Evaluation of a Diagnostic Encyclopedia Workstation for Ovarian Pathology

A. M. VAN GINNEKEN, MD, PHD, J. P. A. BAAK, MD, PHD, W. JANSEN, MD, PHD, and A. W. M. SMEULDERS, PHD

The Diagnostic Encyclopedia Workstation (DEW) is a computer system that provides completely integrated pictorial and textual information as reference knowledge in the field of ovarian pathology. The textual component comprises information per diagnosis such as descriptions of 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 evaluates to what extent the system influences the diagnostic process in efficiency and outcome. Therefore, two groups of six pathologists each, covering a wide 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 rank sum test with respect to five different viewpoints: classification, morphology, clinical consequences, duration of diagnostic process, and consensus among the participants. The results are discussed and it is concluded that classification and morphology showed better results when books were used. The evaluation experiment was, however, very rigid and negatively biased with respect to the DEW system. Positive aspects of the encyclopedia are the easy access to diagnostic and differential diagnostic information and the large set of illustrations. Insight is acquired with respect to existing bottlenecks and how they may be overcome. HUM PATHOL 21:989–997. © 1990 by W.B. Saunders Company.

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] Since some diagnoses are morphologically similar, a pathologist may need reference knowledge to confirm a diagnosis or to find criteria to distinguish difficult cases. Reference knowledge comprises the consultation of experts, documented cases, and, especially, pathology books.

The consultation of books can be very laborious for several reasons. First, in order to keep their size and price reasonable, books cover a limited number of diagnoses with a few (mostly black and white) photographs illustrating each diagnosis. Second, the differentation between morphologically similar diagnoses may be difficult, since differential diagnosis lists and uniquely defined criteria for differentiating among possible diagnoses are scarce. Third, books are diagnosis oriented; access to the information via findings is very limited. Fourth, the field of pathology is so extensive that information in written sources has to be restricted. 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 encyclopedia and does not actively generate diagnostic hypotheses as expert systems do. It has been developed to offer the pathologist easy and flexible access to reference knowledge as laid down in books, extended with a large volume of color illustrations, 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 ovarian pathology, including all common and many rare cases (see next section), a volume that we considered sufficient to test the system's present performance.

The following sections include a short description of the DEW, an explanation of the set-up of the evaluation experiment, a discussion of the results, and our conclusions.

DESCRIPTION OF THE DIAGNOSTIC ENCYCLOPEDIA WORKSTATION

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

The DEW runs on an IBM-AT or compatible machine with a 20-Mb hard disk, 640-Kb internal memory, a Hercules monochrome graphics card, and an RS-232-C serial interface. The computer controls a videodisc player via the serial port.* The monochrome monitor displays textual information and a

From the Department of Medical Informatics, Erasmus University, Rotterdam, and Laboratory of Pathology, Free University, Amsterdam, The Netherlands. Accepted for publication April 4, 1990.

Supported by grant Praeventiefonds PF 28-1207.

Key words: pathology, diagnostic support, computerized atlas, reference knowledge.

Address correspondence and reprint requests to A. M. van Ginneken, MD, Department of Medical Informatics, Erasmus University, PO Box 1738, 3000 DR Rotterdam, The Netherlands.

0046-8177/90/2110-0002$5.00/0

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

video monitor displays the color illustrations from the videodisc. The video signal was taken from photographic slides.

At present, the database contains information concerning 85 ovarian tumors. The tumor classification of the World Health Organization is the basis for the order of the diagnoses in the database.[6] The 85 diagnoses cover 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. Differential diagnostic information is available in the form of differential diagnosis lists and tables that compare pairs of similar diagnoses. The photographs illustrating these diagnoses total approximately 3,000, divided among 158 cases. The illustrations are indexed by diagnosis, case, stain, and magnification.

The user interface of the DEW is mouse driven.[7] The first few screens serve as the table of contents 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. Once a diagnosis has been selected, a window with the microscopic 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 and 2). Small squares in the text are "sense fields," which result in the display of an illustration when selected with the mouse. In this way, the user can call for illustrations of characteristics of a diagnosis that are described in the text preceding the sense field. At the top of the screen, the choices that lead to the selection of the current diagnosis are visible. To the left of the text window is a list of other categories of information about the selected diagnosis, with an indication as to which of them are available. 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.

To support the differentiation between morphologically similar diagnoses, the user can ask for a differential diagnosis list 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.

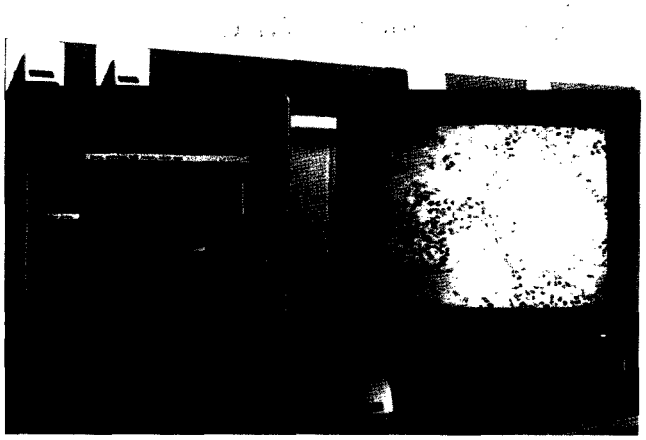

FIGURE 1. An overview of both monitors during a session with the DEW. On display are a microscopy description and an image illustrating a feature of the diagnosis.

MATERIALS AND METHODS

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 a wide 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. 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 archives of the OTC (the Dutch National Ovarian Tumor Committee). However, some of the cases of the OTC archives may have been diagnosed without complete consensus.

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 1 shows the cross-over experiment schematically. Session 1 was always followed by session 2. The OTC diagnoses of the cases included in the experiment are listed in Table 2.

In both sessions the pathologists were offered a list containing the names of the diagnoses covered by the system. Participants were allowed to use the list to find 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, ie, 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 an approximately 15-minute demonstration by the first author, followed by some time for the candidate to become familiar with the system. This required an average of 5 minutes. No time limits were imposed on the participants for the completion of the 13 cases of each session. The cases were offered in a fixed order. During both sessions, the first author (A.M.v.G) observed the participants while making notes of the following: time when the participant started with a case, times of every action of the candidate (looking through the microscope, looking at the list of diagnoses, consultation of

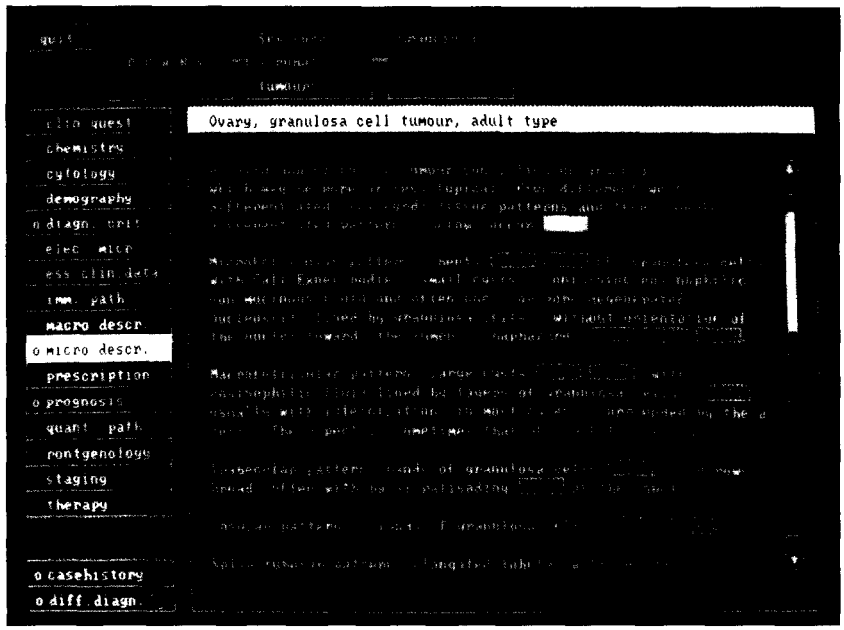

FIGURE 2. Close-up of the screen layout at diagnosis level. The microscopy description is displayed by default. Note the sense fields, the scroll bar at the right, the optional items of information at the left, and the path through the classification tree at the top.

the DEW, or consultation of a book and the chapter that was used), and time when the final diagnosis was made. The observer also recorded whether a diagnosis was made in doubt. During consultation, the DEW system created a log file containing all selections made by the user.

Viewpoints for Evaluation

The experiment permits the evaluation of the following question: Does the diagnostic support of the DEW differ from that 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 learning experience. When the analysis yields equal learning in both groups after the first session the comparison of the DEW system versus written sources is not affected: the relative difference between the books and the system remains the same in both sessions. Whether learning is equal and/or significant can be evaluated with the Wilcoxon rank sum test.

There are several considerations in evaluating 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 morphologic similarity between the diagnoses of the participants and the OTC diagnoses?

What are, as compared with the OTC diagnoses, the

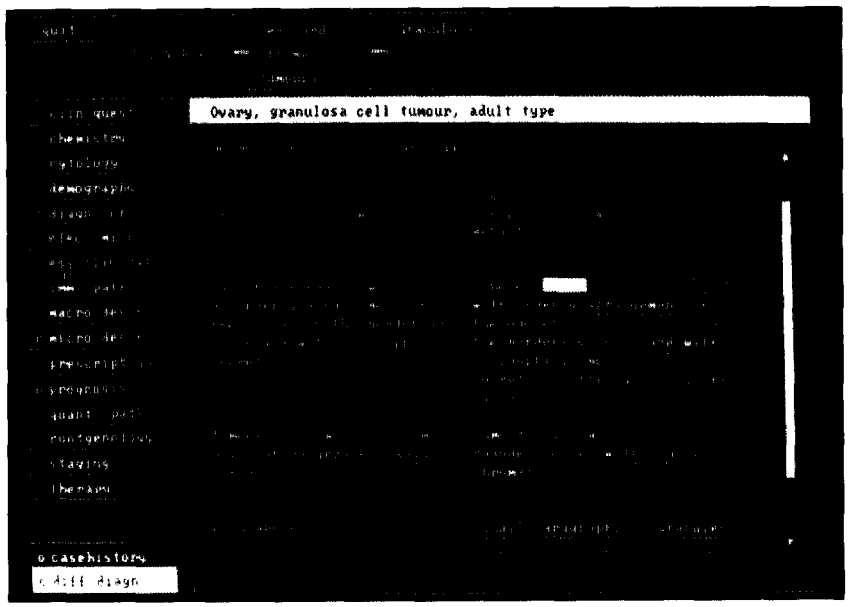

FIGURE 3. A table is shown that lists the common and differentiating features of two morphologically similar diagnoses. Here, "carcinoid" is the selected diagnosis from the differential diagnosis list of the "adult granulosa cell tumor." It is possible to switch directly to "carcinoid" by selection of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

TABLE 1. The Order in Which the Participants Used the Diagnostic Encyclopedia Workstation

	Session 1 (Test Set A)	Session 2 (Test Set B)
Group 1	Books	DEW
Group 2	DEW	Books

Note: Both groups started with slide set A.

clinical consequences of diagnoses that differ from the OTC diagnosis?

Each of these viewpoints represents a different ordering of diagnoses. The ordering in the WHO classification of ovarian tumors is a hybrid mixture in the sense that the division into major diagnostic groups reflects the origin of the tumors, whereas the minor divisions are based on morphologic features. In consequence, some diagnoses show more morphologic similarity with tumors in other diagnosis groups 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. Morphologic similarity, in turn, does not necessarily provide accurate information about the clinical 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 criteria for evaluation (classification, morphology, and consequences), we have also analyzed the degree of consensus among the participants and the efficiency of the books versus the DEW system, based on the time spent on each case.

Statistical Analysis: Scoring

For statistical evaluation of the diagnostic results, a score is assigned to each diagnosis, given by the participants, to express its degree of variance from the "gold standard." A separate score is assigned for each viewpoint of evaluation. The scores expressing the degree of variance from the OTC diagnosis fulfill the properties of the metric concept of "distance."

TABLE 2. The "Gold Standard" Diagnoses of the 13 Test Cases

Diagnoses in the Test Set	Order: Session 1 (Test Set A)	Order: 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

Note: The test cases are listed in the order in which they were presented in sessions 1 and 2. The diagnoses may have been made by the OTC without complete consensus.

The *classification score* is based on the distance between the various levels in the classification tree of ovarian tumors. "Ovary" is the first level, groups and subgroups form intermediate levels, and diagnoses are the end level in 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 first step is to identify the smallest diagnosis group that the diagnosis of the participant and the optimal diagnosis have in common. The next step is to establish if and which of these two diagnoses is closer to the common group. The score is then equal to the difference between the level of the common group and the closer diagnosis. For example, when a participant diagnoses a serous adenofibroma and the OTC diagnosis is a borderline endometrioid tumor, the latter is closer to the common epithelial tumors group (Fig 4). Starting upward from a borderline endometrioid tumor, it is two steps to the common epithelial tumors group. Consequently, the score assigned to the diagnosis of the participant is 2.

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

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 with the OTC diagnosis.

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

The *time score*, used to compare the times required to make a diagnosis, was equal to the number of seconds that elapsed between starting with a case and making the final diagnosis.

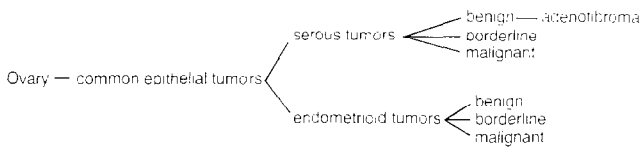

FIGURE 4. When a serous adenofibroma is diagnosed and the OTC diagnosis is a borderline endometrioid tumor, then the former is three levels and the latter is two levels away from the common group "common epithelial tumors," yielding a classification score of 2.

TABLE 3. Differential Diagnosis Lists

Diagnosis	DDs Arranged by Number of Pathologists Mentioning That DD								Total of Different DDs
	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
Homologous 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 transformation	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

Note: All eight experts together mentioned 31 DDs for the OTC diagnosis "insular carcinoid." Of these 31 DDs, 19 were mentioned by one (not necessarily the same) expert, six by two experts, three by three experts, etc. However, none of the DDs was mentioned by all of the experts. Hence, the first column represents minimal consensus and the eighth column represents maximal consensus with respect to morphologic similarity.

Abbreviation: DD, differential diagnosis.

The Wilcoxon rank sum test[11] was used for the statistical analysis. It is important to note that the Wilcoxon rank sum test is used to detect the presence or absence of a significant difference between two small sets of data. It does not provide information about the degree of difference between the two sets. When using the Wilcoxon rank sum test, the null hypothesis was the absence of a significant difference in diagnostic results between the DEW system and the books. As a level of significance, we used 5%. Hence, less than 5% means a rejection of the assumption that the DEW system and the written sources are equivalent. The scores of the participants reveal whether books or the DEW system must be favored.

RESULTS

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.

A significant learning effect was found for both the criterium of classification and that of clinical consequences. From the criteria of both classification and morphology, the books were found to provide better results than the DEW system. It is important to realize that the statistical analysis of the DEW system versus books with respect to clinical consequences is less sensitive than the analyses for classification and morphology. This is due to the fact that there are fewer different therapies available than there are diagnosis names.

As to consensus, there was an unequal learning effect in favor of the DEW. Consequently, only the data of the first session could be used for the evaluation of written sources versus the DEW. It is obvious that no significant difference was found.

During the sessions, neither the books nor the DEW were always consulted. Table 5 shows the number of correct classifications and misclassifications in relation to the use of books or the DEW system. Note that the majority of cases that were diagnosed without the use of reference knowledge were classified correctly as opposed to the cases that were classified with the use of books or the DEW.

Even when the books or the DEW system were used, the participants did not always consult the "correct" diagnosis. With respect to the use of the

TABLE 4. The Average Scores of All of the Pathologists for Each Criterium of Evaluation

Criterium of Evaluation	Average G1		Average G2		Significantly in Favor of:
	Books	DEW	DEW	Books	
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 (Sec)	5,567	5,323	4,681	4,060	—

Note: Good results are reflected by low average scores in the upper three rows, high average scores in the fourth row, and low scores in the last row.

TABLE 5. The Number of Correct Classifications and Misclassifications

	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

Abbreviations: E, equal to gold standard; D, different from gold standard.

DEW, the log files revealed that for only 18 of the 56 misclassifications with the system was the "correct" diagnosis consulted. It was not considered feasible to collect the same information from the sessions with the written sources since this would require allowing the participants to mention every diagnosis they consulted.

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

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 DEW system. In this way, we gained more insight into the weaknesses and strengths of the system. Apart from the use of books or the DEW, all participants mentioned that the sessions differed considerably from the normal diagnostic routine. Since it was not possible to request additional slides of a case or to defer a diagnosis until the next day, participants sometimes felt themselves forced to make diagnoses for which they would not have taken responsibility in real practice. In the following discussion, the positively valued properties of the DEW are discussed first, followed by suggestions for improvement.

The first strength of the DEW is the easy access to the information; only a few mouse clicks are necessary to consult diagnosis information and illustrations. 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' instruction to become acquainted with the DEW.

The second strength of the DEW concerns the availability of a large volume of color illustrations.

TABLE 6. The Number of Diagnoses That Were Made With and Without Doubt

	Certain	Uncertain
Texts	82	5
DEW	87	13

Note: The remaining 125 diagnoses were made without the use of reference knowledge.

The average photographic quality of the illustrations was considered comparable to that of the illustrations found in written sources with the exception of overviews with low contrast, even taking into account that they are displayed on a television monitor. The participants also appreciated the availability of the case illustrations, sorted by diagnosis, stain, and magnification.

The third strength of the DEW is the availability of differential diagnosis lists and criteria, which were regarded as valuable additions when compared with books.

Finally, the majority of the participants considered the system very valuable for training in pathology. Specific topics and their diagnostic problem areas can be studied systematically.

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 may be sufficient to reject a diagnosis. More overviews are needed, especially in the beginning of the diagnosis description. The relative scarcity of overviews is due to the fact that magnifications of × 2.5 or smaller with low contrast require a higher resolution display than a normal video signal can offer. More magnifications of × 10 could be added, and even magnifications of × 2.5 of moderate contrast might be useful for a first impression.

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

Finally, the participants almost unanimously expressed the wish to have access to illustrations sorted by stain and magnification as well as by sense fields. It happened several times that the participants "tried" many sense fields to find an overview or an illustration 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 illustrations on the diagnoses screen will allow access that is more adjusted to the needs of the user.

DISCUSSION

It is not sufficiently satisfactory that the DEW and the books yield equivalent results with respect to

clinical consequences. Since classification is the basis for therapy selection, it is important to strive for optimal classification of diseases.

Role of the Diagnostic Encyclopedia Workstation 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 is the differences in education and experience among experts. Several times, we observed that participants using the same text as a reference diagnosed a case differently. Apparently, they put different emphasis on the morphologic phenomena in the histologic 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 confirmation of a diagnosis: the histologic variability of diagnoses, potential diagnoses together with the criteria to differentiate among them, and differences in clinical consequences among diagnoses under consideration.

Causes of Misdiagnosis

Prior to discussing potential causes of misclassification, it is important to realize that the experiment was negatively biased against the DEW, the main reason being the availability of only one hematoxylin-eosin–stained slide in the majority of the test cases. This posed difficulties in diagnosis making which, in practice, would have been easily solved with the availability of additional stains. The large set of 3,000 illustrations available in the DEW also could not be used to its full advantage. Many cases for which several stains could be obtained were already being used for the videodisc, so we had to accept a selection of the remaining suboptimal cases for the experiment. It is also important to realize that the learning effect of using DEW was significant, which means that the difference between the DEW and written sources may decrease if the DEW were evaluated in another experiment with users who are familiar with its contents. However, the insights gained with respect to the functioning of the DEW are also applicable to situations in which more stains are available. In the following discussion, we concentrate on potential causes of misdiagnosis to gain insight into possible improvements of the DEW.

First, there is the problem of consensus.[12,13] Some diagnoses that we have classified as misdiagnoses on the basis of the "gold standard" may, in fact, be judged correctly by one or more experts. It is interesting to mention the fact that the Sertoli cell tumor in set A received eight different diagnoses. Apparently, the slide showed features that fit several other diagnoses. Other examples 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." Since all participants had the opportunity to reexamine slides after being informed of the OTC diagnosis, it was clear that no major features were overlooked. It is, however, conceivable that the selection of the slides was biased toward relatively benign tumor characteristics.

Second, a possible negative effect on the diagnostic results concerns the illustrations. As the participants mentioned, some of the illustrations did fit their diagnosis, but did not characterize it. Such illustrations serve the purpose of completeness with respect to all possible histologic 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 photographs from other diagnoses. In such cases, it is the combination of illustrations that characterizes the histologic image. However, the use of illustrations from other diagnoses carries the risk of misinterpretation when they show more phenomena than the one(s) they were meant to display. In general, when nonspecific illustrations dominate, insufficient scanning of the available photographs may cause the user to reject the diagnosis too soon.

Third, we observed that in 38 of 56 misdiagnoses made with the aid of the DEW system, the correct diagnosis was not consulted. A possible cause for not consulting the correct diagnosis may be found in the contents of the differential diagnosis lists per diagnosis. The examples in the Appendix and in Table 3 show that experts vary in their opinions on morphologic similarity among diagnoses. Note that for nine diagnoses of the test set, the intersection of the differential diagnosis lists of the eight consulted pathologists is empty. In the same way, the differential diagnosis lists in the DEW differ from those made by the experts. As a consequence, it may be that the user consults a diagnosis that is considered to be morphologically similar to the correct diagnosis by some of the experts but not by the system. The differential diagnosis lists of that diagnosis will, therefore, not help the user to find the correct diagnosis. The scope of the differential diagnostic information is not wide enough. On the other hand, a complete differential diagnosis list will soon become impractically long (if feasible at all), and making tables for all possible combinations is a huge task. A different situation in which the differential diagnosis lists of the DEW 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 differential diagnosis lists, but the fact that unfamiliarity with the case or a misinterpretation of the observations causes the

user to consider the wrong diagnoses. Thus, it depends on the user as to whether the insight emerges that a different entry into the system is necessary. These problems, which also occur when using written sources, need better support.

It is important to keep in mind that mistakes are also made when no reference knowledge is consulted. In general, the search effort of the user is crucial for the diagnostic result; one user may search until a diagnosis that fits moderately with the observations is found, whereas another user may search for the perfect fit. Therefore, it is important to realize that long differential diagnosis lists and large numbers of illustrations, which require extensive scanning effort from the user, may have a negative influence on the diagnostic result.

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. With regard to the differential diagnosis lists, it is laborious to "try" all possibilities. The availability of a few overviews for each diagnosis on a differential diagnosis list offers the possibility of scanning the list prior to making a selection. Generation of differential diagnosis lists based on findings would be a major step forward. However, this requires a formal representation of diagnosis descriptions, in which each finding is separately accessible and its relation to other findings is explicitly known. A method to acquire formal diagnosis descriptions directly from the expert system has been developed and is described.[14]

In the diagnosis texts, the illustrations are only ccessible via sense fields. A more directed scanning of the illustrations would be facilitated if they were also available sorted by diagnosis, laboratory technique, and magnification (as with the slides of the cases). In general, the number of overviews should be increased.

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

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: (1) Written sources offer superior support from the criteria of classification and morphological similarity of diagnoses with the "gold standard", and (2) written sources and the DEW did not differ significantly with respect to the clinical consequences of misdiagnosis, mutual consensus among the participants, and duration of the diagnostic process.

However, it should be kept in mind that the evaluation experiment was tightly controlled and negatively biased against the DEW system; the large set of DEW illustrations could not be used to full advantage and the users were not familiar with the contents of the DEW.

To set goals for improvement of the system's support in the classification process, we analyzed its strengths and weaknesses. From this evaluation, it turned out that strong properties are easy, mouse-driven access to the information, the presence of many color illustrations, and the availability of differential diagnosis lists and criteria. The most prominent aspect to be improved is support in determining the correct set of diagnoses for consideration. Useful extensions include the availability of criteria to differentiate among diagnosis groups, the availability of histologic overviews in differential diagnosis lists, and the possibility of accessing the diagnosis illustrations sorted by stain and magnification in addition to the sense fields in the text. A major step forward would be the generation of differential diagnosis lists based on findings, which requires a formal representation of diagnosis descriptions.

Leaving the diagnostic responsibility with the user, the DEW system is intended to make the diagnostic process less dependent on personal factors such as the user's preexistant 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. In its present state, the system constitutes a valuable tool for training purposes. Provided that the design of the DEW is improved as indicated and its contents are extended to other aspects of pathology, it has the potential of becoming a welcome addition to daily diagnostic practice.

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

APPENDIX. The Columns Represent the Differential Diagnoses as Specified by Each of Eight Gynecopathologists

Test Diagnosis: Mucinous Cystadenocarcinoma, Well-differentiated

Diagnosis	Pathologists								
	1	2	3	4	5	6	7	8	DEW
Serous cystadenocarcinoma, well-differentiated		*		*					*
Serous cystadenocarcinoma, moderately differentiated				*					
Serous adenofibroma, malignant			*						
Mucinous cystadenoma	*	*		*		*	*		
Mucinous cystadenofibroma				*		*			
Mucinous cystadenoma, borderline	*	*	*	*	*	*	*	*	*
Mucinous cystadenofibroma, borderline				*		*			*
Mucinous cystadenocarcinoma, moderately differentiated	*		*	*	*	*	*		*
Mucinous cystadenocarcinoma, poorly differentiated				*					
Endometrioid cystadenoma			*						
Endometrioid cystadenoma, borderline			*						
Endometrioid cystadenocarcinoma, well-differentiated		*	*	*	*	*	*		*
Endometrioid cystadenocarcinoma, moderately differentiated			*	*			*		
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			*				*		

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. New York, NY, Springer Verlag, 1983, pp 6-14

2. Giard RWM: Inflammatoire ziekten van het colon, PhD. thesis, chapter 2. Leiden, The Netherlands, Rijksuniversiteit, 1986

3. Connelly DP, Johnson PE: The medical problem solving process. HUM PATHOL 11:412-419, 1980

4. Jansen W, Baak JPA, Van Ginneken AM, et al: Diagnostic encyclopedia workstation: An interactive system for diagnostic support, in Wamsteker K, Jonas U, van der Veen G, et al (eds): Imaging and Visual Documentation in Medicine. Elsevier, Amsterdam, The Netherlands, 1987, pp 777-780

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

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

7. van Ginneken AM, Smeulders AWM, Jansen W, et al: Design of the Diagnostic Encyclopedia Workstation (DEW). Comp Biol Med (in press)

8. Blaustein A: Pathology of the Female Genital Tract (ed 3). Berlin, Springer Verlag, 1987

9. Scully RE: Tumors of the Ovary and Maldeveloped Gonads. Washington, DC, Armed Forces Institute of Pathology, 1979

10. Fox H, Langley FA: Tumours of the Ovary. London, UK, Heineman, 1976

11. Hills M, Armitage P: The two-period cross-over clinical trial. Br J Clin Pharmacol 8:7-20, 1979

12. Baak JPA, Langley FA, Talerman A, et al: Interpathologist and intrapathologist disagreement in ovarian tumor grading and typing. Analyt Quant Cytol 8:354-357, 1986

13. Baak JPA, Oort J: The case for morphometry in diagnostic pathology, in Baak JPA, Oort J (eds): A Manual of Morphometry in Diagnostic Pathology. Berlin, FRG, Springer-Verlag, 1983, pp 2-5

14. van Ginneken AM, Jansen W, Smeulders AWM, et al: A method for the acquisition of formalized knowledge in pathology. Meth Inform Med 29:182-191, 1990