Clinical Trial of Telepathology As an Alternative Modality in Breast Histopathology Quality Assurance

F. JOEL W.-M. LEONG,1 ANDREW K. GRAHAM,1 PETER SCHWARZMANN,2 and JAMES O’D. McGEE2

ABSTRACT

Telepathology is a potential alternative to conventional histopathology. A clinical trial using a robotic telepathology system was conducted to assess the clinical and technical utility and effectiveness of telepathology in the U.K. breast screening pathology quality assurance program. Eighty-seven cases of breast disease were chosen at random from a series of 192 cases from the U.K. Breast Screening Pathology National Quality Assurance Scheme (NEQAS) collection. There were 20 benign, 23 carcinoma in situ (CIS), and 44 invasive malignant cases. The diagnostic accuracy of telepathology (TP) compared with conventional light microscopic (LM) diagnosis was 98.8%; this included a single case deferred for LM examination. The figure was similar when compared with expert consensus diagnosis (CD). In invasive tumor typing, TP accuracy was 95.4% (42/44 cases), the difference being attributable to slide color fading and would have had no impact on patient management. The accuracy of TP versus LM and expert consensus in tumor grading was 91.3% for carcinoma in situ (21/23 cases), a discordance with no relevance to patient management. TP grading of invasive tumor compared with LM diagnosis, had an accuracy of 86.4% (38/44) with a clinically significant accuracy of 97.7% (43/44). The time taken for TP diagnosis averaged 3.9 minutes per case by the end of the study. This data demonstrates that telepathology diagnostic accuracy is comparable to conventional microscopy and may therefore be envisaged as an alternative to conventional light microscopy for more rapid proficiency testing in breast screening (and perhaps other) quality assurance schemes.

INTRODUCTION

Professional proficiency testing has become an integral part of continuous professional development in all medical specialties in the United Kingdom and North America. Ideally, proficiency should be measured on the same “task” material for each individual. For logistic reasons this is impossible in national schemes. In the United Kingdom, national quality assurance scheme in breast screening pathology comes close to the ideal practical solution. In this scheme, all consultant histopathologists involved in management of breast disease are confidentially assessed twice a year by examining two sets of 12 microscopic slides selected from the spectrum of breast disease. Individual consultant’s diagnoses on

1Nuffield Department of Pathology & Bacteriology, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DU UK.
2University of Stuttgart, Institut fuer Physikalische Elektronik, D-70569 Stuttgart, Germany.
these slides are compared with the consensus diagnosis of 19 histopathologists regarded as expert in breast disease diagnosis. However, with more than 400 pathologists participating in the scheme, having to circulate slides by mail to all participants results in each proficiency round taking 5 months to complete. The sections, although consecutive, are not identical, yet individuals are judged on what is assumed to be present in the section rather than its actuality. Proficiency scores are allocated by reference to the consensus expert diagnosis; these experts similarly view 19 separate sections from the same lesion. The aims of this paper were to investigate the accuracy of diagnosis and utility of telepathology in breast screening pathology. This information is necessary before considering the implementation of telepathology rather than conventional microscopy for proficiency testing in histopathology quality assurance programs.

MATERIALS AND METHODS

Clinical samples

Eighty-seven microscopic sections from breast lesions were chosen at random from the U.K. National Health Service Breast Screening Pathology External Quality Assurance Scheme (NEQAS) collection. The slides were selected from a series of 192 cases used in NEQAS over the past 9 years and had been examined by the pathologist (J.O’D.M.) at least once during this time. Records of his diagnosis were held locally in Oxford and in a database in London. None of the sections had been viewed in the year before this study began.

The diagnostic criteria used in telepathology (TP) diagnosis, tumor grading, and typing were those of the U.K. NHS Breast Screening Pathology Group. The conventional light microscopic (LM) diagnosis and consensus diagnosis of 19 expert U.K. breast pathologists were compared with the TP and LM diagnosis after the TP experiments were completed. No clinical details (e.g., age, sex, duration of history, etc.) were available for any of the cases.

Telepathology

The TP system was developed in the Institut für Physikalische Elektronik of the University of Stuttgart, Germany, and is now available from German Telecom. The technical specifications and design of this have been previously described.

A robotic microscope (Zeiss Axioplan 2) was controlled either with a mouse via the onscreen menu or with the computer keyboard. The pathologist used two 15-inch monitors. One monitor (No. 1) displayed a scanning overview image (objective magnification = 1.25×) of the entire microscopic section accompanied by a menu for manipulating microscope and stage settings including objective selection, focus (automatic or manual), and Köhler illumination. The images were captured via the analogue RGB (red, green, and blue) output of a Sony 1/2-inch, 3-chip, CCD camera. The second monitor (No. 2) displayed the microscopic images at a resolution of 768 × 576 × 24 bit color, and was selected via the low-power overview image (monitor no. 1). Microscope stage control was with a joystick. Each new image was ‘stitched’ almost seamlessly to the preceding one, as if one were actually scanning the slide in a conventional manner. Telecommunication transmission between the robotic microscope and remote station was through four basic rate interface integrated services digital network lines (4 BRI-ISDN) providing a transmission speed of 512 kilobits per second (Kbps).

The TP diagnosis, cancer type and grade, and total time for diagnostic assessment were recorded for each slide. All original biopsy numbers were anonymized to prevent bias. The pathologist had 90 minutes training in operating the system before commencing the study.

RESULTS

The spectrum of cases examined encompassed the majority of breast disease encountered in daily practice. They included 20 benign, 23 carcinoma in situ (CIS), and 44 invasive cancers. The time required for TP diagnosis varied from 1 to 11 minutes. Figure 1 shows the average time for TP diagnosis per
set of five consecutive biopsies over 2 consecutive days. Diagnostic time rapidly diminished from approximately 9 minutes to 3 minutes per biopsy.

TP diagnosis concurred with LM and consensus diagnosis in all but one case (98.9% concordance; 86/87 cases; see Table 1). A TP diagnosis could not be made for that case, which had only 35% of expert pathologists in agreement (consensus diagnosis), indicating the difficulty of the case independent of the diagnostic modality.

There was a difference in tumor type in two cases when comparing TP with both LM and consensus diagnosis (95.4%; 42/44 cases)—both of these were mucoid carcinomas, the staining of which had faded, rendering appreciation of the mucinous areas difficult. The TP diagnosis in each instance was ductal carcinoma, type not otherwise specified.

In two of the twenty-three cases of carcinoma in situ, there was a difference in grading—intermediate versus high grade (91.3% concordance; 21/23 cases). In comparison with the LM diagnosis, six of the 44 invasive lesions showed a difference in grading of not more than one grade (86.4% concordance; 38/44 cases; see Table 1). Grading is not subject to consensus diagnosis within NEQAS.

**DISCUSSION**

This study was designed to compare diagnostic accuracy of TP in breast disease diagnosis with conventional microscopy. The results demonstrate the clinical utility of TP. Of 87 breast biopsies studied, the overall diagnostic accuracy was 98.8% and the clinical accuracy was 100%. The single case where a diagnosis could not be reached by TP was clinically difficult because of interpretation of the micro-

### Table 1. Comparison of Telepathology Performance with Conventional Light Microscopy and NEQAS Consensus Diagnosis

<table>
<thead>
<tr>
<th></th>
<th>Diagnostic accuracy for all cases</th>
<th>Tumor typing accuracy for invasive cancers</th>
<th>Tumor grading accuracy for invasive cancers</th>
<th>Tumor grading accuracy for CIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>CSA (%)</td>
<td>Accuracy (%)</td>
<td>CSA (%)</td>
</tr>
<tr>
<td>TP and LM</td>
<td>98.8 (86/87 cases)</td>
<td>100</td>
<td>95.4 (42/44)</td>
<td>100</td>
</tr>
<tr>
<td>TP and CD</td>
<td>98.8</td>
<td>100</td>
<td>95.4</td>
<td>100</td>
</tr>
<tr>
<td>LM and CD</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

CD, Consensus (diagnosis based upon NEQAS data); CIS, carcinoma in-situ; CSA, clinically significant accuracy; LM, conventional light microscopy diagnosis; TP, telepathology.
scopic features. In fact, the consensus diagnosis in this case was 35%. Consensus diagnosis in the U.K. NEQAS requires 80% consensus of experts before a biopsy is included in calculations to assess a pathologist’s professional diagnostic competence.

The TP diagnostic accuracy of tumor typing was 95.4% with 100% accuracy for patient management. The two cases of TP diagnostic discordance with conventional LM and consensus diagnosis were attributable to fading of the microscopic sections (prepared in 1994). Even so, the TP diagnosis would not have had any impact on management.

The concordance between TP and conventional diagnosis in tumor grading was 91.3% for CIS. The clinical implications of discordance would have been negligible. Nine invasive cases showed differences in grading when compared with consensus diagnosis. However, in only one instance was there possible clinical implication—the LM grade was 3 compared to 2 by TP; this may possibly have resulted in less aggressive therapy had the diagnosis been made by TP. In fact, nearly all of the apparent consensus discordant grades could be ignored, because in only one of these biopsies was agreement by the consensus panel greater than 80%.

Other studies exclude such cases from TP accuracy calculations. The actual expert consensus range for grading all these cases was 44–83% (mean 65.5%) compared with the concordance between TP and LM diagnosis of 86.4%. Hence, the practical implications of this study are that, on the basis of clinically significant accuracy, telepathology is 100% accurate in overall diagnosis, 100% accurate in tumor typing, and 97.7% accurate in tumor grading.

We recognize that this study involved only a single pathologist and hence the results may not be directly extrapolated to the general pathology community. Further studies involving a larger number of pathologists are now being implemented.

In this study, factors that influenced the speed at which a diagnosis was made include the TP system design and connection bandwidth, the inherent difficulty of the case, and slide quality. Although we do not expect a TP system to perform to the same degree of speed and efficiency as a conventional LM, it should be able to function at a level that is usable and practical. We found the accuracy at least equal to past studies and the clinical utility of the system to be adequate with respect to the time taken for biopsy diagnosis. The range in this study was 1–10 minutes, with an average of 5.5 minutes for the first 40 cases and 3.9 minutes for the remaining 47 biopsies. This was a function of the accumulated experience of the pathologist with the TP system (see Fig. 1). Comparable figures from other diagnostic TP studies average 3.8 minutes per slide, 8.9 minutes per case, and 4.3 minutes per slide, 12.1 minutes per case. In other studies it has been as high as 15 minutes.

It is also important to note that this study employed ‘dynamic’ images that were larger in area by a factor of 3.8 (based upon number of pixels) compared with the TP system used in these studies.

One of the objectives of TP is its implementation for professional quality assurance in breast screening programs in the European Union. Attempts to introduce TP in other areas using static images have met with some dissatisfaction initially due to image quality, but more so due to the receiving pathologist being dependent upon the transmitting pathologist for the diagnostically significant images. Sampling error is a significant factor to the detriment of this approach. Using TP, the pathologist is not given but selects the images on which to make a diagnostic judgement.

The overall conclusions from this study are that robotic TP is capable of the same accuracy and timeliness of diagnosis in breast pathology as conventional microscopy. It is reasonable, therefore, to consider that robotic TP could be used in parallel with (or as a substitute for) conventional microscopy in breast screening pathology QA programs. This modality would have the important attributes of ensuring that all participants in a QA scheme select their own images from the same slide and therefore the views of both participants and experts could be directly compared. Second, the delays and cost of the present method of circulation could be avoided. Third, all diagnoses could be submitted directly to a central database, thus avoiding having to collate mailed paper submissions.
ACKNOWLEDGMENTS

The histological slides were provided by the U.K. Breast Screening Program. This work was supported by grants from the E.U. (EuroPath), Reuters Foundation, and U.K. National Health Service R&D.

REFERENCES


Address reprint requests to:
Professor James O'D. McGee
University of Oxford
Nuffield Department of Medicine
Level 5
John Radcliffe Hospital
Headington, Oxford OX3 9DU
UK
E-mail: james.mcgee@ndm.ox.ac.uk