Treanor D, Quirke P, Magee D, and Bulpitt A, **Tracking with virtual slides: A tool to study diagnostic error in histopathology**, *Histopathology*, Vol. 55, pp37-45, 2009
The definitive version is available at www.blackwell-synergy.com or:
http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2559.2009.03325.x/abstract

Tracking with virtual slides: a tool to study diagnostic error in histopathology

Treanor D (1), Lim C (3), Magee D (2), Bulpitt A (2), Quirke P (1)

(1) Pathology and Tumour Biology, Leeds Institute of Molecular Medicine, University of Leeds
(2) School of Computing, University of Leeds
(3) Department of Gastroenterology, Goodhope Hospital

Sources of support
UK Department of Health

Abstract

BACKGROUND: Diagnostic error in pathology is a significant problem. Studying the reasons for error is difficult because of a lack of data on the diagnostic process - virtual slides allow unsupervised study of diagnosis and error.

METHODS: Software was developed to produce visualisations of the diagnostic track followed by pathologists as they viewed virtual slides. These showed the diagnostic path in 4 dimensions (x, y, time and zoom), areas studied for >1000ms, and included pathologists comments about the areas viewed. The system was used to study 2 trainee and 2 expert pathologists diagnosing 60 Barrett's oesophagus biopsies. Comparisons of the diagnostic tracks showed the reason for errors.

RESULTS: 46 cases had an expert consensus diagnosis. The trainees made errors in 21 and 15 cases respectively, of which 11 and 9 were clinically significant. Errors were made across the whole spectrum of diagnoses from negative to intramucosal carcinoma. Detailed examination of the tracks showed that in all errors there was incorrect interpretation of information; in 3 errors there was an additional failure to identify diagnostic features.

CONCLUSIONS: Tracking with virtual slides is a useful tool in studying diagnosis and error which has the potential for use in training and assessment.

Keywords

Virtual slides; diagnosis; error; education; Barrett's oesophagus; training.

Introduction

Diagnosis in surgical histopathology is a highly subjective process which is prone to error. The underlying reasons for error in pathology have not been extensively studied. This paper describes the use of a new technology – tracking with virtual slides – to study diagnostic error in histopathology.

Pathologists in training are taught to search tissue for diagnostic features. They then combine and interpret the features identified in light of their knowledge to come to a diagnosis. There is considerable scope for error during both the information gathering and interpretation stages.

In the diagnosis of dysplasia in the gastrointestinal tract, for example, even with expert observers only moderate agreement has been reported (kappa values of 0.4)[1-3]. When non-experts are included kappa values as low as 0.24 (fair agreement) have been reported [4]. An incorrect diagnosis of high grade dysplasia or cancer could lead to unneccesary oesophagectomy.

Error in histopathology diagnosis is a complex problem with multifactorial causes, some of which can be minimised with quality assurance and management strategies [5]. Error in the diagnostic process itself is more difficult to address.

Error can be categorised as being due to a failure to see a feature on the slide (e.g. failing to see an area of malignant cells in a biopsy) or to a failure to correctly interpret it. Whereas most trainers would recognise these categories, establishing the relative contribution of each is more difficult. Self-reported analysis of diagnostic error ("debriefing") can be misleading as subjects often do not subsequently recall the entire diagnostic reasoning process. Tools to formally examine diagnostic error in a controlled setting are rarely used.

Tracking with modified microscopes has previously been used to train cytoscreeners in proper screening technique for cervical smears [6]. Usually these systems are used simply to ensure that there has been 100% coverage of the slide with screening, rather than interpreting the diagnostic process itself. What was actually examined on the slide was not recorded.

Eyetracking has long been used in psychology to study the cognitive processes undertaken during visual tasks. Eyetracking devices consist of an infrared camera mounted above a display screen. The subject views visual stimuli on the screen, and their eye movements can be recorded with the camera and then superimposed on the original image. Using eyetracking and static histopathological images Tiersma et al. compared diagnostic patterns of pathologists examining 2 cases of cervical intraepithelial neoplasia

[7], and Krupinski et al. studied the evaluation at low magnification of breast cancer [8]. But eyetracking cannot be used to accurately study diagnostic pathology fully as it requires specialised equipment not easily available in pathology laboratories and, crucially, removes the ability of the subject to use the microscope in a normal way (i.e. with panning and zooming).

This latter problem means that the "serial search" approach adopted during histopathological diagnosis cannot be studied without recording the entire diagnostic process on a whole slide. Crowley et al. addressed this problem using glass slides, by taking a video recording of what the subject viewed down the microscope and correlating it with the subject's running commentary.[9]

Virtual slides address this issue in a more flexible way. A virtual slide is produced by scanning a glass slide at high resolution (up to 0.23 microns per pixel [10]). It can be viewed on a standard personal computer with panning and zooming controlled by mouse and keyboard. By recording the co-ordinates being viewed together with a timestamp, a diagnostic track can be obtained which shows exactly what parts of a slide were viewed. Additionally, virtual slides allow unsupervised tracking of trainees (i.e. a trainer can set several training tasks and review the tracks at a later time), even over the internet, and allow systematic study of diagnosis using structured tasks.

The diagnostic track obtained can be useful for training when replayed as a video [11], but detailed analysis of the data in the track allows comparisons to be made between pathologists and conclusions about the diagnostic process to be drawn. While this may be achieved by painstaking video analysis, virtual slides allow automatic generation of simple tracking visualisations which may be analysed far more quickly.

Software was developed to automatically record and visualise diagnostic tracks using virtual slides for the first time. This software was used in an experiment comparing 2 trainee and 2 expert pathologists examining 60 slides of Barrett's oesophagus in order to study the diagnostic process in detail.

Methods

60 cases of Barrett's oesophagus biopsies were selected from the archives of Leeds General Infirmary. Cases were selected to represent a spectrum of diagnoses from negative for dysplasia to intramucosal carcinoma with a significant number of biopsies showing no dysplasia, in order to more accurately represent the daily practice of a pathologist.

The slides were reviewed for technical quality by a consultant pathologist (DT) prior to scanning. A single representative slide was chosen for each case. They were scanned with an Aperio T3 [10] using a 40x objective lens lens to produce a final resolution image of 0.23 microns per pixel. All of the virtual slide images used are freely available to view online at http://www.virtualpathology.leeds.ac.uk/research/barretts.

4 subjects viewed the slides – 2 trainee and 2 expert pathologists as shown in table 1. Both experts were specialists at a national level in gastrointestinal pathology.

Custom-built software was written to track the trainees and experts (figure 1). The software provided a pannable and zoomable virtual slide image to the subject, and recorded a diagnostic track which included a timestamp, x and y co-ordinates, zoom level (magnification), and the specific pan or zoom action taken every time a pan or zoom action was performed. Subjects were aware that their actions were being recorded and timed, but were asked to view the slides in the same way and at the same speed as they would normally examine a diagnostic case.

In order to record the decision making processes being used by the subjects alongside the track taken, the software prompted the subject to mark one or more diagnostic areas on the slide and add an annotated comment explaining what they thought of that area. Subjects were not permitted to progress to the next case until they had marked at least one area and made a comment on it.

When the subject finished viewing the case they were prompted to choose one of 6 diagnostic categories to apply to the case (see table 2).

To decide whether a diagnostic error had been made, a consensus expert diagnosis was determined for every slide. This was the diagnosis when both experts agreed, or the range of diagnoses when they were within 1 diagnostic category of each other. A decision to exclude cases without a consensus diagnosis was made prior to statistical analysis.

A trainee was judged to have made a "correct" diagnosis if their diagnosis was the same as the single consensus expert diagnosis, or within the range of expert diagnoses when they were within 1 diagnostic category of each other; otherwise they were judged to have made an error. When a diagnostic error was detected it was classified as an undercall or overcall (of dysplasia) and as major or minor (if it would or would not alter treatment respectively). For the purposes of analysis, and based on local practice at our institution, hypothetical treatment categories based on the trainees diagnosis were as follows: diagnosis 1-2 = routine follow up; 3-4 intensive follow up; 5-6 surgical or endoscopic intervention.

The diagnostic track taken was analysed with custom-written software in Matlab [12] to produce a visualisation of the track combined with other information about the diagnosis (such as comments made, number of pauses, and total time taken). For the purposes of analysis a significant pause was considered to be one where the subject viewed the area for 1 second or more at 10x or higher magnification. Figure 2 shows an example of the visualisation produced.

These visualisations were qualitatively analysed by a consultant pathologist (DT) to determine the reason for any diagnostic error. Where necessary, the original virtual slide was consulted to clarify decisions about the track. Errors were classified as being of feature identification or feature interpretation. Errors of identification included examination of the tissue at too low magnification, missing a piece of tissue with diagnostic information, or failure to examine all of the levels adequately (based on examination of the heatmaps or tracks produced) - provided that the experts had examined that area of tissue at an appropriate magnification or had annotated it as important. Errors of interpretation were apparent when the subject had adequately examined all of the relevant tissue or correctly annotated a diagnostic area correctly – but failed to interpret the diagnostic meaning of the area. Statistical analysis was performed using SPSS [13]. Comparison of time taken was performed with nonparametric (Mann Whitney) tests and agreement was measured with Cohen's kappa.

Results

Of the 60 slides included in the study, 14 (23%) were excluded because of a lack of consensus diagnosis – the remaining 46 cases had consensus diagnoses as shown in table 3. Agreement between the 2 experts was 53% (kappa 0.38 +/- 0.07 S.E.) before removal of cases without consensus and 70% (kappa 0.57 +/- 0.09 S.E.) after.

Kappa values comparing trainees with experts are shown in table 4. Trainee G's performance was closer to the experts than trainee D, obtaining fair and moderate agreement (kappa values 0.29 and 0.46) with the two experts overall.

Calculating Kappa values for trainee diagnosis versus the consensus diagnosis was not possible as the consensus diagnosis had more categories than the original 6 categories.

Table 5 shows the number of errors made by each trainee. The trainees made errors in 46% and 33% of cases respectively. Although trainee D made more errors than trainee G, the number of major errors (i.e. one which could be clinically significant) was similar in both (errors in 24% and 20% of cases respectively). Major (clinically significant) undercalls were more common than major overcalls (15% vs. 6% respectively).

Errors were made relatively equally in all diagnostic categories from negative to intramucosal carcinoma (see table 7 below). There was evidence that a subset of 12 cases were more difficult to interpret as both trainees made an error in them (see table 6 below).

Visualisations of the tracks produced revealed the reasons for diagnostic error. For example in figure 2 comparing tracks from both trainees with the 2 experts clearly shows the trainees drawing the incorrect conclusion despite correctly identifying and examining the same abnormal tissue as the experts. Such errors of interpretation were made in all 36 diagnostic errors made – in 3 there was an additional component of failure to identify features on the slide.

Trainees spent longer looking at the slides than experts (Figure 3, median 158s vs. 123s, $P < 0.05$). Although the amount of time spent looking at a slide did not significantly vary depending on the diagnostic category applied, when an error was made, the time spent looking at the slide by the trainees was significantly longer (median 243s vs. 155s, $P < 0.05$). Trainees also spent significantly longer than experts looking at cases which were diagnosed by the experts as negative for dysplasia (median 150s vs. 111s, $P < 0.05$). Trainees spent longer examining the slide at high magnification (greater than 10x magnification) than experts (25% vs. 12% of total time ($P < 0.05$)), though both groups spent a similar amount of time at low magnification (less than 5x) (17% and 13% respectively, P = N.S.) – indicating that experts were more able to make a rapid diagnosis at medium magnification (5 -10x magnification)[1]. Trainees did not spend more time at high magnification when they made an error compared to when they made the correct diagnosis (23s vs. 29s, P = N.S.)

Discussion

The aim of this study was to develop a tool to examine the reasons for diagnostic error and to apply it to study the biopsy diagnosis of Barrett's dysplasia. Novel software was developed using virtual slides to track diagnostic behaviour, visualise diagnostic tracks, and compare them. The study explicitly did not seek to compare the diagnosis on glass slides with that on virtual slides.

---

[1] Strictly speaking, magnification is not the correct term to use with virtual slides, as the size of the image depends on both the resolution of the image and the properties of the monitor used. For simplicity here we refer to "5x" and "10x" magnification respectively. In reality the correct description is 12.5% and 25% zoom respectively, relative to the resolution at which the virtual slides were scanned (100%, using a 40x lens).

The difficulty of this area of diagnostic pathology was underlined by the finding of only fair agreement between two expert pathologists (53% agreement for 60 cases, kappa 0.38).

Trainees made clinically significant errors in 22% of cases. Detailed analysis of tracking information revealed that most errors were due to incorrect interpretation, and none were solely due to failure to identify abnormalities on the slide. This contrasts with cervical screening cytology where it is believed that failure to identify or find features on the slide has a significant contribution to error, and laboratory processes have been developed to rescreen slides in order to minimise this problem [14]. Other studies (using conventional microscopes) have also found so-called errors of search to be a minority cause of error [9].

Both trainees were able to correctly identify areas of concern, but their interpretation of the changes seen was frequently incorrect. For example in figure 2 they were aware of the significance of hyperchromasia and nuclear crowding but failed to realise the severity of these histological changes and their significance – Trainee G comments that it "looks degenerative".

Further information about error can be obtained from timing data. Trainees spent a median of 35 seconds (28%) longer looking at cases than experts ($P < 0.05$). In other studies trainees have been reported to take longer overall to make a diagnosis, be slower to generate hypotheses than experts [9], scan slides more slowly than experts (7.1s v. 4.5s) and examine diagnostic areas for less time [8]. In radiology too, experts have been found to make decisions more quickly (a single eye fixation is enough for experienced radiologists to detect and identify major pathological features with 70% accuracy [15]).

When an error was made, trainees spent 70% longer (103s) looking at the slide than experts– indicating either that there was more diagnostic information to absorb or that they had realised the difficulty of the case and were spending longer examining it. Despite this longer study time, an incorrect conclusion was made. A similar trend has been reported in radiology, where prolonging search beyond a certain time (labelled the "global recognition phase" – i.e. the early impressions of the image) was associated with error. [16]

Differences between trainees and experts may be due to difficulties with information processing. In this study pathologists made an average of 271 pan and 11 zoom actions in the course of examining each slide; Krupinski et al documented expert pathologists making saccadic eye movements 14.5 times during the 4.5 seconds they took to decide

which areas were important in one low magnification image [8]. So large amounts of image information must be processed when viewing slides.

When examining complex image data, experts tend to ignore features that are not relevant to the interpretation. Compared to trainees, experienced radiologists have worse memory of normal radiographs but better memory of abnormal ones– indicating that experts learn to selectively detect abnormalities and ignore normal features in order to reduce the processing burden during image interpretation [17].

Similarly Lesgold hypothesised that perceiving features may interfere with interpretation and diagnosis during training [9]– trainees have not yet learned to ignore irrelevant data so the task of information processing is harder for them.

LIMITATIONS OF THE EXPERIMENT

This study is limited by the small sample size, the possibility that the observation will have altered subjects behaviour, and its use of virtual slides rather than glass slides to examine diagnosis. Given these limitations, however, we believe that tracking with virtual slides is a useful tool in studying diagnostic error and the acquisition of expertise in pathology.

IMPLICATIONS OF THE FINDINGS

Microscopy remains the most cost-effective and accurate way to diagnose many diseases – even in Barrett's oesophagus where there is clear variation in diagnostic performance [18]. We have confirmed that trainees make mistakes due to incorrect interpretation, and surmise that this may be due to an inability to process the information on the slide efficiently. Training strategies could take advantage of this finding to improve diagnosis. One approach has been to formulate heuristics for diagnosis. For example "where there is nuclear pleomorphism and crowding ensure there is no acute inflammation before diagnosing dysplasia". Tracking with virtual slides could be used to objectively identify the diagnostic entities and specific histological appearances which cause error in all areas of pathology.

More complex strategies may take advantage of the fact that much diagnostic reasoning is Bayesian – systems which enforce systematic Bayesian reasoning have had success in improving pathology diagnosis [19-21]. In our study there were more clinically significant undercalls than overcalls, so training strategies which emphasised the importance of features indicating dysplasia could help with diagnosis. The effect of encouraging Bayesian reasoning may be in clarifying cognitive processes and encouraging trainees to filter extraneous data from the problem.

Medical educationalists now believe that simply teaching generalised heuristic (problem solving) skills to trainees is not enough to develop expertise, as much knowledge is content specific[22]. Indeed, emulating experts can be difficult as they do not use one strategy: "Clinicians often unconsciously use multiple, combined strategies to solve clinical problems, suggesting a high degree of mental flexibility and adaptability in clinical reasoning".[23]

With virtual slides experts could be tracked while making diagnoses and the information presented in training material (for example as summarised strategies, narrated videos, or interactive tutorials using virtual slides).

The inability of trainees to filter diagnostically relevant information appropriately leads to error. Therefore providing annotated and classified reference images, to compensate for the lack of mental images of diagnostic categories may be helpful. Almost all pathologists currently refer to images in books to assist with diagnosis – online databases of virtual slides or diagnostic images may assist in diagnosis by providing large datasets for comparison with index cases [24, 25].

Experts progress from interpreting features to pattern matching based on previous experience. If this is true then strategies to train pathologists should also expose trainees to high volumes of material as well as train them to recognise and interpret features - "a critical element of becoming an expert is accruing the vast experience that enables experts to recognize patterns effortlessly most of the time — and to recognize, as well, when the signs and symptoms do not fit a pattern at all"[26]. Again, virtual slides can assist in this by providing easily accessible libraries of cases categorised by diagnostic category [27].

Furthermore, tracking of the trainees prospectively could then be used in e-learning systems to provide feedback and compare with expert tracks. By combining robust reasoning strategies with exposure to many cases, trainees may then progress towards the intuitive and rapid strategies that experts use in pathology diagnosis.

Acknowledgements

Thanks to Prof. Mike Dixon, Dr. Abdul Ganjifrockwala, Dr. Sanjeev Katti, and Dr. Nigel Scott for their participation.

References

1. Montgomery E, Bronner MP, Goldblum JR *et al.* Reproducibility of the diagnosis of dysplasia in barrett esophagus: A reaffirmation. *Hum.Pathol* 2001;**32**;368-378.
2. Ormsby AH, Petras RE, Henricks WH *et al.* Observer variation in the diagnosis of superficial oesophageal adenocarcinoma. *Gut* 2002;**51**;671-676.
3. Lim CH, Treanor D, Dixon MF, Axon AT. Low-grade dysplasia in barrett's esophagus has a high risk of progression. *Endoscopy* 2007;**39**;581-587.
4. Lorinc E, Jakobsson B, Landberg G, Veress B. Ki67 and p53 immunohistochemistry reduces interobserver variation in assessment of barrett's oesophagus. *Histopathology* 2005;**46**;642-648.
5. Nakhleh RE. Patient safety and error reduction in surgical pathology. *Arch Pathol Lab Med* 2008;**132**;181-185.
6. Kamentsky LA, Gershman RJ, Kamentsky LD, Pomeroy BM, Weissman ML. Compucyte corporation. Pathfinder system: Computerizing the microscope to improve cytology quality assurance. *Acta Cytol* 1996;**40**;31-36.
7. 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**;677-680.
8. Krupinski EA, Tillack AA, Richter L *et al.* Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Hum.Pathol.* 2006;**37**;1543-1556.
9. Crowley RS, Naus GJ, Stewart J, III, Friedman CP. Development of visual diagnostic expertise in pathology -- an information-processing study. *J.Am.Med.Inform.Assoc.* 2003;**10**;39-51.
10. Aperio technologies inc., san diego, california.
11. Johnston DJ, Costello SP, Dervan PA, O'Shea DG. Development and preliminary evaluation of the vps replaysuite: A virtual double-headed microscope for pathology. *BMC.Med Inform.Decis.Mak.* 2005;**5**;10.
12. Matlab. Natic, MA: The Mathworks Inc.
13. Statistical package for social sciences. Chicago, Illinois: SPSS Inc.
14. Arbyn M, Schenck U. Detection of false negative pap smears by rapid reviewing. A metaanalysis. *Acta Cytol* 2000;**44**;949-957.
15. Kundel HL, Nodine CF. Interpreting chest radiographs without visual search. *Radiology* 1975;**116**;527-532.
16. Nodine CF, Mello-Thoms C, Kundel HL, Weinstein SP. Time course of perception and decision making during mammographic interpretation. *AJR Am J Roentgenol.* 2002;**179**;917-923.
17. Myles-Worsley M, Johnston WA, Simons MA. The influence of expertise on x-ray image processing. *J Exp.Psychol.Learn.Mem.Cogn* 1988;**14**;553-557.
18. Guidelines for the diagnosis and management of barrett's columnar lined oesophagus. A report of the working party of the british society of gastroenterology. British Society of Gastroenterology, 2005.
19. Heckerman DE, Horvitz EJ, Nathwani BN. Toward normative expert systems: Part i. The pathfinder project. *Methods Inf.Med* 1992;**31**;90-105.
20. Hamilton PW, Montironi R, Abmayr W *et al.* Clinical applications of bayesian belief networks in pathology. *Pathologica* 1995;**87**;237-245.
21. Morrison ML, McCluggage WG, Price GJ *et al.* Expert system support using a bayesian belief network for the classification of endometrial hyperplasia. *J.Pathol.* 2002;**197**;403-414.

22. Norman G. Research in clinical reasoning: Past history and current trends. *Medical education* 2005;**39**;418-427.
23. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. *The New England Journal of Medicine* 2006;**355**;2217-2225.
24. van Ginneken AM, Baak JP, Jansen W, Smeulders AW. Evaluation of a diagnostic encyclopedia workstation for ovarian pathology. *Hum Pathol* 1990;**21**;989-997.
25. Virtual slides at the university of leeds.
26. Norman G. Building on experience--the development of clinical reasoning. *N Engl J Med* 2006;**355**;2251-2252.
27. Treanor D, Waterhouse M, Lewis F, Quirke P. A virtual slide library for histopathology. *Annual Meeting of the Pathological Society of Great Britain and Ireland.* Glasgow, 2007.

Tables

Subject	Age	Experience of pathology	Number of virtual slides seen before study
Expert B	50-60	25 years	< 5
Expert E	60-70	30 years	100
Trainee D	20-30	3 years	< 5
Trainee G	20-30	3 years	< 5

Table 1 Characteristics of pathologists in the study. Both trainees had 3 years experience of pathology and had passed MRCPath part 1 examination. Both experts were senior pathologists specialising in gastrointestinal pathology.

Category	Description
1	Negative
2	Indefinite (Probably negative)
3	Indefinite (Probably dysplastic)
4	Low grade dysplasia
5	High grade dysplasia
6	Intramucosal carcinoma

Table 2 Diagnostic categories used in the study, modified from BSG guidelines for diagnosis of dysplasia in Barrett's oesophagus [18]

Consensus diagnosis	Frequency	Percent
1	18	39
1 to 2	8	17
2	1	2
2 to 3	3	7
3 to 4	1	2
4	5	11
5	4	9
5 to 6	2	4
6	4	9
Total	**46**	**100**

Table 3 Frequency of consensus diagnoses amongst the 46 cases where a consensus diagnosis was reached. The consensus diagnosis of 1 to 6 refers to the six categories in table 2. Cases represented the full spectrum of dysplasia with significant numbers of cases negative for dysplasia in order to better replicate daily practice.

	Expert B		Expert E	
	Agreement (%)	Kappa	Agreement (%)	Kappa
Trainee D	41	0.17 (0.02 – 0.32)	50	0.27 (0.12 – 0.42)
Trainee G	50	0.29 (0.11– 0.47)	63	0.46 (0.28 -0.65)

Table 4 Interobserver agreement between experts and trainees for the 46 cases where there was consensus diagnosis. Trainee G had better agreement with both experts than trainee D, but even so the best agreement achieved (between trainee G and expert E) was only 63%.

Number of errors (%)		Trainee D	Trainee G
Overcall	Major	3 (7%)	3 (7%)
Overcall	Minor	9 (20%)	2 (4%)
Undercall	Major	8 (17%)	6 (13%)
Undercall	Minor	1 (2%)	4 (9%)
Total errors made		21 (46%)	15 (33%)
Correct diagnosis made		25 (54%)	31 (67%)
Total cases		**46 (100%)**	**46 (100%)**

Table 5 Frequency of errors made by trainees. Overcalls and undercalls refer to whether the trainee over or underdiagnosed dysplasia or cancer. Major errors were those which would be clinically significant; minor errors would not generally be clinically significant.

		Error made by G		
		No	Yes	Total
Error made by D	No	22	3	25
	Yes	9	12	21
Total		31	15	46

Table 6 Table comparing incidence of errors between the two trainees. A subset of 12 cases were incorrectly interpreted by both trainees; in an additional 12 cases errors were made by only one of the trainees.

Error made	Consensus diagnosis									
	1	1 to 2	2	2 to 3	3 to 4	4	5	5 to 6	6	Total
Major overcall	3	1		1						5
Minor overcall	3	2	1	1		1	3			11
Major undercall					1	5	3	2	4	15
Minor undercall			1	2					2	5
No error made	3	13		2	1	4	2	2	2	56
Total	36	16	2	6	2	1	8	4	8	92

Table 7 Frequency of errors made by consensus diagnosis. Errors were made across all groups of consensus diagnosis with higher frequency extremes of diagnoses as would be expected.

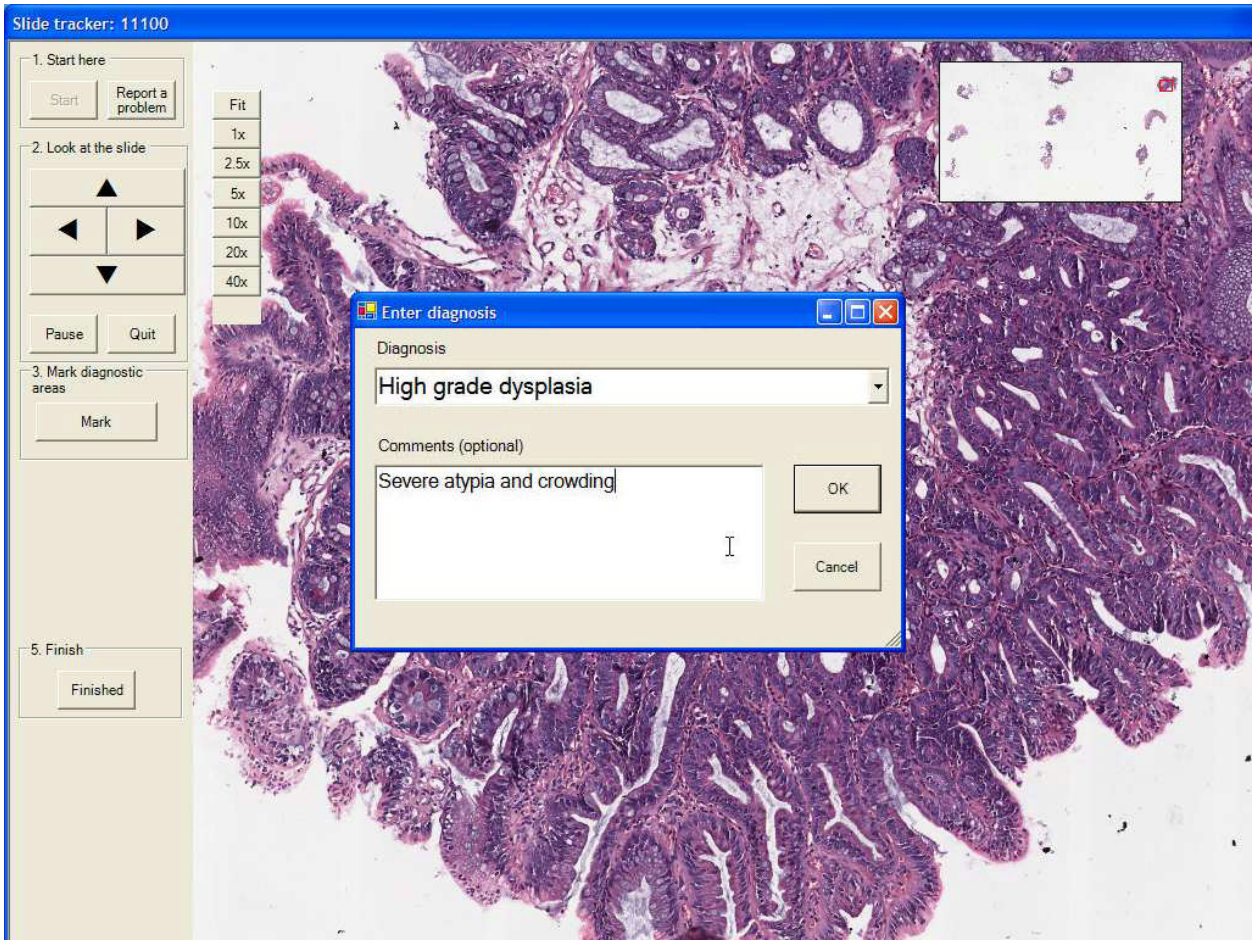

Figure 1

Virtual slide viewing software. The virtual slide is displayed in the centre. The user navigates using the arrow buttons on the left of the screen, keyboard and mouse. They can drag the image to pan around the slide. Pressing one of the numbered buttons zooms to that magnification. A thumbnail of the virtual slide is present in the top right corner of the screen. Clicking on the thumbnail pans the view to the selected part of the slide.

Participants must annotate a diagnostic area of the slide by drawing a box with the mouse. They are prompted to provide a comment or explanation for the area they have annotated. They must annotate at least one area of the slide before they can make a diagnosis and proceed to the next case. The annotated area is marked with a green box. Participants choose one of six diagnoses, and provide comments on the reason for their diagnosis.

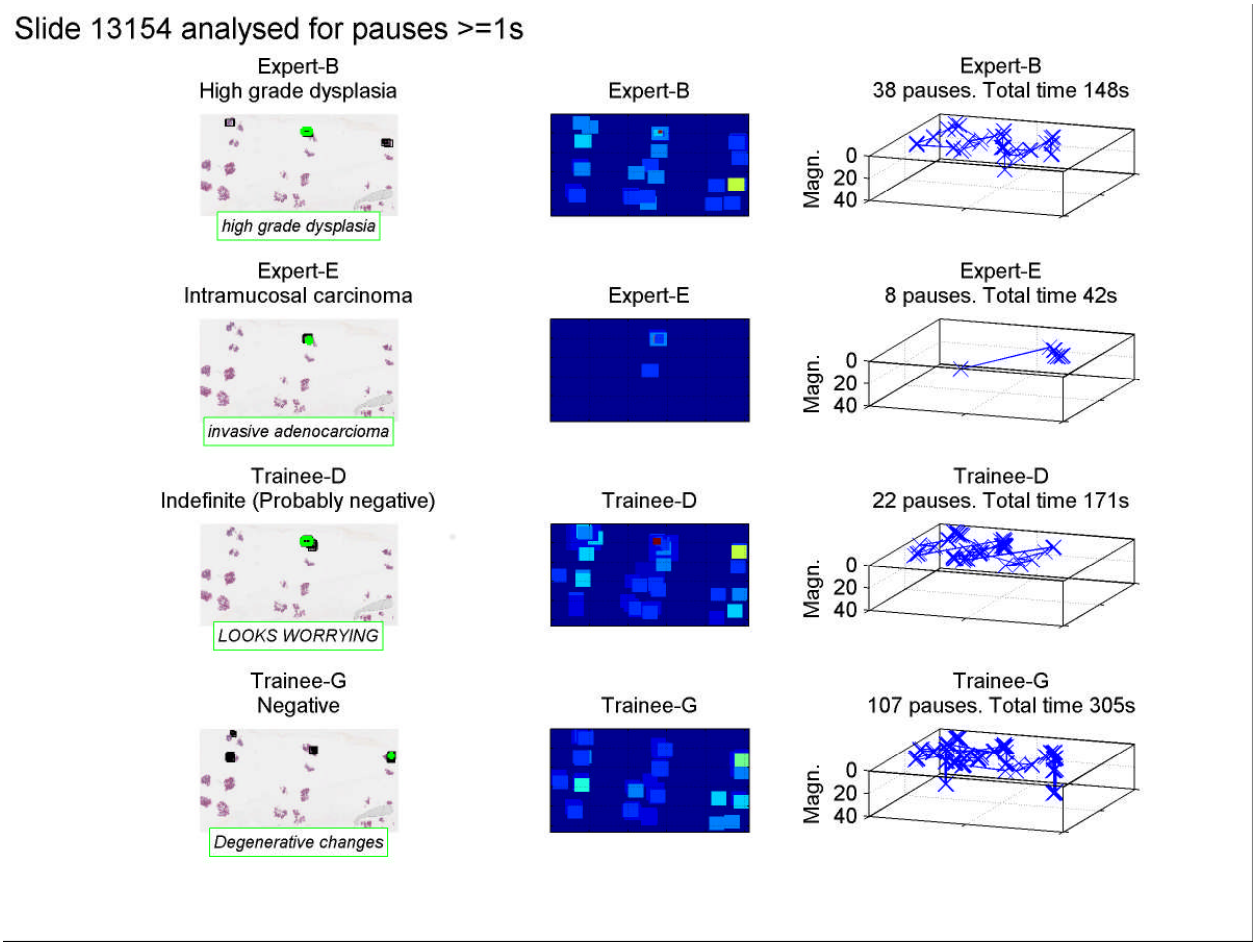

Figure 2

Example of visualisation produced by analysing tracking data for 4 pathologists viewing one slide. 3 graphical representations of the track for each pathologist are present from left to right as follows: (a) a track superimposed on the image to indicate the path followed, (b) a heatmap generated by adding all pauses greater than 1000ms for each x and y pixel of the slide (where colour visualisations of time spent at each point were obtained by multiplying the magnification of the view by the time in seconds), (c) a 3-dimensional plot of the path in x, y, and z (zoom) dimensions.

In this case the consensus expert diagnosis is high grade dysplasia-intramucosal carcinoma. 3 subjects examined all pieces of tissue, but expert E made a rapid decision after examining just 2 pieces of tissue. Both trainees correctly identified the topmost biopsy as being abnormal, but both underestimated the seriousness of the histological changes and erroneously failed to diagnose dysplasia.

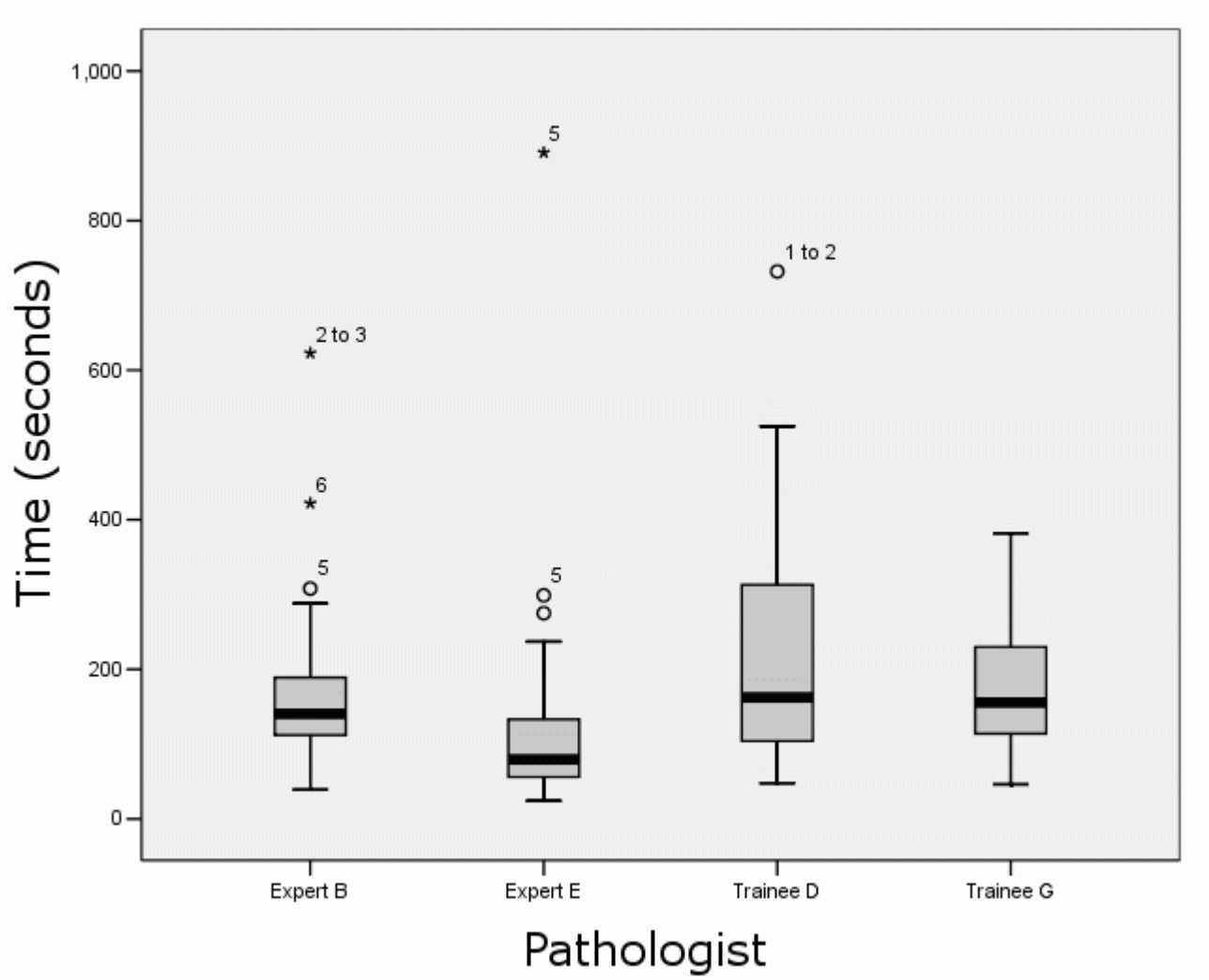

Figure 3

Boxplot of time taken to reach a diagnosis for the 4 pathologist subjects. Boxes show 25th - 75th centiles, error bars show 95% confidence intervals, lines show median. In general the trainees took longer to make a diagnosis than experts (mean 188 seconds vs. 141 seconds, $P < 0.05$)