Computer-based Diagnostic Support Systems in Histopathology: What Should They Do?

G.T.N. Burger1,3, A.M. Van Ginneken3, H. Hollema2

1Dept. of Pathology, SAZINON Foundation, Bethesda Hospital Hoogeveen; 2Dept. of Pathology, University Hospital Groningen; 3Dept. of Medical Informatics, Erasmus University Rotterdam.

Abstract

In many fields in histopathology diagnosis making is notoriously difficult. We explored the diagnostic process in the area of CL and related disorders. The field of cutaneous lymphomas (CL) and borderline lesions is complex and representative for the type of diagnostic problems encountered. A review of diagnostic support systems in diagnostic pathology revealed that the usefulness of these systems has been disappointing. We contribute this to the fact that these systems target only part of the diagnostic process.

A tentative model of diagnosis making in pathology is presented. This model assumes a two-step process, from observation to feature recognition and from features to diagnosis.

In a retrospective study of existing skin biopsy pathology reports we assessed detail and scope of the histological descriptions. In a second, prospective study, a pathologist panel described a set of 16 skin biopsies using a standard set of descriptors.

The retrospective study showed a large variability in the nature and details of described features, whereas the prospective study showed lack of consensus regarding both feature descriptions and diagnostic category. Both studies provide an indication that lack of consensus in feature recognition may be an important contributor to lack of consensus at the diagnostic level.

Diagnostic expert systems target the step from feature to diagnosis. Evidently different input into such systems produces different output. We conclude that support of the feature recognition step can contribute to better diagnostic consensus because of more uniform interpretation of observations.

Keywords:
Pathology; Computers; Diagnosis; Cutaneous lymphomas

Introduction

In the past, the diagnosis-making strategies of physicians in general 1 and more recently those of pathologists 2-4 have been studied. In the field of histopathology, as in other medical disciplines, there are several areas where differential diagnosis frequently poses a problem and where misdiagnosis may have severe prognostic consequences for the patient. Such fields are melanomas5, mammary lesions, cervical dysplasias 6, and cutaneous lymphomas (e.g. 7-9). The aforementioned research led to the expectation that computer-based diagnostic systems would make a significant contribution to clinical practice 10. A classical diagnostic expert system contains a knowledge base, in which relationships between diseases and clinical and/or laboratory findings are formally represented. The inference engine generates uses these relationships to suggest a diagnosis or a ranked list of possible diagnoses on the basis of entered observations. Several expert systems involved the field of pathology 11-13.

This kind of systems has not found widespread use in actual practice situations for several reasons:

Scope. The variety of diseases with which a general pathologist is confronted is larger than the scope of a diagnostic expert system. It is therefore difficult to know whether the absence of a disease in the differential diagnosis (DD) is a result of the inference process or that the disease is not covered by the system’s knowledge base.

Maintenance. Keeping an expert system’s knowledge base up to date is a very labor-intensive task that requires frequent validation. Furthermore, expansion of a knowledge base poses a challenge of keeping it consistent.

Inference. In order to produce a ranked list of disease entities the system uses heuristic expressions of uncertainty or probabilistic parameters. With heuristic strategies it is not always transparent how observations contribute to the ranking of the disease entities. Probabilistic models often
lack values for a posteriori probabilities as an expression of interdependencies. The requirement of normalization is therefore often violated.

Another problem is the role of prevalence in probabilistic models. The pathologist is near the root of the diagnostic tree, where the list of possible diagnoses should have been significantly narrowed down in the preceding medical screening process. He is supposed to make an unequivocal diagnosis that represents the best fit, as opposed to a ranked order of possible diagnoses. Thus, the ranking of possible diseases should primarily be based on sensitivity and specificity rather than on prevalence.

Stand-alone. Although an expert system may offer the option of report generation, they are not designed as an efficient reporting tool for routine use. With stand-alone systems data entry implies extra effort which is often unacceptable for the time-pressured physician. Furthermore, the benefit of use depends on the initiative of the user. When an electronic reporting system can feed the expert system, effort is reduced and discrepancies between observations and diagnoses can be actively detected.

Diagnostic expert systems produce diagnostic suggestions on the basis of entered findings. Different sets of observations are likely to produce different diagnostic output. The diagnostic process also involves the interpretation of observations in histological findings. To appreciate the potential benefit of computerized diagnostic support tools, insight in the full diagnostic process is crucial.

Steps in the diagnostic process

In the diagnostic process in pathology, we can discern two main steps: 1) observation - description, 2) description - diagnosis. These steps may produce different results. As shown in Figure 1, scenario A represents the most desirable situation: 2 pathologists reviewing the same slide, observe the same set of features ‘f1’ and arrive at the same diagnosis ‘D1’. In scenario B there is dissent on step 2: The same set of features is observed, but each observer arrives at another diagnosis ‘D1’ and ‘D2’. In scenario C different feature sets ‘f1 and ‘f2’ are observed, but these different sets lead both observers to the same conclusion ‘D1’. In scenario D, a diagnosis is made from the slide by ‘pure’ pattern recognition, as described by other authors: description is a mere justification of an already made diagnosis. In this scenario, the pathologist might solely report features of which he knows that they should have been present in the slide given diagnosis ‘D1’. The most common result – two observers report feature sets ‘f1’ and ‘f2’ and arrive at two different diagnoses ‘D1’ and D2 – is not presented.

The currently available interactive diagnostic support systems all support step 2. Hence, if much of the diagnostic dissension can be attributed to step 1, this support is no longer opportune.

To our knowledge there are no published reports that clarify for which weaknesses there are not yet satisfactory support strategies and what type of support would be needed to solve them. We present our studies to date that led to increased insight in the diagnostic process. These studies focus on reporting as a tangible reflection thereof. We will then discuss our view on the potential of IT for improvement of diagnostic consensus.

Materials & Methods

It has been our goal to elucidate the pathologist’s diagnostic process in terms of the aforementioned two steps (1: from observation to description; 2: from description to diagnosis). In essence, a histopathology report consists of 4 sections:

- Clinical findings
- Macroscopy
- Microscopy
- Conclusion.

When one regards the histopathologic report as a reflection of the diagnostic process, the microscopy section is the result of step 1 (Fig. 1 in introduction): the visual observations made in the slide are made explicit here. Step 2 corresponds with the report conclusion, which is made from the observations described in the microscopy section.

Two studies were performed: a retrospective study comprising existing pathology reports; and a prospective study involving a comparison of histological descriptions using a set of predefined options.

Inventory of existing reports

We retrieved 268 skin biopsy reports of 53 patients registered by the Dutch Cutaneous Lymphoma Working Group from PALGA (the Dutch pathology registration database). Reports were furthermore categorized by the conclusion of the report:

- Benign
- Pseudolymphoma
Chapter 13: Evaluation

- Suspect for malignancy (CL)
- Malignant

When a report could not be categorized in this way it was omitted from the study.

The microscopy section of the reports was searched for features relevant to the diagnosis of a cutaneous lymphoma. A score was made for each feature mentioned (presence as well as absence). Multiple mention of a feature in a report was counted only once.

The "index of description" (IOD) of reports in a given category was calculated as: \[ \frac{\sum \text{features described in the reports in that category}}{\sum \text{reports in that category}} \times 22 \] (22 is the total number of features listed by Evans).

Reporting using form-based data entry

We asked 29 pathologists to examine 16 slides of CL and related lesions (4 malignant (CL), 6 benign, and 6 uncertain malignant lesions) using a questionnaire. The malignant and the cases suspect for malignancy were selected from the LWCL (Dutch Cutaneous Lymphomas Committee) database. The benign cases came from the archives of the pathology dept. of the University Hospital Groningen. There was a consensus diagnosis by the LWCL with appropriate patient follow-up for all cases from the LWCL database. The questionnaire contained histological and cytological features relevant to diagnosis making in CL according to the literature and to our own experience.

The questionnaire consisted of two sections. The first part was mandatory and involved a set of 27 features for each of which the participants had to indicate the presence or absence in the slide at hand. The second part applied only if the lesion was not felt to be completely benign; several predefined options concerning cytological details could then be selected.

In the questionnaire, the participating pathologists were invited to assign the nature of the lesion at hand to one of 3 diagnostic categories: malignant, benign, suspect for malignancy (these categories representing a very crude form of diagnosis). Interobserver consensus on features and diagnostic category was calculated as Cohen’s \( \kappa \). Correlation between consensus on features and diagnostic categories was calculated using the Spearman rank correlation test.

Results

Inventory of existing reports

We excluded 4 reports because no diagnostic category could be derived from the conclusion. The scoring of histological features in 264 reports of lesions of patients suspected for CL can be summarized as shown in Table 1.

<table>
<thead>
<tr>
<th>Diagnostic category</th>
<th>Features</th>
<th>Reports</th>
<th>IOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>260</td>
<td>50</td>
<td>0.24</td>
</tr>
<tr>
<td>Pseudolymphoma</td>
<td>212</td>
<td>30</td>
<td>0.32</td>
</tr>
<tr>
<td>Suspect</td>
<td>532</td>
<td>78</td>
<td>0.31</td>
</tr>
<tr>
<td>Malignant</td>
<td>764</td>
<td>106</td>
<td>0.33</td>
</tr>
<tr>
<td>Total</td>
<td>1768</td>
<td>264</td>
<td>0.30</td>
</tr>
</tbody>
</table>

The reports investigated were incomplete regarding cell details: except for nuclear shape, there is no cytological description in more than half of reports with a suspect or malignant conclusion. There is hardly any description of cytoplasm (< 0.13) in any diagnostic category. The histology of epidermis and the infiltrate (its location and pattern, as well as the cells forming the infiltrate) is frequently described. The index of description seems to be slightly less for reports with a benign conclusion.

Reporting using form-based data entry

18 pathologists returned the questionnaire for all slides. One pathologist left too many items blank to include his results in this study, leaving 17 pathologists. Of the 16 slides, 4 (IDs 1,2,3,11) were obviously underdiagnosed (LWCL consensus suspect for malignancy, but pathologist conclusion benign), 1 (ID 5) was overdiagnosed. In 4 slides (IDs 1,5,6,9) there was obvious dissension about the nature of the lesion. The other slides did not pose significant problems. The diagnostic category results for the cases discussed above are shown in Table 2.

<table>
<thead>
<tr>
<th>Slide ID</th>
<th>LWCL diagnosis</th>
<th>No diagnosis</th>
<th>Mal.</th>
<th>Ben.</th>
<th>Suspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Suspect MF</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Suspect MF</td>
<td>1</td>
<td>13</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MF (uncommon)</td>
<td>1</td>
<td>12</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Fungal infection?</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>Very suspect MF</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>MF</td>
<td>7</td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Suspect CL (SS)</td>
<td>1</td>
<td>14</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

The average \( \kappa \) value for diagnostic category for all cases was 0.30.

Of the 27 histological features whose presence or absence had to be scored in every slide there was good consensus (\( \kappa = 0.75 \)) on 3 features; reasonable consensus (\( \kappa \) between 0.40 and 0.75) on 8 features; poor consensus on the remaining 16 features.

Spearman rank correlation coefficients of agreement on observed features vs. agreement on diagnostic category...
were as follows (only those features with significant correlation at the 0.05 level in one or more of the categories are shown):

Table 3 - Spearman rank correlation coefficient of agreement on histological features vs. agreement on diagnostic category index, shown as coefficient (significance level).

<table>
<thead>
<tr>
<th>Diagnostic category</th>
<th>Mal.</th>
<th>Ben.</th>
<th>Suspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pautrier abscesses</td>
<td>0.575</td>
<td>-0.429</td>
<td>-0.217</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.098)</td>
<td>(0.419)</td>
</tr>
<tr>
<td>Selection of atypical</td>
<td>0.756</td>
<td>-0.636</td>
<td>0.131</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.008)</td>
<td>(0.629)</td>
</tr>
<tr>
<td>Lymphocytes</td>
<td>-0.526</td>
<td>0.574</td>
<td>-0.197</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.020)</td>
<td>(0.465)</td>
</tr>
<tr>
<td>Lymfocytic vasculitis</td>
<td>-0.601</td>
<td>0.554</td>
<td>-0.037</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.026)</td>
<td>(0.893)</td>
</tr>
<tr>
<td>Perivascular infiltrate</td>
<td>0.513</td>
<td>-0.644</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>(0.042)</td>
<td>(0.007)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>Invasive pattern of infiltrate</td>
<td>0.609</td>
<td>-0.492</td>
<td>0.112</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.053)</td>
<td>(0.679)</td>
</tr>
<tr>
<td>Tumor Formation by infiltrate</td>
<td>-0.157</td>
<td>0.324</td>
<td>-0.534</td>
</tr>
<tr>
<td></td>
<td>(0.562)</td>
<td>(0.221)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Infiltrate location high in dermis</td>
<td>0.609</td>
<td>-0.492</td>
<td>0.112</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.053)</td>
<td>(0.679)</td>
</tr>
</tbody>
</table>

The second part of the questionnaire did not permit reliable statistical analysis because 11 pathologists filled it out for only 5 cases. The results indicated poor or average agreement for 43 out of 60 features. Consensus on the other 17 features was mainly on the absence of those features.

Discussion

It is well known that there is a considerable inter- and intra-observer variation in diagnosis making in many fields of pathology. It has been argued that a histological diagnosis is made more or less on the basis of pattern recognition, and that verbalization of observations in the slide causes great loss of information 22. On the other hand, other authors argue that diagnostic dissenion is caused by applying different diagnostic criteria, or by different interpretation of terms used in defining those diagnostic criteria 5.

The inventory of existing reports in the domain of cutaneous lymphomas addresses step 1 in the diagnostic process: the completeness and uniformity of description. This study shows that pathologists do not adhere to a standard way of describing a cutaneous lesion. There is a strong variability in description of observed features. Thus, the spontaneously reported information is little compared to what is described as important in the literature. It appears that pathology reports are a very sparse reflection of the pathologist’s diagnostic process: a pathologist may describe what he observes and then make a diagnosis, or determine the diagnosis straight from the slide and then enumerate those observations that led him to his decision. The latter approach would not be suitable as input for a diagnostic support system, since it would highly bias its outcome toward what the pathologist already thought.

The second study again points to the variability of the diagnostic process in step 1: even with mandatory feature description there is little consensus on more than half of the features that had to be described. When consensus was good, it was often about the absence of a feature. Also, there is little consensus on the diagnostic category of the lesions, let alone for the regular much more refined level of diagnosis. So, mandatory description improves completeness in description but does not generate a satisfactory consensus on step 1 of the diagnostic process.

It is interesting to note the relatively high consensus on the absence of features: this points toward a tendency to overlook uncommonly encountered abnormalities.

The results from the questionnaire strongly suggest that the general histological impression of a slide greatly influences the a priori chance of observing certain histological details: the more suspect, the more they see.

As an indication of the relevance of lack of consensus on observed features we found 11 cases (each including the first and follow-up reports) where malignancy was unnecessarily missed at an early stage: features indicative of malignancy were reported without proper reflection in the diagnostic conclusion.

What then can be done using IT to improve consensus in diagnostic pathology? As shown with the questionnaire and as reported in the literature, predefined options improve completeness 25. An electronic reporting system with structured data entry would be effective for completeness as well as a more uniform use of terminology 16, 24, 25.

Better consensus on feature recognition is highly desirable as this is the basis for diagnostic reasoning for both humans and expert systems. Optimal diagnostic support should target both step 1 and step 2 of the diagnostic process. We believe that consensus in feature recognition would benefit from an electronic reporting system, based on structured data entry with reference pictures associated with histological features. Furthermore, if such a reporting system is integrated with a diagnostic expert system, diagnostic support no longer solely depends on user initiative and is less labor-intensive.

To test if structured reporting combined with reference pictures improve feature recognition we are currently investigating the consistency of observation of features. In this study several panels of pathologists are matching the histological features in a given slide with reference pictures.

References

[3] Smeulders AW, Van Ginneken AM. An analysis of pathology knowledge and decision making for the
Chapter 13: Evaluation


Address for correspondence

GTN Burger
Bethesda Hospital
Dr GH Amshofweg 1
7909 AA Hoogeveen
The Netherlands
Tel.+31 (528) 286632
Fax+31 (528) 286299