J Pathol Inform

Research Article

Diagnostic performance on briefly presented digital pathology images

Joseph P Houghton[1], Bruce R Smoller[2], Niamh Leonard[3], Michael R Stevenson[1], Tim Dornan[1]

[1]Centre for Medical Education, Queen's University Belfast, Belfast BT9 7BL, [3]Department of Histopathology, St. James's Hospital, James's Street, Dublin 8, Ireland, [2]Department of Pathology and Laboratory Medicine, University of Rochester Medical Center, Rochester, New York, USA

E-mail: *Dr. Joseph P Houghton - Joseph.Houghton@qub.ac.uk
*Corresponding author

Received: 01 July 15 Accepted: 24 July 15 Published: 28 October 2015

Abstract

Background: Identifying new and more robust assessments of proficiency/expertise (finding new "biomarkers of expertise") in histopathology is desirable for many reasons. Advances in digital pathology permit new and innovative tests such as flash viewing tests and eye tracking and slide navigation analyses that would not be possible with a traditional microscope. The main purpose of this study was to examine the usefulness of time-restricted testing of expertise in histopathology using digital images. **Methods:** 19 novices (undergraduate medical students), 18 intermediates (trainees), and 19 experts (consultants) were invited to give their opinion on 20 general histopathology cases after 1 s and 10 s viewing times. Differences in performance between groups were measured and the internal reliability of the test was calculated. **Results:** There were highly significant differences in performance between the groups using the Fisher's least significant difference method for multiple comparisons. Differences between groups were consistently greater in the 10-s than the 1-s test. The Kuder–Richardson 20 internal reliability coefficients were very high for both tests: 0.905 for the 1-s test and 0.926 for the 10-s test. Consultants had levels of diagnostic accuracy of 72% at 1 s and 83% at 10 s. **Conclusions:** Time-restricted tests using digital images have the potential to be extremely reliable tests of diagnostic proficiency in histopathology. A 10-s viewing test may be more reliable than a 1-s test. Over-reliance on "at a glance" diagnoses in histopathology is a potential source of medical error due to over-confidence bias and premature closure.

Key words: Digital pathology, expertise, overconfidence bias, premature closure, time-restricted test

Access this article online
Website:
www.jpathinformatics.org
DOI: 10.4103/2153-3539.168517
Quick Response Code:

INTRODUCTION

Histopathology is, at its core, a visual discipline. The cornerstone of accurate tissue diagnosis is a pathologist viewing and correctly interpreting a microscopic image. A number of other important skills such as clinicopathological correlation are also important, but visual pattern recognition is the critical element. The gradual acquisition of pattern recognition skills, while

This article may be cited as:
Houghton JP, Smoller BR, Leonard N, Stevenson MR, Dornan T. Diagnostic performance on briefly presented digital pathology images. J Pathol Inform 2015;6:56.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2015/6/1/56/168517

progressing from a novice style to an expert style and maintaining these skills throughout a professional career, is essential. Very few studies have attempted to assess visual memory or pattern recognition skills in histopathology.

A widely supported theory explaining how we visually examine images to identify key information proposes that there are two pathways.[1-5] These two systems run in parallel are fluid and communicate with each other. First, a nonselective pathway, which has alternatively been described as holistic, automatic, Gestalt-like, *coup d'oeil*, top-down, thin-slicing or subconscious searching, involves a global (at a glance) impression of the image. The second pathway is the selective pathway and involves careful screening for specific findings. This has also been called conscious, analytic, or bottom-up searching. Studies in neuroimaging which support this two-pathway model have shown that object identification maps to regions in the occipitotemporal cortex whereas global identification maps to other regions in the brain.[6-9]

There are two broad approaches to testing these pathways. The nonselective pathway can be tested using time-restricted tests whereas the selective pathway can be tested using the eye-tracking equipment. A small number of studies have examined differences in performance in time-restricted tests in radiology and pathology.[10-15] Most of these studies have shown superior accuracy by experts. In other words, experts tend to make correct diagnoses quickly and accurately. Performance in time-restricted tests can, therefore, be used as a marker of expertise.

Identifying new and more robust assessments of proficiency/expertise in histopathology (finding new "biomarkers of expertise")[16] is desirable for several reasons (for example, as a serial assessment tool of histopathologists in training) and the main purpose of this study was to examine the usefulness of time-restricted tests in histopathology using digital images.

METHODS

Ethical approval for the study was granted by the Queen's University Belfast Research Ethics Committee, School of Medicine Dentistry and Biomedical Sciences Ref: 14.47v2, 3/11/2014.

Participants (*n* = 56; mean ± standard deviation [SD] age: 33.2 ± 10.3 years; 31 males and 25 females) formed three groups with increasing levels of experience. Novices (Group 1) were 19 third-year undergraduate medical students who were midway through a core pathology module and who, in addition, had just completed an elective in pathology (mean ± SD age: 23.1 ± 2.9 years; 13 males and 6 females). Intermediates (Group 2) were 18 trainee histopathologists/residents (mean ± SD age: 31.3 ± 3.4 years; 6 males and 12 females). Within this group, six trainees had fewer than 2 years' experience viewing histopathology slides; seven trainees had between 2 and 4 years' experience, and five trainees had > 4 years' experience. Experts (Group 3) were 19 practicing consultant histopathologists (mean ± SD age: 45.1 ± 8.0 years; 12 males and 7 females) with a mean ± SD length of experience in consultant practice of 12.2 ± 8.2 years. The consultant group was a heterogenous mixture of general and specialist histopathologists. Trainees and consultant histopathologists were based in Belfast City Hospital and Royal Victoria Hospital, Belfast, and St. James' Hospital, Dublin.

Stimuli consisted of 20 digital histological images from teaching archives [Table 1]. An example of an image used is presented in Figure 1. These were general pathology cases from four different anatomical sites chosen to represent a full range of difficulty ranging from normal histology to challenging cases. In order to achieve this, we referred to the competency framework

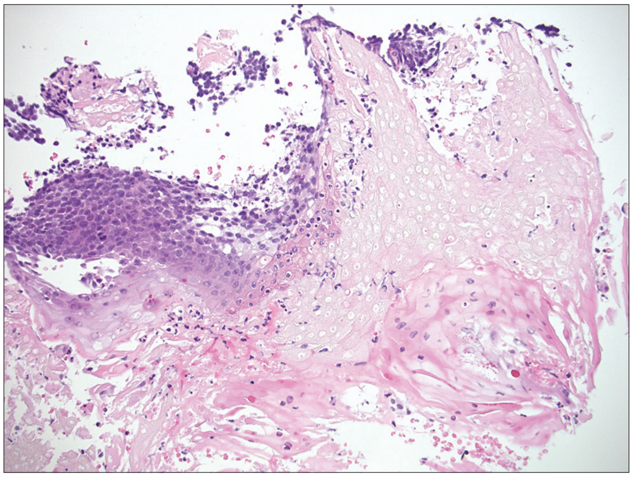

Figure 1: Pilomatrixoma (benign skin adnexal tumor) in which there are uniform basaloid cells on the left and "ghost cells" on the right, high magnification

Table 1: Twenty cases from four anatomical sites representing a broad range of difficulty

Level	Skin	Gynecology	Head and neck	Gastrointestinal
1	Pilar cyst	Normal fallopian tube	Fibroepithelial polyp	Acute appendicitis
2	Molluscum contagiosum	Endometrial polyp	Oral mucosal ulcer/granulation tissue	Hemorrhoids
3	Glomus tumor	Menstrual endometrium	Temporal arteritis	Diverticulosis
4	Actinic keratosis	Cervical intraepithelial neoplasia III	Granular cell tumor	Barrett's metaplasia
5	Pilomatrixoma	Adenomatoid tumor	Ameloblastoma	Colonic spirochaetosis

for graded responsibility for Specialist Registrars in Histopathology and Cytopathology published by the Joint Committee on Pathology Training of the Royal College of Pathologists (UK).[17] This categorizes cases into four increasing levels of complexity, where one is the lowest and four is the highest. In addition, we introduced level 5, which is not in the original document, but includes cases that would be considered more difficult than level 4. The magnification used was tailored to each individual case; if the diagnosis was based primarily on an architectural feature (e.g., diverticulosis) a low-power magnification was used, whereas if the diagnosis was based on a cytological feature (e.g., Barrett's metaplasia) a high-power magnification was used. All test material including the clinical history, image quality, and correct diagnosis for each case was verified by two experienced consultant histopathologists who did not participate in the study.

The experiment was carried out in a seminar room where single representative fixed photographs (.jpg) of each diagnostic entity were displayed on a white screen using an overhead projector. For each case, a brief clinical summary was provided to participants. The clinical information was deliberately brief and only included age, gender, and the site of biopsy and did not give any further information from which participants could guess the correct answer. Each image was displayed for 1 s followed by a 20 s pause during which participants recorded their diagnoses. Responses were written down in free-text format rather than using a multiple-choice format in order to reduce the likelihood of guessing the correct answer. Immediately after the 1-s tests were complete, the test was repeated as before using the same images; however, the images were now displayed for 10-s. The rationale for choosing 1 s/10 s timings was that 1 s represented a brief glance, whereas 10 s was chosen to represent a longer but still challenging exposure. The candidates' answers were marked manually, and their scores were presented using a box and whisker plot graph. Differences in performance between individual groups were analyzed using the Fisher's least significant difference method for multiple comparisons. The Kuder–Richardson formula 20 (KR-20) internal reliability coefficient was calculated using Statistical Package for the Social Sciences (SPSS, IBM SSPS Statistics 21) for both tests as quality measure of internal reliability. As a general principle, KR-20 coefficients of between 0.7 and 0.9 are considered good and coefficients of >0.9 are considered excellent.

RESULTS

The range of participants' scores for the 1-s and 10-s tests is presented in Figures 2 and 3, respectively. For all groups (including consultants), accuracy was higher at 10-s.

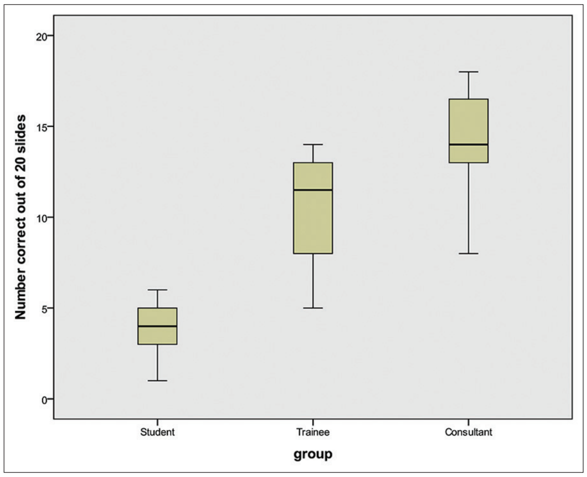

Figure 2: Box and whisker plots with the ends of the whiskers representing the maximum and minimum scores for all participants for the 1-s test

Table 2 illustrates highly significant differences in performance between groups using the Fisher's least significant difference method for multiple comparisons. Differences between groups were consistently greater with the 10-s than the 1-s test.

KR-20 internal reliability coefficients were very high for both tests: 0.905 for the 1-s test and 0.926 for the 10-s test.

The range of incorrect answers (over both tests) for the five cases that the consultants found most difficult is presented in Table 3. In some cases, the incorrect answers are major diagnostic errors; for example interpreting a benign tumor as malignant. The magnification of the image that was selected did not influence the likelihood of error.

CONCLUSIONS

In this study, we analyzed performance in 2 time-restricted tests, at 1-s and also at 10-s. Both of these tests demonstrated very high degrees of internal reliability 0.905 and 0.926, respectively. Intuitively, we had expected that a 1-s test would be more discriminating than a 10-s test because it has been suggested that extreme time restriction can expose differences in ability that would otherwise be undetectable. However, the results of this study did not corroborate this; in fact, the 10-s test had a marginally superior reliability coefficient and differences among groups were consistently greater.

There was a broad range of ability in performance within each group. A strong performance in the student group could potentially identify candidates with a natural talent for pattern recognition who could be suited to a career in pathology. The broad range of trainees most likely reflected the broad range of experience

Table 2: Fisher's least significant difference method for multiple comparisons, 1-s and 10-s tests

Group (I)	Group (J)	Mean difference (I-J)	SE	Significant[a]	95% CI for difference[a]	
					Lower bound	Upper bound
1-s test						
Student	Trainee	−6.71*	0.82	0.000	−8.35	−5.07
	Consultant	−10.63*	0.81	0.000	12.25	−9.02
Trainee	Student	6.71*	0.82	0.000	5.07	8.35
	Consultant	−3.92*	0.82	0.000	−5.56	−2.29
Consultant	Student	10.63*	0.81	0.000	9.02	12.25
	Trainee	3.92*	0.82	0.000	2.29	5.56
10-s test						
Student	Trainee	−7.97*	0.85	0.000	−9.67	−6.26
	Consultant	−12.21*	0.84	0.000	−13.89	−10.53
Trainee	Student	7.97*	0.85	0.000	6.26	9.67
	Consultant	−4.24*	0.85	0.000	−5.95	−2.54
Consultant	Student	12.21*	0.84	0.000	10.53	13.89
	Trainee	4.24*	0.85	0.000	2.54	5.95

Dependent variable: Number correct out of 20 slides based on estimated marginal means. *The mean difference is significant at the 0.05 level, [a]Adjustment for multiple comparisons: Least significant difference (equivalent to no adjustments). SE: Standard error, CI: Confidence interval

Table 3: Range of incorrect answers submitted by consultants for the five most difficult cases

Correct answer	Incorrect answers
Pilomatrixoma	Squamous cell carcinoma, basal cell carcinoma, branchial cleft cyst, nevus
Oral mucosal ulcer/granulation tissue	Granular cell tumor, Langerhans' cell histiocytosis, necrotizing sialometaplasia, granulomatous inflammation
Adenomatoid tumor	Clear cell carcinoma, angiomyolipoma, prolapsed fallopian tube, adenofibroma, female adnexal tumor of probable Wolffian origin, mucinous cystadenoma, lipoma
Ameloblastoma	Adamantinoma, radicular cyst, dentigerous cyst
Colonic spirochaetosis	Normal colonic mucosa, lymphocytic colitis, collagenous colitis

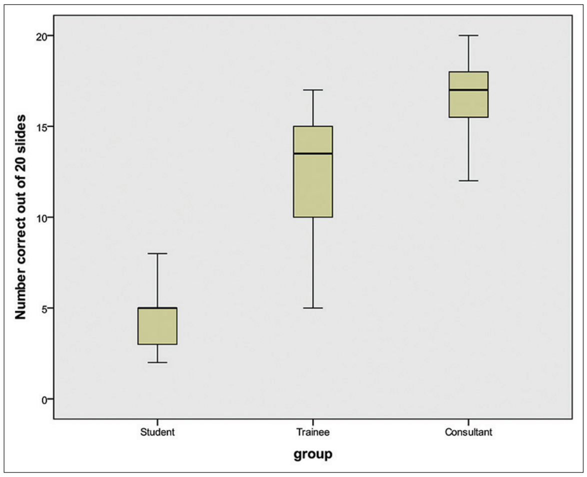

Figure 3: Box and whisker plots with the ends of the whiskers representing the maximum and minimum scores for all participants for the 10-s test

in this subgroup. Among the consultant group, the study highlighted truly expert performances; at the very top was a consultant who scored 18/20 at 1-s and 20/20 at 10-s. The lowest consultant scores most likely represented the loss of general pathology skills among pathologists who had worked as subspecialists for many years.

With respect to the utility formula of assessment,[18] the time-restricted tests used in this study scored highly due to high reliability coefficients were highly valid in that they were an assessment of a key diagnostic skill, could potentially have a positive impact on learning, and were extremely brief and cost effective. We did not formally survey the participants afterward with regards to acceptability but a 10-s test appeared to be more acceptable than a 1-s test.

We propose that time-restricted testing could be used in a number of situations; as a formative assessment during training to track trainees longitudinally to record the acquisition or failure of acquisition of skills with a view to accelerate this progression but also to quickly identify, at an early stage, trainees in difficulty who may require additional support or who may not be suited to a career in histopathology; to demonstrate maintenance of expert skills throughout a career; to assess pathologists who chose to re-train in a new subspecialty; to compare training programs nationally and internationally to identify best practice.

Of course, time-restricted testing only assesses one skill, the ability to make rapid, and accurate assessments of histopathological images. While some cases are "spot/at a glance diagnoses" others require careful screening of slides to identify a subtle or scanty feature; for example, in cervical cytology or prostate chippings. In this study, we only looked at expertise in rapid image identification, whereas differences in experience with regard to screening histological slides requires analysis of pathologists' searching strategies which can be assessed using eye tracking studies or search maps recorded when pathologists view digital slides.[19-26] We suggest that performance in time-restricted tests could be incorporated with performance in other tests such as eye tracking to provide a global overview of an individual pathologist's expertise.

Being a safe and competent practitioner also requires attention to detail, an ability to correlate clinical and pathological information, a commitment to audit, quality improvement, and continuous professional development. Some pathologists may perform well in a time-restricted test but may not be committed to these other attributes.

An interesting finding in this study was the remarkably high diagnostic accuracy of experts of 72% at 1-s and 83% at 10-s, which could have implications for routine practice. With increasing experience, experts can diagnose so many cases within the first few seconds that there is potentially an associated increased risk of medical error due to overconfidence bias and premature closure.[27] During routine busy practice, there is a constant tension between rapid diagnosis and cautious decision making and it is important that histopathologists are aware that over-reliance on "at a glance" diagnoses is a potential source of medical error. In this study, there were a number of examples where the consultant's first impression was that of a malignant tumor when the lesion was benign. It is likely that some of these discrepancies were due to specialists giving opinions on cases with which they were unfamiliar but there are numerous examples in surgical pathology where benign lesions can closely mimic malignant lesions, and an interesting further study could involve specifically focusing on these problematic areas. In addition, in a small number of cases, there is double pathology, and it is important not to miss the second abnormality having identified the first more obvious pathology.

A potential weakness in our study was using a consultant/expert group that was a mixture of generalists and specialists and, while all of the experts had originally trained in general pathology, some had since subspecialized in areas not included in this study. If all the experts had been practicing general pathologists, the expert group might have performed better. Furthermore, inclusion of 19 novices of very low ability could have artificially inflated the high KR-20 reliability coefficient[28] and it would, therefore, be useful to carry out similar tests with trainees and/or consultant histopathologists. Another potential weakness is that interpretation of the images at 10-s may have been influenced by the 1-s test due to a direct/repetition priming effect.[29]

Acknowledgments

JH, MS, and TD designed the study. MS provided the statistical analysis. BS, NL, and TD provided manuscript critiques. All authors approved the final version.

Financial Support and Sponsorship

Nil.

Conflicts of Interest

There are no conflicts of interest.

REFERENCES

1. Drew T, Evans K, Võ ML, Jacobson FL, Wolfe JM. Informatics in radiology: What can you see in a single glance and how might this guide visual search in medical images? Radiographics 2013;33:263-74.
2. Nodine CF, Kundel HI. The cognitive side of visual search in radiology. In: O'Regan JK, Levy-Schoen A, editors. Eye Movements from Physiology to Cognition. Amsterdam: Elsevier; 1987. p. 573-82.
3. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: Gaze-tracking study. Radiology 2007;242:396-402.
4. Swensson RG. A two-stage detection model applied to skilled visual search by radiologists. Percept Psychophys 1980;27:11-6.
5. Wolfe JM, Võ ML, Evans KK, Greene MR. Visual search in scenes involves selective and nonselective pathways. Trends Cogn Sci 2011;15:77-84.
6. Aguirre GK, Zarahn E, D'Esposito MT. Neural components of topographical representation. Proc Natl Acad Sci U S A 1998;95:839-46.
7. Epstein R, Graham KS, Downing PE. Viewpoint-specific scene representations in human parahippocampal cortex. Neuron 2003;37:865-76.
8. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 2001;293:2425-30.
9. Epstein R, Kanwisher N. A cortical representation of the local visual environment. Nature 1998;392:598-601.
10. Kundel HL, Nodine CF. Interpreting chest radiographs without visual search. Radiology 1975;116:527-32.
11. Carmody DP, Nodine CF, Kundel HL. Finding lung nodules with and without comparative visual scanning. Percept Psychophys 1981;29:594-8.
12. Oestmann JW, Greene R, Kushner DC, Bourgouin PM, Linetsky L, Llewellyn HJ. Lung lesions: Correlation between viewing time and detection. Radiology 1988;166:451-3.
13. Mugglestone MD, Gale AG, Cowley HC, Wilson AR. Diagnostic performance on briefly presented mammographic images. Proc SPIE 1995;2436:106-15.
14. Evans KK, Georgian-Smith D, Tambouret R, Birdwell RL, Wolfe JM. The gist of the abnormal: Above-chance medical decision making in the blink of an eye. Psychon Bull Rev 2013;20:1170-5.
15. Jaarsma T, Jarodzka H, Nap M, van Merrienboer JJ, Boshuizen HP. Expertise under the microscope: Processing histopathological slides. Med Educ 2014;48:292-300.
16. Weinstein RS. View master – An expert eyes digital pathology's future. CAP Today 2007;21:5-12.
17. Available from: https://www.rcpath.org/Resources/RCPath/Migrated%20Resources/Documents/A/A_competency_based_framework_for_graded_responsibility_arranged_by_level_of_competence.pdf. [Last cited on 2015 Jun 12].
18. Van Der Vleuten CP. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ Theory Pract 1996;1:41-67.
19. Tiersma ES, Peters AA, Mooij HA, Fleuren GJ. Visualising scanning patterns

of pathologists in the grading of cervical intraepithelial neoplasia. J Clin Pathol 2003;56:677-80.

20. Krupinski EA, Tillack AA, Richter L, Henderson JT, Bhattacharyya AK, Scott KM, *et al.* Eye-movement study and human performance using telepathology virtual slides: Implications for medical education and differences with experience. Hum Pathol 2006;37:1543-56.
21. Treanor D, Lim CH, Magee D, Bulpitt A, Quirke P. Tracking with virtual slides: A tool to study diagnostic error in histopathology. Histopathology 2009;55:37-45.
22. Roa-Peña L, Gómez F, Romero E. An experimental study of pathologist's navigation patterns in virtual microscopy. Diagn Pathol 2010;5:71.
23. Krupinski EA, Weinstein RS. Changes in visual search patterns of pathology residents as they gain experience. Proc SPIE 2011;7966:79660P.
24. Mello-Thoms C, Mello CA, Medvedeva O, Castine M, Legowski E, Gardner G, *et al.* Perceptual analysis of the reading of dermatopathology virtual slides by pathology residents. Arch Pathol Lab Med 2012;136:551-62.
25. Raghunath V, Braxton MO, Gagnon SA, Brunyé TT, Allison KH, Reisch LM, *et al.* Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images. J Pathol Inform 2012;3:43.
26. Brunyé TT, Carney PA, Allison KH, Shapiro LG, Weaver DL, Elmore JG. Eye movements as an index of pathologist visual expertise: A pilot study. PLoS One 2014;9:e103447.
27. Croskerry P. Achieving quality in clinical decision making: Cognitive strategies and detection of bias. Acad Emerg Med 2002;9:1184-204.
28. Schuwirth LW, Van Der Vleuten CP. How to design a useful test: The principles of assessment. In: Swanwick T, editors. Understanding Medical Education: Evidence, Theory and Practice. UK: Wiley-Blackwell; 2010. p. 195-207.
29. Forster KI, Davis C. Repetition priming and frequency attenuation in lexical access. J Exp Psychol Learn 1984;10:680-98.