NIHR | National Institute for Health and Care Research

Health Technology Assessment

Extended Research Article

Variation within and between digital pathology and light microscopy for the diagnosis of histopathology slides: blinded crossover comparison study

David RJ Snead,[1,2*] Ayesha S Azam,[1,2] Jenny Thirlwall,[2] Peter Kimani,[2] Louise Hiller,[2] Adam Bickers,[3] Clinton Boyd,[4] David Boyle,[4] David Clark,[5] Ian Ellis,[5,6] Kishore Gopalakrishnan,[1] Mohammad Ilyas,[5,6] Paul Kelly,[4] Maurice Loughrey,[4,7] Desley Neil,[8] Emad Rakha,[5,6] Ian SD Roberts,[9] Shatrughan Sah,[1] Maria Soares,[9] YeeWah Tsang,[1] Manuel Salto-Tellez,[7,10] Helen Higgins,[2] Donna Howe,[2] Abigail Takyi,[1] Yan Chen,[5] Agnieszka Ignatowicz,[11] Jason Madan,[2] Henry Nwankwo,[2] George Partridge[5] and Janet Dunn[2]

[1]Histopathology, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK
[2]Warwick Medical School, University of Warwick, Coventry, UK
[3]Pathlinks, Northern Lincolnshire and Goole NHS Foundation Trust, Lincoln, UK
[4]Institute of Pathology, Belfast Health and Social Care Trust, Belfast, Northern Ireland, UK
[5]Histopathology Department, Nottingham University Hospital NHS Trust, Nottingham, UK
[6]School of Medicine, University of Nottingham, Nottingham, UK
[7]Centre for Public Health, Queen's University, Belfast, Northern Ireland, UK
[8]Department of Cellular Pathology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[9]Department of Cellular Pathology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[10]Integrated Pathology, Institute for Cancer Research London, London, UK
[11]Institute of Applied Health Research, University of Birmingham, Birmingham, UK

[*]Corresponding author david.snead@uhcw.nhs.uk

Published July 2025
DOI: 10.3310/SPLK4325

Scientific summary

Variation within and between digital pathology and light microscopy for the diagnosis of histopathology slides: blinded crossover comparison study

Health Technology Assessment 2025; Vol. 29: No. 30
DOI: 10.3310/SPLK4325

NIHR Journals Library www.journalslibrary.nihr.ac.uk

Scientific summary

Background

There is considerable interest in the development of digital pathology (DP) as a means of reporting histopathology samples. The flexibility that electronic distribution of the reporting workload permits is seen as an important development to improve quality and efficiency of histopathology, which is currently a major cause of delay in many cancers as well as many other chronic disease pathways. Previous studies have not reported on cancer screening samples and include few large (1000 plus cases) multisite studies. In addition, some studies have shown there may be important differences in the way pathologists report cases on DP compared with light microscopy (LM), particularly with reference to identifying bacteria, grading dysplasia, recognising calcium oxalate crystals or small nodal metastases. Concerns over the quality of evidence supporting DP in cancer screening samples led to an embargo on the use of the technology for reporting these samples pending further data, which remains in place.

Additional interest lies in understanding how transformational change of this character will be seen by pathologists and laboratory technicians and how it may impact on existing laboratory workflow. The change to DP requires capital investment in slide scanning equipment, workstations, computer servers and networking infrastructure, all of which will place considerable strain on already overstretched information technology resources. Therefore, there is considerable interest in how these investment costs may be offset by improved efficiency in the service, particularly in whether DP provides any advantage to the speed of reporting slides over conventional LM. Finally, since in radiology, which as a diagnostic imaging modality shares some parallels with DP, the use of eye-tracking studies has led to an understanding of how poor examination technique can contribute to errors in screening images, we were interested to learn if similar approaches may be relevant to DP.

Objectives

The primary objective was to estimate intra-pathologist agreement between reports issued on DP in comparison to LM. The secondary objectives were to estimate inter-pathologist agreement for LM reports, estimate inter-pathologist agreement for DP reports and compare diagnosis confidence for LM and DP reports.

A qualitative study to understand the views of pathologists and technicians on the impact DP on laboratory practice was conducted before and during the study. A health economics study analysed measurements made on how long the reporting of cases took using DP in comparison to LM, and an eye-tracking study examined different pathologists' examination techniques using DP.

Methods

The main study was a multicentre validation comparison study, with a blinded crossover design measuring intraobserver variability of pathologists' diagnoses of histopathology samples using LM and DP and interobserver variability measuring pathologists' diagnoses on LM and DP against consensus ground truth (GT). Pathologists recorded confidence of diagnoses made on a seven-point Likert scale, and recorded the time taken to report the cases. A questionnaire survey was undertaken examining the viewpoints of a range of pathologists and laboratory technicians at the start and during the course of the study. Eye tracking of pathologists was undertaken on a subset of study cases examining the technique used by different pathologists.

Equipment and training

Two DP systems, designed for high-throughput scanning of histopathology slides, were used to digitise the slides which were stored on image repositories provided by the manufacturers. Whole slide images were examined on

computer workstations matching the manufacturer's specifications, via internet-enabled connections to the servers. All pathologists received training on the use of the DP system which followed the Royal College of Pathologists best practice guidance. Reporting of cases was carried out by pathologists blinded to the reference diagnosis, the reports of other pathologists and, in the second view, from their initial report of the case. All annotations made on the slides were hidden from the other pathologists. The 6-week gap between viewings was managed independently by the trial management team.

Samples

In breast, gastrointestinal (GI) and skin, the majority (80%) of cases were recruited from the laboratories taking part in the study as sequential cases. These were enriched by 20% of cases from conditions or sample types deemed to be either moderately difficult or difficult. Renal sequential cases were used without enrichment for difficulty.

Arbitration

Reports issued by the study pathologists were compared by independent reviewers blinded to the pathologists and modality. Differences detected were reviewed by an independent arbitration team blinded to the pathologist and modality into differences which alter management of the patient (clinically significant) and those which would not (clinically insignificant). The arbitration team included clinical colleagues to assist in deciding if differences were significant or not.

Primary and secondary end points

The primary end point was intra-pathologist clinical management concordance (CMC) meaning identical diagnoses plus differences which do not affect patient management in LM compared to DP. Secondary end points were inter-pathologists' CMC of LM and DP compared to GT, the level of complete concordance between the two modalities and pathologists, and the pathologist's rating of the confidence of their diagnosis on a seven-point Likert scale.

Sample size

Target recruitment was 2000 cases: 600 cases each for breast, skin and GI specialties, and 200 cases for renal. Sample size adequacy was based on getting a precise estimate [narrow confidence interval (CI)] for percentage CMC. The mix for routine cases, moderately difficult to read cases and difficult cases was assumed to be 70%, 20% and 10%, respectively. Percentage CMC for routine and difficult cases were assumed to be respectively 98.8% (Snead *et al.*, 2016) and 55% (based on 40–70% range found in literature), and 75% for moderate cases (mid-point between routine and difficult). Consequently, the overall percentage CMC was assumed to be 90%. There were four LM versus DP comparisons arising from four pathologists diagnosing each case and intraclass correlation (ICC) was assumed to be 0.8 so that the design effect was [1 + ICC (comparisons per case – 1)] = 3.4. We took 2400 (600 × 4) breast reports to correspond to 705 (2400/3.4) independent reports to give a margin of error of 2.2%. So, precision was high while analysing breast, skin and GI specimens separately. Due to smaller sample size, for renal, the margin of error was 3.1%.

Statistical analysis

An intra-pathologist agreement was estimated by computing percentage LM versus DP CMC using a random-effects (RE) logistic regression model with crossed RE terms for pathologist and case. A logistic regression model was used because the outcome was binary, whether there was CMC between LM and DP diagnoses or not, and the RE terms were crossed because within a specialty, each pathologist reported all cases, and each case was reported by all four pathologists so that there was no nesting. The percentage CMC obtained was referenced to 98.3%, the pooled

percentage CMC in a recent meta-analysis. Inter-pathologist agreement for LM reports was estimated by computing ICC from a RE logistic regression model with crossed RE terms for pathologist and case with the outcome being whether there is LM versus GT CMC. The ICC to quantify inter-pathologist agreement for DP reports was computed using a similar model. Diagnosis confidence level was one of seven consecutive integer scores. Therefore, because diagnosis confidence scores could not be assumed to be normally distributed, a RE generalised Poisson model with crossed RE terms for pathologist and case was used to analyse the scores. Rate ratio of LM and DP Poisson mean rates from this model was used to make inferences on the difference between LM and DP diagnosis confidence. The analysis was performed on all cases and repeated in subgroup analysis by specialty, case difficulty and by screening.

Inclusion/exclusion criteria

The majority of cases were chosen from sequential histopathology cases within the relevant specialty group, entering the recruiting laboratories. Samples with either broken or missing slides, or with missing clinical data were excluded; oversized slides from cases with megablocks were excluded. Cases where a prior sample was important to the interpretation of the study sample were also excluded.

Results

A total of 2024 cases were included in the study. These comprised 608 breast, 607 GI, 609 skin and 200 renal. Cancer screening samples from the breast cancer screening service numbered 207 (34%) and there were 250 (41%) samples from the large bowel cancer screening programme. Overall, the primary end-point LM versus DP comparisons showed CMC levels were 99.95% (95% CI 99.91 to 99.97). Similar results were observed within specialties groups, namely, breast 99.40% (95% CI 99.06 to 99.62); GI 99.96% (95% CI 99.89 to 99.99); skin 99.99% (95% CI 99.92 to 100.0); and renal 99.99% (95% CI 99.57 to 100.0). Within cancer screening cases, overall CMC was 98.96% (95% CI 98.42 to 99.32), breast 96.27 (94.63 to 97.43), large bowel 99.93 (99.68 to 99.98).

Pathologists recorded high levels of confidence in all specialty groups, with higher confidence seen in LM compared with DP, although this was not statistically significant.

The qualitative study showed there were a range of views expressed on the impact of DP. In order to achieve wide acceptance, it is important DP needs to integrate seamlessly into the laboratory workflow. The advantages DP offers will not be realised if on implementation pathologists and/or technicians have to constantly move between systems to complete tasks or if networking speed impacts the systems performance. The need for accurate data on the benefits of DP is likely to be important in helping laboratories make the decision to transition to DP. Successful implementation requires careful planning avoid the many potential pitfalls.

The health economics study showed no clear advantage with either modality, but clear evidence about pathologists' speed in reporting with DP improved over the course of the study. While there are likely to be considerable benefits in transitioning to DP, the differences in time taken to report cases between the two modalities appear very small and probably insignificant.

The eye-tracking study showed that a collection of data relating to slide examination is feasible and there was a clear correlation between experience and diagnostic accuracy. There were differences in examination technique between experienced and less experienced pathologists, with the latter showing greatly more efficient slide examination, and more use of low and intermediate power, with targeted use of high-power objectives. Experienced pathologists were quicker to recognise features and move on than less experienced pathologists.

Conclusions

This is the first study to comprehensively examine intra-pathologist and inter-pathologist variability using LM and DP compared to a consensus GT on the same set of slides, and the first study to examine cancer screening samples. The results show pathologists give equivalent results with either modality in all the areas studied. No trends to favour either modality were identified, even concerning the identification of small objects such as the detection of *Helicobacter pylori* or the grading of dysplasia.

The study provides definitive evidence that pathologists provide equivalent results when using DP as they would using LM. However, pathologists did show a trend to increased confidence using LM compared with DP which did not reach statistical significance, but which may reflect the improved resolution of this modality, and/or increased familiarity with LM.

This is also the first study to assess DP as a means of reporting native and transplant renal biopsies, including assessing fluorescence-stained slides. This is a potentially transformational technology for this specialty. Renal biopsies represent a small-volume, highly complex area of diagnostic pathology. Providing adequately trained pathologists to serve the needs of these patients across the country is a major challenge, even more so to provide cover out of hours which is needed to support the care of renal transplant recipients in need of urgent assessment. The results, particularly from the renal biopsy cases which demand fine resolution for interpretation, suggest DP is very likely to be suitable for other specialty areas with similar demands, such as haematopathology and neuropathology. Furthermore, immunofluorescence-stained sections are non-permanent. DP provides a potential solution for all these challenges, enabling difficult cases to be shared with experts many miles distant from the host laboratory, thereby providing the basis for a more resilient service 24/7. Finally, DP images provide, for the first time, a permanent record of the fluorescence-stained sections performed as a routine on native biopsies.

Study registration

This study is registered as ISRCTN14513591.

Funding

This award was funded by the National Institute for Health and Care Research (NIHR) Health Technology Assessment programme (NIHR award ref: 17/84/07) and is published in full in *Health Technology Assessment*; Vol. 29, No. 30. See the NIHR Funding and Awards website for further award information.

Health Technology Assessment

ISSN 2046-4924 (Online)

Impact factor: 3.5

A list of Journals Library editors can be found on the NIHR Journals Library website

Launched in 1997, *Health Technology Assessment* (HTA) has an impact factor of 3.5 and is ranked 30th (out of 174 titles) in the 'Health Care Sciences & Services' category of the Clarivate 2022 Journal Citation Reports (Science Edition). It is also indexed by MEDLINE, CINAHL (EBSCO Information Services, Ipswich, MA, USA), EMBASE (Elsevier, Amsterdam, the Netherlands), NCBI Bookshelf, DOAJ, Europe PMC, the Cochrane Library (John Wiley & Sons, Inc., Hoboken, NJ, USA), INAHTA, the British Nursing Index (ProQuest LLC, Ann Arbor, MI, USA), Ulrichsweb™ (ProQuest LLC, Ann Arbor, MI, USA) and the Science Citation Index Expanded™ (Clarivate™, Philadelphia, PA, USA).

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta.

Criteria for inclusion in the *Health Technology Assessment* journal

Manuscripts are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

Health Technology Assessment (HTA) research is undertaken where some evidence already exists to show that a technology can be effective and this needs to be compared to the current standard intervention to see which works best. Research can evaluate any intervention used in the treatment, prevention or diagnosis of disease, provided the study outcomes lead to findings that have the potential to be of direct benefit to NHS patients. Technologies in this context mean any method used to promote health; prevent and treat disease; and improve rehabilitation or long-term care. They are not confined to new drugs and include any intervention used in the treatment, prevention or diagnosis of disease.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

This article

The research reported in this issue of the journal was funded by the HTA programme as award number 17/84/07. The contractual start date was in November 2023. The draft manuscript began editorial review in July 2023 and was accepted for publication in April 2024. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' manuscript and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this article.

This article presents independent research funded by the National Institute for Health and Care Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care.

This article was published based on current knowledge at the time and date of publication. NIHR is committed to being inclusive and will continually monitor best practice and guidance in relation to terminology and language to ensure that we remain relevant to our stakeholders.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Newgen Digitalworks Pvt Ltd, Chennai, India (www.newgen.co).