D_1 \cap D_2$$
$$D_2' = D_2 \backslash D_1 \cap D_2$$
$$D_1 \cap D_2$$

Let the corresponding conditional probabilities of evidence e for a given hypothesis be as follows:

$$\begin{aligned} P(e|D_1') &= P(e|D_1) \\ P(e|D_2') &= P(e|D_2) \\ P(e|D_1 \cap D_2) &= P(e|D_1) + (1 - P(e|D_1)) * P(e|D_2) \end{aligned}$$

Applying Bayes' rule to each of the five hypotheses (D_1,D_2,D_1',D_2', $D_1 \cap D_2$), gives the following quantitative relationship for the conditional probabilities of the hypothesis, given evidence e:

$$P(D_1'|e) < P(D_1|e) < P(D_1' \cup (D_1 \cap D_2)|e)$$

An illustration is given in Figure 1. In Appendix A an example of this explanation is given in the setting of pathology. As all examples in this paper, it does not reflect reality, but is constructed to illustrate a particular topic. The condition for disease hypotheses to be mutually exclusive implies that all possible combinations of two or more diseases are introduced as separate (thus "new") disease hypotheses.

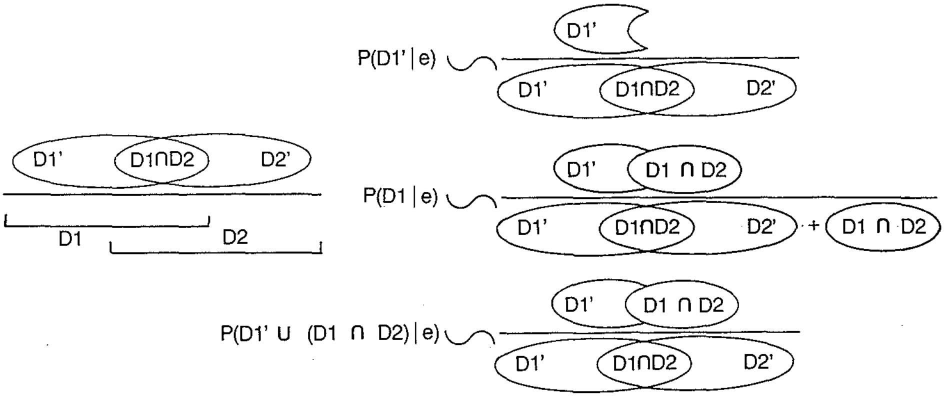


Figure 1. Diagram, showing the quantitative relation between $P(D_1'|e)$, $P(D_1|e)$ and $P(D_1' \cup (D_1 \cap D_2)|e)$.

The affect of several pieces of evidence on the probability of D_i is computed by sequential application of Bayes' rule. A new piece of evidence e is combined with already known evidence E as follows:

$$P(D_i|e \cup E) = \frac{P(e|D_i \wedge E) * P(D_i|E)}{\Sigma_j \; P(e|D_j \wedge E) * P(D_j|E)}$$

When pieces of evidence are dependent or their independence is unknown, probabilities of the form $P(e|D_i)$ are to be known for every subset of e. This requires a huge amount of probabilistic data. For this reason it is desirable to work with pieces of evidence, which are independent of each other. However, the mutual exclusiveness of diseases and the independence of findings is often not explicitly known and difficult to determine. In probabilistic terms the independence of evidence means:

$$P(e \mid D_i \wedge E) = P(e \mid D_i)$$
$$P(E \wedge e \mid D_i) = P(e \mid D_i) * P(E \mid D_i)$$

How additional evidence is combined can be shown using the correct example of Appendix A.

Given the fact, that the findings ripening follicles and lymphocytes are independent of another, the conditional probability of a teratoma given these two findings can be computed using the following formula:

$$P(D_1' \mid e_1 \wedge e_2) = \frac{P(e_1 \mid D_1') * P(e_2 \mid D_1') * P(D_1')}{\Sigma_i \; P(e_1 \mid D_i) * P(e_2 \mid D_i) * P(D_i)}$$

where:

D_1'	= teratoma
D_2'	= inflammation
$D_1 \cap D_2$	= teratoma and inflammation
e_1	= ripening follicles
e_2	= lymphocytes

Using the specified values, we get:

P(teratoma | ripening follicles $\wedge$ lymphocytes) = 0.05

The mutual independence of both evidence and diseases leads to a substantial reduction of probabilistic data as $P(e \mid D_i)$ has to be known for single pieces of evidence only. But even these data are scarce and difficult to acquire reliably. Given the amount and required precision of the necessary data, the approach of Bayes might be appropriate when a small part of the pathology spectrum is considered.

4.2.2 Certainty factor model of MYCIN

The well-known MYCIN [2] system consists of rules, basically of the form : if A then B. These rules are used in the reasoning system to (partially) confirm or disconfirm a hypothesis, given a set of evidential data.

To deal with uncertainties MYCIN uses certainty factors (CF's), which are assigned to the if-part of a rule to represent the belief in the findings as well as to the rule as a whole. In the latter case the CF represents the reliability of the knowledge the rule contains. The reasoning system does not make a distinction between these two CF values. A CF lies in the interval [-1,1], where 1 represents total belief and -1 total disbelief. When the reasoning system combines CF values, each CF value acts as a belief update, i.e. the CF value decreases or increases the belief of a (preliminary) conclusion. According to the new formulas of Heckerman [3] CF's can be expressed in terms of probabilities as follows:

$$CF(D|e) = \begin{cases} \dfrac{P(D|e) - P(D)}{P(D|e) * (1 - P(D))} & \text{if } P(D|e) \geq P(D) \\[2ex] \dfrac{P(D|e) - P(D)}{P(D) * (1 - P(D|e))} & \text{if } P(D|e) < P(D) \end{cases}$$

CF(D | e) represents the measure of belief in hypothesis D, given evidence e. Note, that $CF(D|e) = - CF(\neg D|e)$, expressing that evidence confirming D will act as disconfirmation of its negation.

Two conditions have to be met to apply the CF model. Firstly, the hypotheses has to be mutually exclusive and exhaustive and secondly, the pieces of evidence have to be independent of each other. The latter means, that the degree to which each piece of evidence influences the measure of belief of a hypothesis does not depend on the presence of previously acquired evidence: $CF(D|e) = CF(D|e \wedge E)$. The necessity of the first condition is explained in [3].

The following example gives an illustration in pathology. Use of hypotheses, which are not mutually exclusive leads to infringement of the requirement, that CF values are independent of previously acquired evidence. The example is not meant to reflect reality, but is constructed to offer insight. It is given in terms of probabilities, since the requirement implies that $P(e|D) = P(e|D \wedge E)$. Given are three diagnoses and two findings:

D_1 = teratoma
D_2 = inflammation
D_3 = post-menopausal ovary
e_1 = ripening follicles
e_2 = normal progesteron levels

Furthermore we have:

$P(D_1) = 0.30$
$P(D_2) = 0.60$
$P(D_3) = 0.40$
$P(e_1 | D_1) = 0.70$
$P(e_2 | D_1) = 0.80$
$P(e_1 | D_2) = 0.60$
$P(e_2 | D_2) = 0.80$
$P(e_1 | D_3) = 0.01$
$P(e_2 | D_3) = 0.05$

Given is the hypothesis "no inflamed ovary". From this it can be computed that:
P(normal progesteron levels | no inflammation) = 0.37

If ripening follicles happen to be observed in the ovary previously, we have:
P(normal progesteron levels | no inflammation) = 0.80

The difference between these two CF values is due to the fact, that the presence of ripening follicles reduces the hypothesis "no inflammation" to the hypothesis "teratoma". In fact, the difference results from the fact, that "no inflammation" is a mixture of hypotheses. It is thus illustrated, why hypotheses have to be mutually exclusive.

Therefore each hypothesis has to be uniquely defined, which implies, that the negation of each hypothesis has to be defined as a separate hypothesis. The negation of a hypothesis is in fact a set containing all hypotheses minus one and it might be difficult to assign a CF value to such a hypothesis. When CF values are not known, they are assigned, according to the rule $CF(D | e) = - CF(\neg D, e)$.

The combinatorics for CF's distinguish parallel combination from sequential combination [4]. The firs combination applies to the situation, where two rules have the same "then" part. The latter is used when the "then" part of a rule acts as the conditional part of an other rule. The formula for parallel combination is as follows:

$$CF(D|e_1 \wedge e_2) = \begin{cases} CF(D|e_1)+CF(D|e_2) - CF(D|e_1) * CF(D|e_2) & \text{if both CF's} \geq 0 \\[2ex] \dfrac{CF(D|e_1)+CF(D|e_2)}{1 - \min(|CF(D|e_1)|,|CF(D|e_2)|)} & \text{if one of both CF's} < 0 \\[2ex] CF(D|e_1)+CF(D|e_2) + CF(D|e_1) * CF(D|e_2) & \text{if both CF's} < 0 \end{cases}$$

In the formula for sequential combination, e_1 acts as evidence for e_2 and e_2 on its turn for D:

$$CF(D,e_1) \begin{array}{l} = CF(e_2,e_1) * CF(D,e_2) \quad \text{if both CF's} \geq 0 \\[1ex] =- CF(e_2,e_1) * CF(D,\neg e_2) \quad \text{if } CF(e_2,e_1) < 0 \end{array}$$

4.2.3 Dempster - Shafer theory

Dempster - Shafer's theory [5] is pre-eminently apt for working with sets of hypotheses, e.g. several diseases. Starting from the hypothesis set of all diseases under consideration, the acquisition of more patient data may lead to a reduction in the size of this initial hypothesis set. The complete set of mutual exclusive and exhaustive diseases will be called D_{tot}. In Dempster-Shafer's theory each subset of Dtot constitutes a separate hypothesis.

Based on a piece of evidence e, each subset Di is assigned a measure of belief in the form of a BPA (= basic probability assignment). This BPA is in the interval [0,1] and all BPA values sum to 1. The distribution of belief, based on e, over each subset D_i is performed by a function m_e, which fulfills the following conditions:

$$\begin{array}{lll}
m_e(\{\varnothing\}) & = 0 & \\
m_e(D_i) & = 0 & \text{when e does not support } D_i \\
m_e(D_j) & \in \ <0,1> & \text{when e supports } D_j \\
m_e(D_{tot}) & = 1 - \ R_i m_e(D_i) & \text{for all } D_i, \text{ supported by e}
\end{array}$$

As a consequence, $m_e(D_i) = 1$ if there is only one hypothesis D_i, for which e is sufficient evidence. When e does not support D_i then $m_e(D_i)$ will not default to a value, reflecting the a priori likelihood of D_i.

When there is evidence against hypothesis set D_i, a BPA is assigned to the negation ($\neg D_i$) of D_i. The quantity of belief, which cannot be explicitly assigned to any other subset D_j is represented by $1 - \Sigma_i m_e(D_i)$ for all $i \neq j$. As D_{tot} holds all hypotheses, it is evident that the hypothesis set accounting for this remaining belief is part of D_{tot} and therefore entails belief in D_{tot}. Consequently, if no evidence is available $m_e(D_{tot}) = 1$ and $m_e(D_i) = 0$ for all i.

A BPA is explicitly assigned to one subset D_i from D_{tot}. It does not result in the assignment of BPA values to its supersets, nor does it include BPA values given to its subsets.

Since belief in a subset Di entails belief in all subsets of Dtot containing D_i, the total amount of belief for D_i, based on evidence e, is represented by the Belief function, $B_e(D_i)$, as follows:

$$B_e(D_i) = m_e(D_1) + \ldots + m_e(D_k) \text{ where } D_j \subset D_i \text{ for } j = 1 \,..\, k$$

Unlike a function m_e, $B_e(D_i)$ sums all portions of belief committed to the subsets of D_i. Thus $m_e(D_i)$ can be 0 where $B_e(D_i) <> 0$. As a consequence $B(D_{tot}|e) = 1$. Usually $B_e(D_i) + B_e(\neg D_i) < 1$. The information contained in the Belief function for a given subset D_i is expressed by the belief interval for D_i: $[B_e(D_i), 1 - B_e(\neg D_i)]$ Note that $1 - B_e(D_i) - B_e(\neg D_i)$ represents all belief, committed to the subsets of D_{tot}, which intersect D_i, but are no subsets of D_i.

Again, also for the theory of Dempster-Shafer, all hypotheses in D_{tot} have to be mutually exclusive and exhaustive. If they are not exhaustive, it is not correct to assign remaining belief to D_{tot}. When the hypotheses are not mutually exclusive, we have a problem similar to that in the CF - model of MYCIN.

When two pieces of evidence (e_1 and e_2) support hypothesis set D with corresponding functions m_{e1} and m_{e2}, the combined function m_{e1e2} is defined as follows:

$$m_{e1e2}(D) = \Sigma_{ij}(m_{e1}(D_i) * m_{e2}(D_j)) \text{ for i and j where } D_i \cap D_j = D$$

Combinations of more than two pieces of evidence are dealt with by applying the rule repeatedly. Like every m function m_{e1e2} has to meet the conditions, that it sums to 1 over all subsets of D_{tot} and assigns the value 0 to the empty set. These conditions are not met when there are subsets D_i and D_j for which $m_{e1}(D_i) > 0$, $m_{e2}(D_j) > 0$ and the intersection of D_i and D_j is the empty set. This problem is solved as follows:

$$m_{e1e2}(\{\emptyset\}) = 0$$

$$m_{e1e2}(D) = \frac{\Sigma_{ij}(m_{e1}(D_i) * m_{e2}(D_j)) \quad \text{for i,j} \quad \text{where } D_i \cap D_j = D}{\Sigma_{kl}(m_{e1}(D_k) * m_{e2}(D_l)) \quad \text{for k,l} \quad \text{where } D_k \cap D_l \neq \{\emptyset\}}$$

In Appendix B an example is given to illustrate how this combinatoric is applied in practice.

It is important for the preservation of information, that BPA values are assigned to hypothesis sets as small as possible. Belief in a set entails belief in its supersets, which comes to expression in the values of the Belief function. The more refinement in BPA assignments the smaller are the corresponding Belief intervals.

In the following example m_{e1} and m'_{e1} both are based on evidence e_1 and m_{e2} is based on evidence e_2. Given are the following values for m_{e1}, m'_{e1} and m_{e2}:

$m_{e1}(D_1)$	$= 0.5$	$m'_{e1}(D_1 \cup D_2)$	$= 0.7$
$m_{e1}(D_2)$	$= 0.2$	$m'_{e1}(D_{tot})$	$= 0.3$
$m_{e1}(D_{tot})$	$= 0.3$		

$m_{e2}(D_1)$	$= 0.4$
$m_{e2}(D_2)$	$= 0.3$
$m_{e2}(D_{tot})$	$= 0.3$

When m_{e1} and m'_{e1} are combined with function m_{e2} then m_{e1e2} differs from m'_{e1e2} in that $m_{e1e2}(D_1) > m'_{e1e2}(D_1)$. This is due to the fact, that m'_{e1} leaves the combined BPA values for D_1 and D_2 unchanged when it is combined with m_{e2}. The additional information with respect to D_1 and D_2, explicitly present in m_{e1} is not used. In m_{e1e2} all BPA values are explicitly assigned to D_1, D_2 and D_{tot}, whereas in m'_{e1e2} there are some BPA values assigned to $D_1 \cup D_2$. These latter values do not add to the total belief in D_1 and D_2 respectively, resulting in a relatively larger Belief interval. This way it is illustrated how loss of explicit information affects the reliability of a concluded hypothesis.

As the intersections of many subsets are not empty, it will however be difficult for a pathologist to make a proper selection of the hypotheses to assign belief to.

Intuitively, a BPA resembles probability. The difference however is, that BPA values can be assigned to hypothesis sets instead of only to singleton hypotheses as is the case in a probability density function. Furthermore, probabilities, which cannot be explicitly assigned are not divided among the remaining possibilities, but committed to the complete set D_{tot}.

The amount of data needed to work with the theory of Dempster -Shafer is equal to the number of subsets of Dtot, multiplied by the number of pieces of relevant evidence. It might be difficult to have all BPA values sum to 1, especially when a piece of evidence supports many hypotheses.

4.2.4 Fuzzy sets

Fuzzy set theory [6] is developed as an extension to statistical pattern recognition to classify a patient state, described by features, to a class (diagnosis group) when the class definitions are not exact. As disease definitions are sometimes intrinsically vague and not well separated, the theory is worth to be considered here. By a hierarchical ordering of all diagnoses in groups of diagnoses, intermediate classes are defined to which the patient state can be assigned first. Each diagnosis(group) serves as a separate disease hypothesis. With the availability of new observational evidence, the patient state can, directly or via subsequent intermediate classes, be assigned to one diagnosis.

The reliability of the assignment of a patient state to a diagnosis(group) is expressed in a DOM (degree of membership) value for each relevant combination of a piece of evidence and a diagnosis(group). A DOM value lies in the interval [0,1], where 0 represents total belief in non-membership, whereas 1 denotes total belief in membership.

The disease hypotheses need not necessarily be mutually exclusive and DOM values assigned to them do not have to sum to 1. This is attractive [7] because a patient state can be made member of more than one diagnosis(group), even with high DOM values for each of them. For example, a patient can present with signs corresponding to left ventricular hypertrophy. When medical inquiries or the clinical discourse make more data available, the DOM values for members of this group of diseases can increase until a more specific disease hypothesis gains preference over left ventricular hypertrophy.

The combinatoric for DOM values is simple: The combined DOM for a hypothesis is equal to the smallest DOM value for that hypothesis:

$$\mathrm{DOM}(D \mid e_1 \wedge \ldots \wedge e_k) = \min \{\mathrm{DOM}(D \mid e_1), \ \ldots, \mathrm{DOM}(D \mid e_k)\}$$

Each time when DOM values are combined, no default values are used; only available evidence affects the combined DOM value.

The application of fuzzy theory requires an amount of data equal to all relevant combinations of a piece of evidence and a disease hypothesis. Since DOM values do not have to sum to 1, they are no real probabilities and therefore cannot be interpreted as such. Their values have to be approximated by experts. Tests are necessary to evaluate the correctness of such approximations and may aid in the development of a mapping from uncertainties in reality to DOM values in the fuzzy set model.

For its applicability in medicine it should be noted that this combinatoric seems not to allow for the expression of the predictive value of available evidence with respect to a particular disease hypothesis. Take for example the following DOM values for disease hypothesis D, based on e_1 and e_2 respectively as available evidence:

$$\mathrm{DOM}(D \mid e_1) = 0.7 \qquad \mathrm{DOM}(D \mid e_2) = 0.9$$

The combined DOM value for D will be 0.7. However, DOM(D | e_2) is larger than DOM(D | e_1) supposedly because e_2 constitutes evidence with a higher predictive value for D than does e_1. If one, nevertheless, takes DOM(D | e_1) as the new combined DOM value, the absence of e_1 is implicitly regarded as evidence against hypothesis D. This will not always be so. Therefore it is not correct to take the smallest DOM value as a rule.

The following example will show how the combinatoric is applied and how this may lead to a wrong conclusion. Assume, that the following two diseases are under consideration together with four available findings:

D_1 = teratoma
D_2 = inflammation
e_1 = ripening follicles
e_2 = lymphocytes
e_3 = hairs
e_4 = abdominal swelling

The DOM values assigned to the combinations of these findings and the diseases are as follows:

$\mathrm{DOM}(D_1 | e_1) = 0.6$
$\mathrm{DOM}(D^1 | e_2) = 0.5$
$\mathrm{DOM}(D_1 | e_3) = 0.8$
$\mathrm{DOM}(D_1 | e_4) = 0.8$

$\mathrm{DOM}(D_2 | e_1) = 0.5$
$\mathrm{DOM}(D_2 | e_2) = 0.9$
$\mathrm{DOM}(D_2 | e_3) = 0.2$
$\mathrm{DOM}(D_2 | e_4) = 0.5$

The resulting DOM values will be 0.5 for teratoma and 0.2 for inflammation. In case of inflammation, it is not unreasonable that the presence of hairs has much influence on the combined DOM value since this finding disfavors the diagnosis inflammation. However it is open to discussion whether the influence of a specific finding, such as the presence of lymphocytes may be neglected, as is the case here. As to the diagnosis teratoma, the situation is different. A DOM value of 0.5 with respect to the finding lymphocytes reflects the low predictive value of this finding. The presence of hairs and abdominal swelling strongly favors the diagnosis teratoma since hairs constitute a finding with a high predictive value. Here, it of questioning have been exhausted. In the last case Pathfinder will list the remaining disease hypotheses and show why none of them could be concluded. To this respect the system shows which discriminating information could not be obtained. How disease profiles are used in the scoring algorithm is illustrated in the example of Appendix C.

A few notes concerning the reasoning system of Pathfinder are made. First, causality is not represented adequately in the knowledge base [9]. In the disease profile factors predisposing for a disease and findings resulting from a disease are not treated differently. Assume, that A predisposes for disease D, then D is accepted as an explanation for A. The system will take no further action to find a cause for A, which is not correct. Second, the system does not take into account the interdependency of manifestations [9]. A disease-profile may very well contain manifestations A and B such that A implicates B. Disease hypotheses, which have these manifestations in their profile will receive scores, which are too high. Two findings, which are dependent of each other, should not add as much to the score as they would when independent.

Though these first two points are mentioned to make the insight in the reasoning strategy of Pathfinder more complete, they are more relevant for the functioning of Internist than of Pathfinder [8]. In pathology, a histological slide may show more than one abnormality, but they do not necessarily have a causal relation. Their cause can be one or more underlying diseases. As long as the histologic abnormalities are spatially separated in the section, each abnormal area generally poses a diagnostic problem on its own. Mixed histologic disorders may occur such as tumor and inflammation, but in general they are not a frequently occurring phenomenon in pathology.

Thirdly, evidence against a previously concluded diagnosis will not result in the removal of that diagnosis from the list of concluded diagnoses. Finally, we find it illogical that in each cycle, manifestations, which remain unexplained within one problem area, are set aside for additional cycles. Then, they can only play a part in the conclusion of lower ranked diagnoses. Whatever score a disease-hypothesis receives, it can never become higher in rank than diagnoses, which have already been concluded and this may be undesirable.

As to the final point it can be explained how the strategy can lead to wrong results. Manifestation A is on the disease-profile of hypothesis D_1 and D_1 happens to become the first concluded diagnosis. Subsequently, A is removed from the manifestation list. Suppose that A is full evidence against disease-hypothesis D_2. Since

A is no longer present on the manifestation list, D_2 might very well be concluded as the second diagnosis, based upon the remaining manifestations in its disease-profile. An example of this follows.

> We take the previous example (Appendix C), where the first concluded diagnosis was teratoma. Assume, that additional evidence becomes available about the patient: low progesteron levels, amenorrhoea and atrophy of the ovaries. All these findings support the diagnosis post-menopausal ovary and it is very well possible, that the reasoning system will conclude post-menopause as the second 'diagnosis'. However, this conclusion would not be correct as the presence of ripening follicles makes post-menopausal period very unlikely. Here, the removal of this finding has lead to an incorrect conclusion.

It seems a better strategy to remove a concluded diagnosis and competitors from the Master-DD list as opposed to the removal of explained manifestations from the manifestation list. With a strategy as proposed, hypotheses having findings in common with a concluded diagnosis without being a competitor of that diagnosis, have a more realistic chance to become the following concluded diagnosis. Such hypotheses do no longer depend for their score on a subset of the manifestations in their profile. In addition no manifestation will incorrectly become part of the category: unknown but present in the disease-profile. However, findings that have already been explained by one or more concluded diagnoses should be marked as such and their importance should be reduced. Present as such, they can still act as counter-evidence, at the same time causing no extra action by the system to explain them.

4.3 ANALYSIS

4.3.1 What aspects of uncertainty are expressed?

We will consider two important aspects of uncertainty:

- The predictive value of a finding: the probability of a diagnosis given the finding
- The sensitivity of a finding: the probability of the finding given the diagnosis

In probability theory, $P(D \mid e)$ expresses the predictive value, whereas $P(e \mid D)$ expresses the sensitivity of a finding. Evidence in favor of a hypothesis is expressed by the fact evidence with respect to D is unknown. Belief which cannot explicitly be assigned to one or more hypotheses is assigned to the complete set D_{tot}. Consequently, unknown uncertainties do not influence the value of B(D). In fuzzy set theory and Pathfinder, unknown uncertainties receive no default values as in the previously discussed models. As DOM values do not have to sum to a particular value, it is not necessary to assign default values. Therefore, existing DOM values are not influenced. In Pathfinder, unexplained manifestations add negatively to the score of that hypothesis, depending on their importance. Though importance cannot be regarded as an expression of uncertainty, it influences the effect of evoking strengths and frequencies by decreasing the hypothesis score.

It can be concluded, that unknown uncertainties, wether or not assigned default values, do not update existing belief in a hypothesis except for Pathfinder. In Pathfinder, unexplained findings have a negative effect on the hypothesis score, i.e. reduce the chance for a hypothesis to become a concluded diagnosis.

4.3.3 How is evidence in favor of a hypothesis combined with evidence against that hypothesis?

With respect to this topic two aspects are important:

- Does the presence of pieces of evidence with an opposite effect on the certainty of a diagnosis remain visible throughout the combination process?
- Is the net effect of the combination of pieces of evidence with an opposite effect on the certainty of a diagnosis proportionate to their weight?

The first aspect is important, since it makes a difference wether a diagnosis is concluded on positive evidence alone or the combination of both positive and negative evidence. The second aspect reflects the desire that all pieces of evidence contribute to the certainty of a diagnosis proportionally. When e_1 is evidence in favor of hypothesis D and e_2 pleads against it (with e_1 and e_2 independent), then Bayes rule gives:

$$P(D \mid e_1 \wedge e_2) = \frac{P(e_1 \mid D) * P(e_2 \mid D)}{P(e_1) * P(e_2)} * P(D)$$

This quotient determines wether $P(D|e_1 \wedge e_2)$ will increase or decrease. The new probability of D does not reveal separately the positive and negative influence of evidence, leading to the change of P(D). When the quotient exceeds the value 1, the combined evidence of e_1 and e_2 supports D, though e_1 and e_2 do not both support D. An equal amount of evidence in favor of a hypothesis D and against it does not differ from the absence of evidence since in both cases the existing belief in D remains unchanged.

In the CF-model evidence in favor of a hypothesis is represented by a positive CF value as opposed to a negative CF value for evidence against a hypothesis. When CF's with opposite signs are combined, the CF with the largest absolute value will not only determine the sign of the new combined CF, but is also proportionally slightly favored over the other CF value since the denominator of the combination formula is less than 1. Here again, information with respect to the presence of a combination of positive and negative evidence is lost. As in Bayes, the presence of an equal amount of positive and negative evidence is indistinguishable from the absence of knowledge as the CF value in both cases is 0.

In the strategy of Dempster-Shafer a hypothesis can only receive BPA values based upon evidence in favor of that hypothesis. If evidence against a hypothesis becomes available, only its negation will receive a BPA, but not D. In Dempster-Shafer's strategy positive and negative evidence remain recognizable in the belief values and are very well distinguished from the absence of evidence. In addition it is obvious that each piece of evidence adds to D or the negation of D proportionate to its weight.

In fuzzy logic DOM values assigned to evidence in favor of a hypothesis range from 0 to 0.5, whereas DOM values corresponding to positive evidence range from 0.5 to 1. The combination rule for DOM values prescribes that the combined DOM value is equal to the lowest or highest value of the composite DOM values. Assume, that the lowest DOM value is decisive. The correctness of this combination rule is questionable when positive evidence is absent for a hypothesis as it implicitly acts against that hypothesis. However, this rule has another consequence: a low DOM value, resulting from negative evidence overrules all evidence present with higher DOM values. Evidence in favor of and against a hypothesis is not at all treated proportionally, nor does it remain recognizable in the resulting DOM value wether the evidence is positive, negative or a combination of both.

In Pathfinder, positive manifestations add positively to the hypothesis score. Negative manifestations, i.e. evidence against the hypothesis add negatively. A decrease of the hypothesis score may also stem from unexplained manifestations, depending on their importance. Thus, the reasoning strategy of Pathfinder not only attempts to find the most likely diagnosis, but takes into account the consequence of a conclusion as well. However, positive manifestations are favored, since the increase in score due to positive manifestations is larger than the decrease resulting from similar negative manifestations. The resulting hypothesis score does not reveal wether positive and negative evidence is present in combination.

It can be concluded, that the models of Bayes, MYCIN and Dempster-Shafer treat positive and negative evidence proportionally, but only the latter separates the influence of positive and negative evidence. The fuzzy set model and Pathfinder do not treat evidence proportionally and do not separate evidence in favor of and evidence against a diagnosis. In Pathfinder hypothesis selection is not only based on evidence, but influenced by clinical consequences as well.

4.3.4 Is it possible to reason with uncertainties assigned to hypothesis sets?

In pathology the pathologist usually focusses on a decreasing number of diagnoses as more patient data become available. In an initial stage, when little is known about a patient, it is attractive when a reasoning system can establish which disease or combination of diseases is most likely. Each time when new evidence is gathered the hypothesis set should become smaller.

Using Bayes' rule or the CF model, only one disease is assumed to be present. When little information about the patient is available, many probabilities and CF's receive default values respectively. Then, the reliability of a conclusion is poor and as more data become available the concluded disease may change many times before the system concentrates on one disease hypothesis. A pathologist will not easily understand such a jumpy behavior, which also obscures the line of questioning.

When using strategies like Dempster-Shafer, fuzzy logic and to some extent Pathfinder, it is possible to zoom in from the conclusion of a large hypothesis set based upon little data to a smaller and more specific set of hypotheses as more data becomes available. Hypothesis sets, serving as intermediate conclusions, contain diseases which are related on the basis of one ore more common properties. These hypothesis sets do not necessarily correspond to disease groups as found in pathological classification trees. As a consequence, additional patient data are asked on a general groups level first and a specific, detailed level later. A better understandable line of questioning is the result.

4.3.5 What part do uncertainties play in the selection of questions?

In discussing the topic of question selection our present emphasis will be on how uncertainties may play a part in the strategy of question selection. The knowledge base, upon which the reasoning system operates, contains the relations between findings and diagnoses. These relations enable the system to select questions, appropriate for a particular diagnosis or group of diagnoses. To assess the usefulness of a particular question, the predictive value of the possible answers, the likeliness of them to occur and their importance have to be considered. The more specific a finding is for a disease, the more important it is to acquire that evidence. It is obvious when two pieces of evidence have the same predictive value, that the one most likely to occur is the most attractive one to ask. The crucial issue is: which question is most likely to give maximum information gain?

In probability theory and in Pathfinder a finding is likely to occur when both the values for P(e) and frequency are high. P(e) can be computed from sensitivities and a priori probabilities or approximated by an expert. Sensitivity and predictive value are not necessarily related. A finding occurring in almost every patient with a certain diagnosis and not in patients without that diagnosis, has both a high sensitivity and predictive value for that diagnosis. If, on the other hand, the same finding is also often seen in patients without that diagnosis, than the finding still has a high sensitivity, but a low predictive value. In the CF - model, fuzzy sets and the theory of Dempster-Shafer, sensitivity is not expressed in their parameters for uncertainty. Question selection strategies, using these models depend on other sources for the information on frequency of findings. A well-known strategy for selecting the question from which maximum information gain may be expected is the use of the entropy formula. However, this formula requires Bayesian probabilities to be available.

It is important for a system to decide when it stops to ask questions; when no appropriate questions remain or when the remaining available questions have too little predictive value. A user may expect from the reasoning system to ask questions, which enable this system to make progress in its diagnosis selection. A question which will not bring about any change in the systems conclusion should not be asked. The CF model of MYCIN stops asking questions, when the remaining appropriate rules have CF's of 0.2 or less. Nothing is mentioned about negative CF values. Pathfinder will not ask questions with importance values of 2 or less. When the task of the reasoning system is to purely generate a list of possible diagnoses by descending certainty, importance values should only serve as a threshold when to stop question selection. For the remaining models, threshold values for stopping question selection can be defined.

Predictive value and to a lesser extent the occurrence of findings are important for question selection. All five models express some form of predictive value in their uncertainties, but only the models of Bayes and Pathfinder explicitly express sensitivities. When not present in a model, the subtlety of question selection can be improved by defining the thresholds for question selection and parameters to express the occurrence of findings.

4.3.6 What are the conditions for application?

Prior to mentioning which condition has to be met by which model, the conditions will be discussed.

The condition of mutually exclusive hypotheses implies, that only one diagnosis is assumed to be present. In reality, several diagnoses may be present in a patient. Therefore, the condition is difficult to realize since it requires that all possible combinations of diagnoses be defined as separate diagnoses. Though the condition might limit the applicability of models with this condition in many fields of medicine, this is less true for pathology. Multiple diseases play a less important part in pathology: abnormalities in patient tissue are generally spatially separated, each abnormal area posing its own diagnostic problem. As a consequence, application of a model, which has to meet the condition of mutually exclusive hypotheses, may require a limited number of "newly defined" mixture diagnoses, or leads to a minor violation of the condition.

The condition of independence of evidence implies, that the effect of a new piece of evidence on the certainty of any diagnosis is not influenced by the presence of previously acquired evidence. This condition is formulated in some models for the purpose of simplification and modularity. It is however difficult to satisfy. Firstly, there is not enough information or knowledge available to determine for each possible piece of evidence its dependency relation with any other piece of evidence. Secondly, even if it were possible to determine which pieces of evidence are independent of each other, it is still questionable wether enough evidence remains to differentiate between all separately defined diagnoses. It may seem convenient to have a model, which does not require independence of evidence. If, however, interdependence of evidence is completely ignored, some diagnoses may be disproportionately favored, due to the fact that each piece of evidence adds to the certainty as much as it would if it were independent of other evidence.

Bayes' rule, the CF-model of MYCIN and the theory of Dempster-Shafer all require that the hypotheses are mutually exclusive. Only the latter two require independence of evidence as well.

4.4 CONCLUSION

To gain insight in the suitability for pathology of the five strategies mentioned above, we now concentrate on those topics, which are important for the choice or design of a reasoning system and which differ considerably in the five strategies. The topics are: the condition of mutual exclusive hypotheses and independence of evidence, the explicit expression of aspects of uncertainty, the feasibility of data acquisition, reasoning with hypothesis sets and the suitability of the combinatorics.

The condition of mutual exclusive hypotheses, which has to be met by the models of Bayes, MYCIN and Dempster-Shafer, seems not to be a major disadvantage when a well defined part of the pathology spectrum is considered. The condition of independence of evidence cannot be easily met, but the omission of the condition is not desirable either. It is attractive when a model allows for the expression of interdependence of evidence. The latter is possible in Bayes, but of little practical use as long as the data are not available. It is interesting to study how the models of Bayes, MYCIN and Dempster-Shafer respond quantitatively to changes in the dependence of evidence.

The degree to which knowledge about uncertainties can be made explicit, depends on which aspects of uncertainty are expressed in uncertainty parameters. All five models allow the expression of predictive value, but only Bayes and Pathfinder can explicitly express sensitivity. Explicit knowledge about the predictive value of findings is important for an efficient reasoning and question selection strategy. Sensitivities or a priori occurrences of findings have to be expressed explicitly for the question selection strategy, but this is not necessarily so for the reasoning strategy (MYCIN, fuzzy sets and Dempster-Shafer). For the purpose of question selection it is a disadvantage for a system when its uncertainties do not explicitly express sensitivity or occurrence.

For the acquisition of uncertainties, statistical data are an important source of information, which can be used directly for the application of Bayes' rule or the computation of CF values, as well as indirectly for approximation of uncertainties in the remaining three models. Detailed statistical information is scarce in pathology especially the occurrence of findings, which implies that uncertainties have to be approximated by experts or be given default values. Approximations of uncertainties are difficult even by experts when rare diseases or combinations of diseases are concerned. It is questionable, whether an expert can reliably approximate the prevalence of diseases in the order of two or three per thousand. This difference may seem small, but in the reasoning system the difference is relative and thus 50 percent.

When experts approximate uncertainties, they have to make a proper mapping from uncertainty in reality to uncertainties as used in a reasoning system. Therefore, the expert must be able to form an idea of these uncertainties. The parameters of these five systems are of a different nature: BPA's and DOM values resemble probabilities, CF's represent belief updates and the parameters of Pathfinder correspond to probabilities, but are expressed on a small scale. Uncertainties are likely to be easier to approximate, when their expressions have more similarities with true probabilities. When no statistical data are available, it may be more difficult to approximate a CF value, than a DOM value. In addition, the more aspects of uncertainty are incorporated in one parameter, the more difficult it is to approximate the value for that parameter. Thus it is to be expected, that the parameters of Bayes and Pathfinder are easier to approximate than the parameters of the other three models, as they express different aspects of uncertainty in different parameters.

The strategies of Dempster-Shafer and fuzzy sets allow for reasoning with hypothesis sets as opposed to Bayes and MYCIN. An initial set of possible diseases may be reduced as more patient data become available. This is an attractive characteristic as it corresponds to the general diagnostic procedure. In Pathfinder, group discriminate mode allows reasoning along a tree of predefined diagnostic groups, but not with any subset of the complete set of diagnoses.

As to the applicability of the combinatorics in pathology, it seems a drawback of fuzzy set theory, when - for example - the lowest DOM value is decisive. Not only negative specific evidence, but also nonspecific evidence overrules positive specific evidence. This implies, that diseases tend to receive less belief than would be expected. Though less prominent than in fuzzy set theory, the models of Pathfinder and MYCIN do not combine pros and cons proportionate to their weight. In Pathfinder, findings with an unknown relation to a diagnosis are not indifferent to the certainty of that diagnosis. Though the models of Bayes, Dempster-Shafer and MYCIN lack these disadvantages, only application of Bayes' rule guarantees a conclusion with minimal overall error.

	BAYES	MYCIN	D|S	FUZZY	PATHF
Expression uncertainties					
sensitivity (or analogous concept)	+	-	-	-	+
predictive value (or analogous concept)	+	+	+	+	+
Unknown uncertainties					
indifference	+	+	+	+	-
Pro's and con's					
visibility	-	-	+	-	-
proportionate weighing	+	-	+	-	-
Sets	-	-	+	+	-
Conditions					
not necessarily mutually exclusive	-	-	-	+	+
evidence not necessarily independent	+	-	-	+	+

Table 3. An overview of the properties of the five reasoning systems as discussed in the analysis.

It can be concluded, that an important disadvantage of the CF -model and the theory of Dempster-Shafer lies in the restrictive condition of independence of evidence. These two models together with Bayes' rule require diagnoses to be mutually exclusive, a condition which might only be slightly violated when a small part of the pathology spectrum is considered. Only Bayes and Pathfinder explicitly express more than one aspect of uncertainty in their parameters, which serves not only in data acquisition but in question selection as well. The strategies of Dempster-Shafer, fuzzy set theory and to a lesser extent Pathfinder allow reasoning with hypothesis sets, which is in conceptual accordance with the general diagnostic procedure. As to the handling of unknown data and pros and cons, the combinatorics of Bayes and Dempster-Shafer are slightly preferable over the model of MYCIN and Pathfinder, but largely over the fuzzy combinatorics. Provided the required information is known, Bayes has a completely predictable performance and guarantees a conclusion with minimal overall error. Table 3 gives a schematic overview of the topics as discussed in the analysis.

ACKNOWLEDGEMENTS

We thank Johan van der Lei for providing literature and useful discussions. Joop Kuik is acknowledged for increasing our insight in probability theory.

REFERENCES

[1] Shortliffe EH and Buchanan BG, A Model of Inexact Reasoning in Medicine, in: Buchanan BG and Shortliffe EH, Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison-Wesley, 1984, pp. 233-262.

[2] Buchanan BG and Shortliffe EH, Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison-Wesley, 1984.

[3] Heckerman DE, Probabilistic Interpretations For MYCIN's Certainty Factors, in: Kanal LN and Lemmer JF (eds.), Uncertainty in Artificial Intelligence, Elsevier Science Publishers, North Holland, 1986, pp. 167-195.

[4] Buchanan BG and Shortliffe EH, Uncertainty and Evidential Support, in: Buchanan BG and Shortliffe EH, Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison-Wesley, 1984, pp. 209-232.

[5] Gordon J and Shortliffe EH, The Dempster-Shafer theory of evidence in: Buchanan BG and Shortliffe EH, Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison-Wesley, 1984, pp. 272-292.

[6] Zvieli A and Chen PP, Entity-Relationship modeling and Fuzzy databases, Second Conference on data engineering, Los Angeles, 1986.

[7] Talmon JL, Pattren recognition of the ECG, Ph.D. dissertation, Department of Medical Informatics, Free University, Amsterdam, 1983.

[8] Horvitz EJ, Heckerman DE, Nathwani BN and Fagan LM, Diagnostic Strategies in the Hypothesis-Directed PATHFINDER System, in: Proceedings of The First Conference on Artificial Intelligence Application, IEEE Computer Society, 1984, pp. 630-636.

[9] Miller RA, Pople HE and Myers JD, INTERNIST-I, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine, N Eng J Med 307(8), 1982, pp. 468-476.

[10] Horvitz EJ, Heckerman DE, Nathwani BN and Fagan LM, The Use of a Heuristic Problem-Solving Hierarchy to Facilitate the Explanation of Hypothesis-Directed Reasoning, MEDINFO 86 Washington D.C., IFIP-IMIA, North Holland, 1986, pp. 27-31.

Appendix A

In general, the probability, that a patient has a particular disease has to be specified, denoted by P(disease) and also the probability of that disease given a finding, denoted by P(disease | finding). The example has two parts. In the first part, the hypotheses are not mutually exclusive. The second part is a correct version of the same example to show the difference between the wrong and the correct application. For the first part of the example, the following diseases and findings are given:

D_1 = teratoma
D_2 = inflammation
e_1 = ripening follicles
e_2 = lymphocytes

with the following probabilities:

$P(D_1) = 0.40$
$P(D_2) = 0.70$
$P(e_1 | D_1) = 0.70$
$P(e_1 | D_2) = 0.60$
$P(e_2 | D_1) = 0.10$
$P(e_2 | D_2) = 0.90$

Note, that the diseases teratoma and inflammation of the ovary are not mutually exclusive, which is expressed by the fact, that their a priori probabilities do not sum to 1. Now, the a priori probabilities of ripening follicles and lymphocytes, denoted by P(finding), can be computed using the formula:

$$P(e) = P(e_1 | D_1) * P(D_1) + P(e_1 | D_2) * P(D_2)$$

$P(e_1) = 0.70$
$P(e_2) = 0.67$

Application of Bayes' rule gives:

$P(D_1 | e_1) = 0.40$
$P(D_1 | e_2) = 0.06$
$P(D_2 | e_1) = 0.60$
$P(D_2 | e_2) = 0.94$

Here, we see, that the finding ripening follicles is a very non-specific finding with respect to the diagnoses teratoma and inflammation. Their a priori probabilities are not changed and a little changed respectively. On the contrary, the finding lymphocytes is very specific for inflammation, raising the probability of the diagnosis from 0.70 to 0.94. However, the example is not correct as the two diseases are not mutually exclusive. How this affects the probability values will become clear when the same example is given in a correct way. Note, that the combination of teratoma and inflammation has become a separate diagnosis. Now, we have three mutually exclusive diagnoses and the following values:

$P(D_1') = 0.30$
$P(D_2') = 0.60$
$P(D_1 \cap D_2) = 0.10$
$P(e_1 | D_1 \cap D2) = 0.88$
$P(e_2 | D_1 \cap D_2) = 0.91$
(weighted values according to the formula:

$$P(e|A \cap B) = P(e|A) + (1 - P(e|A)) * P(e|B))$$

The remaining values remain unchanged· When calculations are made, similar to those in the previous example, we get the following values:

$P(D_1'|e_1) = 0.32$
$P(D_1 \cap D_2|e_1) = 0.13$
$P(D_1' \cup (D_1 \cap D_2)|e_1) = 0.45$

These values serve to illustrate that:

$$P(D_1'|e_1) < P(D_1|e_1) < P(D_1' \cup (D_1 \cap D_2)|e_1)$$

In other words: if diseases are not mutually exclusive and the conditional probability has to be computed of a particular diasease given a piece of evidence, the resulting value deviates from the true value in the following way:

- It is too large when it intends to represent the conditional probability of the pure disease, given the evidence.
- It is too small when it intends to represent the conditional probability of the disease with and without combinations with other diseases, given the evidence.

Appendix B

The following example will show how the theory of Dempster-Shafer combines m functions. Given are the following disease hypotheses and findings:

D_1 = teratoma
D_2 = inflammation
D_3 = post-menopausal ovary

e_1 = ripening follicles
e_2 = lymphocytes

With respect to these findings, we have the following m functions:

$m_{e1}(D_1)$	$=0.60$	$m_{e2}(D_1)$	$=0.10$
$m_{e1}(D_2)$	$=0.40$	$m_{e2}(D_2 \cup D_3)$	$=0.80$
$m_{e1}(D_3)$	$=0.00$	$m_{e2}(D_{tot})$	$=0.10$

	m_{e2} (D_1) 0.10	m_{e2} $(D_2 \cup D_3)$ 0.80	m_{e2} (D_{tot}) 0.10
m_{e1} (D_1) 0.60	D_1 0.06	ø 0.48	D_1 0.06
m_{e1} (D_2) 0.40	ø 0.04	D_2 0.32	D_2 0.04
m_{e1} (D_3) 0.00	ø 0.00	D_3 0.00	D_3 0.00
m_{e1} (D_{tot}) 0.00	D_1 0.00	$D_2 \cup D_3$ 0.00	D_{tot} 0.00

Table 4. Table, showing how the values of m_{e1} and m_{e2} are combined to compute the combined function m_{e1e2}.

How m_{e1e2} is computed is shown in Table 4. Note, that $\Sigma_i m_{e1e2}(D_i) < 1$. According to the correction rule, all combined m values have to be divided by:

$$\Sigma_{kl} m_{e1}(D_k) \; * \; m_{e1}(D_l) \quad \text{where} \quad D_k \cap D_l \neq \{\emptyset\}$$

This value is 0.48. The correct function m_{e1e2} now has the following values:

$m_{e1e2}(D_1)$	$=0.25$	$m_{e1e2}(D_2 \cup D_3)$	$=0.00$
$m_{e1e2}(D_2)$	$=0.75$	$m_{e1e2}(D_{tot})$	$=0.00$
$m_{e1e2}(D_3)$	$=0.00$		

As a result:

$B(D_1)$	$=0.25$	$B(D_2 \cup D_3)$	$=0.75$
$B(D_2)$	$=0.75$	$B(D_{tot})$	$=1.00$
$B(D_3)$	$=0.00$		

The Belief interval of D_2 (inflammation) based on the given evidence is: [0.75 , 0.75]. This indicates, that no positive BPA value is assigned to hypotheses, which are no subset of D_2, but intersect D_2.

Appendix C

How Pathfinder's knowledge base is structured and how the reasoning system functions is described well in the literature. Here, a small example will be given to illustrate the concept of disease profiles and the scoring mechanism.

	teratoma			inflammation			post-menopausal ovary		
	F	E	I	F	E	I	F	E	I
ripening follicles	3	1	1	3	1	1			1
absent ripening follicles			3			3	5	4	3
lymphocytes	2	1	4	4	3	4			4
hairs	4	4	4			4			4
fever	2	1	3	4	3	3	1	1	3
low progesterone level			2			2	4	2	2
atrophy			2			2	4	3	2
amenorrhoea			3			3	4	3	3
abdominal swelling	3	2	4			4			4
ovary on X-ray	3	3	4			4			4

Table 5. An example of three disease profiles.

Table 5 shows several disease profiles, where the symbols "F","E" and "I" refer to "frequency", "evoking strength" and "importance" respectively. Assume, that the following findings are present:

- ripening follicles
- lymphocytes
- hairs
- abdominal swelling

	teratoma		inflammation		post-menopausal ovary	
ripening follicles	A	4	A	4	C	- 2
lymphocytes	A	4	A	20	C	−20
hairs	A	40	C	−20	C	−20
abdominal swelling	A	10	C	−20	C	−20
		58		−16		−62

Table 6. The computation of the scores of three disease hypotheses.

Based upon this evidence, the score can be computed as shown in Table 6. "A", "B" and "C" are the categories to which the findings belong. It is obvious, that teratoma is the topmost disease hypothesis, but it cannot yet be concluded as the diagnosis since the numerical distance between the scores of teratoma and inflammation is less than 90. Pathfinder would now use confirmation mode, resulting in the selection of questions, which are likely to lead to the acquisition of evidence in favor of the diagnosis teratoma. Such a question could be wether the ovary is visible on X-ray or not. It is likely, that the reasoning system will conclude the diagnosis teratoma. As a result all the explained findings are removed from the manifestation list. Then system will recycle with remaining unexplained findings.

CHAPTER 5

A Method for the Acquistion of Formalized Knowledge in Pathology

Submitted for publication

Astrid M. van Ginneken, Wicher Jansen, Arnold W.M. Smeulders,
Johan van der Lei and Jan P.A. Baak

ABSTRACT

A knowledge base structure is introduced for the acquisition of formalized knowledge in pathology, which is intended to serve as a basis for diagnostic support, including the confirmation of diagnoses, differentiation between diagnoses and the search for potential diagnoses based on findings. Diagnostic support for these three problems requires the availability of reference knowledge, differential diagnosis (DD)-criteria and the generation of DD-lists respectively. In general, textbooks offer reference knowledge by diagnosis name and, therefore, books are not the suitable medium for accessing information with findings as entry. The knowledge base structure can contain formalized knowledge within its context. An important issue of the design is that knowledge is acquired from the expert without the need for a knowledge engineer. The acquisition of knowledge has been made menu-driven in order to control the vocabulary used and to promote unambiguousness and completeness. The paper discusses the considerations underlying the design, followed by a description of the knowledge base structure and the user interface of the knowledge editor. A pilot study with three experts revealed userfriendliness and sufficient expression capability of the knowledge editor. One instance of ambiguity was found. The experts appeared to have adopted different starting points with respect to the selection of the specified diagnostic features.

Potential use of the acquired knowledge is addressed.

5.1 INTRODUCTION

In this paper we describe the structure of a knowledge base, intended to contain formalized pathology knowledge. The formalized knowledge may serve as a basis for diagnostic support in pathology. Crucial in the design of the knowledge base is the knowledge editor, allowing experts in pathology to fill the knowledge base.

Although an increasing number of expert system shells become commercially available for the construction of expert systems, they are seldom suited for direct use by experts in the field of application. The problem is the desire to keep a shell general and flexible on the one hand, and sufficiently simple to operate on the other. The result is a reduced expression capability. Shells, which combine both flexibility and a large expression capability have a complex user interface and require programming skills of the user [1]. Shells with simple, self explanatory user interfaces are usually rigid and suitable for applications in well-circumscribed and restricted domains [2,3]. Many expert and diagnostic-support systems in the field of pathology have required programming effort, admit a limited explicitation of pathology knowledge in formalized format and have been filled with the aid of a knowledge engineer [4,5,6,7]. In the "classical" situation a knowledge engineer interrogates the expert and converts the transferred knowledge into a formalized format, which can be entered in the knowledge base [8,9]. Because of the knowledge engineering bottleneck a different strategy for knowledge acquisition has emerged: learning from well documented cases [10]. However, the learning process in these systems usually is a long-term process, which is dependent on the available material. In addition, experts have to check the validity of the acquired knowledge and they have little or no control over the completeness of the acquired knowledge. Our aim is to make the knowledge engineer superfluous by offering an knowledge editor for acquisition of knowledge directly by the expert. The development of such a knowledge editor requires meta-knowledge engineering as there has to be information about which terms are allowed and which relations among them are legal in the domain. The terms and their relations determine which knowledge can, in principle, be expressed. In other words, the knowledge base must contain information with respect to the syntax, the semantics and the scope of the knowledge it is to contain.

In section 2 the diagnostic process in pathology is briefly discussed. The role of reference knowledge is addressed and the importance of access to that knowledge on the basis of findings. Section 3 deals with the considerations underlying the design of the knowledge base with emphasis on the expression capability and the support of automated knowledge acquisition. Section 4 describes the implementation of the knowledge base and the knowledge editor in the field of ovarian pathology. Section 5 discusses a pilot study with three pathologists. The final section summarizes how the formalized knowledge can be used for diagnostic support and indicates how the knowledge of more than one expert can be combined.

5.2 THE DIAGNOSTIC PROCESS IN PATHOLOGY

In this section the characteristics and problems of the diagnostic process [11] are briefly reviewed with emphasis on those aspects, which are considered suited for improvement by computerized diagnostic support.

The major issue of diagnosis making in pathology is the classification and grading of histologic slides. A histologic diagnosis involves visual interpretation of a slide in the context of the clinical findings about the patient. Diagnostic problems include the confirmation of diagnoses, differentiation between morphologically similar diagnoses and the search for possible diagnoses based on findings. During the diagnostic process these problems usually alternate. The whole body of knowledge which the pathologist uses to confirm a diagnosis or to distinguish a diagnosis from morphologically similar diagnoses will be referred to as "reference knowledge". This reference knowledge is composed of the training received, experience gained during years of practice, books, atlases, journals and the consultation of colleagues. We will call knowledge, obtained from written sources and colleagues, external knowledge. The aim of computerized diagnostic support is to improve the accessibility and utilization of the sources of external knowledge.

Pathology knowledge as laid down in textbooks consists of descriptions and illustrations of diagnoses with emphasis on the macroscopic and microscopic features of these diagnoses. Note that the word "feature" refers to a characteristic of a diagnosis as described in the literature, whereas the word "finding" refers to an observation interpreted by the pathologist. The contents of books are usually ordered according to an accepted diagnostic classification scheme. In other words, a book is diagnosis-oriented: knowledge is stored by diagnosis name. When a pathologist consults reference knowledge concerning a particular diagnosis for confirmation this approach poses no problem. In practice the problem is often reversed: from findings to diagnosis. In textbooks access to diagnoses via presenting features is very limited since the index lists only a small portion of them. Some features, especially the ones which are very specific to a certain diagnosis, are included (such as Call Exner bodies and Reinke crystals). Less specific features, though important in combination with others, are not found in the index as the number of page references would increase rapidly when diagnostic features become less specific. When findings cannot be used as the entry to reference knowledge it may be difficult to select a diagnosis for consultation. A considerable searching effort may be needed to find a diagnosis, which fits the findings [12]. This means that books, though most widely used as source of reference knowledge, are not the most suitable medium to handle the "inverse problem" of diagnosis making. Ideally, diagnostic support should include the generation of a set of potential diagnoses based on findings. This approach may reduce searching efforts and promote efficient use of personal or reference knowledge.

An important part of the knowledge needed for diagnosis making consists of differential diagnostic criteria to tell morphologically related diagnoses apart. Textbooks offer differential diagnosis criteria to a varying extent. Due to the complexity of differential diagnosis extensive differential diagnostic information would result in an enormous volume of redundant information or an almost unwieldy number of references criss-cross in the book. Hence, diagnostic support should also include the availability of differential diagnostic criteria per pair of diagnoses based on a comparison of their features. In summary, diagnostic support should consist of:

- Efficient access to extensive external knowledge by diagnosis name for confirmation of diagnoses [12]
- The generation of a set of potential diagnoses based on observed findings
- The generation of differential diagnostic criteria for an arbitrary pair of differential diagnoses on the basis of diagnostic features

5.3 CONSIDERATIONS FOR THE DESIGN OF THE KNOWLEDGE BASE

The intention of our approach is to provide one source of knowledge, which can serve as a basis for the three kinds of diagnostic support as mentioned at the end of section 2. This approach has the advantage that knowledge has to be acquired only once and, in addition, there is not the problem of inconsistency related to the existence of several sources of knowledge in one domain. However, when diagnostic support is based on features these must be accessible separately. Hence, diagnostic information cannot be stored as plain text. Instead we have chosen to store diagnostic features as separate entities of formalized knowledge.

5.3.1 The formalization of knowledge in pathology

The explicitation of pathology knowledge in formalized features requires consideration of the following topics: vocabulary, unambiguousness, context, uncertainty and completeness. They are subsequently discussed.

Experts do not always use the same terminology. That is, they may use different words to express the same intention (synonymy) or the same word for different intentions (homonymy), which implies a restricted consensus about the meaning of those words. This may lead to ambiguity in the knowledge base.

Unambiguousness is important when the formal knowledge is used by algorithms for the generation of potential diagnoses and differential diagnostic criteria. Such algorithms require both the matching of formalized features among the diagnoses in the knowledge base and the matching of formalized features with findings. It is important to realize that ambiguity may vary with the level of detail of knowledge explicitation. Two pathologists may completely agree about the presence of macrophages in a histological image, yet differ (slightly) in their description of this cell-type. However, detailed descriptions are sometimes essential (certainly in formalized knowledge), for example when atypical tumor cells have to be distinguished from similarly looking benign cells. Focus is not yet on dealing with the problem of consensus but on minimizing ambiguity, which may arise when one expert fills the knowledge base. In general, pathologists understand many more terms then they would actively use. For the purpose of unambiguousness synonyms can be mapped onto one word. This is, however, difficult with homonyms. Therefore, a standard vocabulary has to be defined in consort with several experts, such that the number of homonyms is minimal and the expression capability sufficient.

The explicitation of pathology knowledge requires that features are placed in a proper context. An example of context is the occurrence of certain cell-types related to the tissue structures in which they occur. Context also includes the expression of the spatial architecture of tissue components: cell-type A surrounds the vessels or cell-type A is present in a diffuse pattern.

An important issue in the explicitation of knowledge is the topic of uncertainty. Many diagnoses are not made with 100% certainty. Several kinds of uncertainties are responsible for the absence of complete certainty:

- Uncertainties with respect to findings themselves: the presence of a certain cell-type may be unlikely, possible, probable, or certain.

- Uncertainties as part of the knowledge, i.e. the conclusions based on a finding: when cell-type A is present, diagnosis B may be unlikely, possible, probable, or certain.
- Uncertainty due to lack of consensus about the knowledge: two experts may disagree upon the likeliness of diagnosis B when cell-type A is present.

The expression of uncertainty and strategies to reason with them are extensively discussed elsewhere [13,14]. Accurate numeric expressions of uncertainty are scarce and difficult to obtain. However, as there are no more feature combinations than diagnoses, we have chosen to focus on the qualitative expression of diagnostic information with the possibility to use a limited set of words to indicate in what frequency features occur. These words are never, rarely, sometimes, usually, often, and always. Each of them covers part of the range from 0% to 100%.

Since the relevance of features may vary with each differential diagnostic problem, it is important that the knowledge about a diagnosis is complete as possible. The aim is that the diagnostic information is at least unique for each diagnosis. In addition, what seems trivial to the expert might not be trivial to the general pathologist. To encourage the expert to be as extensive as possible in the explicitation of knowledge, in some way all possibly relevant features should pass successively. In that manner attention is drawn to that part of the expert's knowledge, which he or she would not have made explicit otherwise.

The construction of a knowledge editor enables control of the vocabulary used and the furtherance of completeness during the acquisition of formalized features. For this purpose, the knowledge editor must both make typing superfluous and let all describable items about a diagnosis pass in successive review.

5.3.2 Requirements

The requirements for the knowledge base are not only based on the considerations with respect to the acquisition of formalized pathology knowledge. Though the consultation of reference knowledge is an important part of diagnostic support, the existing applications in pathology [5,15,16,17] do not combine decision support and consultation of knowledge. Consultation requires that the knowledge can be presented as readable text. The text for consultation should be consistent with the formal knowledge. Therefore a "sentence generator" is required to present the formal knowledge in the form of readable sentences. Finally, the structure of the knowledge base should allow for updating without the problem of creating inconsistencies as is a well known drawback of large rule bases [18]. We can now formulate requirements for the knowledge base:

- The vocabulary to be used should be predefined: especially homonyms are to be avoided.
- The structure must allow for the expression of context with respect to the formalized features.
- The formalized knowledge must be unambiguous as much as possible: there should preferably be a unique formalization for each feature and only one place to store it in the knowledge base.
- The structure must allow for the construction of a knowledge editor for knowledge acquisition directly by the expert.
- The structure must allow for the generation of sentences.

5.4 A KNOWLEDGE BASE FOR PATHOLOGY OF THE OVARY

5.4.1 Epitool

To implement the knowledge base we have used the software package Epitool [19], an environment for the development of knowledge software. The package is suited for the development of frame-based applications [20,21]. In the following, those elements of Epitool are mentioned, which are relevant for the implementation of the knowledge base structure.

A *concept* in Epitool is characterized by certain properties called *aspects*. The instantiation of a *concept* results in an *individual* with the properties of that *concept.* In the *individual,* all or part of the *aspects* receive a specific *value* identifying the object.

To illustrate these definitions let the *concept* "cyst" represent the set containing all kinds of cysts. Some examples of *aspects* of "cyst" are its "color", its "appearance" and its "contents". Then, an *individual* is a cyst with *aspect values,* "red" for "color", "smooth" for "appearance" and "fluid" for "contents". Each *aspect* value is an *individual* or a set of *individuals* and for each *aspect* it is specified to which *concept* its *value* must belong. The *aspect* "color" may only have an *individual* of the *concept* "color" as its *value* and "red" is indeed an *individual* of the *concept* "color".

In Epitool *concepts* are organized in a hierarchy, where the hierarchy represents a set-subset relationship. This implies that the aspects of a child *concept* include all *aspects* of its parent in the hierarchy.

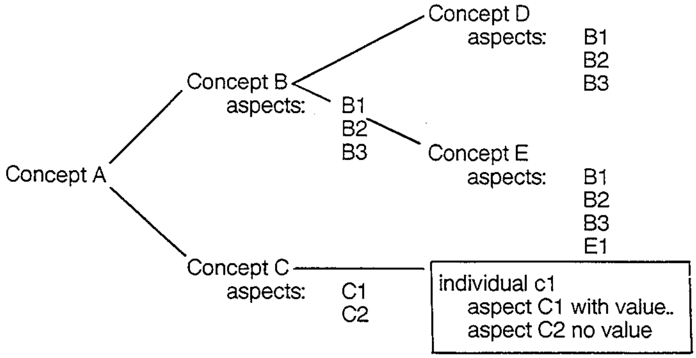


Figure 1. The organization of *concepts, aspects* and *individuals* in an Epitool knowledge tree.

Fig. 1 shows how *concepts, aspects* and *individuals* are hierarchically organized. In this Figure, as well as in the examples to follow, *individuals* are represented by names in a box. In Fig. 1 the *concepts* D and E are children of B and they inherit the *aspects* from B. Note, that *concept* E also has *aspect* E1, in addition to those inherited from *concept* B. In *individuals not* necessarily all *aspects* have a *value* as is visible at *individual* c1.

5.4.2 Design of the knowledge base

In the design of the knowledge base the word "tree" is used to denote an implementation of an Epitool hierarchy. Two trees are in use: a classification tree and a knowledge tree.

The *concepts* at the nodes of the classification tree constitute all diagnosis groups and diagnoses. We have chosen to use the WHO classification for ovarian tumors as the basis for the ordering of the tree [22]. The classification tree plays a secondary role and is used to provide the expert pathologist with a familiar way to select a diagnosis. A part of the classification tree is given in Figure 2.

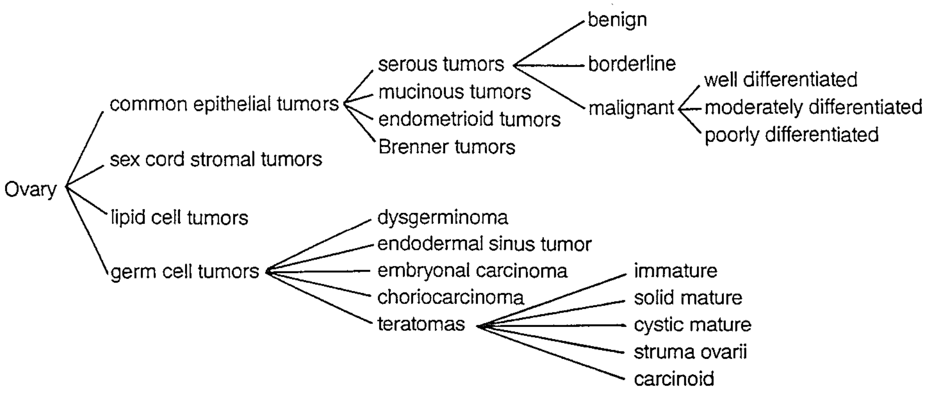


Figure 2. An excerpt of the classification tree on the basis of the WHO classification for ovarian rumors.

The knowledge tree is different from the classification tree: it is meant to hold diagnosis information in the form of formalized features. The naked knowledge tree, i.e. when knowledge is not yet entered, contains meta knowledge: knowledge about the vocabulary and structure of the diagnosis information. When knowledge is entered the knowledge tree grows, its contents represented by numerous *individuals.*

The concepts of the naked knowledge tree are divided into macroscopic and microscopic concepts following common practice. Part of the naked knowledge tree is shown in Fig. 3. The *individuals* in Fig. 3 are not specific for a diagnosis. They represent predefined *individuals,* each of which occurs only once in the knowledge base. They can be used to characterize diagnoses, but are not described themselves in further detail. Note, that in Fig. 3 the *concept* "cysts" has the *aspect* "size" in addition to those inherited from its parent "tissue structures".

The *concepts* and predefined *individuals* in the knowledge tree determine which items and morphological features can, in principle, be formulated about a diagnosis within this system. A feature of a diagnosis is represented in the knowledge tree by the combination of an *individual* and one of its *aspect* values as is shown in Fig. 4a (for example, an *individual* "solid areas" with the *value* "stroma" for its *aspect* "tissue structures"). The feature, represented in this example is: the solid areas have the tissue structure stroma. Both "solid areas" and "stroma" are *individuals* belonging to a particular tumor. As the knowledge tree can hold information with respect to many different diagnoses, a *concept* may very well have *individuals* belonging to more than one diagnosis. The information with respect to aparticular diagnosis is represented by a set of *individuals,* which *are* linked in the proper context by means of their *aspect* values.

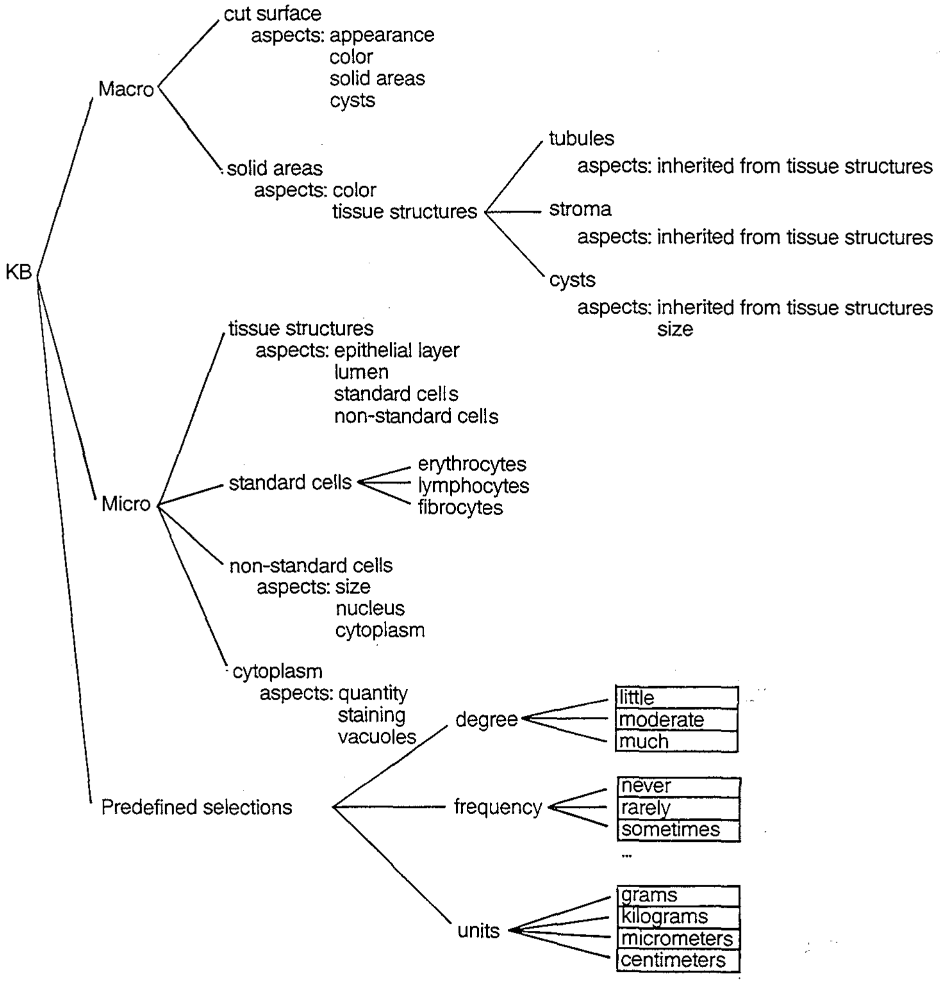


Figure 3. A part of the naked knowledge tree with *concepts*, *aspects* and predefined *individuals*. The naked knowledge tree as a whole defines the syntax, semantics and scope of the knowledge it is to hold.

The knowledge tree, as presented in Fig. 4a, shows the *concepts* and *individuals* in a set-subset hierarchy. Though this presentation of the knowledge tree is suited for the definition of *concepts* and *aspects,* it does not reflect the relations between the *individuals* of one diagnosis. To visualize the formalized knowledge about a single diagnosis, the information can be presented in another way: a tree where the connections represent part-of links between *individuals.* This is shown in Fig. 4b.

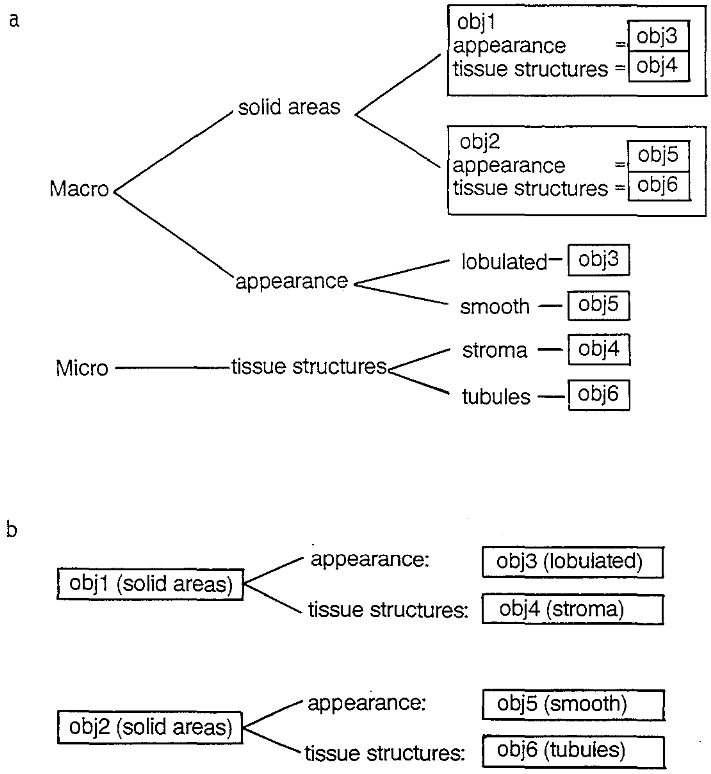


Figure 4. Diagnosis features as represented in the knowledge tree (a) and in their functional context (b).

The design of the knowledge tree enables controlled knowledge acquisition. In the present design when the expert enters a formalized feature, the system creates an *individual* with *aspect* values as specified by the expert. The specification of an *aspect* value may require the creation of an *individual,* which in turn has to be specified by *aspect* values. Since the *concept* of each *aspect value* is known, the *aspects* to characterize a new *individual* are also known. For example, a tumor has the *aspect* "cut surface". The *value* of that *aspect* will be a newly created *individual* of the *concept* "cut surface". The *aspects* of "cut surface" are known (appearance, color, cysts, solid areas etc.) and the new *individual* can subsequently receive *values* for its *aspects. Individuals* are never created from parent *concepts* since these only function as nodes to group *concepts* sharing the same *aspects.* In other words *individuals* are always created from *concepts* which have no child *concepts.*

Aspects may receive predefined *individuals* as their value. Examples of predefined *individuals* are "slight" and "often", which serve as *values* for the *aspects* "degree" and "frequency" respectively. In this way, the user can express how frequent a feature occurs and to what degree it is present. The *aspects* allow for the assignment of more than one *value:* the appearance of a tumor may be both lobulated and shiny.

Each *aspect value* is always worked out down the tree. When the leaves of the tree are reached recursion takes the user back in the direction of the stem. However, the assignment of a new *aspect value* at a higher level reverses the direction again: in practice the user moves up and down the tree many times before the knowledge entry session finishes at the top node of the tree.

In order to achieve uniformity during knowledge acquisition, five *aspect* categories have been defined. Each *aspect* belongs to exactly one of these five categories. Depending on which *aspects* are defined at a *concept,* part or all of the categories are represented. *Aspects* receive their *values* in a sequence determined by the following fixed order of categories:

- parameter aspects
- correlation and reference aspects
- property aspects
- phenomenon aspects
- part-of aspects

The parameter *aspects* express frequency and degree of a feature as well as its spatial occurrence (e.g. focal or dispersed). Correlation and reference *aspects* are used to express correlation with or reference to some other *individual:* when the same tubules are always surrounded by necrosis, these tubules and necrosis are correlated. References are used to avoid redundancy in the knowledge tree: a cell-type, which occurs in the stroma as well as in the epithelium, needs only be described at one position in the tree and can be referenced in others. In other words: one *individual* can serve as the *value* of several reference *aspects.* Property *aspects* characterize *individuals* as a whole, e.g. "color", "size", "appearance" and "weight". Phenomenon *aspects* describe which phenomena might occur associated with an *individual* such as "necrosis" in a "stroma". Finally, part-of *aspects* link an *individual* with its composing parts: the contents of a cyst or the nucleus of a cell. By means of the part-of *aspects* the path through the knowledge tree is from macroscopic to microscopic scale. For example, an *individual* of the *concept* "solid areas" may be characterized by its property *aspect* "color" with the *value* "red" and the phenomenon "necrosis". Its part-of *aspect* "tissue structures" forms a link with the microscopic *individual* "stroma".

The next paragraph gives a part of a knowledge entry session, followed by a description of the internal representation of that knowledge and its feedback to the user.

5.4.3 Knowledge acquisition and representation

Using an example this section addresses the following topics: the acquisition of knowledge, the internal representation of knowledge and the external representation of it to the user.

A part of an exemplatory knowledge entry session is shown in Fig. 5. The entry of knowledge is menu-driven. The order in which the menus appear on the screen depends on the choices of the user. Note, that menus allow for more than one choice, which is visible in the menus "Nucleus" and "Chromatin". Note also, that a choice is followed down the tree before offering a second choice. In the Figure, the choice "chromatin" is worked out first. Having returned to the menu for "nucleus" the additional choice "pattern" is made. The expert may leave a menu by selecting "ready" Ten, that menu is subsequently skipped on the way back. Fig. 5 shows that the user is not returned to the menu with the property *aspects* of "atypical cells" after having completed the menu with the part-of *aspects* of "atypical cells".

References or correlations are specified as follows. As soon as the entry "reference" is selected, the user returns to the first menu, containing the major diagnostic groups of ovary pathology. As usual the user then progresses through the classification tree with the difference that menus only show diagnoses, already entered in the knowledge base. As soon as a selection is made, the *individuals* of the selected diagnosis are displayed as nodes of a tree. One of them can be selected with the mouse. With the selection of an *individual* from the tree, the specification of the reference is completed. Appendix A shows two such diagnosis trees.

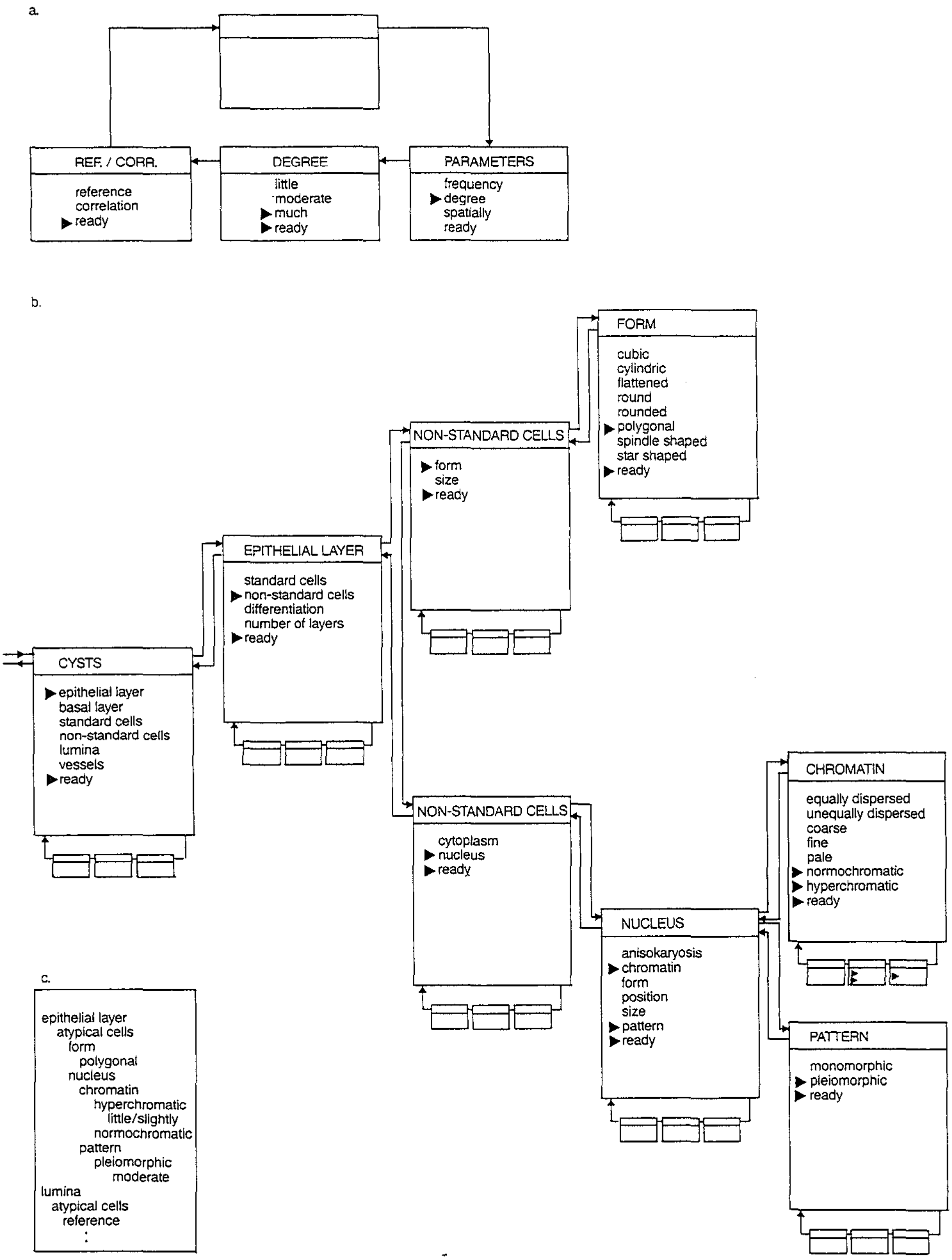


Figure 5. An example of menu-driven knowledge entry. Note, that each choice in a menu (b) is followed by a fixed sequence of menus to allow for the specification of parameter *aspects*, correlations and references when necessary (a). A seperate window (c) displays feedback during knowledge entry.

The formal representation of knowledge in the knowledge base is hidden for the user. There is a one-to-one correspondence between the knowledge entered by the user, and its formal representation. Table 1 shows the formal representation of a part of the knowledge as entered in Fig. 5. Note, that a choice, does not always lead to the instantiation of an *individual.* No *individuals* are created from the parent *concepts* "form" and "chromatin", but from one of their child *concepts* and no *individuals* are created from *concepts,* which have predefined *individuals* only.

Menu selection	Operation in knowledge base
epithelial layer	*Create Individual* of epithelial layer: obj10
ready	- (no operation)
atypical cells	*Create Individual* of atypical cells: obj11 *Set Value* obj10.atypical cells = obj11
ready	-
form	- (No *individual* is created of form,) (as it is a parent *concept.*)
polygonal	*Create Individual* of polygonal: obj12 *Set Value* obj11.form = {obj12} (The *aspect* form may have several values,) (hence the values are stored in a set.)
ready	-
nucleus	*Create Individual* of nucleus: obj13 *Set Value* obj11.nucleus = obj13
ready	-
chromatin	- (No *individual* is created of chromatin,) (as it is a parent *concept.*)
hyperchromatic	*Create Individual* hyperchromatic: obj14 *Set Value* obj13.chromatin = {obj14} (The set represents that the *aspect*) (chromatin may have several values.)
degree	- (No *individual* is created of degree,) (only of its predefined *individuals.*)
little/slightly	*Set Value* obj14.degree = little/slightly
ready	-

normochromatic	*Create Individual* normochromatic: obj15 *Set Value* obj13.chromatin = {obj14, obj15}
ready	-
pattern	- (No *individual* is created of chromatin,) (as it is a parent *concept.*)
pleiomorphic	*Create Individual* pleiomorphic: obj16 *Set Value* obj13.pattern = obj16
ready	-
ready	-
ready	-
ready	-

Table 1 Sequence, showing the relation between the actions of the user and the formal representation of the operations on the knowledge base during knowledge entry as shown in fig. 5.

The form of the feedback to the user depends on the situation. Feedback during knowledge acquisition is straightforward and consists of the chronological display of the choices made by the expert. These choices, however, are displayed with varying margination to clearly visualize their context. Both the feedback and the menu sequence during knowledge acquisition are visible in Fig. 5.

For the purpose of inspection of the knowledge base, feedback has the form of simple sentences, which are composed automatically from the features in the knowledge base by the sentence generator. Each sentence is of one of the following types:

1. The <*individual*> [frequency] <verb> [spatially] [degree] <*individual*>.
2. The <*aspect*> of the <*individual*> [frequency] <verb> [spatially] [degree] <*individual*>.

The first type of sentence is used when the *concept* corresponding with the subject has no child *concepts,* otherwise the second type is used. An expression between "< >" is always present, whereas an expression between "[ ]" is optional. Since the internal *individual* names (objnr) are not suited for external representation of knowledge, their names are replaced by the name of the *concept* of which they are an instantiation. For both types of sentences, the *individual* after the verb expresses an *aspect value*. The verb is automatically inserted, depending on the type of aspect according to the following list:

property *aspect*	:	*"is"* or "are".
phenomenon *aspect*		*"shows"* or "show".
correlation	:	verb corresponds to type of correlation e.g.: "is intermixed with", "resembles".
part-of *aspect*	:	*"has"* or "have".

The *values* of parameter *aspects* are expressed by means of the optional expressions, which are automatically inserted as well. Each *aspect value* is represented by exactly one sentence. When a reference or correlation is present, sentences follow, that give a description of the referenced or correlated *individual.* Since the description of a referenced *individual* serves as a complete description, the reference itself is not represented by a separate sentence. The following sequence of sentences represents the knowledge as entered in Fig. 5. Varying margination is used here as well, to express the context of the sentences.

.
.
The epithelial layer has atypical cells.
The form of the atypical cells is polygonal.
The atypical cells have a nucleus.
The chromatin of the nucleus is slightly hyperchromatic.
The chromatin of the nucleus is normochromatic.
The pattern of the nucleus is moderately pleiomorphic.
.
.

Note, that the first, third and seventh sentences are of the first type, whereas the other ones are of the second type.

5.5 AN EXPERIMENTAL SESSION WITH THREE PATHOLOGISTS

To get insight into the feasibility of formalizing pathology knowledge a pilot study has been conducted in the area of ovarian pathology with three experts. The experts were asked to characterize a dysgerminoma by heart and a Sertoli cell tumor from an existing text. Since none of the experts was familiar with the system the sessions were preceded by an introduction of approximately 20 min. During the experiment the experts were free to ask questions about the use of the system or to comment on it. The experiment was intended to gain an insight in the following questions:

- What is the ease of interaction through the user interface?
- To what degree can pathology knowledge be made explicit?
- What are the differences in extensiveness of the formalized knowledge among experts?
- What are the differences in the contents of the formalized knowledge among experts?
- Is ambiguity present in the knowledge base structure?

The following discussion of these five items is based on observations during the experiment, comments by the participating experts and the formalized knowledge they entered. Appendix A shows two trees, representing the knowledge about a dysgerminoma, entered by two different experts.

As to the operation of the system, none of the experts had difficulties with understanding how to enter knowledge. The order in which issues passed in revue was experienced as natural during knowledge entry by heart. In contrast, when entering formal knowledge on the basis of the existing text, they were faced with the mental strain of the conversion of the ordering of the written knowledge to the ordering of the system as a separate effort.

As concerns the user interface, it was felt as an inconvenience that, initially, the contents of the menus are not known to the user. Therefore, the participants had to try out menus several times in order to see whether they contained the desired property. This occurs especially when a property can be viewed from more than one viewpoint. For example, "smooth" can be interpreted as a touch as well as an appearance. As the knowledge base deliberately supports only one viewpoint for each property it is to be expected that a user soon learns which menus to select. The second remark on the user interface concerns the overview the user has during the entry of knowledge. Only one menu is visible at a time and the direct feedback is displayed in a separate window. When little attention is paid to this information, features may be placed in a wrong context.

The degree to which the knowledge could be made explicit was considered sufficient. None of the experts had the feeling that features were left unspecified because they were too complex or too subtle to express in the syntax of the system. Only at two instances a choice was felt missing in a menu: "necrotic" was missing in the menu for "appearance" and "syncytial cells" in the menu for "standard cell types". Other options will, probably, be missing when other diagnoses are entered, but these gaps can be easily filled by the addition of new, non-overlapping choices.

As expected, the experts differed in the extensiveness of the knowledge bases they created when formalizing by heart. This difference concerns both the presence and absence of features. Two experts mentioned the occasional presence of granulomas in a dysgerminoma, whereas the third expert regarded this feature as non-essential for the diagnosis. Conversely, Schiller-Duval bodies were not mentioned by two of the experts since they are absent in dysgerminomas. However, the third expert considered it important to mention that the presence of Schiller-Duval bodies is sufficient to reject the diagnosis dysgerminoma.

An explanation for these differences may be found in the fact that experts work from different starting points, which are characterized by one or more of the following ingredients:

- Specify all features, which can possibly be present in the diagnosis
- Specify the features, which are essential to confirm the diagnosis
- Specify the features, which are essential to confirm the diagnosis as well as the features, which are sufficient to reject it.

It is important to realize that these starting points have considerable consequences when algorithms are developed to match findings with features. It is beyond the scope of this paper to discuss this in detail but a few examples may illustrate the problems involved. Assume that the first starting point is adopted and that the matching algorithm includes in the DD-list all diagnoses that have a set of features in common, which corresponds to the set of findings specified by the pathologist.

The absence of even one feature of that set will exclude a diagnosis from the DD-list. This is correct when the features specified at each diagnosis are exhaustive and under the assumption that at least one diagnosis should explain for all findings. However, it is questionable whether or not this is feasible. Another problem may arise when the second starting point is adopted in combination with a matching algorithm that requires at least a match of the findings with features which are essential for confirmation. Due to poor quality of the slide, the absence of a certain stain or an omission of the pathologist, an essential finding may not be specified. As a result, diagnoses may be excluded from the DD-list even when highly specific findings are present. In other words: matching based on essential features alone probably does not allow for dealing with findings which are not essential but sufficient for confirmation.

Problems involved are the selection of features which should be marked as essential and whether or not additional categories are needed such as "possibly present" and "sufficient for confirmation". When knowledge would be entered according to the third starting point, matching could be performed in two steps: a match with essential features and a match with features sufficient for rejection. The first step implies problems as mentioned for the second starting point. In summary, the existence of multiple starting points illustrates the need for a careful evaluation of what combination of starting point and matching strategy is most suitable and feasible.

Quantitative differences in the contents of the three knowledge bases concern the degree and frequency of occurrence. Examples from the experiment are: "little granular" versus "non granular" and "usually eosinophilic" versus "sometimes eosinophilic". More qualitative differences occur when experts use different terms to characterize a property of an item, such as "fine chromatin" versus "coarse chromatin", instead of using different degree and frequency modifiers with the same term.

Differences in the contents of knowledge bases is partially due to the problem of limited expert consensus. Very few differences were found in the knowledge entered by heart and those found were subtle quantitative differences. It is yet unclear whether this is a systematic property of the knowledge editor or a coincidental finding. The knowledge bases created on the basis of the text were exactly equal among the experts, except for one feature, which was omitted by one expert.

Ambiguity in the knowledge base structure is present when a single feature can be characterized in more than one way. The experiment revealed the existence of such ambiguities. Lymphocytic infiltration occurs in all dysgerminomas. One expert characterized a number of tissue structures and mentioned an inflammatory infiltrate with lymphocytes separately. The second expert mentioned both the occurrence of lymphocytes as standard cell types in all tissue structures and the inflammatory infiltrate. The third expert characterized several tissue structures as did the first and mentioned the lymphocytes separately as part of the stroma. Leaving out inflammatory infiltrate as a tissue structure would eliminate this specific ambiguity since the feature can be formalized using other *concepts* in the knowledge base. However, ambiguity cannot simply be avoided by leaving out all *concepts,* which can be resolved in other *concepts.* The presence of separate *concepts* for specific features, which can also be formalized using a combination of already existing *concepts,* is sometimes desirable since the level of detail of the formalizations, which would be required otherwise, entails an even greater ambiguity.

5.6 POTENTIAL USE OF THE FORMALIZED PATHOLOGY KNOWLEDGE

Formalized pathology knowledge can be used in several ways, two of which were among the starting points for the design of the knowledge base structure. First, the knowledge can be used for findings-oriented diagnostic support, i.e. the generation of a list of potential diagnoses based on the findings of the user. For this purpose, findings obtained from a patient can be dynamically (temporarily) stored in a structure similar to the static knowledge base.

Then, algorithms can be developed which match the dynamic knowledge with the static knowledge base. By varying the level of detail in the dynamic knowledge, the user can control the scope of the match. In other words, the more detailed the findings, the smaller the set of matching diagnoses.

The same algorithms which match the findings of the user with the static knowledge base, can be used to compare the formalized knowledge of a pair of diagnoses. In that way differential diagnostic criteria can be generated.

Second, the formalized knowledge can be used for consultation of knowledge by diagnosis name. This form of diagnostic support serves two purposes: (1) the confirmation of diagnoses and (2) inspection of the contents of the knowledge base. As consultation requires knowledge to be presented as readable text, a sentence generator has been proposed. In that way, one knowledge base can be used for both the matching of findings and consultation, thereby guaranteeing consistency.

The third possibility concerns the combination of the formalized knowledge of more than one expert. It remains difficult to deal with lack of consensus, but the pilot study revealed that differences in the formalized knowledge are not necessarily due to disagreement among the experts. Many differences can be contributed to variations in the starting points of the experts as to what features are entered into the knowledge base. It is conceivable to combine the knowledge of several experts to a new knowledge base, which should then be more complete than any of its composing parts. Such a process of combining several knowledge bases can, in principle, be made part of the knowledge acquisition software. The newly created knowledge base will have to be offered to a panel of experts to see whether or not the result is valid.

5.7 CONCLUSION

In this paper we have described a knowledge base structure with a knowledge editor for the acquisition and storage of formalized pathology knowledge in its context. The knowledge editor has a menu-driven user interface. By presenting menu options the pathologist automatically uses the proper syntax and vocabulary. The menus appear in the order from macroscopic to microscopic detail to correspond with common practice and to promote completeness.

The results of our pilot study with three experts in which they were asked to formalize knowledge was promising with respect to the following:

- The user interface was quickly understood and easy to use.
- The order in which the menus were presented was experienced as natural when entering knowledge by heart.
- When entering knowledge from a text, the process of formalization was felt as a separate mental effort.
- The expression capability was considered sufficient.
- Ambiguity in the actually formalized knowledge was found for one feature only.

The study also revealed two issues, which require extra attention. First, the experts differed in the extensiveness of their knowledge explicitation. Therefore, the experts should be encouraged to adopt the starting point that the knowledge about a diagnosis should include features that are essential to confirm the diagnosis, but also features that are sufficient to reject the diagnosis. The second issue concerns the problem of ambiguity. A trade-off has to be made between the ambiguity resulting from the presence of *additional* concepts for features, which can in principle be formalized using other *concepts* and the ambiguity resulting from highly detailed formalization.

The advantages of the design over existing expert systems in pathology can be summarized as follows:

- Knowledge acquisition does not require a knowledge engineer.
- The knowledge can be made explicit to a high degree.
- The knowledge is accessible as separate features for findings- oriented diagnostic support.
- The knowledge can be directly used for consultation purposes.
- The knowledge base structure allows for the combination of knowledge of more than one expert.

Though the naked knowledge tree will have to be extended with many more *concepts* to allow for the formalization of pathology knowledge on a large scale, the design provides pathologists with a tool to formalize pathology knowledge in general. Formalized pathology knowledge is a conditio sine qua non for the development of diagnostic support systems based on findings.

ACKNOWLEDGEMENTS

This work was supported by grant PF 28-1207 of the Praeventiefonds. We gratefully acknowledge C. Kooijman M.D. Ph.D. and J.A.J. Spaas M.D. for their participation in the pilot study.

REFERENCES

[1] Gevarter B. The Nature and Evaluation of Commercial Expert System Building Tools. Computer May 1987:24-41.

[2] Vittal J. Languages for use in artificial intelligence research. The Second International Conference on Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer-Chicago Illinois, 1988.

[3] Salzman GC, Krall RB, Marinuzzi JG. Knowledge Engineering Software: High End Tools. The Second International Conference on Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer - Chicago Illinois, 1988.

[4] Potter B and Ronan SG. Computerized dermatopathologic diagnosis. J Am Acad Dermatol 1987;17:119-131.

[5] Horvitz EJ, Heckerman DE, Nathwani BN, Fagan LM. Diagnostic Strategies in the Hypothesis-Directed PATHFINDER system. In: Proceedings of The First Conference on Artificial Intelligence Applications: IEEE Computer Society, 1984: 630-636.

[6] Fuezesi L, Fuezesi S. Expert System for Inflammatory Skin Diseases. The Second International Conference on Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer-Chicago Illinois, 1988.

[7] Dhawan AP. An expert System for early Detection of Melanoma using Knowledge-Based Image Analysis. The Second International Conference on Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer - Chicago Illinois, 1988.

[8] Musen MA, van der Lei J. Knowledge engineering for clinical consultation programs: modeling the application area. Meth Inform Med. 1989;28:28-35.

[9] Gaines BR. An overview of knowledge-acquisition and transfer. Int J Man-Machine Studies 1987;26:453-472.

[10] Hunter L and Silbert JA. Progress Report on IVY: A Learning System for Intelligent Information Retrieval in Pathology. Artificial Intelligence in Medicine Workshop, March 1987.

[11] Langley FA, Baak JPA, Oort J. Diagnosis making: Error sources. In: Baak JPA, Oort J. eds. A Manual of Morphometry in Diagnostic Pathology: Springer Verlag, 1983:6-14.

[12] van Ginneken AM, Baak JPA, Jansen WJ, Smeulders AWM. Evaluation of a Diagnostic Encyclopedia Workstation for Ovarian Pathology. Submitted for publication in Hum Pathol.

[13] van Ginneken AM, Smeulders AWM. Reasoning in uncertainties: an analysis of five strategies and their suitability in pathology. Accepted for publication in Anal Quant Hist, 1988.

[14] Gammerman A, Creaney N. Modelling of Uncertainty in Expert Systems. Second International Expert Systems Conference - Oxford: Learned information,1986:265-274.

[15] Chang E, McNeeley M, Gamble K. Strategies for choosing the next test in an expert system. In: Proceedings of the American Association of Medical Systems and Informatics Congress - Bethesda Md: American Association of Medical Systems and Informatics, 1984: 198-202.

[16] Alvey PL and Greaves MF. Observations on the development of a high performance system for leukemia diagnosis. In: Bramer MA ed. Research and Development in Expert Systems III, Cambridge Unversity Press, 1986:99-110.

[17] Donovan RM, Nagel M, Goldstein E. An expert system-based leukocyte classification instrument. The First International Conference on Artificial Intelligence Systems (Expert Systems) as Diagnostic Consultants for the Cytologic and Histologic Diagnosis of Cancer - Universal City CA, 1987.

[18] Hayes-Roth F. Rule-based systems. In: Communications of the ACM 1985;28:921-932.

[19] Epitool Development Environment, Copyright 1986 by Epitec AB, S:t Larsgatan 12, S-582 24 Linkoping, Sweden.

[20] Fikes R, Kehler T. The role of frame-based representation in reasoning. Communications of the ACM 1985;28:904-920.

[21] Minsky M. A framework for representing knowledge. In: Winston P.M. ed. The psychology of computer vision, McGraw-Hill, 1957:211-277.

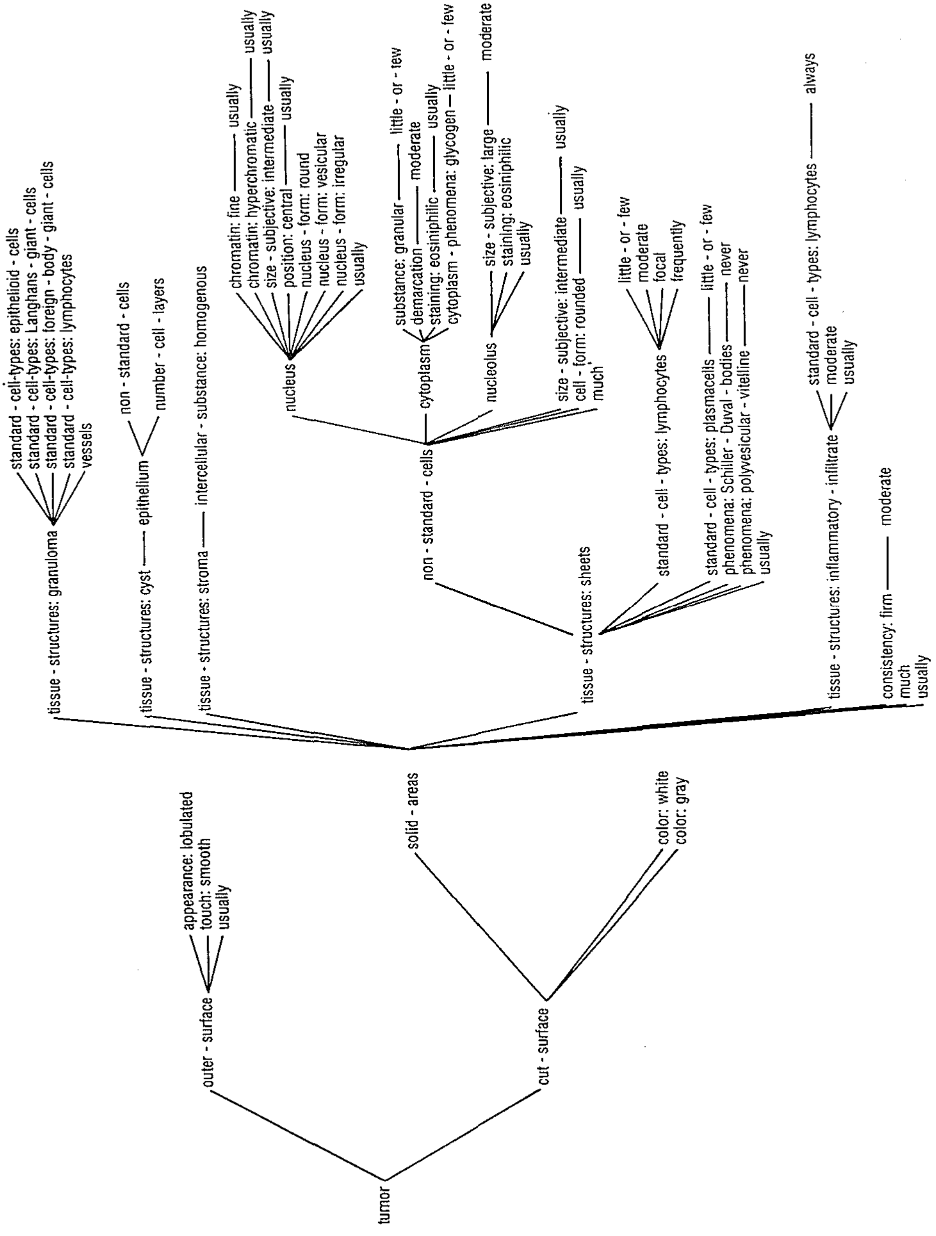
[22] Serov SF, Scully RE, Sobin LH. Histological typing of ovarian tumors. In: International Histological Classification of Tumors 9, Geneva: World Health Organization, 1973.

Appendix A

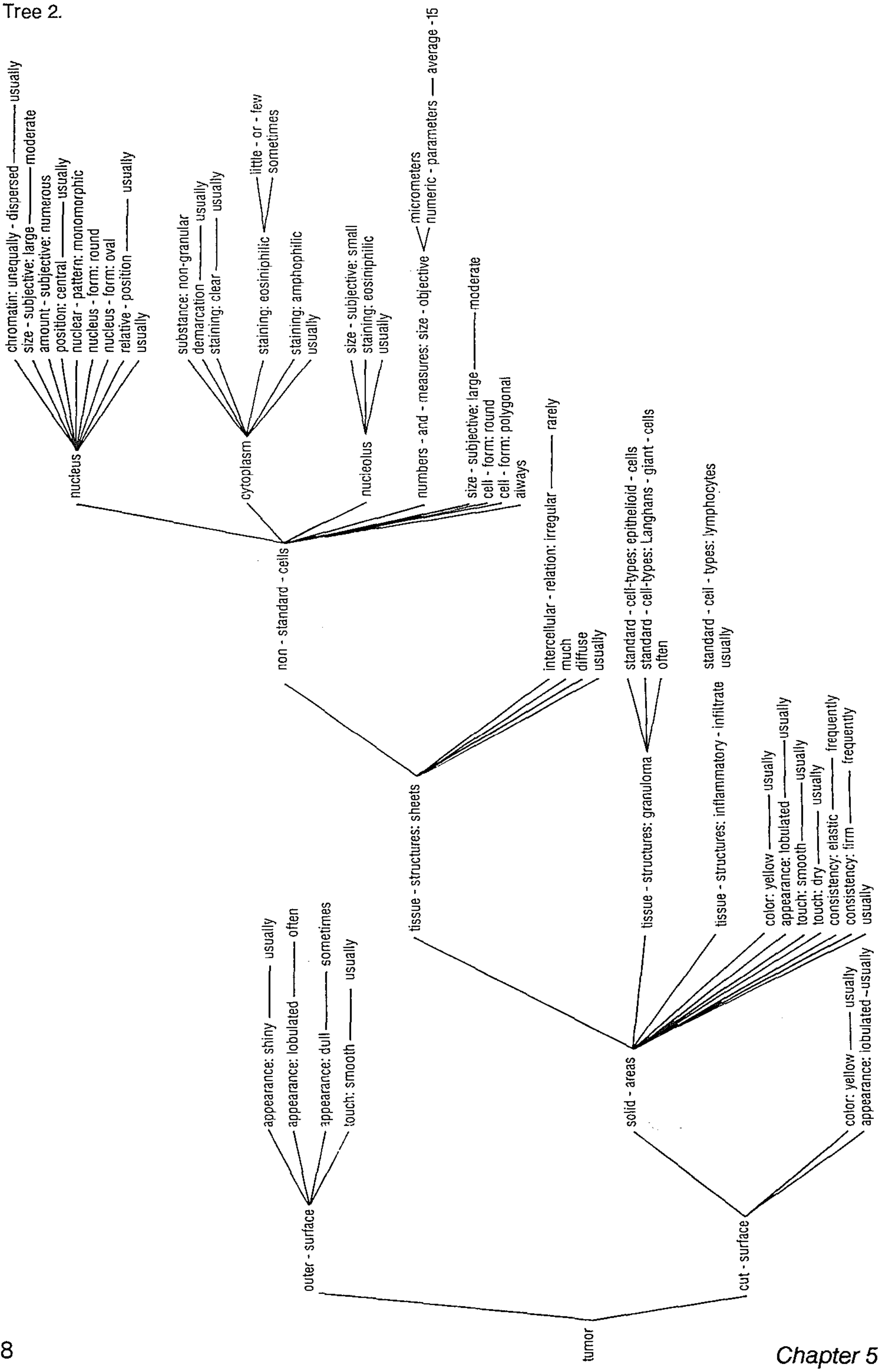
Two trees, each representing formalized knowledge about the diagnosis dysgerminoma in its functional context. When specifying a reference to an *individual* the user can select a node from this tree. Nodes representing values of parameter aspects cannot be referenced.

The two trees reflect several kinds of differences in the formalized knowledge. Many differences in extensiveness reflect differences between the pathologists as to what features should be mentioned. An example is the fact that the first tree mentions more tissue structures than the second. Furthermore, the second tree contains features, which (may) fit with a dysgerminoma, whereas the first tree also contains features, which should not be present in the diagnosis. Differences in the characterization of corresponding features are found in the form of the nuclei, the granularity of the cytoplasm, and several values of parameter *aspects.* Finally, the two trees show the presence of ambiguity in the naked knowledge tree: the presence of inflammation can be characterized by the presence of an inflammatory infiltrate, by mentions of lymphocytes in the tissue structures or by a combination of both.

Tree 1.


tumor
outer - surface
appearance: lobulated
touch: smooth
usually
cut - surface
solid - areas
color: white
color: gray
tissue - structures: granuloma
standard - cell-types: epithelioid - cells
standard - cell-types: Langhans - giant - cells
standard - cell-types: foreign - body - giant - cells
standard - cell-types: lymphocytes
vessels
tissue - structures: cyst
epithelium
non - standard - cells
number - cell - layers
tissue - structures: stroma
intercellular - substance: homogenous
tissue - structures: sheets
non - standard - cells
nucleus
chromatin: fine
usually
chromatin: hyperchromatic
usually
size - subjective: intermediate
usually
position: central
usually
nucleus - form: round
nucleus - form: vesicular
nucleus - form: irregular
usually
cytoplasm
substance: granular
little - or - few
demarcation
moderate
staining: eosiniphilic
usually
cytoplasm - phenomena: glycogen
little - or - few
nucleolus
size - subjective: large
moderate
staining: eosiniphilic
usually
size - subjective: intermediate
usually
cell - form: rounded
usually
much
standard - cell - types: lymphocytes
little - or - few
moderate
focal
frequently
standard - cell - types: plasmacells
little - or - few
phenomena: Schiller - Duval - bodies
never
phenomena: polyvesicular - vitelline
never
usually
tissue - structures: inflammatory - infiltrate
standard - cell - types: lymphocytes
always
moderate
usually
consistency: firm
moderate
much
usually

Tree 2.


tumor
outer - surface
appearance: shiny — usually
appearance: lobulated — often
appearance: dull — sometimes
touch: smooth — usually
cut - surface
color: yellow — usually
appearance: lobulated –usually
solid - areas
tissue - structures: sheets
tissue - structures: granuloma
tissue - structures: inflammatory - infiltrate
color: yellow — usually
appearance: lobulated — usually
touch: smooth — usually
touch: dry — usually
consistency: elastic — frequently
consistency: firm — frequently
usually
non - standard - cells
intercellular - relation: irregular — rarely
much
diffuse
usually
standard - cell-types: epithelioid - cells
standard - cell-types: Langhans - giant - cells
often
standard - cell - types: lymphocytes
usually
nucleus
chromatin: unequally - dispersed — usually
size - subjective: large — moderate
amount - subjective: numerous
position: central — usually
nuclear - pattern: monomorphic
nucleus - form: round
nucleus - form: oval
relative - position — usually
usually
cytoplasm
substance: non-granular
demarcation — usually
staining: clear — usually
staining: eosiniphilic
little - or - few
sometimes
staining: amphophilic
usually
nucleolus
size - subjective: small
staining: eosiniphilic
usually
numbers - and - measures: size - objective
micrometers
numeric - parameters — average -15
size - subjective: large — moderate
cell - form: round
cell - form: polygonal
always

CHAPTER 6

Summary and conclusions

6.1 THE PROBLEM

As mentioned in Chapter 1, an important task of a pathologist is the classification and grading of tissue abnormalities by visual examination of histologic and cytologic slides in the context of the clinical history of the patient. The interpretation of the images by the pathologist may involve consultation of textbooks and colleagues for reference. A pathology diagnosis covers a range of tissue abnormalities. Due to differences in education and experience, experts do not always use the same criteria for defining such a range and they may also differ in their interpretation of findings. As a result, experts may disagree about the diagnosis of a case.

Techniques, some of them involving computer applications, have been developed to enhance consistency in pathology diagnosis. With respect to their goals we divide these techniques into two main categories: (1) increasing objectivity in the acquisition and interpretation of diagnostic data, and (2) promoting the accessibility and utilization of reference knowledge for classification and grading. The first category includes morphometry, image processing, and statistical pattern recognition, whereas the second category encompasses the computerization of patient archives, the storage of pictures on optical discs, and decision support systems. Though these techniques and computer applications perform their specific tasks adequately, they share one or more of the following limitations:

(1) Limited applicability and scope
(2) Too much emphasis on decision support based on findings solely
(3) Poor facilities for knowledge acquisition

6.2 FIRST PART OF RESEARCH

6.2.1 Consultation of Knowledge

The first part of the research, described in this thesis, concerns the second limitation. The question was addressed whether it is feasible to increase the quality and efficiency of the diagnostic process in pathology by offering an interactive consultation system for routine use. Such a system should consist of extensive diagnostic information, integrated with a large amount of illustrated patient cases. To that end the so-called Diagnostic Encyclopedia Workstation (DEW) has been developed which is described in Chapter 2. At present, the DEW contains 85 ovarian diagnoses, illustrated by 3,000 pictures, which was sufficiently large to allow for evaluation of the prototype. The system runs on an IBM-compatible PC connected to a videodisc player. The information is accessed via a self-explicable mouse-driven user interface.

6.2.2 Evaluation

In Chapter 3, the use of the DEW versus the use of textbooks for the consultation of reference knowledge in the domain of ovarian pathology has been evaluated in a cross-over experiment with two groups of pathologists. Statistical analysis of the evaluation showed the following results:

- Textbooks yielded better results with respect to the classification of diagnoses and the morphological approximation of the correct diagnosis.
- The DEW and the textbooks differed, though not significantly, in favor of the books with respect to the clinical consequences of diagnoses and the mutual consensus among the participants.

To these conclusions it should be added, however, that the experiment was negatively biased for the DEW: due to the limited availability of suitable test cases, the large set of different stains could not be used to its full advantage.

The evaluation revealed that differential diagnostic support, though more extensive in the DEW than in books, was not yet sufficiently complete for all cases of the test set. This is not surprising, as the differential diagnosis lists of the experts differed considerably. However, for each diagnosis of the test set the lists showed clustering around a subset of the differential diagnoses. Differential diagnostic information for systems like the DEW should include at least these subsets.

The results of the evaluation by the participants can be summarized as follows:

- Strong properties were the user interface, the large set of pictures and the differential diagnostic information.
- Suggestions for improvement concerned the relative scarcity of overview pictures and that diagnostic information at group level was not yet available.

When a pathologist was confronted with an unfamiliar case, it proved to be very difficult to find an appropriate entry into the system. This was as expected since the DEW was designed for the consultation of information by diagnosis name, following common practice. However, it is conceivable to expand the DEW with a list of indexed keywords, which can be used to zoom in on a set of diagnoses for consideration.

The DEW as a computerized encyclopedia has several advantages over books. It can contain much more extensive pictorial and textual diagnostic information than a book and even has the potential to cover the entire field of pathology. In that respect it is worth mentioning that the system has well-developed facilities for filling the database. The information in the database is highly structured and can be accessed via several entries, making consultation more flexible. Especially, the possibility to combine information at different locations is crucial in offering differential diagnostic information. In addition, updates of the information can be distributed at smaller intervals and lower costs than new editions of books. By combining patient material from well-known experts on a video disc, these valuable collections can be used to the benefit of a large community of pathologists.

6.3 SECOND PART OF RESEARCH

6.3.1 Formalization of Knowledge

The second part of the research concerns the problem of knowledge acquisition. It addressed the question whether a tool can be developed for the acquisition and storage of formalized pathology knowledge, directly by the expert. The resulting knowledge base is aimed to serve the search for diagnoses based on findings in addition to consultation of information by diagnosis name.

Prior to designing a structure for the storage of formalized knowledge, it is important to decide if and how uncertainty will be dealt with, as uncertainty at the level of both the observations and the knowledge plays an important part in the diagnostic process. Therefore, in Chapter 4 several ways to express uncertainty and strategies for its propagation in conclusions have been compared. The Bayesian model appears to be the most powerful and predictable strategy. However, the main problems involved in reasoning with uncertainties are differences in mental conception of parameters to express uncertainty, scarcity of probabilistic data and the conditions for application of the strategies. These problems increase when the domain becomes larger. On the basis of these considerations the decision was made to restrict the expression of uncertainty to an indication of the frequency of occurrence of features in relation to diagnoses.

For the purpose of the acquisition and storage of pathology knowledge, in Chapter 5, a knowledge base structure has been introduced, which can contain such knowledge in the form of formalized features and their mutual relations. The latter are important for the expression of tissue architecture. A knowledge editor provides the expert with a menu-driven user interface to fill the knowledge base without the intervention of a knowledge engineer.

6.3.2 Evaluation

A pilot study has been carried out with three pathologists, who were asked to formalize two diagnoses: one by heart and one from a text. The aim was to gain more insight in the process of knowledge formalization, adaptations to be made in the system, and the potential use of the acquired knowledge. The experiment yielded the following preliminary conclusions:

- The sequence of actions was experienced as natural when entering knowledge by heart. When entering knowledge from the text, the formalization was experienced as a separate mental effort.
- The expression capability was sufficient.
- Experts adopt one or a combination of different starting points when specifying diagnostic features: (1) possibly occurring, (2) necessary for confirmation, and (3) sufficient for rejection.

Two major problems remain: the minimization of ambiguity and the limited consensus among experts. The first is that each single feature, ideally, can be characterized in one and only one way and at only one location in the knowledge base. Ambiguity, however, cannot be avoided completely: consensus decreases when formalizations become more detailed, but some features require much detail to be adequately characterized. The second problem might be dealt with in the future when methods become available for combining knowledge of more than one expert. It is important to realize that these two problems are intrinsic to pathology knowledge in particular and medical knowledge in general. Therefore, they should not be interpreted as limitations of the knowledge base design.

With the described tools for the interactive acquisition of formalized pathology, knowledge experts can directly build knowledge bases, which together can cover the whole field of pathology. The large expression capability of the knowledge base with respect to the characterization of diagnostic features and their mutual repationships minimizes the loss or distortion of knowledge. In addition, the knowledge base can serve as a basis for different kinds of diagnostic support, thereby eliminating the need for separate sources of knowledge for each kind of diagnostic problem. This is important as knowledge has to be acquired only once and updates of the knowledge base do not entail inconsistencies, which might occur when several knowledge bases were involved for the same domain. Finally, the formalized knowledge can easily be integrated with pictures on a videodisc, which creates, as compared to the DEW, the extra facility to access pictures by subject.

6.4 FUTURE RESEARCH

Several questions remain for further research. These are briefly discussed at the end of Chapter 5.

First, which level of detail is most appropriate for what parts of the knowledge in order to minimize ambiguity?

Second, is it feasible to combine knowledge bases created by more than one expert to a new knowledge base more complete than any of its composing parts? In that case, the new knowledge base has to be presented to a panel of experts to get insight into the feasibility of acquiring a more valid knowledge base.

Third, findings can be matched with features in the knowledge base to generate a differential diagnosis on the basis of those findings. Then, the matching strategy should be adapted to the starting point, which an expert adopts when selecting features for formalization. Hence, the question is: what is the most suitable combination of starting point and matching strategy in the determination of a set of diagnostic hypotheses based on findings?

6.5 CONCLUSION

Systems like the DEW and tools for the interactive formalization of knowledge open the way to powerful up-to-date diagnostic support, which is characterized by a large scope, versatility and flexibility. Integration with highly specialized systems and techniques can further extend the applicability from general diagnostic problems to highly specific ones. When this research is continued and experts are willing to participate on a broad scale, the object in view will eventually be achieved.

Samenvatting en conclusies

HET PROBLEEM

Zoals genoemd in Hoofdstuk 1, bestaat de taak van de patholoog-anatoom voor een belangrijk deel uit het classificeren en graderen van weefsel afwijkingen door middel van microscopisch onderzoek van cytologische en histologische preparaten. Bij het interpreteren van de beelden raadpleegt de patholoog regelmatig referentie kennis in de vorm van tekstboeken en collega's. In de pathologie bestrijkt een diagnose vaak een meer of minder groot gebied van histologische afwijkingen. Door verschillen in opleiding en ervaring gebruiken pathologen niet altijd dezelfde criteria voor het afgrenzen van diagnoses en kunnen zij ook verschillen in het interpreteren van bevindingen. Het gevolg daarvan is dat experts van mening kunnen verschillen over de diagnose van een casus.

Er zijn technieken ontwikkeld, al of niet met gebruik van computers, om tot een meer consistente diagnostiek te komen in de pathologie. Met betrekking tot hun doelstellingen verdelen wij deze technieken in twee hoofd categorieen: (1) het bevorderen van objectiviteit in het verkrijgen en interpreteren van diagnostische gegevens, en (2) het bevorderen van efficient gebruik van referentie kennis voor het typeren en graderen van afwijkingen. Tot de eerste categorie behoren morfometrie, beeldbewerking en statistische patroonherkenning. De tweede categorie omvat de automatisering van patient archieven, de opslag van visueel materiaal op beeldplaat en diagnostiek ondersteunende computer systemen. Hoewel deze technieken en computer toepassingen geschikt zijn voor hun doel, hebben ze allemaal een of meer van de volgende beperkingen:

(1) Beperkte toepasbaarheid en reikwijdte
(2) Teveel nadruk op diagnostische ondersteuning, die alleen op bevindingen is gebaseerd
(3) Beperkte mogelijkheden voor kennis verwerving

EERSTE DEEL ONDERZOEK

Raadplegen van Kennis

Het eerste deel van het onderzoek, zoals beschreven in dit proefschrift, betreft de tweede beperking. De vraag werd gesteld of het mogelijk is om de kwaliteit en de efficientie van het diagnostisch proces in de pathologie te verbeteren door het aanbieden van een raadpleegsysteem voor dagelijks gebruik. Zo'n systeem zou dan een grote hoeveelheid diagnostische informatie moeten bevatten, geintegreerd met een groot aantal geillustreerde patient geschiedenissen. Voor dit doel werd het Diagnostisch Encyclopaedisch Werkstation (DEW) ontwikkeld, hetgeen beschreven is in Hoofdstuk 2. Op dit moment bevat het DEW 85 diagnoses uit de ovariumpathologie, geillustreerd met 3.000 beelden, hetgeen voldoende was voor een evaluatie van het prototype. Het systeem draait op een IBM-compatibele PC, die verbonden is met een beeldplaatspeler. De informatie is toegankelijk via een zichzelf verklarend gebruikers interface.

Evaluatie

In Hoofdstuk 3 is in een cross-over experiment met twee groepen van pathologen een evaluatie gemaakt van het gebruik van het DEW tegenover het gebruik van boeken voor het raadplegen van kennis in de ovariumpathologie. Statistische analyse van de evaluatie leverde de volgende resultaten op:

- De tekstboeken gaven betere resultaten dan het DEW met betrekking tot het classificeren van diagnoses en het morfologisch benaderen van de juiste diagnose.
- Het DEW en de tekstboeken verschilden, hoewel niet significant, in het voordeel van de boeken met betrekking tot de klinische consequenties van diagnoses en de consensus van de deelnemers onderling.

Wat deze conclusies betreft moet worden vermeld, dat de opzet van het experiment nadelig was voor het DEW: door de beperkte beschikbaarheid van geschikt testmateriaal kon de grote verzameling verschillende kleuringen slechts ten dele worden benut.

De differentiatiaal-diagnostische ondersteuning, hoewel in het DEW uitgebreider dan in boeken, bleek niet voldoende compleet voor alle diagnoses van de test set. Dit wekte geen verbazing, aangezien de differentiaal-diagnose lijsten van de experts ook aanzienlijk verschilden. Toch bleek er voor iedere diagnose uit de test set een deelverzameling van differentiaal-diagnoses te zijn, die door de meerderheid van de experts werden genoemd. Differentiaal-diagnostische informatie voor systemen als het DEW zou dan ook minstens op zulke deelverzamelingen gebaseerd moeten zijn.

Van de kant van de deelnemers gaf de evaluatie de volgende resultaten:

- Sterke eigenschappen waren gebruikers-vriendelijkheid, de grote verzameling beelden en de differentiaal-diagnostische informatie.
- Suggesties voor verbetering betroffen de relatieve schaarste van overzichts beelden en de nog niet aanwezige informatie op het niveau van diagnose groepen.

Het bleek een groot probleem voor een patholoog om bij een onbekend beeld een geschikte ingangsdiagnose te vinden voor het raadplegen van het systeem. Dit probleem was verwacht omdat het DEW ontworpen was om geraadpleegd te worden met diagnose namen als ingang, zoals bij boeken gebruikelijk is. Het is echter denkbaar om het DEW uit te breiden met een lijst van geindexeerde sleutelwoorden, die dan gebruikt kunnen worden om gericht in te zoemen op een verzameling diagnoses voor overweging.

Als gecomputeriseerd systeem heeft het DEW verscheidene voordelen ten opzichte van boeken. Het kan een veel grotere hoeveelheid visuele en tekstuele informatie bevatten dan een boek en het kan in principe zelfs het volledige terrein van de pathologie dekken. Wat het laatste betreft is het vermeldenswaard dat het DEW over een goed ontwikkeld programma beschikt voor de invoer van gegevens. De informatie in de database is sterk gestruktureerd en de toegang tot die informatie is flexibel omdat daarvoor meerdere ingangen mogelijk zijn. Vooral de mogelijkheid om informatie te combineren vanuit verschillende locaties in de database is van fundamenteel belang bij het ondersteunen van differentiele diagnostiek. Bovendien kunnen nieuwe edities worden verspreid met kleinere intervallen en tegen lagere kosten dan het geval is bij boeken. Door het patienten materiaal van erkende experts te combineren op beeldplaat, kunnen waardevolle verzamelingen door een veel bredere groep van pathologen worden benut.

Tweede deel Onderzoek

Het Formaliseren van Kennis

Het tweede deel van het onderzoek betreft het probleem van kennis verwerving. De vraag werd gesteld of een systeem kan worden ontwikkeld voor het verwerven en opslaan van geformaliseerde pathologie kennis, direct door de expert. Het verkregen kennisbestand is niet alleen bedoeld voor het raadplegen van diagnostische informatie op diagnosenaam, maar ook voor het genereren van een differentiele diagnose op basis van bevindingen.

Alvorens een struktuur te ontwerpen voor het opslaan van geformaliseerde kennis, is het belangrijk om te bsluiten of en hoe er met het fenomeen onzekerheid wordt omgegaan. Onzekerheid speelt namelijk een belangrijke rol in het diagnsotisch proces: zowel op het niveau van de observaties als op het niveau van de kennis. Daarom zijn in Hoofdstuk 4 een aantal modellen, om onzekerheid uit te drukken en te verwerken in conclusies, met elkaar vergeleken. Het model van Bayes lijkt het meest krachtig en voorspelbaar. De grootste problemen bij het redeneren met onzekerheden zijn het feit dat experts verschillende voorstellingen hebben van de parameters waarin onzekerheden worden uitgedrukt, de schaarste van statistische gegevens en de voorwaarden waaraan voldaan moet worden om de modellen te mogen toepassen. Deze problemen worden groter naarmate het domein van toepassing groter wordt. Op grond van deze overwegingen werd besloten om het uitdrukken van onzekerheid te beperken tot een aanduiding van de frequentie, waarmee een bepaald verschijnsel bij een bepaalde diagnose wordt aangetroffen.

Voor het verwerven en opslaan van pathologie kennis is in Hoofdstuk 5 een struktuur geintroduceerd voor een kennis bestand, dat kennis kan bevatten in de vorm van geformaliseerde verschijnselen in hun onderlinge samenhang. Die samenhang is van belang voor het uitdrukken van weefselstrukturen. Een menu-gestuurde kennis editor biedt de expert de mogelijkheid om zonder hulp van een systeemdeskundige het kennis bestand te vullen.

Evaluatie

In een verkennend onderzoek werden drie pathologen gevraagd om twee diagnoses te formaliseren: een uit het hoofd en een van een tekst. Het doel was om meer inzicht te krijgen in het proces van kennis formalisatie, noodzakelijke aanpassingen aan het systeem en mogelijkheden om de geformaliseerde kennis te gebruiken. Het experiment leverde de volgende resultaten op:

- Bij het invoeren van kennis uit het hoofd werd de volgorde van handelingen als natuurlijk ervaren. Dit in tegenstelling tot het invoeren op basis van een tekst, waarbij het formaliseren als een aparte mentale inspanning werd ervaren.
- De uitdrukkingskracht van het systeem was voldoende.
- Experts hanteren verschillende uitgangspunten bij het specificeren van diagnostische verschijnselen: (1) mogelijk voorkomend bij de diagnose, (2) noodzakelijk voor bevestiging van de diagnose en (3) voldoende voor verwerping van de diagnose.

Twee duidelijke problemen doen zich voor: het bevorderen van eenduidigheid en de beperkte consensus tussen experts onderling. Het eerste betekent dat, in ideale zin, ieder verschijnsel op precies een manier kan worden uitgedrukt en op slechts een plaats in het kennis-bestand kan worden opgeslagen. Absolute eenduidigheid is niet haalbaar: de consensus neemt af wanneer beschrijvingen meer gedetailleerd worden, maar sommige verschijnselen vereisen een gedetailleerde beschrijving. Met het tweede probleem kan het systeem in de toekomst misschien rekening houden als methoden beschikbaar komen om kennis van meerdere experts te combineren. Het is belangrijk, dat men zich realiseert dat beide problemen niet alleen inherent zijn aan pathologie kennis, maar aan medische kennis in het algemeen. Daarom moet men deze problemen niet interpreteren als beperkingen van het ontwerp van het kennis-bestand.

Met behulp van het beschreven systeem voor het interactief verwerven van geformaliseerde pathologie kennis kunnen experts zelf kennis-bestanden bouwen, die samen het gehele terrein van de pathologie kunnen dekken. Door de grote uitdrukkingskracht van het systeem met betrekking tot het beschrijven van verschijnselen en hun onderlinge samenhang, treedt slechts een gering verlies en weinig vervorming van de kennis op. Bovendien kan het kennis-bestand dienen voor meerdere soorten van diagnostische ondersteuning, zodat niet voor ieder soort van diagnostisch probleem een aparte kennis bron nodig is. Dit is belangrijk omdat de kennis nu slechts eenmaal verworven hoeft te worden en veranderingen niet de inconsistenties met zich meebrengen, die kunnen optreden bij het gebruik van meerdere kennis-bestanden in hetzelfde domein. Tenslotte kan de geformaliseerde kennis gemakkelijk worden geintegreerd met beelden op een beeldplaat, hetgeen vergeleken bij het DEW zelfs de extra mogelijkheid biedt om beelden op te vragen op onderwerp.

TOEKOMSTIG ONDERZOEK

Een aantal vragen blijven liggen voor verder onderzoek. Deze zijn kort besproken in Hoofdstuk 5.

Ten eerste, welke maat van detail is het meest geschikt voor welke delen van de kennis om een zo eenduidig mogelijk kennis-bestand te krijgen?

Ten tweede, is het haalbaar om kennis-bestanden van meerdere experts te combineren tot een nieuw kennis-bestand, dat completer is dan de samenstellende delen? In dat geval moet het nieuwe kennis-bestand ter beoordeling worden aangeboden aan een panel van experts om inzicht te krijgen of het mogelijk is om op deze wijze een volwaardiger kennis-bestand te krijgen.

Ten derde, bevindingen kunnen worden vergeleken met geformaliseerde diagnostische verschijnselen in het kennis-bestand om tot een differentiele diagnose te komen op basis van die bevindingen. Bij die vergelijking moet men rekening houden met het standpunt dat de expert bij het formaliseren van die kennis heeft gehanteerd. De vraag is dan ook: hoe moeten uitgangspunt en vergelijking op elkaar worden afgestemd om een goed differentiaal-diagnostisch resultaat te krijgen?

CONCLUSIE

Systemen als het DEW voor het raadplegen van kennis en mogelijkheden voor het interactief verwerven van kennis openen de weg naar krachtige diagnostische ondersteuning, die gekarakteriseerd is door een grote reikwijdte, veelzijdigheid en flexibiliteit. De toepasbaarheid kan verder worden uitgebreid van algemene diagnostische problemen naar meer specifieke taken door integratie met sterk gespecialiseerde technieken en systemen. Als experts bereid zijn om op grote schaal bij te dragen aan de voortzetting van dit onderzoek zal het beoogde doel uiteindelijk worden bereikt.

Curriculum vitae

Astrid van Ginneken was born on June 5, 1956 in 's Gravenhage, The Netherlands. She completed grammar school in 1974 in Amsterdam. In 1975 she moved to Venezuela with her parents, where she passed the first year examination of "Exact Sciences" in summer 1976 at the Universidad Simon Bolivar in Caracas. In fall of the same year she returned to Holland and started with mathematics, main subject informatics, at the Free University in Amsterdam. After her bachelors degree in 1978 she went to medical school at the same university. In 1979-1980 she gave lessons in anatomy as a student assistent. After recieving her M.D. in 1985 Astrid started as a Ph.D. student at the Department of Medical Informatics. The first part of the research, reported on in this thesis, was carried out at the Free University in Amsterdam and the second part at the Erasmus University in Rotterdam. Since summer 1987 she participates in research on killer whales as volunteer staff member of Orca Survey at The Center for Whale Research in Washington State. Upon finishing the Ph.D. project, Astrid van Ginneken continues research as scientific staff member of the Department of Medical Informatics at the Erasmus University.