NIHR | National Institute for Health and Care Research

Journals Library

Check for updates

Health Technology Assessment

Volume 29 • Issue 30 • July 2025

ISSN 2046-4924

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

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

DOI 10.3310/SPLK4325

NIHR | National Institute for Health and Care Research

Health Technology Assessment

Check for updates

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

This report should be referenced as follows:

Snead DRJ, Azam AS, Thirlwall J, Kimani P, Hiller L, Bickers A, *et al.* Variation within and between digital pathology and light microscopy for the diagnosis of histopathology slides: blinded crossover comparison study. *Health Technol Assess* 2025;**29**(30). https://doi.org/10.3310/SPLK4325

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).

Abstract

Background: Digital pathology refers to the conversion of histopathology slides to digital image files for examination on computer workstations as opposed to conventional microscopes. Prior to adoption, it is important to demonstrate pathologists provide equivalent reports when using digital pathology in comparison to bright-field and immunofluorescent light microscopy, the current standard of care.

Objective: A multicentre comparison of digital pathology with light microscopy for reporting of histopathology slides, measuring variation within and between pathologists on both modalities.

Design: A blinded crossover 2000-case study estimating clinical management concordance (identical diagnoses plus differences not affecting patient management). Each sample was assessed twice by four pathologists (once using light microscopy, once using digital pathology, the order randomly assigned and a 6-week gap between viewings). Random-effects logistic regression models, including crossed random-effects terms for case and pathologist, estimated percentage clinical management concordance. Findings were interpreted with reference to 98.3% concordance (Azam AS, Miligy IM, Kimani PKU, Maqbool H, Hewitt K, Rajpoot NM, Snead DRJ. Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis. *J Clin Pathol* 2021;**74**:448–55. https://doi.org/10.1136/jclinpath-2020-206764).

Setting: Sixteen consultant pathologists, four for each specialty, from six National Health Service laboratories. Experience ranged from 3 to 35 years. Some were early adopters of digital pathology, but the majority were new to digital pathology.

Interventions: Eight viewings per sample (four pathologists with light microscopy and with digital pathology), culminating in a consensus ground truth, enabling measurement of agreement within and between readers. Samples enrolled reflected routine practice, included cancer screening biopsies, and were enriched for areas of difficulty [e.g. dysplasia (7, 10, 11)]. State-of-the-art digital pathology equipment designed for diagnosis, and holding either Conformité Européene or Food and Drug Administration approval, was used.

Main outcome: Intra-pathologist variation between reports issued on digital pathology and light microscopy, inter-pathologist variation against ground-truth diagnosis using light microscopy and digital pathology.

Secondary outcomes: Pathologist-recorded reporting times, along with their confidence in diagnosis, analysis of eye-tracking evaluating examination techniques, and a qualitative study examining attitudes of pathologists and laboratory staff to digital pathology adoption.

Results: Two thousand and twenty-four cases (608 breast, 607 gastrointestinal, 609 skin, 200 renal) were recruited, with breast and gastrointestinal including screening samples [207 (34%) breast, 250 (41%) gastrointestinal]. Overall, in light microscopy versus digital pathology comparisons, clinical management concordance levels were 99.95% (95% confidence interval 99.91 to 99.97). Similar results were observed within specialties [breast: 99.40% (95% confidence interval 99.06 to 99.62); gastrointestinal 99.96% (95% confidence interval 99.89 to 99.99); skin 99.99% (95% confidence interval 99.92 to 100.0); renal 99.99% (95% confidence interval 99.57 to 100.0)], and within screening cases [98.96% (95% confidence interval 98.42 to 99.32), breast 96.27% (94.63 to 97.43), gastrointestinal 99.93% (95% confidence interval 99.68 to 99.98)]. Reporting time between digital pathology and light microscopy was similar, but pathologists became faster on digital pathology with familiarity. Pathologists recorded high levels of confidence in their diagnosis with light microscopy, significantly higher than digital pathology.

Limitations: Cytology cases and specialty groups outside those tested were not examined. The study used two digital pathology scanning systems. Other systems available on the market were not tested.

Conclusions: Clinical management concordance levels between the two modalities exceed the reference 98.3% in breast, gastrointestinal, skin and renal specialties, and pooled breast and large bowel cancer screening cases. Subgroup analysis of clinically significant differences revealed a range of differences including areas where interobserver variability is known to be high, which were distributed between reads performed with both platforms and without apparent trends to either.

Future work: The use of digital pathology for cytology samples remains an area for further research.

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.

Contents

List of tables

List of figures

List of abbreviations

AI	artificial intelligence
BCC	basal cell carcinoma
CC	complete concordance
CMC	clinical management concordance
CMV	cytomegalovirus
DCIS	ductal carcinoma in situ
DP	digital pathology
EQA	external quality assessment
FDA	Food and Drug Administration
GI	gastrointestinal
GT	ground truth
H&E	haematoxylin and eosin
HRA	Health Research Authority
HSV	herpes simplex virus
HTA	Health Technology Assessment
IBD	inflammatory bowel disease
ICC	intraclass correlation
IMS	Philips Image Management System
IRAS	Integrated Research Application System
ISRCTN	International Standard Randomised Controlled Trial Number
LM	light microscopy
MDT	multidisciplinary team
NIHR	National Institute for Health and Care Research
PPI	patient and public involvement
RD	reference diagnosis
RE	random effects
REC	Research Ethics Committee
RF	research fellow
SCC	squamous cell carcinoma
SSC	Study Steering Committee
WCTU	Warwick Clinical Trials Unit
WSI	whole slide images

Plain language summary

Pathologists use a microscope to examine tissue samples, called light microscopy. This enables them to make the diagnosis, give information on treatment and provide prognosis to clinicians. The reports made by pathologists are interpretations of what the slides are showing, and this can be extremely difficult, so differences in interpretation between pathologists occur quite often. Asking colleagues' opinion on cases is one of the best ways of recognising and reducing these differences in interpretation. Digital pathology is a process of converting microscope slides to computer image files.

Digital pathology allows some advantages to pathologists, namely: to move cases easily between pathologists, for example to get a case seen by the next available pathologist who can report it, to view the cases at any location, for example work from home or report cases for a distant laboratory with a shortage of pathologists, to confer easily with multiple colleagues on difficult cases, and to rapidly check diagnoses made on previous samples the patient may have had. As a result, digital pathology could potentially lead to safer more efficient working.

In order to use digital pathology in practice, we need to know pathologists produce equivalent reports as compared to light microscopy.

This study compared light microscopy with digital pathology in examining 2024 samples from breast, gastrointestinal, including cancer screening samples, skin and kidney.

Most samples recruited (80%) were routine, but at least 20% of cases were challenging cases.

Pathologists worked in teams examining the same series of cases twice once through light microscopy and once through digital pathology with viewings separated by 6 weeks and the order randomised.

Differences in reports were arbitrated to establish if they would have changed treatment (significant) or not (insignificant). Pathologists reviewed all the significant differences to decide the ground truth. Statistical analysis measured the agreement between light microscopy and digital pathology in comparison to a reference point of 98.3% agreement derived from a previous study.

The results show an agreement overall of 99.95%. Specialty groups showing: breast 99.4%, gastrointestinal 99.96%, skin 99.99% and renal 99.99%. Cancer screening cases showed overall agreement was 98.96%, and in breast 96.27% and large bowel 99.93%.

In comparison to ground truth, the differences between pathologists were similar with light microscopy and digital pathology. Analysis showed the differences detected occurred in entities known to produce differences in interpretation between pathologists.

The study shows that pathologists give equivalent diagnoses when using digital pathology as they would using light microscopy. The differences detected are those you would expect to see in any event due to interpretable nature of examining these samples.

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.

Chapter 1 Introduction

Background

Histopathological diagnosis is an integral component of a patient care pathway as it provides clinicians with understanding of a patient's disease at the cellular level. This vital diagnostic information aids the process of therapeutic decision-making for patients. The mainstay of this diagnostic process remains light microscopic (LM) examination of tissue sections.

Increasing workload remains a global problem for pathology laboratories due to advances around early detection of cancer, improved life expectancy, expanding cancer screening programmes, more molecular tests and allied ancillary tests.[1–5] In this context, the most efficient use of limited cellular pathology workforce has become increasingly vital, to maintain standard of care and patient safety.[6]

Scanning histopathology slides with high-resolution digital cameras, and stitching these images together, enable pathology slides to be examined with computer workstations analogous to radiology workstations. The process of using digital whole slide images (WSI) as a means of examining pathology slides has been termed digital pathology (DP) and has experienced exponential growth over the past decade.[7] It offers several potential benefits to pathology departments. DP allows the pathologist to view the slides remotely from their site of production, thereby allowing work to be moved easily between pathologists, to assist flow, provide multidisciplinary, expert or out-of-hours' review, or review where patients move between sites for treatment.[8–10] DP thereby provides almost limitless flexibility of workload allocation, including distribution to remote sites – a factor exploited by many laboratories in response to the COVID-19 pandemic.[11] Digital slide archives can also provide instant retrieval of prior slides saving laboratory technician time and transport costs and enable analysis of pixel data contained in the images to be exploited to develop aids to improve diagnostic accuracy, efficiency and enhanced prognostic and predictive information.[12,13] DP can therefore benefit patients in several ways; increased efficiency leads to quicker results and freeing up budget to spend in other ways, for example through introduction of artificial intelligence (AI) algorithms to assist pathologists give more accurate and reproducible results. Indeed, the digitisation of slides permits these algorithms to be used at all. The ability to share cases with colleagues or refer cases to hub centres where patients will be treated and to review prior samples quickly and easily will all help reduce delays and increase overall accuracy, thus giving patients the benefit of more reliable accurate results. DP hitherto has been used extensively for teaching and external quality assessment (EQA),[14] but its use in routine reporting of slides has only been delivered recently in a small number of laboratories.[15–20]

Any novel technology replacing an existing standard requires definitive evidence of comparable accuracy, as well as evidence of gains in efficiency and cost-effectiveness. Multiple studies have assessed comparison of LM to DP; although many contained small numbers of cases (< 1000), there have been some large-scale studies aimed at providing evidence for clinical adoption.[2,15,21] A recent meta-analysis demonstrated high concordance rates between the digital and glass readings in these studies.[22] However, the majority (92%) of those studies were performed at a single institution, and without enrichment for challenging cases or samples from cancer screening programmes, leading to concerns over the use of DP in this setting. Additionally, to date, no studies have evaluated the accuracy of DP for samples from medical renal biopsies or immunofluorescence slides, a specialty comprising highly complex and low-volume samples where DP may prove to have important benefits in providing improved access to specialist expertise. The impact of DP on productivity and efficiency has not been studied, so the return on investment is unknown preventing business case development. Reluctance among pathologists to use DP is partly based on a lack of comprehensive multicentric evidence proving that DP is safe to use for primary diagnosis. Reluctance to change may also be due to the fear of adopting unfamiliar technology.

Pathological assessment of histopathology slides depends on the interpretation of histomorphological features of the sample in light of the clinical setting and is subject to both inter- and intraobserver variation. The studies comparing DP with LM published to date lack rigorous assessment of both inter- and intraobserver variation, making an assessment of equivalence between the two platforms difficult.

In this study,[23,24] we performed a multisite comparison of pathologists' ability to provide reports on cases using LM and DP with WSI. This included a mix of routine cases, cancer screening samples, biopsies and resections, as well as complex cases known to contain challenging lesions. The study included breast, gastrointestinal (GI), skin and renal specialties with consultant pathologists experienced in reporting these samples. The primary aim was the intraobserver agreement for pathologists' diagnoses using DP as opposed to LM using a cohort of cases where we could also measure the interobserver variability.

Existing research

In order to identify relevant literature pertaining to studies that compared DP with LM for diagnostic purposes, a comprehensive systematic review and meta-analysis were performed to analyse the existing evidence on safety and reliability of DP.[22] The search encompassed databases such as PubMed, MEDLINE, EMBASE, Cochrane Library and Google Scholar, with a focus on publications from 2013 to August 2019. The search protocol was designed to identify studies comparing DP with LM for diagnostic purposes (primary or secondary). Studies involving other uses of DP (education, image analysis, research) as well as those primarily involving immunohistochemistry and frozen section slides were not included. To ensure the inclusion of all potentially relevant articles that may not have been identified in the aforementioned search, a supplementary manual search was conducted. This manual search involved forward citation tracking and a thorough examination of the references cited in the included studies. A comprehensive data extraction protocol was developed based on the Cochrane Effective Practice and Organization of Care template.

This review adhered to the guidelines outlined by the Preferred Reporting Items for Systematic Review and Meta-Analyses. The review protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) and was assigned registration number CRD42019145977 in the PROSPERO database, which is maintained by the Centre for Reviews and Dissemination at the University of York, England. The evaluation of individual study quality and the assessment of potential bias were conducted using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS 2) tool, as recommended.

The initial systematic literature search yielded a total of 994 records. Following the removal of duplicate results, the abstracts of 828 records were assessed for eligibility. Through this eligibility screening process, 45 research studies were identified for full-text review. Ultimately, 25 of these articles met the criteria for inclusion in the review. The results showed a range of DP versus LM percentage agreement from 92.3% to 100%, with the majority (23/25) having concordances above 95%. The pooled percentage agreement for overall concordance was 98.3% [95% confidence interval (CI) 97.4 to 98.9] across 25 studies included in the meta-analysis. Our findings also brought attention to the potential diagnostic challenges encountered on the digital platform. A categorisation system was devised to group all diagnostic discrepancies as major or minor, considering their clinical significance. This approach successfully identified reported challenges associated with DP. The valuable insights derived from these data were employed in this study to guide the selection and enrichment of the study sample.

Research objectives

Primary objective

- The primary objective of this study was to compare pathologists' diagnoses made by LM with the same pathologists' diagnoses of the same sample (intraobserver concordance) using DP.

Secondary objectives

- Compare DP with LM in reporting of histopathology slides to measure variation between pathologists on both modalities (interobserver concordance).
- Measure the confidence in pathologists' diagnoses made with LM and DP.

Substudies objectives

- Explore the difference in speed of reporting of cases using LM in comparison to PD to assess any likely costs and benefits associated with DP compared with LM using a health economic evaluation.
- Explore the existing views of the pathology staff (pathologists and technicians) and the impact of introducing DP (migration from LM to DP) on pathologists and laboratory workforce with a view to understanding the barriers and facilitators to DP.
- Determine how the study pathologists examine DP images of different pathology modalities to establish how the techniques used to examine these images contribute to error in their interpretation.

Chapter 2 Methods

Study design

The study was designed incorporating principles outlined in documents published by the Royal College of Pathologists and the College of American Pathologists.[25,26] This multicentre validation comparison study used a blinded crossover design which compared pathologists' diagnoses of histopathology samples using LM and DP. Teams of four pathologists across each of the four specialty groups independently analysed the same cases and slides using both the LM and DP modalities, separated by a 'washout' period of a minimum of 6 weeks (*Figure 1*). The order of the platform used was randomised. For each read, pathologists were provided with identical clinical details and macroscopic descriptions, and were blinded to the original report, their previous read and the reports of other pathologists. All pen marks on the slides were removed and annotations made using the digital platform were not visible to any other pathologist. All reports on each of the study cases were compared for differences by a team of reviewers. All differences detected were sent for arbitration.

The Health Research Authority (HRA; National Health Service, UK) approved the study protocol and all subsequent amendments. The co-ordinating centre (UHCW NHS Trust Coventry) was responsible for leading all aspects of the study. The Study Steering Committee (SSC) included an independent chair, the chief investigator and patient representatives, which provided oversight of the study.

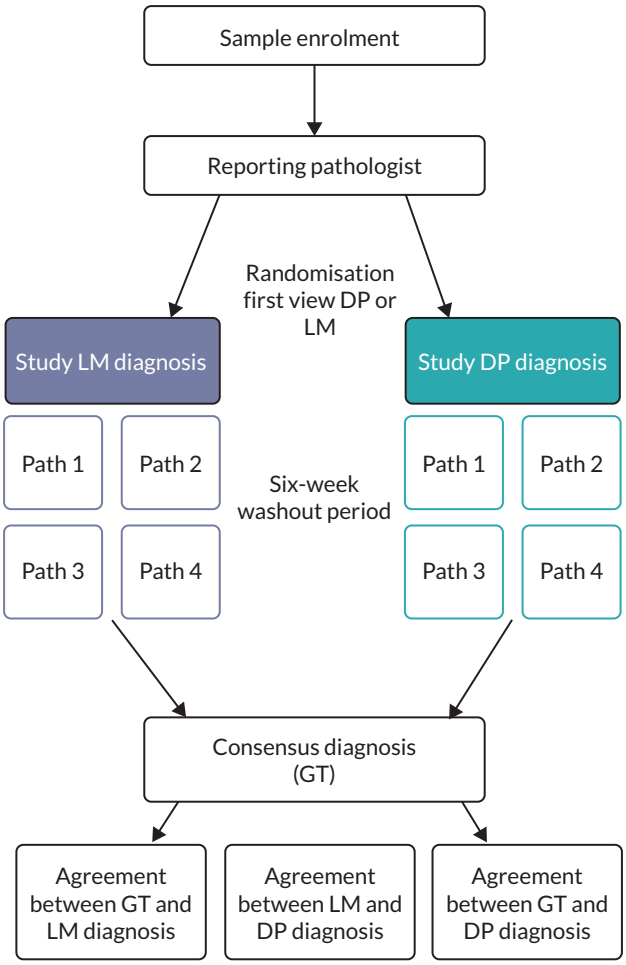

FIGURE 1 Study design. GT, ground truth; Path, pathologist.

Ethics and research and development approvals

The study protocol and amendments were reviewed and approved by HRA and Research Ethics Committee (REC). International Standard Randomised Controlled Trial Number (ISRCTN) number 14513591, Integrated Research Application System (IRAS) number 258799. The study raised minor ethical concerns, namely the use of patient data and slides without consent, and the potential for identifying a discrepancy in reporting which may be relevant to ongoing care of the patient. In regard to the latter point, all clinically significant differences between the ground truth (GT) and reference diagnosis (RD) were notified back to the originating trust for assessment as to whether they had any relevance to the care of the patient.

Protocol amendments

There were two non-substantial amendments on the study protocol; see details in *Table 1*.

Sponsorship

University Hospitals Coventry and Warwickshire NHS Trust is the sponsor of the study.

Patient and public involvement

A patient and public involvement (PPI) group was formed at the design concept and members advised on the trial design, areas of pathology to be included in the study, the plan of how the validation of digital microscopy should be conducted, the size of the study and the approach to enrichment with difficult samples, the practicalities around using prospective and retrospective samples, and the ethical considerations around the lack of patients' consent. The group stated the importance of feeding back to pathologists at sites any important clinical discrepancies found in the study which could result in a change of patient management. It was agreed that any discrepancies which would not result in a change of clinical management need not be reported back.

TABLE 1 Protocol amendments

Protocol version	Protocol date	Details of amendment
V1.0	29 May 2019	N/A
V2.0	8 October 2020	Non-substantial amendment 2 (NSA2) Updates to 'Sample pathway' section, including: • Addition of Coventry to the list of sites enrolling breast cases • Removal of quota of cases for each site to enrol • Clarification that we will be recruiting at least 200 cancer screening biopsies for each of the breast and GI specialties (Previous wording: a total of 600 samples including 200 cancer screening biopsies) • Clarification of processes, including provision of pseudonymised reference reports for renal cases Removal of Kappa statistic (KS) measure from 'Statistical analysis' section Removal of the paragraph regarding the breast radiology PERFORMS scheme from the 'Health economics evaluation' section Update to timelines for the qualitative study Update to 'Data shared with third parties' section to include sharing of digital images (and relevant descriptive variables) for reference and teaching purposes Update to 'Ethical approval' section to confirm that REC approval is in place (previously it was thought that the study did not require REC approval) Update to 'Study timetable and milestones' section
V3.0	21 September 2022	Non-substantial amendment 6 (NSA6) 'Contact details' have been amended to reflect changes within the study team Update to 'Introduction' section to provide further description of the report comparison process Update to 'Data shared with third parties' section to provide details of PathLake and how data will be shared with PathLake, including timelines and usage of study data Update to 'Study timetable and milestones' section, dates have been amended to reflect delays caused by COVID-19 pandemic. Additional milestones including SSC meeting and revised study end date Update to 'Publication' section to include plans to measure the impact of the study by sending out surveys to other centres in the UK

The group agreed to support the study throughout including monitoring and dissemination, with two members co-opted onto the SSC, one member assigned to make a patient video and all members commenting on progress reports and the final report. The group agreed to help with the dissemination of results to include infographics and a patient summary.

Cases

A total of 2000 histopathology samples was targeted reflecting routine laboratory workload across the four subspecialty areas: breast, GI tract, skin and renal – including cancer screening biopsies (GI and breast) and enriched with 20% cases considered either difficult or moderately difficult to report. Each subgroup agreed the parameters of cases to be included in the moderately difficult and difficult categories prior to recruitment (*Table 2*). Study pathologists were provided with a detailed sample selection protocol which was aided by the findings of a comprehensive literature review to identify cases known to present challenging diagnoses on DP and also identify potentially difficult diagnoses that had not been studied in the previously published literature.[22] All renal biopsy samples were deemed to be difficult due to the inherently complex nature of these biopsies, and all were procured from a single specialist centre. In the remaining groups, cases were enrolled from each of the five participating sites, according to an agreed target number.

For the breast, GI and skin specialties, consecutive cases reflected the bulk of the routine workload of a pathologist.

For breast cases, it was agreed that special stains (like Congo red, Giemsa, Gram, ZN, Warthin starry or fungal stains) would be included; however, immunocytochemistry for microorganisms [herpes simplex virus (HSV), cytomegalovirus (CMV)], very small tumour deposits in sentinel nodes were not included. The sample sites included breast, axillary node, or nipple biopsies. Cases with histological diagnosis of lymphoproliferative disorders and re-excision specimens were excluded. Estrogen receptor (ER) and progesterone receptor (PR) slides were included (if available) but human epidermal growth receptor 2 (Her2) slides were not included. Moderately difficult cases included excision samples for cancer resections including wide local excisions, mastectomies, cavity shaves or lymph nodes excision. Difficult cases included diagnostically challenging biopsies with an anticipated increased degree of interobserver variability and identified areas of difficulty in digital modality. Examples included diagnostic biopsies for lesions with calcium oxalate (Weddellite calcification), diagnostic category B3/B4, low-grade ductal carcinoma in situ (DCIS), atypical hyperplasia, sclerosing and papillary lesions, and lymph nodes with micrometastasis (without immunocytochemistry).

For GI cases, special stains/immunohistochemistry were not included. The biopsy sites included the oesophagus, stomach, small bowel, colon and anal canal. Appendices and gallbladders were also included. Polyp biopsies included screening as well as non-screening cases. The case mix represented normal histology, inflammatory, benign or malignant

TABLE 2 Distribution of samples as per specialty, sample size, enrolment site and level of difficulty

Specialty	Sample size	Enrolling sites	Case mix
Skin	600	Coventry Belfast Lincoln	480 – Routine samples including biopsies and BCC excisions 60 – Moderately difficult samples including complex cancer resections for SCC, malignant melanoma, Merkel cell carcinoma, lymph node excisions 60 – Difficult diagnoses, for example, challenging melanocytic and spindle cell lesions, microorganisms, amyloid, subtle dysplasia, early invasive malignancy etc.
Breast	600	Nottingham Belfast Lincoln Coventry	480 – Routine biopsies including at least 200 breast cancer screening cases 60 – Moderately difficult samples including complex cancer resections (wide local excision or mastectomy) 60 – Difficult diagnoses, for example, micrometastasis, B3 lesions, B4 lesions, atypical hyperplasia, low-grade DCIS, sclerosing and papillary lesions or calcium oxalate (Weddellite calcification) etc.
GI	600	Belfast Coventry Nottingham	480 – Routine upper and lower GI biopsies including at least 200 bowel cancer screening cases 60 – Moderately difficult samples including complex cancer or non-cancer resections 60 – Difficult diagnoses, for example, polyp cancers, microorganisms, inflammatory bowel disease, minimal change colitis, graft vs. host disease etc.
Renal	200	Oxford	Native and transplant renal biopsies including immunofluorescence

diagnoses. Multipart cases were included in the study. Cases with histological diagnosis of lymphoproliferative disorders were excluded. Moderately difficult cases included excision samples for malignant (oesophagectomy, gastrectomy, colectomy, anterior resection/abdomino-perineal resection (AR/APER) and neuroendocrine tumour excisions) and benign [excision for inflammatory bowel disease (IBD), diverticulosis, ischaemic bowel, etc.] diagnoses. Difficult cases included diagnostically challenging biopsy cases with an anticipated increased degree of interobserver variability and identified areas of difficulty on digital modality. For the purpose of adequate case mix, there were three categories: (1) dysplasia or malignant diagnoses including oesophageal dysplasia, polyp cancer, focal tumour invasion, goblet cell carcinoid, etc. (2) inflammatory or infectious conditions including minimal change colitis, granulomatous inflammation, challenging cases of inflammatory bowel disease, microorganisms (giardia, HSV, CMV, fungi, *Helicobacter pylori*, etc.); and (3) others including graft versus host disease, amyloid, etc.

For skin cases, special stains (like Congo red, Giemsa, Gram, ZN, Warthin starry or fungal stains) were included; however, immunocytochemistry for microorganisms (HSV, CMV), very small tumour deposits in sentinel nodes were excluded. Examples of routine cases included seborrheic keratosis, cysts, and biopsy samples for common benign and malignant diseases, and excision specimens for basal cell carcinomas (BCCs). Cases with clinical impression or histological diagnosis of lymphoproliferative disorders and re-excision specimens were not included. Moderately difficult cases included excision samples for non-BCC malignancies, for example, squamous cell carcinoma (SCC), melanoma, adnexal carcinoma, Merkel cell carcinoma as well as lymphadenectomy specimen associated with melanoma or SCC. Difficult cases included biopsy samples for diagnostically challenging cases, with an anticipated increased degree of interobserver variability, such as inflammatory dermatosis and identified areas of difficulty on digital modality such as viral and bacterial infections, and difficult melanocytic lesions.

All renal cases were deemed difficult. The case mix was a mixture of native (requiring special stains and immunofluorescence) and transplant renal biopsies (requiring special stains and immunocytochemistry).

All slides were included for all biopsies and most resection samples. For some larger breast and GI resection samples (> 10 blocks), pathologists had the option to select representative slides deemed sufficient to provide all the essential diagnostic details in the synoptic report. All the available stains including haematoxylin and eosin (H&E), specials, immunohistochemistry and immunofluorescence were included in the study except GI where only H&E stains were included.

Only cases that had been reported by the participating laboratory and discussed at the relevant multidisciplinary teams (MDTs; if applicable) were suitable for the study. Cases with broken, missing or oversized slides and slides with overhanging coverslips or excess mounting material on the surface were excluded. Cases received from non-participating centres for second opinion and cases where knowledge of previous sample(s) was essential for the diagnosis were also excluded.

Sites and pathologists

Sixteen consultant (specialist) pathologists from six NHS cellular pathology laboratories (Coventry, Lincoln, Nottingham, Belfast, Birmingham and Oxford) participated in reporting study samples across four specialty teams (breast, GI, skin and renal). Consultant experience ranged from 3 to 35 years. All pathologists were subspecialised and reported cases in the study belonging to areas of expertise in their routine practice. Four pathologists were early adopters of DP and used this modality for most of their routine reporting. The remainder were new to DP.

Sample enrolment/recruitment procedure

Skin, breast and gastrointestinal samples

All participating sites identified and enrolled previously reported consecutive samples from their respective laboratory archive systems. The glass slides were retrieved along with the corresponding reports and anonymised by removing personal identifiers. Slides were then dispatched to the co-ordinating centre and sample details (including patient age, patient gender, difficulty level, number of slides, clinical details, macroscopic description, RD and reference report) were entered onto the study database. Upon receipt of slides at the co-ordinating centre, the cases were checked

and recruited into the study by a team of research technicians. Each case was allocated a study number and the slides relabelled with this number and bar coded.

Enrolment/recruitment of renal biopsies

All renal biopsies were enrolled at Oxford University Hospitals NHS Foundation Trust. Consecutive renal biopsies in native and transplant groups from patients providing generic consent for research were enrolled. Sample details were entered onto the study database as part of the enrolment process. Following enrolment, each case was allocated a unique study number. The slides were relabelled with the study number and barcoded.

The research team closely monitored the case enrolment process, ensuring that any samples found to be ineligible during the study were promptly replaced.

Equipment/training/reporting

Whole slide image scanners and digital pathology workstations

The research slides were scanned using WSI scanners from two different vendors. The skin, GI and breast slides were scanned at the co-ordinating centre using Food and Drug Administration (FDA)-approved Philips IntelliSite Pathology Solution (Philips, the Netherlands) that included a Philips Ultra-Fast Scanner (UFS 1.8, IVD-CE), an Image Management System (IMS 3.3.1) and a display monitor. Once digitised at 40× magnification (average scan time 60 seconds), the WSIs were stored locally at the co-ordinating centre (network connection: 1 GB/second bandwidth) in two HP DL380 iron servers with a net 24 TB storage capacity. All participating sites were provided viewing access to images on the firewall-enabled server via a secure network (Secure Sockets Layer) connection.

In addition, a dual function (bright-field and fluorescent) 3DHISTECH PANNORAMIC SCAN II was used to digitise renal slides (at 40× magnification) at the slide source (Oxford) due to specific storage requirements for fluorescent slides. These WSIs were stored on a vendor-provided secure cloud-based server accessible to all renal pathologists via image viewing system – 3DHISTECH CaseCentre v2.9.

All reporting pathologists were provided with standardised Conformité Européene (CE) marked HP workstations (Z4) comprising a Dual-core @3GHz CPU (Microsoft Windows Server v2012 R2 SP1, RAM 3 GB with upgraded graphics cards) and Philips 27″ display monitors (resolution 1920 × 1200; brightness > 300 cd/m^2; contrast 1000 : 1).

Slide scanning protocol

A team of research technicians at both scanning sites were provided comprehensive training to operate the system. Any pen marks on the slides were removed before scanning. To ensure quality control of the digital images, appropriate quality check measures were performed on each image by the technicians. Slides were rescanned if the image showed any artefacts or on request by the first DP reporting pathologist. Any flagged images during the course of the study were reviewed by the research fellow (RF) to confirm that the scanned image quality was adequate for the diagnosis and a record of scanning issues and re-scan request log was kept by the technicians.

Pathologists' training

All participating pathologists completed training on the use of the DP IMS. Pathologists who were not using DP as part of their regular practice followed an initial training programme in DP supervised by a Pathology Research Fellow and following the Royal College of Pathologists best practice recommendations.[25] Briefly, training involved creating a training slide set comprising at least 30 previously reported samples for each pathologist reflecting the routine workload for the specialty. The case mix included a range of sample types including biopsies, resections, simple and complex cases and a variety of immunohistochemistry and special stains. Each pathologist was provided with the glass slides and digital slides along with the clinical details, and findings were recorded on a spreadsheet. Pathologists underwent review sessions with the Pathology Research Fellow to review experience and confidence. If needed, additional training slides were provided until the pathologists felt confident, and they could examine slides effectively using the DP platform. The training was ratified according to the college guidance and was approved for continuing professional development (CPD) credits.

Reporting of samples

Blinded to the original RD, each pathologist reported the same study sample twice: once using DP and once using LM. For each reading, pathologists consulted the same clinical details and macroscopic description on the database. LM was conducted using the microscopes used for routine diagnostic work. DP was done using the workstations provided and via access to the study servers. Reporting of some case types was done using proformas agreed by the group, including cancer screening samples based on UK NHS Breast and Bowel Cancer Screening programme requirements. Reporting of cancer resections followed proformas based on the Royal College of Pathologists minimum data sets. Free-text reporting was used where proformas were not available.

The annotations and measurement tools available on the DP systems were permitted but hidden from fellow pathologists. Additional comments regarding the slide quality or the need or preference for additional work were recorded. Pathologists reported cases in isolation using their normal range of reference material including textbooks and internet resources. They did not confer on cases.

Pathologists manually recorded the time taken to interpret the slides and enter the report and diagnosis on the database (reporting time) for each reading using a digital timer. In addition, pathologists were also asked to record their diagnostic confidence for each report on a seven-point Likert scale, from least confident to most confident.[25]

Review, arbitration and consensus process

Following completion of all readings, the reports were compared. Any variations between reports were forwarded to the arbitrator for adjudication.

Three arbitrating pathologists (who were not involved in reporting of the cases) decided whether the differences identified would have resulted in differences in patient management (clinically significant) or not (clinically insignificant). In uncertain cases, this decision was referred to a consulting clinician who had agreed to assist with the arbitration. The comparison and arbitration teams were blinded to the reporting modality, pathologists and participating sites, and did not have access to the slides.

All cases were analysed as a whole rather than by parts. For example, a case with a clinically significant discordance in a single part was labelled as discordant.

Arbitration table: reference guide for arbitration

A reference guide (arbitration table – see *Tables 3–6*) was formulated for each specialty to allow consistent decision-making during the arbitration process. It was utilised by the report comparison and arbitration teams to classify the variations in DP and LM reports (diagnoses) in relation to the GT diagnosis.

TABLE 3 Recurring differences in breast and the decision regarding which were clinically insignificant and those which were clinically significant

Category	Ground truth vs. study diagnosis (DP or LM)	Complete agreement (0) clinically unimportant difference (1)/clinically important difference (2)
Biopsy diagnosis	B1 vs. B2	1 or 2 (context dependent)
	B2 vs. B3 with/without atypia	2
	B3 with atypia vs. B3 without atypia	2
	B4 vs. B5	2
	B5 vs. any non-malignant diagnosis	2

continued

TABLE 3 Recurring differences in breast and the decision regarding which were clinically insignificant and those which were clinically significant (*continued*)

Category	Ground truth vs. study diagnosis (DP or LM)	Complete agreement (0) clinically unimportant difference (1)/clinically important difference (2)
DCIS	Intermediate grade with necrosis vs. intermediate grade without necrosis	1
	Intermediate grade without necrosis vs. low grade with necrosis	1
	Low/intermediate grade without necrosis vs. high grade with necrosis	2
	DCIS vs. DCIS with microinvasion	2
DCIS (excision)	High grade vs. low grade in the presence of invasive component	1
	Presence vs. absence of DCIS in the presence of invasive component	1
Phyllodes tumour	Benign/borderline phyllodes vs. fibroadenoma	1
Phyllodes tumour (excision)	Benign vs. borderline vs. malignant	1
Tumour subtype	Pure lobular vs. other types (other than basal like)	2
	Basal like vs. other types	2
	no special type (NST) vs. papillary type	1
	Pure ductal vs. mixed NST	1
	Pure NST vs. mixed NST	1
Tumour grade	Grade 1 vs. grade 2	1
	Grade 2 vs. grade 3	1 or 2 (context dependent)
	Grade 1 vs. grade 3	2
Vascular invasion	Presence vs. absence of vascular invasion	1 or 2 (context dependent)
Tumour stage	Stage 1 vs. stage 2 stage 3	2
Excision status	Margin positive vs. negative	2
Microcalcifications	Calcification vs. no calcification (if biopsy done for calcification	1 or 2 (context dependent)
Receptor status	ER positive vs. ER negative	2
	PR positive vs. PR negative	2
Lymph nodes	Micrometastasis vs. macrometastasis	2
	Metastatic carcinoma vs. metastatic melanoma	2

TABLE 4 Recurring differences in GI and the decision regarding which were clinically insignificant and those which were clinically significant

Category	Ground truth vs. study diagnosis (DP or LM)	Complete agreement (0)/clinically unimportant difference (1)/clinically important difference (2)
Colon biopsy	Hyperplastic polyp vs. sessile serrated lesion	2
	Low-grade dysplasia vs. high-grade dysplasia in polyps (even if focal)	2
	Tubular adenoma vs. tubulovillous adenoma	1
	Focal active colitis vs. no active colitis	1 or 2 (context dependent)
	Presence vs. absence of invasive malignancy	2

TABLE 4 Recurring differences in GI and the decision regarding which were clinically insignificant and those which were clinically significant (*continued*)

Category	Ground truth vs. study diagnosis (DP or LM)	Complete agreement (0)/clinically unimportant difference (1)/clinically important difference (2)
Oesophageal biopsy	Presence vs. absence of Barrett's metaplasia	2
	Indefinite for dysplasia vs. definite for dysplasia	1 or 2 (context dependent)
	Low-grade vs. high-grade dysplasia	2
Gastric biopsy	Chronic gastritis vs. reactive gastritis	0
	Presence vs. absence of *H. pylori*	2
	Low-grade dysplasia vs. high-grade dysplasia	1 or 2 (context dependent)
	Flat dysplasia vs. adenoma/polyp	1 or 2 (context dependent)
	Presence vs. absence of intestinal metaplasia	2
Duodenal biopsy	Presence vs. absence of significant intraepithelial lymphocytosis	2
	Presence vs. absence of giardia trophozoites	2
Any biopsy	Presence vs. absence of amyloidosis	2
	Presence vs. absence of microorganisms	2
	Presence vs. absence of granulomas	2
All cancer resections	Margin positive vs. negative	2
	Presence vs. absence of lymph node metastasis	2
	Presence vs. absence of extramural vascular invasion	1 or 2 (context dependent)
	Stage 1 vs. stage 2 vs. stage 3 vs. stage 4 cancer	2

TABLE 5 Recurring differences in skin and the decision regarding which were clinically insignificant and those which were clinically significant

Category	Ground truth vs. study diagnosis (DP or LM)	Complete agreement (0)/clinically unimportant difference (1)/clinically important difference (2)
Biopsy	High-risk vs. low-risk BCC	2
	BCC vs. basosquamous carcinoma	1
	Actinic keratosis vs. SCC in situ vs. Bowen's disease	1
	Bowenoid actinic keratosis vs. Bowen's disease	0
	Bowen's disease vs. SCC	2
	Dysplastic naevus vs. bordering on melanoma in situ	1
	Melanoma in situ vs. invasive melanoma	2
	Benign naevus vs. melanoma	2
	Dermatofibroma vs. dermatofibrosarcoma protuberans	2
Resection for BCC	High-risk vs. low-risk BCC (completely excised)	1
	High-risk vs. low-risk BCC (close of involved margin)	2
	Variations in margin clearance: > 1 mm vs. margin < 1 mm	1
	Variations in margin clearance: > 0.5 mm vs. margin < 0.5 mm	2

continued

TABLE 5 Recurring differences in skin and the decision regarding which were clinically insignificant and those which were clinically significant (*continued*)

Category	Ground truth vs. study diagnosis (DP or LM)	Complete agreement (0)/clinically unimportant difference (1)/clinically important difference (2)
Resection for SCC	Superficially invasive SCC vs. SCC in situ (completely excised)	1
	Keratoacanthoma like SCC vs. keratoacanthoma (completely excised)	1
	Variations in subtype	1
	Well vs. moderately differentiated	1
	Variations in margin clearance; all readings > 1 mm	1
	Variation in margin clearance; ≥ 1 mm vs. < 1 mm	2
	Presence vs. absence of amyloidosis	2
	Presence vs. absence of microorganisms	2
	Presence vs. absence of granulomas	2
Resections for BCC or SCC	Variations in staging	1 or 2 (context dependent)
	Presence vs. absence of perineural invasion	1 or 2 (context dependent)
	Presence vs. absence of vascular invasion	1 or 2 (context dependent)
	Maximum tumour dimension: readings within criteria for the same stage	0
	Variations in tumour thickness: if within the same range, that is, > 6 mm, < 6 mm	0
	Variations in tumour thickness: > 6 mm vs. < 6 mm	2
	Variations in margin clearance; if within the same range, that is, 1–5 mm, > 5 mm	0
Resections for malignant melanoma	Variations in Breslow thickness without altering pathology tumour stage (pT) stage	0
	Ulceration vs. no ulceration	2
	Variations in margin clearance: if within same range, that is, 1–5 mm, > 5 mm	0
	Variations in margin clearance; all readings > 1 mm	1
	Variation in margin clearance; ≥ 1 mm vs. < 1 mm	2

TABLE 6 Recurring differences in renal and the decision regarding which were clinically insignificant and those which were clinically significant

Category	Ground truth vs. study diagnosis (DP or LM)	Complete agreement (0)/clinically unimportant difference (1)/clinically important difference (2)
No evidence of rejection	Additional comments on interstitial fibrosis and tubular atrophy, acute tubular injury, ATI, arteriosclerosis and other non-rejection pathologies	1
	Complement factor 4d (C4d) +ve (without rejection) vs. C4d −ve	1
	Stating Banff subclassification vs. not stating subclassification	0
	Additional comments on presence or absence of Focal segmental glomerulosclerosis	0 or 1
	Calcineurin inhibitor toxicity vs. no calcineurin inhibitor toxicity	1
	antibody mediated rejection vs. donor vascular disease and tubular changes in keeping with calcineurin inhibitor toxicity	2

TABLE 6 Recurring differences in renal and the decision regarding which were clinically insignificant and those which were clinically significant (*continued*)

Category	Ground truth vs. study diagnosis (DP or LM)	Complete agreement (0)/clinically unimportant difference (1)/clinically important difference (2)
Immunoglobulin A (IgA) nephropathy	Oxford score documented vs. not documented	2
	IgA nephropathy vs. acute tubular injury	2
FSGS	Subtype documented vs. not documented	1
Amyloidosis	Stating subtype of amyloid	0
Vasculitic GN	Documentation of healed crescent vs. no documentation	1
	Documentation of necrotising crescent vs. no documentation	2

FSGS, Focal segmental glomerulosclerosis

Ground truth

The GT diagnosis for each case was compiled using the eight study reports and the initial reference report. Where the eight study reports agreed, this was accepted as the GT (*Figure 2*, row 4). In cases where one or more study reports showed a clinically significant difference from the other reports, the WSI and all the reports (study and reference reports) were reviewed by the study pathologists reporting that case group and a consensus view of the GT was agreed (see *Figure 2*, row 6).

The GT for each case was also compared with the RD generating three potential outcomes: complete agreement, clinically insignificant difference, or clinically significant difference. Clinically significant differences between GT and RD were notified back to the recruiting site.

Feeding back discordant results to the clinical teams

Where the consensus diagnosis showed a clinically significant difference to the RD, the recruiting centre was notified of the consensus diagnosis.

Outcomes

Each case was reported by four pathologists on two occasions, using LM and using DP, resulting in eight reports per case. Each of the eight reports were compared to the GT diagnosis, and each of the four pathologist's LM and DP reports were also compared, giving 12 comparisons per case, as shown in *Table 7*. A comparison was recorded as complete agreement, clinically unimportant difference (difference which would not affect patient management) or clinically important difference (difference which would affect patient management).

Primary outcome

The primary outcome, clinical management concordance (CMC), categorised a comparison as CMC (identical diagnoses plus differences not affecting patient management, i.e. complete agreement or clinically unimportant difference based on *Table 7*) or not. The primary objective of the study was to estimate intraobserver intermodality CMC. The estimate was obtained by computing percentage CMC using all pathologists' LM–DP reports comparisons, that is, comparisons in column 3 ('LM = DP') in *Table 7*. Secondary objectives were the inter-pathologist comparison within LM and DP diagnoses obtained by computing intraclass correlation (ICC) based on CMC for LM = GT comparisons in column 4 and ICC based on CMC for DP = GT comparisons in column 5, respectively.

Secondary outcomes

Secondary outcomes were the inter-pathologist comparison of LM and DP diagnoses against GT and complete concordance (CC), categorising a comparison as CC (identical diagnoses, i.e. complete agreement in *Table 7*) or not (i.e. clinically important difference or clinically unimportant difference in *Table 7*). Another secondary outcome was diagnosis confidence, a pathologist's rating of the confidence of their diagnosis on a seven-point Likert scale. In the statistical

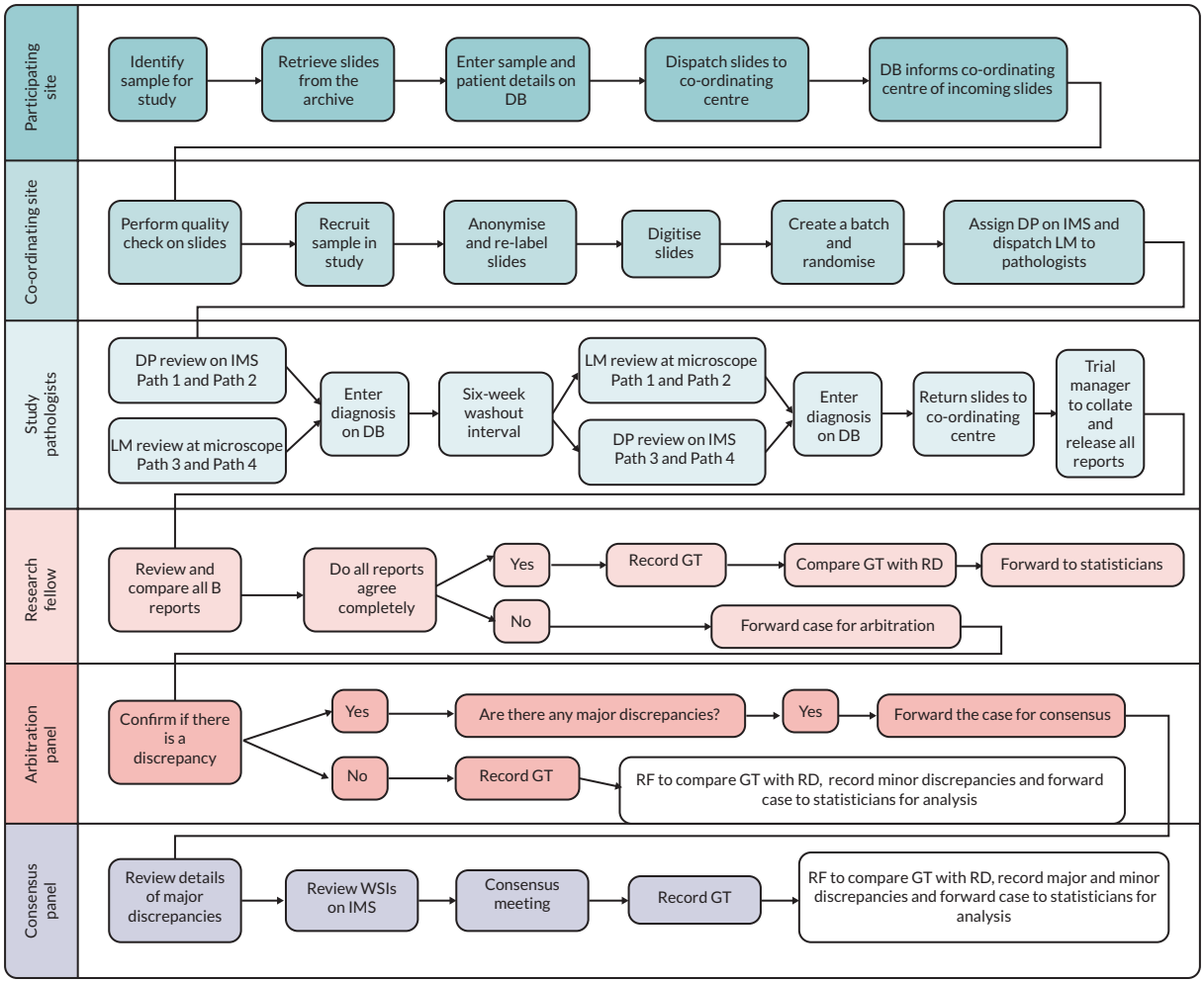

FIGURE 2 Overall study workflow, reports review, arbitration and consensus process. DB, database.

TABLE 7 Reports' comparisons template for one case

Case ID	Pathologist ID	LM = DP	LM = GT	DP = GT
1	1	0 = Complete agreement	0 = Complete agreement	0 = Complete agreement
1	2	1 = Clinically unimportant difference	1 = Clinically unimportant difference	0 = Complete agreement
1	3	1 = Clinically unimportant difference	0 = Complete agreement	1 = Clinically unimportant difference
1	4	2 = Clinically important difference	2 = Clinically important difference	1 = Clinically unimportant difference

analysis, scores for the highest and lowest confidence are 0 and 6, respectively. Pathologists gave confidence ratings for both their LM and DP diagnoses.

Sample size

The target sample size was 2000 cases consisting of 200 renal cases and 600 cases from each of the other three specialties breast, skin and GI. It was anticipated that within a specialty, approximately 10% would be difficult to diagnose cases, 10–20% moderate cases and the rest would be routine cases. In the actual study, if necessary, the plan was to undertake enrichment of cases to have approximately at least 10% in each difficulty level category.

Adequacy of the sample size was based on the precision for the percentage LM versus DP (LM–DP) CMC. It was assumed that percentage CMC for routine cases would be 98.8%.[15] In the literature, percentage CMC for difficult cases ranged from 40% to 70%, and so it was assumed that in this study it would be 55%. We assumed that the percentage CMC for moderate cases would be 75% (a percentage CMC that is between percentage CMC for difficult and routine cases). Therefore, the weighted percentage CMC within this study, with the anticipated 10% difficult, 20% moderate and 70% routine proportions, would be 90%.

The four LM–DP comparisons for each case, arising from each case reported by four pathologists, are likely to be correlated. Accounting for the expected percentage CMC, we assumed that ICC for routine, moderate and difficult cases to be 0.9, 0.7 and 0.4, respectively, so that considering anticipated case difficulty mix, we took overall ICC to be 0.8. Consequently, the design effect was assumed to be [1 + ICC (comparisons per case − 1)] = 3.4.

High precision of percentage LM–DP CMC was desired in the subgroup analysis by specialty, and consequently, in the primary analysis including 2000 cases. For each of the breast, skin and GI specialties, because each case was reported by four pathologists on both LM and DP, there were 2400 (600 × 4) LM–DP comparisons to estimate percentage LM–DP CMC. Dividing 2400 LM = DP comparisons by the design effect, we considered the 2400 comparisons to be equivalent to 705 (2400/3.4) independent (a setting where a case is reported by one pathologist only) comparisons. With 705 independent comparisons, the margin of error [1.96 × standard error (SE), where SE formula is given below] is 2.2% so that precision is high while analysing breast, skin and GI cases separately. For renal cases, the sample size is smaller (800 comparisons that are equivalent to 800/3.4 = 235 independent comparisons), giving the margin of error as 3.1% which is bigger but that still gives high precision.

Where n is the equivalent number of independent comparisons and p = 0.9 (90%) is the percentage CMC that was expected in the study, the SE was computed as follows:

$$SE = 100 \times \sqrt{\left(\frac{p(1-p)}{n}\right)}.$$

Sample batching

When enough samples per specialty were received by the co-ordinating centre, it batched the samples using the Warwick Clinical Trials Unit (WCTU) database system. The system then allocated a unique batch number and randomised the batch by selecting which pathologists would view using LM first, and which would view using DP first.

Statistical methods

Database and data processing

Study database

A bespoke electronic research database was developed for this study. All pathologists and research staff accessed the database and its role-defined functionality by means of username and password. Functionality included sample enrolment, recruitment, batching, randomisation, sample tracking, pathologists' reporting, comparison of reports and arbitration.

Statistical analysis

Statistical analysis was performed as per the pre-specified study's statistical analysis plan, which had been pre-approved by the SSC prior to final data lock. Categorical characteristics of patients and cases such as difficulty levels of cases and sex were summarised using counts and percentages in different category levels. For age of patients, minimum, maximum, median, quartiles, mean and standard deviation (SD) were reported.

Clinical management concordance data corresponding to LM versus DP comparisons were summarised by reporting the count and percentage of cases where:

- All four LM versus DP comparisons, corresponding to four pathologists reporting each case, were concordant.
- Three of the four comparisons were concordant.
- Two of the four comparisons were concordant.
- Only one of the four comparisons was concordant.
- All four comparisons were not concordant.

To compute percentage CMC, which was used as a measure of the intraobserver intermodality CMC (primary objective), and corresponding 95% CI, a random-effects (RE) logistic regression model was fitted. A logistic regression model was fitted because the outcome, whether or not there was CMC between LM diagnosis and DP diagnosis, was binary and the model enables estimating percentage CMC. There were two crossed RE terms corresponding to case and pathologist. The RE term for case was included because each case was reported by four pathologists and the corresponding four LM–DP comparisons can be correlated while the RE term for pathologist was included because each pathologist reported multiple cases and the LM–DP comparisons by the same pathologist can be correlated. The two RE terms were crossed because within a specialty, each pathologist reported all cases, and each case was reported by all four pathologists and so there was no nesting of cases within a pathologist or pathologists within a case. The RE model was fitted using the 'gamm4' package[27] in R (The R Foundation for Statistical Computing, Vienna, Austria) statistical program.[28] The intercept from the RE model was transformed to obtain percentage LM–DP CMC and 95% CI.

Clinical management concordance data corresponding to LM versus GT comparisons were analysed as described above for CMC data corresponding to LM versus DP comparisons. The percentage LM–GT CMC obtained from the RE model quantified the probability of LM diagnosis agreeing with GT. Additionally, the ICC from this model has LM inter-rater agreement interpretation. Let σ^2_{path} and σ^2_{case} be the RE estimates for pathologist and case, respectively. The ICC defined as:

$$ICC = \frac{\sigma^2_{case}}{\sigma^2_{case}+\sigma^2_{path}+\pi^2/3'}.$$

is the correlation of LM diagnoses for a case regardless of the pathologist (i.e. correlation of LM diagnoses for a case between any two random pathologists). From the expression of ICC, high values of the ICC imply that most of the observed variation in diagnoses is explained by variation between cases and not variation between pathologists or random variation. Therefore, the ICC is a measure of inter-rater agreement of LM diagnoses. The 95% CI for ICC was the 2.5th percentile and 97.5th percentile of 500 bootstrap ICC estimates.

Clinical management concordance data corresponding to DP versus GT comparisons were analysed as those for CMC data corresponding to LM versus GT comparisons. The percentage DP–GT CMC obtained from this analysis was compared to the percentage LM–GT CMC described above to assess whether one of the two modalities (LM and DP) was substantially better in giving diagnosis that agrees with GT. Additionally, ICC from this analysis was compared to the ICC obtained from LM versus GT analysis to assess whether interobserver agreements for the two modalities are similar.

The approach used to analyse CMC data was the same that was used to analyse CC data.

Within a modality, diagnosis confidence scores (0 = highest confidence and 6 = lowest confidence) were summarised by reporting the number and percentage of diagnoses that were given to each of the possible seven scores. To make confirmatory inference, diagnosis confidence scores were analysed using a model for count data (Poisson model) because, with the narrow range of the scores from 0 to 6, data could not be assumed to be normally distributed. Consequently, to compare LM and DP, rate ratio of the mean Poisson rate for LM diagnosis confidence scores and the mean Poisson rate for DP diagnosis confidence scores was used. A RE generalised Poisson model with crossed RE terms for case and pathologist was fitted using the 'glmmTMB' package[29] in R statistical program. A generalised Poisson model was used because exploration of the data indicated they were under-dispersed compared to what would be expected

for a Poisson model. Specifically, Conway–Maxwell–Poisson distributions using a parameterisation that enables mean rates to be compared were used.[30]

The above set of analyses was performed using all the eligible cases recruited in the study and repeated in pre-specified subgroup analyses. Subgroup analyses consisted of:

- Analysing cases from the different specialties separately.
- Analysing cases by difficulty levels.
- Analysing the screening cases only.

Chapter 3 Results

Screening and recruitment

Figure 3 shows the pathway for all skin, breast and GI samples; *Figure 4* shows the pathway for all renal samples, which were scanned by the recruiting site, Oxford.

Between July 2019 and July 2021, a total of 2065 cases were enrolled (*Table 8* and *Figure 5*).

Sample characteristics

A total of 2024 cases were included in the analysis (see *Figure 5*). They comprised 608 breast, 607 GI, 609 skin and 200 renal cases that came from 1271 (62.8%) females and 753 (37.2%) males from birth up to 96 years old (*Table 9*). Percentages of screening cases for breast and GI cases were 34.0 and 41.2, respectively.

Results

Primary outcome results

Clinical management concordance

For 1784/2024 (88.1%) cases, all four study pathologists had LM versus DP CMC, and there was only one case where there was no CMC for any of the four pathologists (*Table 10*). By specialty, LM versus DP CMC was highest for renal cases while it was lowest for breast cases. CMC for LM versus GT comparisons was similar to CMC for DP versus GT comparisons (see *Table 10*).

Using a RE logistic regression model on the 8096 LM versus DP comparisons over all 2024 cases showed percentage CMC, corresponding to the study's primary objective of estimating intraobserver LM versus DP agreement, was 99.95 (95% CI 99.90 to 99.97) (*Table 11*, row 1 in column 2). LM versus GT percentage CMC and DP versus GT percentage CMC were both 99.95% (see *Table 11*, columns 3 and 4) illustrating an identical level of agreement with GT from both modalities. The interobserver agreements for LM and DP were also similar with both ICCs being 0.91 (columns 4 and 5).

In subgroup analyses, compared to the analysis of all cases, there were noticeably lower LM versus DP percentage CMCs for moderately difficult cases (95.34, 95% CI 93.09 to 96.89), difficult cases excluding renal cases (96.78, 95% CI 94.27 to 98.22) and breast screening cases (96.27, 95% CI 94.63 to 97.43) (see *Table 11*). However, these are all above 95%. In all subgroup analyses, LM versus GT percentage CMC and DP versus GT percentage CMC were similar illustrating similar levels of agreement with GT from both modalities (columns 3 and 4). Also, in all subgroup analyses, LM interobserver agreement and DP interobserver agreement had overlapping 95% CIs of the ICCs (columns 4 and 5) showing comparable interobserver agreements for the two modalities.

Secondary outcomes results

Complete concordance

For 1500/2024 (74.1%) cases, all four study pathologists had LM versus DP CC and there were only five cases where there was no CC for any of the four pathologists (*Table 12*). As expected, CC rates are lower than CMC rates (see *Table 10* vs. *Table 12*). Like CMC, LM versus DP CC was highest for renal cases and lowest for breast cases while LM versus GT CCs were similar to corresponding DP versus GT CCs.

Based on all cases, using a RE logistic regression, percentage LM versus DP CC was 95.74 (95% CI 94.39 to 96.77) (*Table 13*). Similar agreement rates were obtained when looking at percentage LM versus GT CC and percentage DP versus GT CC (95.13% and 94.95%, respectively) illustrating similar levels of agreement with GT from both modalities. The ICCs for LM and DP were also similar, with overlapping 95% CIs.

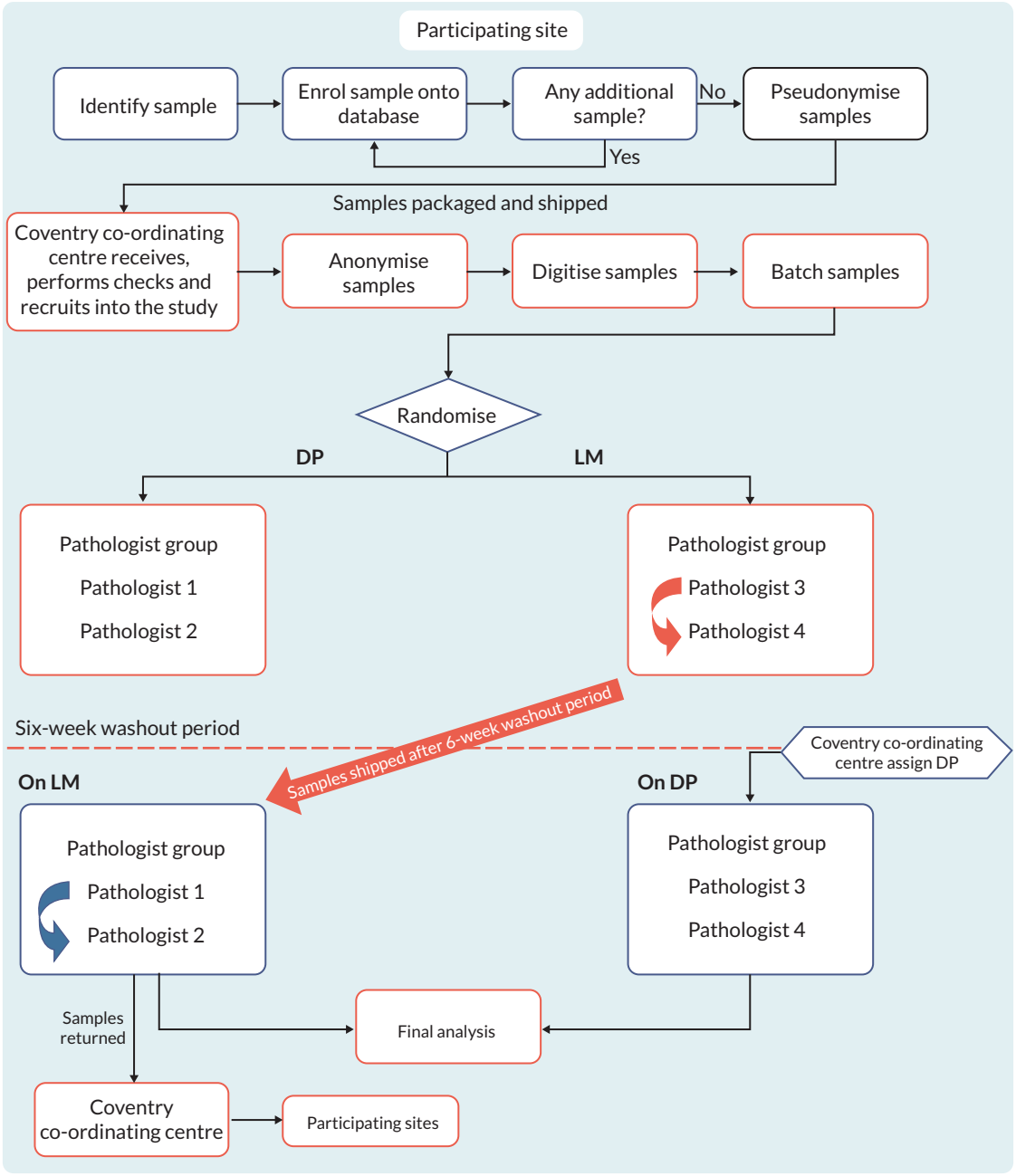

FIGURE 3 Pathway for skin, GI and breast samples.

In subgroup analyses, compared to the analysis of all cases, there were noticeably lower LM versus DP percentage CC for moderately difficult cases, difficult cases that do not include renal cases, all breast cases and breast screening cases (see *Table 13*). In all subgroup analyses, LM versus GT percentage CC and DP versus GT percentage CC were similar, and 95% CIs overlapped. Similarly, in all subgroup analyses of interobserver agreements, LM ICCs and DP ICCs were comparable, with overlapping 95% CIs.

Diagnosis confidence

Diagnosis confidence rating was not recorded in only 2/8096 LM reads and 3/8096 DP reads. Overall, pathologists had the highest confidence in 88.3% of their LM reads compared to 87.5% for DP reads (*Table 14*). There was noticeable variation across the specialty. Confidence was substantially higher for GI and skin diagnoses than for breast and renal diagnoses, with confidence for renal diagnoses markedly lowest. Overall, compared to confidence in the LM diagnoses, confidence in the DP diagnoses were very slightly lower. By specialty, the slightly lower confidence in DP diagnoses

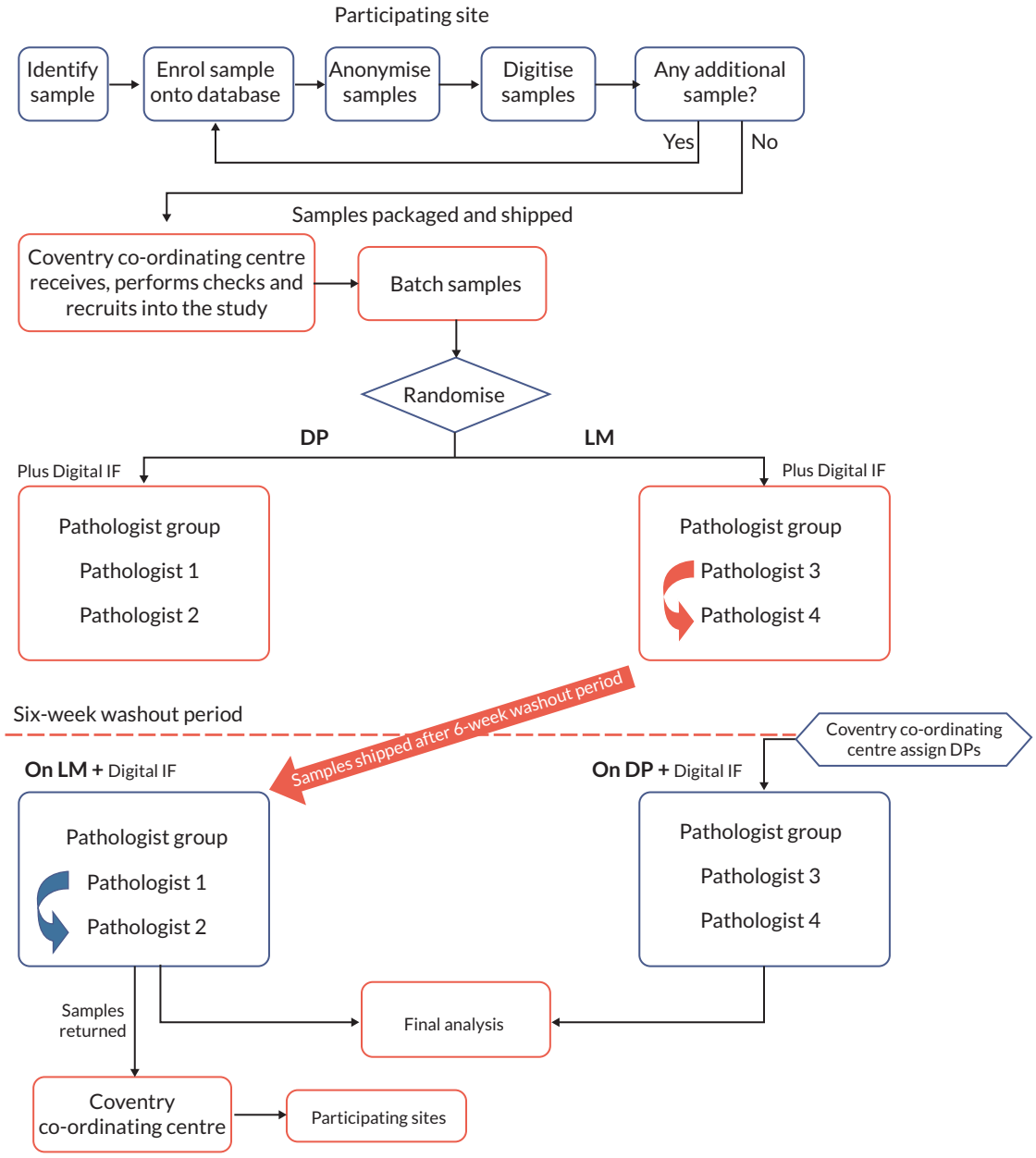

FIGURE 4 Pathway for renal samples.

compared to LM diagnoses is apparent for breast, GI and renal diagnoses but the difference is not apparent for skin diagnoses.

Using RE generalised Poisson model, overall, the mean confidence rate for LM diagnoses scores was close to zero reflecting high confidence (mean rate = 0.065, 95% CI 0.040 to 0.105) (*Table 15*, column 2). Confidence for DP diagnoses was lower than confidence for LM diagnoses (rate ratio = 0.92, 95% CI 0.85 to 1.00; $p = 0.053$) (see *Table 15*, column 3). For the LM diagnoses, in the subgroup analyses, compared to overall analysis including all cases, diagnosis confidence is lower for breast cases, renal cases and decreases with increasing difficulty to report cases (column 2). However, focusing on rate ratios which compare LM with DP, like the results including all cases, the results by specialty show a non-significant trend to less confidence with DP diagnoses, except in skin diagnoses where the difference was in the opposite direction (column 3). In routine cases, pathologists were significantly less confident with DP diagnoses compared with LM diagnoses (rate ratio = 0.86, 95% CI 0.76 to 0.98; $p = 0.024$), but this difference was not apparent in moderate or difficult to diagnose cases.

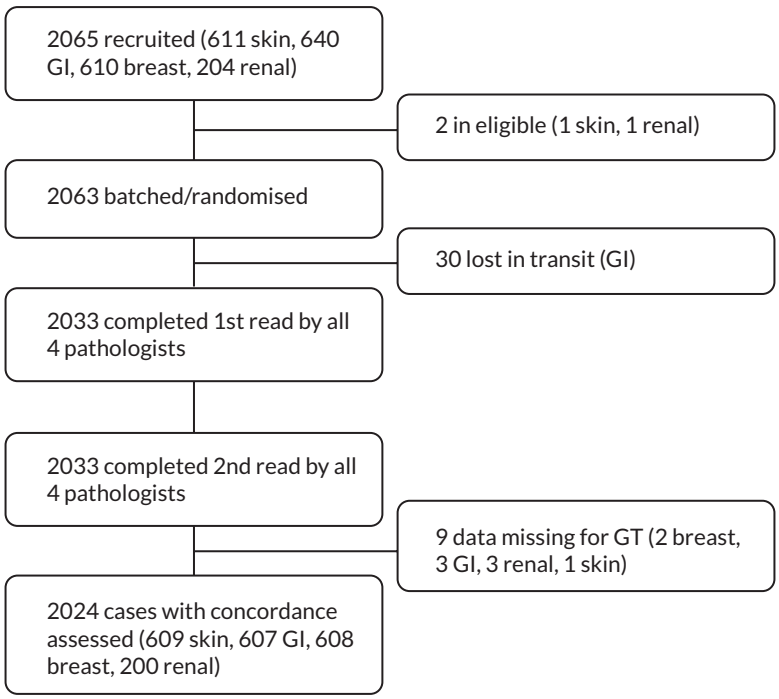

FIGURE 5 Consolidated Standards of Reporting Trials diagram.

TABLE 8 Cases by specialty, difficulty and recruiting site

Specialty	difficulty	Site: Coventry		Belfast		Lincoln		Nottingham		Oxford	All sites totals	
Breast	Routine	35	70%	122	72%	147	84%	182	85%		489	80%
	Moderate	3	6%	19	11%	22	13%	10	5%		54	9%
	Difficult	12	24%	28	17%	5	3%	23	11%		68	11%
	Total	50		169		174		215			608	
GI	Routine	162	80%	159	81%			156	75%		477	79%
	Moderate	18	9%	14	7%			21	10%		53	9%
	Difficult	23	11%	23	12%			31	15%		77	13%
	Total	203		196				208			607	
Skin	Routine	197	75%	189	87%	98	75%				484	79%
	Moderate	29	11%	6	3%	22	17%				57	9%
	Difficult	36	14%	22	10%	10	8%				68	11%
	Total	262		217		130					609	
Renal	(No difficulty)									200	200	
	Total									200	200	
Total		515		582		304		423		200	2024	

TABLE 9 Characteristics of patients and cases

Characteristic	All cases (*N* = 2024)	Breast (*N* = 608)	GI (*N* = 607)	Skin (*N* = 609)	Renal (*N* = 200)
Sex, n (%)					
Male	753 (37.2)	2 (0.3)	355 (58.5)	280 (46.0)	116 (58.0)
Female	1271 (62.8)	606 (99.7)	252 (41.5)	329 (54.0)	84 (42.0)
Age of patients (years)					
Minimum–maximum	< 1–96	18–94	< 1–89	1–96	19–96
Mean (SD)	58.0 (17.11)	54.8 (15.01)	59.5 (15.18)	60.0 (20.34)	56.9 (16.52)
Median (LQ–UQ)	59 (48–71)	54 (46–65)	62 (55–71)	63 (45–77)	57.5 (43–71)
Difficulty level, n (%)					
Routine	1447 (71.5)	486 (79.9)	477 (78.6)	484 (79.5)	0 (0)
Moderate	164 (8.1)	54 (8.9)	53 (8.7)	57 (9.4)	0 (0)
Difficult	413 (20.4)	68 (11.2)	77 (12.7)	68 (11.2)	200 (100)
Screening cases, n (%)					
Yes	NA	207 (34.0)	250 (41.2)	NA	NA
No		401 (66.0)	357 (58.8)		

LQ, lower quartile; UQ, upper quartile.

TABLE 10 Summary of the reports' comparisons data

Outcome	All cases (*N* = 2024)	Breast (*N* = 608)	GI (*N* = 607)	Skin (*N* = 609)	Renal (*N* = 200)
Clinical management concordance (primary outcome) summary					
LM and DP diagnoses concordance, n (%)					
All four comparisons concordant	1784 (88.1)	494 (81.2)	532 (87.6)	567 (93.1)	191 (95.5)
Three in four comparisons concordant	170 (8.4)	76 (12.5)	56 (9.2)	30 (4.9)	8 (4.0)
Two in four comparisons concordant	55 (2.7)	29 (4.8)	18 (3.0)	7 (1.1)	1 (0.5)
One in four comparisons concordant	14 (0.7)	8 (1.3)	1 (0.2)	5 (0.8)	0 (0)
All four comparisons discordant	1 (0.0)	1 (0.2)	0 (0)	0 (0)	0 (0)
LM and GT diagnoses concordance, n (%)					
All four comparisons concordant	1769 (87.4)	501 (82.4)	513 (84.5)	562 (92.3)	193 (96.5)
Three in four comparisons concordant	164 (8.1)	70 (11.5)	59 (9.7)	30 (4.9)	5 (2.5)
Two in four comparisons concordant	62 (3.1)	25 (4.1)	22 (3.6)	13 (2.1)	2 (1.0)
One in four comparisons concordant	27 (1.3)	12 (2.0)	11 (1.8)	4 (0.7)	0 (0)
All four comparisons discordant	2 (0.1)	0 (0)	2 (0.3)	0 (0)	0 (0)
DP and GT diagnoses concordance, n (%)					
All four comparisons concordant	1763 (87.1)	508 (83.6)	503 (82.9)	560 (92.0)	192 (96.0)
Three in four comparisons concordant	167 (8.3)	62 (10.2)	63 (10.4)	34 (5.6)	8 (4.0)
Two in four comparisons concordant	64 (3.2)	23 (3.8)	30 (4.9)	11 (1.8)	0 (0)
One in four comparisons concordant	25 (1.2)	15 (2.5)	7 (1.2)	3 (0.5)	0 (0)
All four comparisons discordant	5 (0.2)	0 (0)	4 (0.7)	1 (0.2)	0 (0)

TABLE 11 Summary of the CMC analysis using RE logistic regression models

Cases[a]	%CMC (95% CI)			ICC coefficient	
	Intraobserver LM vs. DP agreement	LM vs. GT agreement	DP vs. GT agreement	LM interobserver agreement	DP interobserver agreement
Overall (*n* = 2024)	99.95 (99.90 to 99.97)[b]	99.95 (99.91 to 99.97)	99.95 (99.91 to 99.97)	0.91 (0.89, 0.92)	0.91 (0.89, 0.93)
By specialty					
Breast (*n* = 608)	99.40 (99.06 to 99.62)	99.76 (99.54 to 99.87)	99.88 (99.73 to 99.95)	0.83 (0.60, 0.89)	0.88 (0.77, 0.91)
GI (*n* = 607)	99.96 (99.89 to 99.99)	99.92 (99.80 to 99.97)	99.89 (99.74 to 99.95)	0.90 (0.83, 0.93)	0.89 (0.77, 0.93)
Skin (*n* = 609)	99.99 (99.92 to 100.0)	99.99 (99.93 to 100.0)	99.98 (99.91 to 100.0)	0.94 (0.92, 0.95)	0.93 (0.92, 0.95)
Renal (*n* = 200)	99.99 (99.57 to 100.0)	100 (99.24 to 100.00)	99.18 (97.84 to 99.69)	[c]	[c]
By difficulty level					
Routine (*n* = 1447)	99.98 (99.94 to 99.99)	99.98 (99.94 to 99.99)	99.98 (99.94 to 99.99)	0.93 (0.91, 0.94)	0.93 (0.91, 0.94)
Moderate (*n* = 164)	95.34 (93.09 to 96.89)	93.91 (90.95 to 95.94)	94.24 (91.41 to 96.17)	0.53 (0.36, 0.78)	0.53 (0.36, 0.76)
Difficult excluding renal (*n* = 213)	96.78 (94.27 to 98.22)	97.78 (96.11 to 98.74)	98.40 (97.14 to 99.11)	0.42 (0.13, 0.53)	0.62 (0.24, 0.77)
Difficult including renal (*n* = 413)	99.84 (99.62 to 99.93)	97.63 (96.02 to 98.60)	97.68 (96.00 to 98.67)	0.33 (0.14, 0.90)	0.33 (0.17, 0.91)
Screening cases only					
Breast (*n* = 207)	96.27 (94.63 to 97.43)	97.57 (96.18 to 98.47)	98.23 (97.03 to 98.94)	0.53 (0.33, 0.87)	0.59 (0.35, 0.88)
GI (*n* = 250)	99.93 (99.68 to 99.98)	99.97 (99.78 to 100.0)	99.98 (99.83 to 100.0)	0.93 (0.89, 0.96)	0.94 (0.90, 0.96)
Breast and GI (*n* = 457)	98.96 (98.42 to 99.32)	99.87 (99.68 to 99.95)	99.89 (99.71 to 99.96)	0.88 (0.67, 0.92)	0.89 (0.73, 0.93)

a *n* is the number of cases. Each case is reported by four pathologists and so the number of comparisons in the analysis is 4*n*.
b Primary objective intraobserver intermodality CMC.
c Unable to be estimated reliably due to too few cases of discordance between LM and GT reports and between DP and GT reports (see *Table 10*).

TABLE 12 Summary of the reports' comparisons data

Outcome	All cases (*N* = 2024)	Breast (*N* = 608)	GI (*N* = 607)	Skin (*N* = 609)	Renal (*N* = 200)
Complete concordance (secondary outcome) summary					
LM and DP diagnoses concordance, n (%)					
All four comparisons concordant	1500 (74.1)	362 (59.5)	447 (73.6)	515 (84.6)	176 (88.0)
Three in four comparisons concordant	356 (17.6)	148 (24.3)	123 (20.3)	68 (11.2)	17 (8.5)
Two in four comparisons concordant	123 (6.1)	71 (11.7)	30 (4.9)	16 (2.6)	6 (3.0)
One in four comparisons concordant	40 (2.0)	23 (3.8)	7 (1.2)	9 (1.5)	1 (0.5)
All four comparisons discordant	5 (0.2)	4 (0.7)	0 (0)	1 (0.2)	0 (0)
LM and GT diagnoses concordance, n (%)					
All four comparisons concordant	1438 (71.0)	388 (63.8)	375 (61.8)	499 (81.9)	176 (88.0)
Three in four comparisons concordant	365 (18.0)	133 (21.9)	145 (23.9)	73 (12.0)	14 (7.0)

continued

TABLE 12 Summary of the reports' comparisons data (*continued*)

Outcome	All cases (*N* = 2024)	Breast (*N* = 608)	GI (*N* = 607)	Skin (*N* = 609)	Renal (*N* = 200)
Two in four comparisons concordant	154 (7.6)	61 (10.0)	61 (10.0)	25 (4.1)	7 (3.5)
One in four comparisons concordant	57 (2.8)	23 (3.8)	22 (3.6)	10 (1.6)	2 (1.0)
All four comparisons discordant	10 (0.5)	3 (0.5)	4 (0.7)	2 (0.3)	1 (0.5)
DP and GT diagnoses concordance, n (%)					
All four comparisons concordant	1420 (70.2)	381 (62.7)	367 (60.5)	493 (81.0)	179 (89.5)
Three in four comparisons concordant	362 (17.9)	136 (22.4)	140 (23.1)	72 (11.8)	14 (7.0)
Two in four comparisons concordant	179 (8.8)	67 (11.0)	74 (12.2)	32 (5.3)	6 (3.0)
One in four comparisons concordant	50 (2.5)	23 (3.8)	19 (3.1)	7 (1.1)	1 (0.5)
All four comparisons discordant	13 (0.6)	1 (0.2)	7 (1.2)	5 (0.8)	0 (0)

TABLE 13 Summary of the CC analysis using RE logistic regression models

	%CC (95% CI)			ICC coefficient	
Cases[a]	Intraobserver LM vs. DP agreement	LM vs. GT agreement	DP vs. GT agreement	LM interobserver agreement	DP interobserver agreement
Overall (*n* = 2024)	95.74 (94.39 to 96.77)	95.13 (94.09 to 96.00)	94.95 (93.55 to 96.06)	0.37 (0.30, 0.50)	0.35 (0.28, 0.48)
By specialty					
Breast (*n* = 608)	89.72 (88.15 to 91.10)	91.98 (90.57 to 93.20)	91.07 (89.61 to 92.34)	0.36 (0.28, 0.44)	0.32 (0.24, 0.40)
GI (*n* = 607)	94.62 (92.48 to 96.18)	90.88 (88.76 to 92.64)	90.34 (88.03 to 92.24)	0.32 (0.22, 0.40)	0.33 (0.24, 0.42)
Skin (*n* = 609)	99.86 (99.71 to 99.94)	99.78 (99.58 to 99.89)	99.79 (99.60 to 99.89)	0.84 (0.58, 0.92)	0.86 (0.64, 0.92)
Renal (*n* = 200)	99.94 (99.64 to 99.99)	99.97 (99.70 to 100.0)	99.96 (99.69 to 100.0)	0.93 (0.87, 0.96)	0.91 (0.83, 0.94)
By difficulty level					
Routine (*n* = 1447)	96.05 (94.47 to 97.19)	95.88 (94.98 to 96.62)	95.81 (95.12 to 96.40)	0.40 (0.29, 0.56)	0.45 (0.29, 0.55)
Moderate (*n* = 164)	87.88 (83.93 to 90.96)	80.89 (76.54 to 84.60)	81.65 (77.17 to 85.42)	0.27 (0.14, 0.39)	0.31 (0.18, 0.43)
Difficult excluding renal (*n* = 213)	91.88 (87.81 to 94.67)	91.14 (87.02 to 94.04)	88.88 (83.63 to 92.60)	0.16 (0.06, 0.28)	0.18 (0.07, 0.30)
Difficult including renal (*n* = 413)	93.82 (91.46 to 95.56)	93.39 (90.37 to 95.51)	92.52 (88.69 to 95.13)	0.29 (0.16, 0.43)	0.22 (0.11, 0.33)
Screening cases only					
Breast (*n* = 207)	88.60 (85.65 to 91.00)	88.18 (84.80 to 90.89)	87.79 (84.96 to 90.14)	0.23 (0.12, 0.35)	0.21 (0.11, 0.33)
GI (*n* = 250)	95.15 (93.17 to 96.58)	92.52 (89.87 to 94.51)	93.88 (91.29 to 95.73)	0.30 (0.15, 0.46)	0.41 (0.24, 0.59)
Breast and GI (*n* = 457)	92.17 (89.76 to 94.06)	90.54 (88.30 to 92.39)	91.02 (89.08 to 92.64)	0.27 (0.17, 0.36)	0.31 (0.20, 0.42)

a *n* is the number of cases. Each case is reported by four pathologists and so the number of comparisons in the analysis is 4*n*.

TABLE 14 Summary of diagnosis confidence levels

Modality	All reports (*N* = 8096)	Breast reports (*N* = 2432)	GI reports (*N* = 2428)	Skin reports (*N* = 2436)	Renal reports (*N* = 800)
LM, n (%)					
0[a]	7144 (88.3)	2090 (86.0)	2265 (93.3)	2244 (92.1)	545 (68.1)
1	713 (8.8)	254 (10.4)	146 (6.0)	134 (5.5)	179 (22.4)
2	180 (2.2)	66 (2.7)	15 (0.6)	41 (1.7)	58 (7.2)
3	40 (0.5)	11 (0.5)	0 (0)	13 (0.5)	16 (2.0)
4	7 (0.1)	4 (0.2)	0 (0)	2 (0.1)	1 (0.1)
5	5 (0.1)	3 (0.1)	0 (0)	1 (0.0)	1 (0.1)
6[b]	5 (0.1)	3 (0.1)	1 (0.0)	1 (0.0)	0 (0)
DP, n (%)					
0[a]	7079 (87.5)	2044 (84.1)	2249 (92.7)	2256 (92.6)	530 (66.3)
1	754 (9.3)	289 (11.9)	152 (6.3)	122 (5.0)	191 (23.9)
2	195 (2.4)	78 (3.2)	24 (1.0)	37 (1.5)	56 (7.0)
3	47 (0.6)	15 (0.6)	0 (0)	16 (0.7)	16 (2.0)
4	7 (0.1)	1 (0.0)	2 (0.1)	1 (0.0)	3 (0.4)
5	2 (0.0)	1 (0.0)	0 (0)	0 (0)	1 (0.1)
6[b]	9 (0.1)	3 (0.1)	0 (0)	3 (0.1)	3 (0.4)

a Highest confidence.
b Lowest confidence.

TABLE 15 Comparison of diagnosis confidence data using RE generalised Poisson models

Data included	LM mean rate (95% CI)[a]	Rate ratio (95% CI), *p*-value
All the data (all cases in study) (*n* = 2024)[b]	0.065 (0.040 to 0.105)	0.92 (0.85 to 1.00), 0.053
Subgroup analysis by specialty		
Breast cases (*n* = 608)[b]	0.091 (0.061 to 0.134)	0.90 (0.78 to 1.02), 0.108
GI cases (*n* = 607)[b]	0.027 (0.013 to 0.057)	0.87 (0.71 to 1.07), 0.189
Skin cases (*n* = 609)[b]	0.003 (0.001 to 0.008)	1.04 (0.86 to 1.25), 0.701
Renal cases (*n* = 200)[b]	0.300 (0.206 to 0.439)	0.91 (0.79 to 1.05), 0.208
Subgroup analysis by difficulty level		
Routine cases from all specialties (*n* = 1447)[b]	0.022 (0.013 to 0.037)	0.86 (0.76 to 0.98), 0.024
Moderate cases from all specialties (*n* = 164)[b]	0.105 (0.069 to 0.160)	1.32 (1.00 to 1.75), 0.052
Difficult cases from all specialties (*n* = 413)[b]	0.214 (0.158 to 0.289)	0.92 (0.82 to 1.02), 0.124
Difficult cases excluding renal cases (*n* = 213)[b]	0.182 (0.124 to 0.267)	0.92 (0.78 to 1.09), 0.357
Subgroup analysis of screening cases		
Combined breast and GI screening cases (*n* = 457)[b]	0.034 (0.014 to 0.082)	0.87 (0.70 to 1.07), 0.176
Breast screening cases (*n* = 207)[b]	0.099 (0.060 to 0.165)	0.84 (0.67 to 1.05), 0.119
GI screening cases (*n* = 250)[b]	0.002 (0.000 to 0.103)	1.00 (0.60 to 1.66), 0.994

a DP mean rate is obtained by dividing LM mean rate by rate ratio in the last column.
b *n* is the number of cases. Each case is reported by four pathologists on both LM and DP and so the number of rows for each case in the analysis is 8*n*. In the entire database, only 5 reports (out of 16,192 reports) had missing diagnosis confidence data.

TABLE 16 Errors recorded in two or more instances in breast, GI and skin specialties

Breast difference type	All	LM vs. GT	DP vs. GT	LM vs. DP	Screening cases
Tumour type	56	37	37	39	13
B2 vs. B3	48	37	29	30	12
B2 vs. B3 with atypia	35	20	19	31	16
B2 vs. B1	26	15	18	19	15
B3 with atypia vs. B5a	16	13	8	11	10
B5a vs. B5a mi	12	5	11	8	11
B3 with atypia vs. B3 no atypia	8	8	8	0	5
B5a vs. B5b	8	5	5	6	5
B3 with atypia vs. B3	7	2	5	7	2
B4 vs. B5a	3	3	2	1	2
B2 vs. B4	2	2	1	1	1
B2 vs. B5a	2	2	1	1	2
DCIS vs. no DCIS	2	1	1	2	
Missed lymphoma	2	2	1	1	
Missed melanoma	2	1	0	1	
Total	229	153	146	158	94
GI difference type	**All**	**LM vs. GT**	**DP vs. GT**	**LM vs. DP**	**Screening cases**
HP vs. SSL	37	29	26	21	31
LGD vs. HGD	32	22	28	14	14
Tumour stage	13	10	9	6	1
Normal vs. HP	12	10	9	5	10
Missed *H. pylori*	8	6	7	3	
TA vs. SSL	7	7	5	3	7
Normal vs. BA2	5	4	5	1	
TA vs. TA LGD	4		4	4	4
Inflammation NOS vs. IBD	4	4	3	1	1
Inflammation vs. LGD	3	3	3		
Quiescent vs. active colitis	3	3	3		
GI difference type	**All**	**LM vs. GT**	**DP vs. GT**	**LM vs. DP**	**Screening cases**
Inflammation vs. indefinite for dysplasia	3	2	2	2	
Gastritis vs. amyloidosis	2	2	2		
Normal vs. fundic polyp indefinite for dysplasia	2	2	2		
Quiescent vs. IBD NET	2	2	2		
Reactive vs. TA	2	2	2		
Reported incorrect case	2	2	2	1	2
TA vs. HP	2	2	2	2	

TABLE 16 Errors recorded in two or more instances in breast, GI and skin specialties (*continued*)

Breast difference type	All	LM vs. GT	DP vs. GT	LM vs. DP	Screening cases
Tumour type	2	2	2		
Barrett's vs. indefinite for dysplasia	2	1	2	1	
Normal vs. IEL	2		2	2	
TA vs. polyp cancer	2		2		
Normal vs. non-specific inflammation	2	2	1	1	
Inflammation vs. IM	2	1	1	2	
Total	155	118	126	69	70
Skin difference type	**All**	**LM vs. GT**	**DP vs. GT**	**LM vs. DP**	
BCC with high-risk component vs. BCC	18	9	11	16	
MM vs. naevus	11	8	6	8	
SCC margin involvement vs. no margin involvement	10	8	7	5	
BCC vs. SCC	6	3	6		
SCC vs. AK or IEC	6	5	5	2	
Breslow thickness	5	3	4	3	
Blue naevus vs. atypical naevus	5	4	3	5	
KA vs. SCC	5	5	3	2	
In situ vs. invasive melanoma	5	4	2	4	
Melanoma margin involvement	4	3	4	1	
Adenoid cystic carcinoma vs. benign adnexal tumour	2	1	2	1	
DFSP vs. DF	2	1	2	1	
Herpes vs. alternative inflammatory lesions	2	1	2	1	
Lichenoid keratosis vs. compound naevus	2	2	2	0	
Bowen's disease vs. stasis	2	2	1	1	
Metastatic melanoma vs. benign node	2	2	1	1	
Viral wart vs. polyp	2	1	1	2	
Total	89	62	62	53	

AK, actinic keratosis; B1–B5, NHS Breast Screening Programme pathology category classification 1–5 (a, in situ; b, invasive; mi, microinvasive carcinoma); BA2, Barrett's metaplasia; DF, dermatofibroma; DFSP, dermatofibrosarcoma protuberans; HGD, high-grade dysplasia; HP, hyperplastic polyp; IEC, intraepidermal carcinoma; IEL, intraepithelial lymphocytosis; IM, intestinal metaplasia; KA, keratoacanthoma; LGD, low-grade dysplasia; MM, malignant melanoma; NET, neuroendocrine tumour; SSL, sessile serrated lesion; TA, tubular adenoma.

Chapter 4 Qualitative substudy

Introduction

The provision of WSI is seen as essential for high-quality and cost-effective health care. However, adoption of DP in the UK has been relatively slow, with the published research focusing mainly on technical applications in single hospital sites[31,32] and integration into reporting and laboratory management systems.[33–35] The barriers and facilitators to implementation of DP are well acknowledged in the international literature,[36,37] but practical learning from the UK is limited. Few studies have examined the perceived barriers and facilitators to DP uptake and implementation, even fewer explored the evolution of these as pathologists embark on using and transition to digital slides. The qualitative component of this study, therefore, aimed to explore the barriers and facilitators to DP and experiences of main study participants using and/or transitioning from glass to digital slides.

Methods

This was a qualitative longitudinal study with semistructured interviews and focus groups with pathologists and laboratory staff taking part in the main validation study. Data were collected and analysed by a trained qualitative researcher with over 10 years' experience in qualitative health research.

Sample and data collection

Twelve consultant pathologists and three laboratory staff took part in the qualitative substudy.

Data collection was undertaken in three phases:

1. Pilot study: individual interviews and focus groups with data collection between December 2018 and April 2019.
2. Beginning of the main study: individual interviews with data collected between August 2020 and January 2021.
3. 12–15 months into the main study: individual interviews with the same participants as in phase 2 with data collected between August and October 2021.

In the pilot study, participants were asked, as appropriate to experience, about their willingness to adopt DP into routine practice, benefits and limitations of the conventional microscope and current workflows, and the potential and impact of DP on workflows and organisational processes. The first and second interviews as part of the main study covered similar topics but focused on experiences of using DP, transitioning from glass to digital slides and DP implementation. The interview guide for laboratory staff was similar, with some adjustments to the focus and order of questions. Participants were asked to discuss both their own experiences and those of their colleagues in their respective departments.

The whole corpus of data included two focus groups (with 4–7 participants in each) conducted face-to-face and 34 interviews conducted either face-to-face or on Zoom/Teams (Zoom Video Communications, San Jose, CA, USA).

On average, focus groups lasted between 1 and 1.5 hours, and interviews around 45 minutes. All focus groups and interviews were audio-recorded and later transcribed by a professional transcription company. Written and verbal consent was taken before each focus group and interview. All participants were assigned a participant identification number to protect anonymity and were given the opportunity to contact the study team if they wished to read the transcripts.

Data analysis

The principles of longitudinal qualitative research as a flexible methodology guided the analysis process.[38] Data for all time points were coded separately with the help of NVivo (QSR International, Warrington, UK) qualitative data analysis software (NVivo Qualitative Data Analysis Software; 2015). For each data set time point, a sample of transcripts was

read to identify examples of benefits and limitations of conventional microscope, expressions of challenges of adopting DP into routine practice and reflections on the potential impact on current workflows and organisation of care. An initial set of codes derived both inductively from the data and deductively based on the broad aims of the substudy was created and discussed among wider project team members. This coding framework was then applied to all remaining data.

The analysis was concurrent with data collection, including continuous processes of coding and categorisation, with project team members meeting regularly to discuss the emerging findings and their interpretations.[39]

Results

The overarching message from the data was that adopting and implementing DP on a larger scale might need time and resources to test out the approaches and gain experience with digital technology. When introducing or scaling up DP, several barriers and enablers may also need to be considered, such as appropriate training, local validation and implementation plans to share knowledge and experiences, depending on how comfortable individual pathologists and the wider pathology team are with the technology.

Table 17 presents main barriers and facilitators in the three stages of the study, as perceived by the participants.

Pilot study

In the pilot phase, most participants were cautiously optimistic about the potential benefits of DP and perceived the wider deployment in the NHS as inevitable. The majority of pathologists and laboratory staff emphasised the need for time and resources to test out the approaches and experience the technology.

There was some difference in excitement for implementation and scale up of DP. Those who were already using digital slides in their everyday practice believed that the lack of familiarity with digital slides, a lack of communication about what digitalisation entailed, and overstating the benefits were to blame for the opposition to go digital within the larger pathology community. All participants agreed that perceived usefulness was an important reason for implementing DP. However, pathologists argued that for DP to be a valuable working tool, it needs to be more than merely digitising the slides.

TABLE 17 Main barriers and facilitators to DP implementation

Data collection phase	Barriers	Facilitators	Example quotes from the data
Pilot phase	• Resistance to go digital among the broader pathology community • Lack of communication around what digitalisation means and overstating the benefits • Perceived lower quality of digital image • Lack of compatibility of digital technology and current reporting systems • Lack of established workflows • Initial costs on laboratories, pathologists, and other staff • Perception that digitalisation will be limited and the scale-up of technology halted without the necessary resources	• Time and resources to test out the approaches and experience digital technology • Perception that DP is useful in everyday work • Training and validation process, being able to learn from more experienced colleagues • Ability to work collaboratively to adapt the workflows • Long-term, multi-year approach to adoption and scale up	'I am just not willing to give up the microscope, not yet'. 'We need to roll out something and let people experience it'. 'If you're just replacing the microscope with a computer screen, where is the benefit?' 'Money is available for the trail blazers and then, even if they do really well, by the time you're getting into the second and third wave, the money's petered out, you know, you'd be lucky if you get a new PC and an old digital camera'.

continued

TABLE 17 Main barriers and facilitators to DP implementation (*continued*)

Data collection phase	Barriers	Facilitators	Example quotes from the data
Beginning of the main study	• Perception that it takes time to set up and deploy DP, and get different people on board • Lack of an appreciation of the difficulties involved in DP implementation and scale-up • View that the hybrid approach to reporting will introduce variations in the processes and practices • Perception that the actual benefits of DP will not be immediate • Technical and interoperability challenges, lack of compatibility of digital technology and reporting systems • Broader challenges that relate to change and digital transformation in health care	• Creating space and making time to experience the technology, becoming familiar with the technology • Local validation processes and implementation/scale-up plans • Understanding the value, benefits and risks of DP • Sharing knowledge and experiences among pathologists and departments transitioning to digital slides	'With the pressure and encouragement to reduce our working hours and interactions within the department, digital pathology has had to be welcomed by everybody'. 'The COVID crisis was a huge encouragement for us to become more familiar with the technology but if you don't feel completely safe to report entirely on a digital system, then you have to go to the department to do spot checks'. 'You have all these hospitals that are just going about their business relatively secretly. All these people are doing digital in a different way and they're not speaking to each other about this'. 'The quality of the digital images is fantastic. The irony of everything is if we were starting to do digital pathology tomorrow, I would be looking at this fantastic image on my £10,000 monitor, and then picking up one of these old things to generate the report'.
12–15 months into the main study	• Technical and interoperability challenges, lack of integration of DP into existing reporting systems • Perception that efficiency and cost-saving benefits of digitalisation are not likely to directly impact the individual pathologists	• Appropriate technology and good-quality equipment to enable viewing of slides • Flexibility and individual preference when working with digital slides • Ability to seek diagnostic reassurance on difficult or unusual cases	'The software needs to be sufficiently sophisticated to give me an efficient workflow, you know, all the technology, the mouse, the way of controlling needs to work well, to allow me to be as efficient'. 'The main issue with digital pathology is how the workflow goes. Most of us share this experience, which is we still prefer glass, it gives you a more solid notion of your workflow, your workload and how to manage it'. 'There are weaknesses both with light microscopy and fluorescence digitally and so long as you're aware of those and know when to defer to glass slides then I think it is probably going to be safe to report digitally'.

Lower digital picture quality and compatibility of digital technologies and reporting systems appeared to be major issues for pathologists new to DP. These pathologists saw the significance of training and validation, and viewed their participation in the study as an opportunity to learn from and benefit from the knowledge of their more experienced colleagues.

In individual interviews, a lack of established workflows was seen as an important obstacle to the implementation and widespread use of DP. Pathologists and laboratory staff shared their concerns about changing and adapting how they work. There was a recognition that for the DP to succeed, everyone will need to work collaboratively to build a shared understanding of what they are trying to achieve, and, from there, adapt the workflows.

Access to funding for initial hardware, software and training was raised as an issue among all focus group participants. Concerns were expressed, for example, about digitalisation being limited and technological scaling-up being halted due to a lack of resources. The general consensus was that if DP was to be implemented on a large scale for primary

diagnosis, departments and laboratories needed to be allowed to take a long-term approach. The participants acknowledged that the DP gains will not be immediate, but there will be costs on laboratories, pathologists, and other staff. This was thought to create tensions, particularly in the context of wider buy-in and the lack of financial support for DP adoption.

Beginning of the main study

The beginning of the main study coincided with the worldwide COVID-19 pandemic. Participants talked about changing attitudes towards DP. Pathologists saw more openness to change and interest in digitalisation among their colleagues as a result of the pandemic, whereas laboratory staff commented on the increased buy-in and motivation for going digital. All participants stressed the importance of the pathology workforce feeling confident in using DP.

Pathologists reported on their improved familiarity with DP during their interviews. Most people believed that DP was as good as LM after engaging with it and learning how to use it. We heard, again, that creating space and making time to experience the technology was important. Participants discussed ongoing work in their departments and laboratories on how to implement or scale up DP. They also highlighted perceived challenges and shared their concerns about the impact of transitioning to digital slides.

Some pathologists involved in strategic planning and management emphasised the time commitment required to set up and deploy DP and bring diverse groups of people on board. They observed that there was frequently a lack of understanding of the difficulties, frustrations and pressures associated with DP adoption. Those pathologists whose departments started using digital slides reflected on the new ways of working and highlighted that it takes time for a new style of working to become habitual.

Among all participants, there was a recognition that migration to digital slides cannot happen instantaneously, and a hybrid glass–digital working will be required. Concerns were raised about the potential variations in the processes and practices among different pathologists and across different laboratories. In this regard, understanding the value, benefits and risks of DP was emphasised as key to establishing DP as a new way of working. More experienced pathologists were generally more sceptical about the benefits of for improving efficiency and reporting times, recognising that the often-cited gains of DP will not be immediate. Those pathologists who were at the beginning of their DP journey highlighted the importance of sharing knowledge and experiences among pathologists and departments transitioning to digital slides.

Technical and interoperability problems were viewed as hurdles to successful implementation and pathologists' confidence in using DP. The need of compatibility of digital technologies and reporting systems was emphasised again.

Compared to the pilot phase, participants' views began shifting towards broader structural and organisational barriers that impact DP implementation and scale up. They noted, for example, that the broader challenges that relate to change and digital transformation in health care can and often will complicate the adoption of DP.

12–15 months into the main study

In this phase of the main study, adaptation of DP to personal and local needs became a dominant theme in the interviews. Pathologist explained that choosing a suitable technology and good-quality equipment to enable viewing of slides is important if DP was to assist their routine practice. Most pathologists also felt that the system must function in an optimal manner to support their daily work. Similarly, to the other two data collection phases, integration of DP into existing reporting systems seemed to remain one of the key considerations for pathologist when discussing their experiences of using and transitioning to digital slides. Discontentment with having to move between different computer screens, frustration about low internet bandwidth in their respective NHS Trusts and outdated equipment in the departments were mentioned as factors discouraging the use.

At the same time, the importance of having strategies in place to adapt to the local and individual needs was highlighted. One such method was to seek diagnostic reassurance in difficult or atypical cases. Participants commented on the hybrid glass–digital slide method, but with a noticeable shift in their concern regarding potential variances in processes and practices when compared to the first round of interviews. They acknowledged that in some specialties

there will always be cases where they will want to look at the glass slides. Whereas more experienced pathologists saw their participation in the study as supporting what they already knew about the advantages and disadvantages of DP, others reported engaging with new technology and increasing their willingness to use it, despite limitations.

Finally, there was an acknowledgement that, in the long run, DP deployment and larger scale-up must progress towards realising the benefits of financial savings and, ultimately, strengthening the resilience of pathology as a profession. However, several more experienced pathologists expressed concern that DP would not be a panacea because the efficiency and cost-saving benefits of digitalisation would not immediately affect individual pathologists.

Discussion

The aim of this qualitative study was to investigate the barriers and facilitators of DP, as well as the perspectives and experiences of pathologists and laboratory staff before and during DP implementation. We identified various challenges and facilitators to adoption, reflecting participants' opinions and experiences while using and transitioning to digital slides.

In the early phases of the study, participants were cautiously hopeful about DP and its potential benefits. Their perspectives emphasised the importance of time and resources to test out the ways and experience the technology. They talked about wider resistance to go digital among the broader pathology community and initial set-up costs as important impediments. As participants gained first-hand experience with DP in the months that followed, their opinions began to move to considerations of broader organisational and structural constraints that impact acceptance and implementation. Their perspectives reflected the need to consider the investment of time and resources in setting up and deploying DP and bringing different people on board, variations in processes and practices when implementing DP, and a recognition that the frequently cited benefits of DP will not be immediate. In the final round of interviews, participants began to place a higher priority on flexibility and individual preference when it came to using and implementing DP. Considerations such as having suitable technology and acceptable quality equipment to view slides, familiarity with the systems and processes, and acceptance that digital slides had limits worked both for and against DP adoption and deployment.

The difficulties associated with implementing new technologies and managing change in healthcare organisations are well acknowledged in the literature.[40–42] The analysis of participants' perspectives and experiences in this qualitative substudy revealed the tensions that may exist in the context of using and implementing DP. There may be various types of barriers and facilitators to DP uptake and use. This qualitative study found some variation in how pathologists with varying levels of familiarity with DP viewed the technology and reached differing conclusions about its utility. While similarities are important, differences between individuals with and without experience with DP show that all groups may have a unique viewpoint on the implementation problems that should be considered.

Furthermore, the range of barriers and facilitators, as well as their varying importance at different stages of the study, highlight the need to considering the varied barriers and facilitators as fluid and negotiated through participants' experiences of adopting and using DP.[43,44] Individual participant attitudes and concerns, as well as the technology itself and the structural and organisational factors around implementation, all seemed to have an impact on the experience of using and transitioning to digital slides. Pathologists and laboratory staff offered a variety of perspectives, with some barriers and facilitators recurring and persisting throughout the duration of study. Some barriers were recognised only later, as pathologists and laboratory staff grew more acquainted with the technology. However, based on participants' reflections from the three stages of the study, it appears that providing opportunities for first-hand experience with the system and understanding the value, benefits and risks of DP may be critical to overcoming some of the barriers. The process of adoption and scale-up of technology in health care is not straightforward.[45] It may be necessary to engage with pathologists and laboratory staff to better understand how the various stages of implementation shape distinct barriers and facilitators to wider adoption into routine practice.

Finally, our findings resonate with existing implementation theories and frameworks such as Normalization Process Theory,[46] Technology Acceptance Model[47] and Organizational and Practice Theories.[45,46] According to our participants, the perceived usefulness of DP and how it integrated into their everyday practice and the larger organisational context were some of the main considerations for effective adoption and implementation. Our findings are also consistent with Liberati *et al.*'s implementation framework,[43] implying that in order for DP to be embraced, it is necessary to investigate not only the barriers and facilitators, but also the implementation processes.

Study limitations

The study had some limitations. Firstly, this was a qualitative study undertaken as part of a larger project and the generalisability of results may therefore be limited. Secondly, only people who volunteered to participate in the main validation research were included in this study's sample. They already had preconceived ideas about DP as a result of their participation in the main study. Comparisons with larger research and the use of large samples will be required to strengthen the findings of this study. Third, the choice of methods of data collection to explore the barriers, facilitators and experiences of transitioning from glass to digital slides may have had an impact. Investigating implementation challenges through observations may provide further insights. Finally, we used a longitudinal lens to investigate participants' experiences of using and transitions from glass to digital slides. The findings presented here can help to broaden understanding of how the transition process is perceived, but further study is needed to understand how various factors impact the perspectives and opinions of people adopting and implementing DP. In particular, a better understanding of the factors and processes involved in the scale-up and embedding of DP is routine practice is needed.

Conclusions

When adapting and scaling up DP, several challenges and facilitators may need to be considered. Successful deployment may need to include careful consideration of general attitudes towards technology implementation, pathologists' perceptions of DPs usefulness, and time and resources to test out the approaches and experience digital technology first-hand.

Chapter 5 Health economics substudy

Introduction

Digital pathology is the use of high-throughput slide scanners to digitise diagnostic histopathology slides that are reported by pathologists on computer workstations as opposed to a conventional LM.

Digital Pathology Trial was designed around teams of four pathologists, blind to the original diagnoses, all examining the same series of samples, using both LM and DP. The modality each pathologist reports on first was randomised for each batch of samples, and there was a minimum of a 6-week washout period between LM and DP viewings. Reports were scrutinised by a trained Research Fellow independent of the reporting pathologists. Differences detected were classified by an independent pathologist into major (will alter the clinical management) or minor (which will not). The original diagnosis served as the RD. The GT for each sample was decided on conclusion of the readings, by consensus of the study pathologists, using multi-head LM viewing if needed and taking into account the RD.

This allowed a comparison of each pathologist's performance on LM and DP against the GT. The study examined 2028 complete histopathology samples including 600 samples each of breast, GI (including 200 cancer screening samples) and skin, and 200 renal samples taken for native, or transplant related renal disease (four samples were removed from the main study due to incomplete clinical and macroscopic data). The renal biopsies included immunofluorescence for detection of immunoglobulin deposits. The population of samples enrolled were a combination of sequentially selected samples and those from study centres' archives. Except for renal samples, all other samples were enriched with about 10% moderately difficult and at least 10% difficult samples. The main study objective was to compare pathologists' diagnoses made by assessment of LM of breast, GI, skin and renal samples, with the same pathologists' diagnoses of the same samples (intrarater reliability) using DP.

The use of DP allows for remote reporting of histopathology samples from the laboratory. Widespread implementation of DP allows for more efficient matching of workloads to pathologist capacity in hospitals and targeting of difficult samples to experienced pathologists to reduce error.[8] It also allows for the creation of digital pool of histopathology samples to avoid blockages at individual pathologist level due to temporary unavailability. Furthermore, it may lead to a reduction in the human resources required for the management of diagnostic histopathology slides and MDT meetings.[48] Beyond the potential efficiency gains that may accrue due to DP, questions remain as to whether reporting time is affected by DP technology. It is also important to know whether digitised histopathology slides are of sufficient quality to ensure diagnostic accuracy equivalent to bright-field and immunofluorescent LM currently used.

The aims of the study are:

1. To explore the impact of DP on the time taken for diagnosis.
2. To explore the costs and benefits of implementing DP, and its impact on throughput in the laboratory.

The objectives of the study are:

I. To estimate the effect of DP technology on histopathology reporting time across the whole sample.
II. To estimate the time taken for diagnosis for each pathologist and across specialties.
III. To explore interaction effects between characteristics of samples and diagnostic technology on time taken for histopathology reporting.
IV. To explore potential learning effects following continued use of DP.
V. To estimate the costs associated with implementing DP.
VI. To explore the impact of DP on resource use and productivity within the laboratory; and the incremental costs per additional sample reported following digitisation.
VII. To explore the impact of DP on diagnostic accuracy.

The following sections describe the methods and results of estimating reporting time. The costs and benefits associated with implementing DP are currently being explored using a discrete event simulation model and are not reported here.

Methods

The difference in time taken for diagnosis using standard LM and DP was estimated using unadjusted and adjusted regression analysis. The adjusted analysis controlled for a number of variables to isolate the effect of technology on reporting time. Variables adjusted for includes technology first used (i.e. whether a DP or LM was first used for diagnosis before the 6-week washout period), order of reading (i.e. whether the observation is the first or second read), pathologist, confidence in diagnosis, complexity (difficulty level), number of slides and specialty. Each sample generated eight different observations, hence a hierarchical generalised linear model was used to allow for clustering at the sample level.[49] The statistical models were used to investigate the effects of technology on time taken for diagnosis across the whole sample, in each specialty and for each pathologist. Interaction effects were fitted to explore whether:

I. There was a learning effect with the use of DP technology (i.e. whether pathologist reporting time improved as they became more familiar with DP).
II. Reporting time is affected by technology for cases with relatively higher number of slides.
III. Reporting time differs between technology for more complex samples.

To explore learning effects associated with continued with DP, the dates within which each pathologist reported their first sample and last sample was divided into 10 quantiles. Model 2 (interaction models) was fit to the data to quantify any learning effect associated with increasing familiarity with DP technology. The mean reporting time for each technology and the mean difference in reporting time for observations falling within each quantile of reported date were estimated.

The general form of the statistical model is shown in *Appendix 1*. All models were fitted using the meglm Stata command (StataCorp LP, College Station, TX, USA'.

Estimating reporting time for digital pathology and light microscopy

For each fitted model, we obtain margins of derivatives of reporting time with respect to technology. Various distributions were explored for the response variable and statistical models were selected based on convergence and log likelihood.

Results

Table 18 shows the breakdown of enrolled samples by complexity and specialty levels. Two thousand and twenty-eight samples (including four samples removed from the main study due to incomplete clinical and macroscopic data) were recruited and enrolled in the study (608 breast samples, 610 GI samples, 201 renal samples and 609 skin samples).

TABLE 18 Breakdown of enrolled samples by complexity and specialty

	Routine samples	Moderate samples	Difficult samples	Total
Breast	486	54	68	608
GI	480	53	77	610
Renal	0	0	201	201
Skin	484	57	68	609
Total	1450	164	414	2028

Sixteen pathologists (four per group) read each sample twice generating 16,219 observations. There were 4861 breast observations, 4878 GI observations, 1608 renal observations and 4872 skin observations.

Completeness of data

Table 19 shows the completeness of data on reporting time by specialty and diagnostic technology. Of 16,219 observations, data were missing in 38 observations (0.23%). Two observations reported an unrealistic reporting time of 0 and was excluded from the analysis. Further investigation revealed a frustration by the reporting pathologist on the number of slides in the case. Due to the very low levels of missingness (< 0.5% missingness across the study), only observations with complete data were used for the analysis.

Time taken to report samples using digital pathology and light microscopy technology

Pathologists spent an unadjusted mean time of 4.89 minutes reporting samples using LM technology and 5.01 minutes reporting samples using DP technology. The mean unadjusted difference (DP–LM) of 7 seconds is not statistically significant at a significance level of 95% (CI: 16.7 seconds to −2.82 seconds). Reporting time ranged from 4.02 seconds to 65 minutes for samples reported using LM technology and 3 seconds to 66.48 minutes for samples reported using DP technology. *Table 20* shows the mean and range of reporting time, and the median and interquartile range of number of slides per sample, by specialty and complexity.

Number of slides per specialty

The number of slides per case varied across the specialties, about 76% of breast samples had five slides or less and over 90% of GI and skin samples had five slides or less. GI and skin samples both had a median of two slides per sample. Breast samples had a median of 3 slides per sample, while renal samples had a median of 10 slides per sample as shown in *Table 20*. *Figure 6* shows the distribution of the number of slides per sample by specialty.

Statistical modelling of reporting time

Table 21 reports the estimates of reporting time across the whole study, by specialty and for each pathologist. The statistical model was adjusted for all pre-specified variables. Pathologists have statistically similar times reporting samples using both technology (mean: 1.4 seconds; 95% CI −2.4 seconds to 5.4 seconds). Examining the time taken for diagnosis by specialty, pathologists were 35 seconds faster reporting skin samples using DP compared to LM. They took longer reporting breast, renal and GI samples on DP compared to LM as shown in *Table 20*.

Within the specialties, there was no clear advantage of either technology on reporting time. In breast samples, two of four pathologists had statistically similar reporting time while the other two were slower using DP. Similarly in GI samples, two of four pathologists had statistically similar reporting times on both technology while the other two were slower reporting samples on DP. In renal samples, one pathologist had statistically similar reporting time on DP and LM, while the other three pathologists were slower on DP. In skin samples, one pathologist had statistically similar reporting times on both DP and LM, one pathologist was about 6 seconds slower on DP, and two pathologists were faster on DP as shown in *Table 21*.

Some pathologists were relatively slower reporting samples on DP, while others were relatively faster reporting samples on DP. Within breast samples, pathologist II took about 1.85 minutes longer to report samples on DP, while technology had no effect on pathologist I and pathologist III. Most pathologists reporting renal samples were considerably slower

TABLE 19 Completeness of reporting time by specialty and modality

	DP (*n* = 8109)	LM (*n* = 8110)	Total (*n* = 16,219)
Breast (*n* = 4861)	2419 (99.55%)	2426 (99.79%)	4845 (99.67%)
GI (*n* = 4878)	2435 (99.84%)	2436 (99.88%)	4871 (99.87%)
Renal (*n* = 1608)	799 (99.38%)	802 (99.75%)	1601(99.56%)
Skin (*n* = 4872)	2431 (99.79%)	2432 (99.84%)	4863 (99.82%)
	8084 (99.69%)	8096 (99.84%)	16,180 (99.76%)

TABLE 20 Range, unadjusted mean reporting time and number of slides by specialty and difficulty levels

	Number of slides	DP (*n* = 8109)		LM (*n* = 8110)	
	Median (IQR)	Range (minutes)	Mean (SD)	Range (minutes)	Mean (SD) (minutes)
Routine samples					
Breast (*n* = 3873)	3 (1–4)	0.45–50	3.79 (2.93)	0.07–49	3.33 (2.37)
GI (*n* = 3836)	2 (1–2)	0.33–26	3.86 (3.01)	0.42–55	3.68 (3.09)
Skin (*n* = 3864)	1 (1–2)	0.05–48	2.01 (2.28)	0.07–33.97	2.56 (2.82)
Moderate samples					
Breast (*n* = 431)	6 (3–9)	1.27–45	8.89 (6.84)	1.00–38.20	7.15 (5.33)
GI (*n* = 421)	6 (4–14)	1.25–59.17	13.56 (10.52)	1.75–65	13.64 (10.76)
Skin (*n* = 455)	4 (2–6)	0.63–28.38	6.58 (5.13)	0.27–40.47	8.59 (7.25)
Difficult samples					
Breast (*n* = 541)	5 (3–8)	0.5–40	6.83 (5.84)	0.08–60	6.22 (6.09)
GI (*n* = 613)	2 (1–4)	1.02–57	7.16 (5.95)	1.02–56.17	6.74 (5.63)
Renal (*n* = 1601)	10 (9–13)	3.25–66.48	12.21 (6.33)	3.38–41.5	11.03 (5.33)
Skin (*n* = 544)	3 (1–8)	0.27–36.17	6.66 (6.19)	0.27–38.37	7.99 (7.79)
Aggregate					
Level of difficulty					
Routine (*n* = 11,573)	2 (1–3)	0.05–50	3.22 (2.88)	0.07–55	3.19 (2.81)
Moderate (*n* = 1309)	5 (3–8)	0.63–59.17	9.59 (8.28)	0.27–65	9.74 (8.51)
Difficult (*n* = 3299)	9 (3–11)	0.27–66.48	9.47 (6.71)	0.08–60	8.94 (6.33)
Specialty					
Breast (*n* = 4845)	3 (2–5)	0.45–50	4.58 (4.18)	0.07–60	4.00 (3.60)
GI (*n* = 4872)	2 (1–3)	0.33–59.17	5.11 (5.39)	0.42–65	4.93 (5.45)
Renal (*n* = 1601)	10 (9–13)	3.25–66.48	12.21 (6.33)	3.38–41.5	11.03 (5.33)
Skin (*n* = 4863)	2 (1–2)	0.05–48	2.96 (3.78)	0.07–40.47	3.73 (4.83)

on DP compared to LM, with pathologist IV taking as much as 3 minutes longer to report renal samples on DP. However, renal pathologist I had statistically similar time reporting samples on both DP and LM. Conversely, in skin samples, pathologists were generally faster reporting samples using DP compared with LM with pathologist II being about 4 minutes faster reporting skin samples on DP compared to LM while pathologist IV was 6 seconds slower reporting skin samples on DP.

Is there a learning effect following continued use of digital pathology?

Most pathologists use standard LM technology for routine histopathology diagnosis; hence they may be faster reporting samples on a more familiar technology. However, as pathologists become more familiar with using DP, there may be a learning effect leading to faster reporting time. *Table 22* reports the results of the model estimates for the 1st quantile (start), 5th quantile (mid-point) and 10th quantile (end). *Figures 7–9* report graphically the predicted mean reporting time of samples reported using DP and LM technology in observations falling within each quantile, alongside the mean difference at each quantile with 95% CIs. Estimates falling above the reference line at 0 on the right side of *Figures 7–9* indicate that pathologists were slower reporting samples on DP compared with LM while observations falling below the

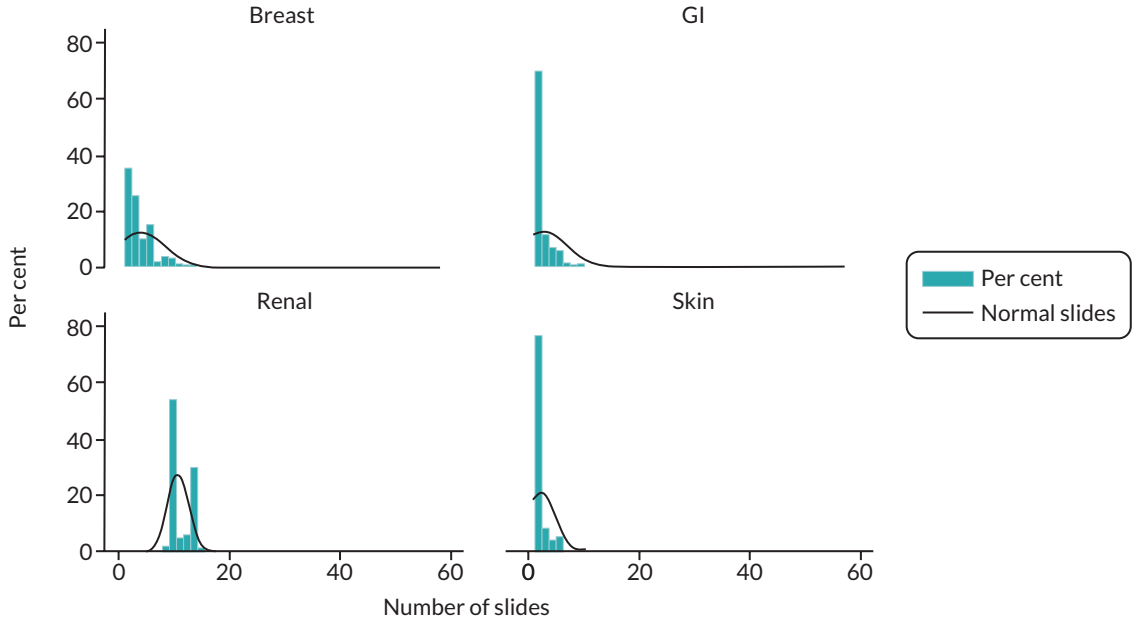

FIGURE 6 Number of slides by specialty.

TABLE 21 Parameter estimates[a] for effects of technology on time taken for diagnosis (model 1)

Whole sample	Mean (SE)
DP	5.47 minutes (0.10)
LM	5.49 minutes (0.10)
Mean difference[b]	1.4 seconds (−2.4 seconds to 5.4 seconds)
Specialty	***DP–LM Mean difference (95% CI)***
Breast	27 seconds (22 seconds to 33 seconds)
GI	18 seconds (11 seconds to 24 seconds)
Renal	58 seconds (43 seconds to 74 seconds)
Skin	−35 seconds (−41 seconds to −30 seconds)
Pathologist	
Breast	
Pathologist I	6 seconds (−2 seconds to 14 seconds)
Pathologist II	1.85 minutes (1.6 minutes to 2.1 minutes)
Pathologist III	2 seconds (−5 seconds to 10 seconds)
Pathologist IV	19 seconds (10 seconds to 27 seconds)
GI	
Pathologist I	2 seconds (−2 seconds to 7 seconds)
Pathologist II	2 seconds (−2 seconds to 7 seconds)
Pathologist III	44 seconds (30 seconds to 57 seconds)
Pathologist IV	12 seconds (6 seconds to 19 seconds)

TABLE 21 Parameter estimatesa for effects of technology on time taken for diagnosis (model 1) (*continued*)

Whole sample	Mean (SE)
Renal	
Pathologist I	25 seconds (−64 seconds to 14 seconds)
Pathologist II	30 seconds (19 seconds to 42 seconds)
Pathologist III	67 seconds (48 seconds to 86 seconds)
Pathologist IV	3 minutes (2.32 minutes to 3.52 minutes)
Skin	
Pathologist I	−17 seconds (−24 seconds to −10 seconds)
Pathologist II	−3.8 minutes (−3.7 minutes to 3.1 minutes)
Pathologist III	−1 second (−8 seconds to 6 seconds)
Pathologist IV	6 seconds (3 seconds to 8 seconds)

a Estimates calculated as marginal predicted mean effect of technology on reporting time.
b Estimates calculated as marginal predicted mean difference in reporting time between technology.

reference line indicate pathologists were faster reporting samples on DP compared with LM. Estimates were derived using Model 2 in *Appendix 1*.

There was a clear 'learning effect' across the whole sample as shown in *Table 22* and *Figure 7*. Pathologists were about 30 seconds slower on DP compared with LM in reporting samples at the start of the study (i.e. observations read within the 1st reported date quantile). However, this difference gradually reduced over time and at the end of the study (i.e. 10th date quantile); they were significantly faster reporting samples on DP as shown in *Figure 7*.

We explored whether the learning effect differed by levels of case complexity. For routine samples, pathologists were 11 seconds slower reporting samples on DP compared with LM, but this difference gradually reduced and towards the end of the study, and they were 21 seconds faster reporting samples on DP compared with LM as shown in *Figure 8*.

The learning effect was more apparent in moderate and difficult samples where pathologists were about 1.5 minutes slower reporting samples on DP compared with LM at the start of the study but were 50 seconds faster reporting samples on DP compared with LM at the end of the study as shown in *Figure 9*.

Is digital pathology more efficient in samples with higher number of slides?

The impact of increasing numbers of slides on the difference in reporting time speed between the modalities varied between specialty groups as shown in *Table 23*. For each unit of slide increase, breast pathologists were 2 seconds slower on DP compared with LM while GI pathologists had similar reporting times on both DP and LM. For each unit increase in the number of slides, pathologists reporting renal and skin samples were 12 seconds faster reporting samples on DP compared with LM.

However, as seen in *Table 23* and *Figure 6*, renal cases had a relatively higher number of slides per sample compared to other specialties. Hence, any effect between technology used for diagnosis and number of slides is likely to be better reflected within renal samples.

Renal pathologists were on aggregate slower reporting samples on DP compared with LM as seen in *Table 20*. However, they became slower reporting samples on LM as the number of slides increased. In contrast, an increase in the number of slides had no impact on renal pathologists reporting time on DP technology as shown in *Figure 10*. Reporting times were statistically similar in renal samples with 11 or more slides.

TABLE 22 Mean estimated time to diagnosis over the 3-year reporting period

	DP (95% CI)	LM (95% CI)	Marginal mean difference (DP–LM) 95% CI
All samples			
Start	6.7 minutes (6.3 minutes to 7.0 minutes)	6.2 minutes (5.8 minutes to 6.5 minutes)	30 seconds (16 seconds to 43 seconds)
Mid-point	5.1 minutes (4.9 minutes to 5.3 minutes)	5.0 minutes (4.8 minutes to 5.2 minutes)	7 seconds (−5 seconds to 19 seconds)
End	4.6 minutes (4.4 minutes to 4.9 minutes)	5.2 minutes (4.9 minutes to 5.4 minutes)	−32 seconds (−43 seconds to −20 seconds)
Routine samples			
Start	3.9 minutes (3.7 minutes to 4.2 minutes)	4.1 minutes (3.9 minutes to 4.4 minutes)	11 seconds (−1 second to 23 seconds)
Mid-point	3.0 minutes (2.9 minutes to 3.2 minutes)	3.0 minutes (2.9 minutes to 3.1 minutes)	3 seconds (−6 seconds to 11 seconds)
End	2.9 minutes (2.7 minutes to 3.0 minutes)	3.2 minutes (3.1 minutes to 3.4 minutes)	−21 seconds (−29 seconds to −13 seconds)
Moderate and difficult samples			
Start	11.3 minutes (10.5 minutes to 12.1 minutes)	9.8 minutes (9.1 minutes to 10.5 minutes)	87 seconds (54 seconds to 120 seconds)
Mid-point	9.5 minutes (9.0 minutes to 10.1 minutes)	9.5 minutes (9.0 minutes to 10.1 minutes)	0 seconds (−38 seconds to 38 seconds)
End	8.8 minutes (8.2 minutes to 9.4 minutes)	9.6 minutes (9.1 minutes to 10.2 minutes)	−50 seconds (−83 seconds to −17 seconds)

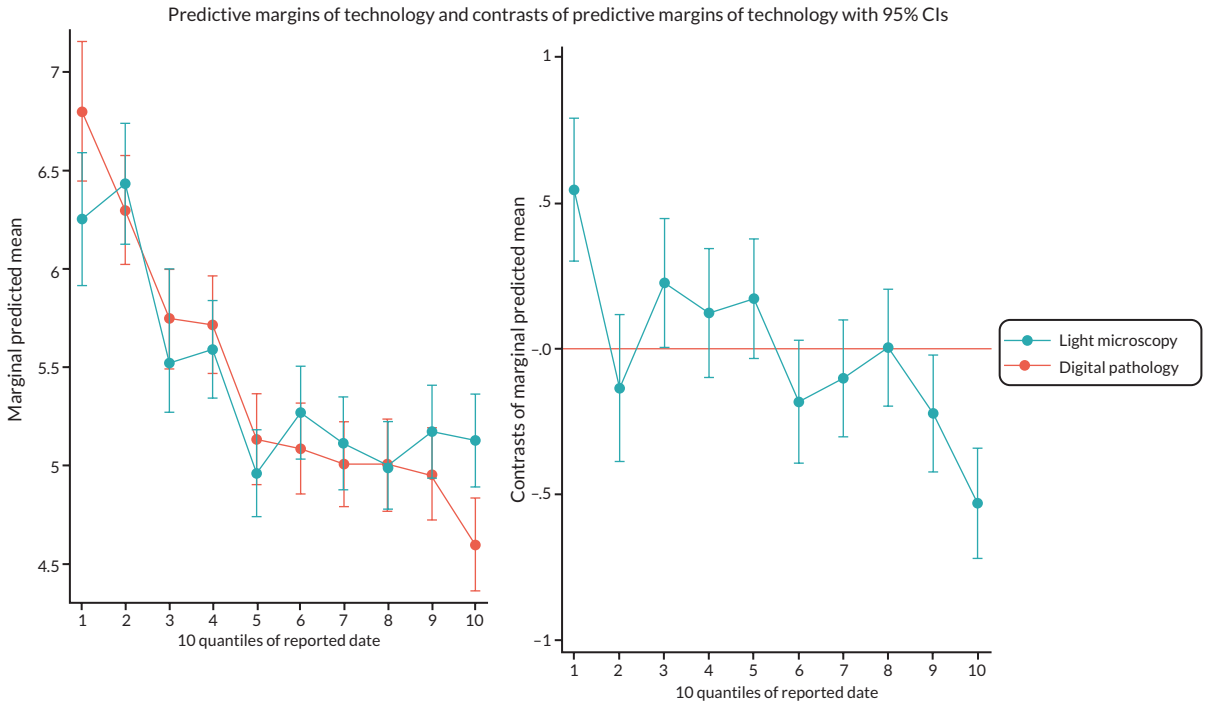

FIGURE 7 Marginal predicted mean reporting time per technology used and contrast of marginal predicted mean for all observations.

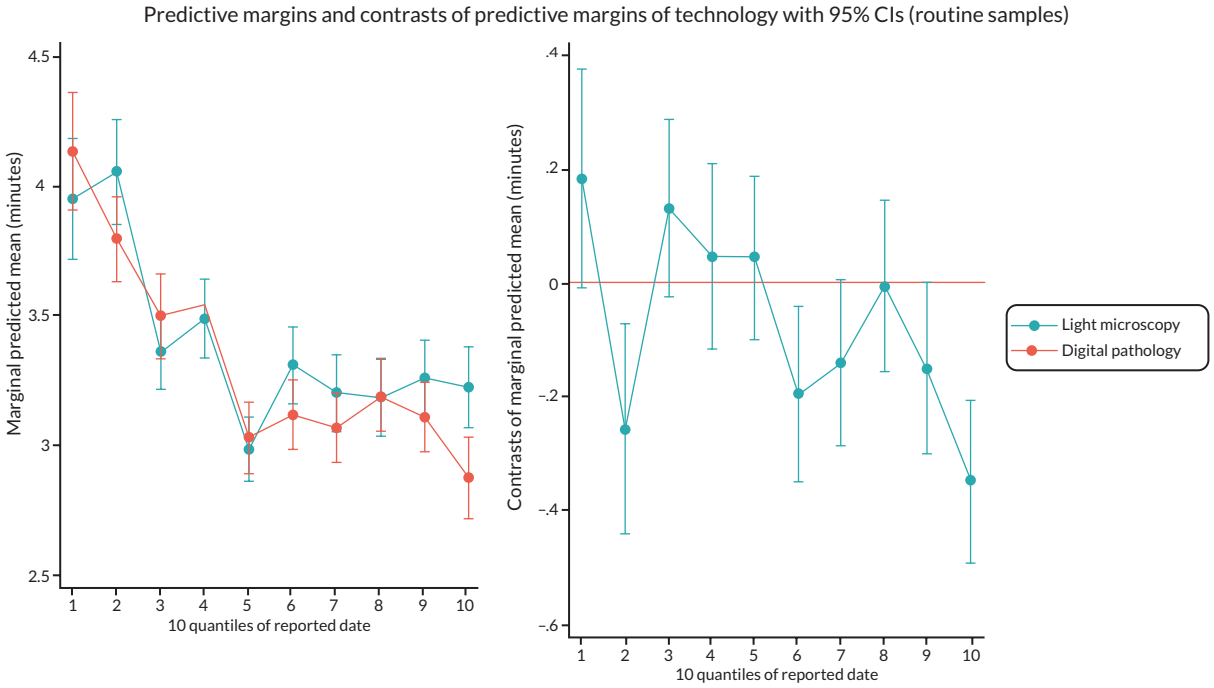

FIGURE 8 Marginal predicted mean reporting time per technology used and contrast of marginal predicted mean for routine samples.

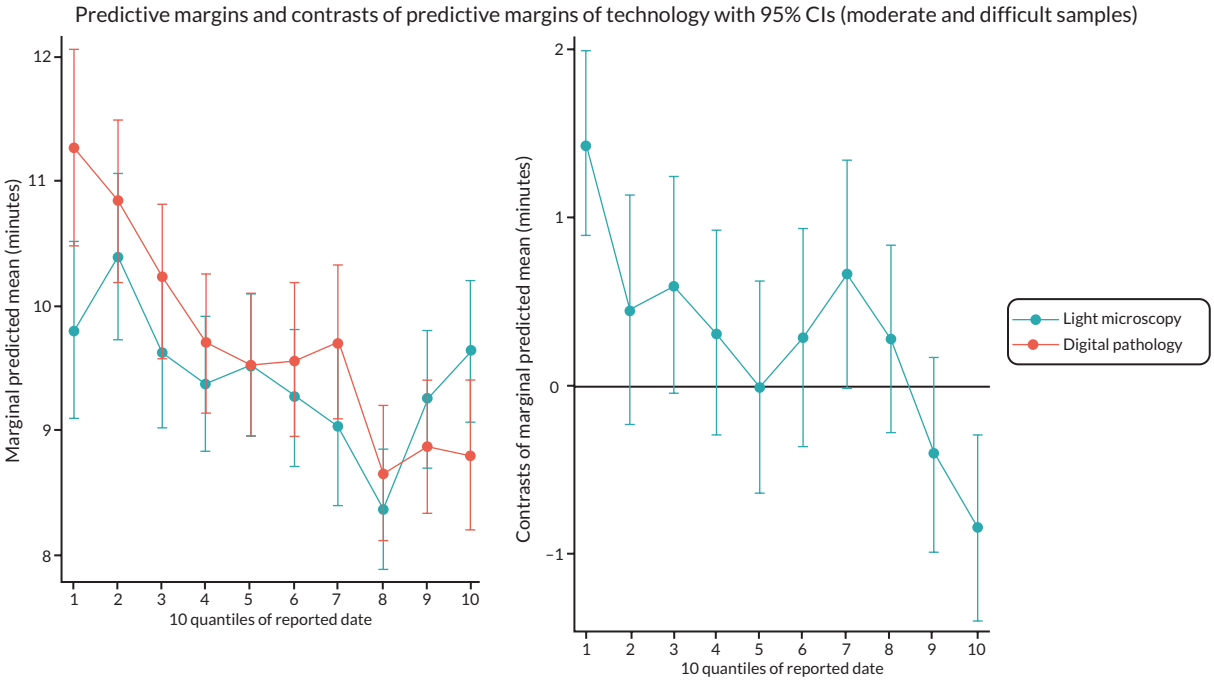

FIGURE 9 Marginal predicted mean reporting time per technology used and contrast of marginal predicted mean for moderate and difficult samples.

Is digital pathology efficiency affected by the level of sample complexity?

Overall, there were little or no differences in reporting time on DP and LM for routine and moderate samples. However, pathologists were about 34 seconds slower (95% CI 22 seconds to 46 seconds) reporting difficult samples on DP compared with LM as seen in *Table 24*.

TABLE 23 Estimates of interaction between pathologist and number of slides with 95% CI

	Estimates (95% CI)
All observations	3 seconds (1 second to 4 seconds)
Breast	2 seconds (0 second to 4 seconds)
GI	0 second (−2 seconds to 2 seconds)
Renal	−12 seconds (−23 seconds to −1 second)
Skin	−12 seconds (−15 seconds to −8 seconds)

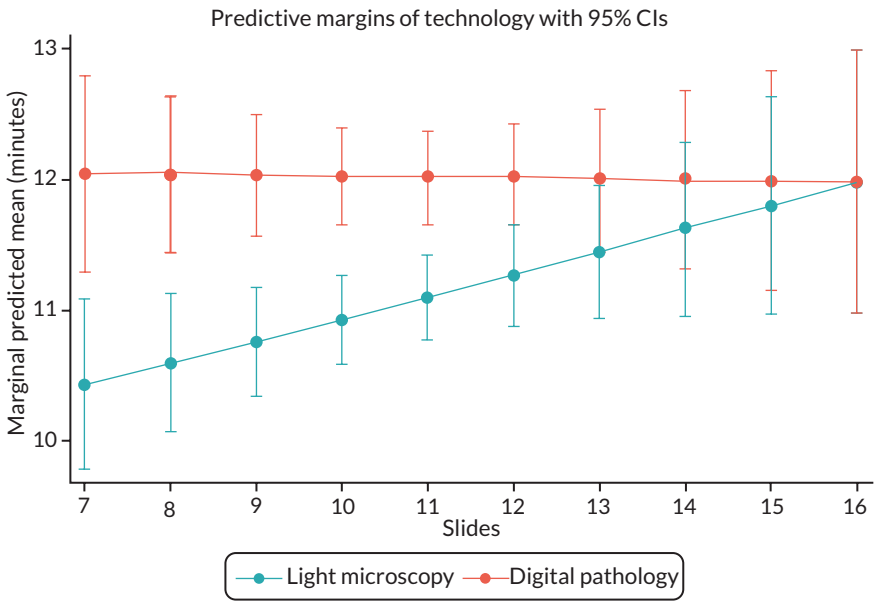

FIGURE 10 Interaction between technology and number of slides.

TABLE 24 Differences in reporting time between technology by difficulty level

Difficulty level	Mean difference (95% CI)
Routine	−4 seconds (−7 seconds to −2 seconds)
Moderate	−6 seconds (−29 seconds to 16 seconds)
Difficult	34 seconds (22 seconds to 46 seconds)

Discussion

This report details the statistical analysis of histopathology reporting time at levels of specialty and pathologist. The results show as pathologists became more experienced using DP within the study, they became faster reporting samples on DP compared with LM. While overall there was no clear benefit for either technology within each specialty, this may be due to any subtle changes being masked by changes inherent in familiarity with the technique. In specialties where pathologists were on average, faster using either DP or LM, some had no statistically significant difference in reporting time between each technology. Indeed, the 'learning effect' associated with continued use of DP was evident across all levels of difficulty but were more apparent in difficult cases. Furthermore, using DP to report cases with small number of slides may not lead to reduced reporting time. However, there are potential time savings in samples with a higher number of slides, when pathologists become fully familiar with the DP system.

While DP may be a more efficient way to report tissue samples, the time savings in histopathology reporting time are very small. This is not surprising since the central part of the task, that is interpretation of morphological features presented in a case, remains precisely the same for both modalities. However, the benefits of implementing DP go beyond potential efficiency gains in histopathology reporting and include non-reporting processes in the laboratory such as handling, filing, locating misplaced slides and storage of slides, all of which are beyond the scope of this study. Furthermore, substantial upfront costs following digitisation may be offset by potential reduction in labour requirements relating to non-reporting tasks. It is important that an economic evaluation of the benefits of implementing DP accounts for all processes in the laboratory that can be affected by digitisation.

Future workstream in this substudy will implement discrete event simulation model to explore the impact of implementing DP on slide throughput in the laboratory. The potential costs and disbenefits of clinically significant differences in diagnosis due to technology will be investigated.

Chapter 6 Eye-tracking substudy

Some text in this chapter has been reproduced with permission from Sudin *et al*.[50]

Introduction

Digitising the pathology workflow has the opportunity to transform clinical pathology by facilitating rapid slide sharing, remote reporting and second opinion acquisition, as well as its potential for pathology education and training and its compatibility with AI systems to assist with detection and diagnosis.[8,13,51] Further to this, replacing the microscope with a computer workstation set-up easily facilitates the passive tracking and capture of valuable information during the diagnostic process, including data on slide zooming and panning, mouse–keyboard usage, and eye movement tracking. These data sources can be analysed to objectively investigate search and interpretation strategies, which could further our understanding of the nature of medical errors during pathology slide reporting.[50,52–54] Additionally, utilising tracking technology in pathology could be invaluable for the development and evaluation of AI support systems as well as a useful tool for benchmarking and assessment for pathology training and quality assurance programmes.[55,56]

Eye tracking has been employed frequently in medicine to understand medical image interpretation tasks – with most of these studies conducted in the field of radiology, a discipline that has been fully digital for over two decades.[57–59] From this body of work, medical errors during radiology image interpretation have been more comprehensively understood and categorised, and interpretation behaviours associated with greater experience and diagnostic performance have been identified.[59,60] Using this understanding in radiology, research groups have been investigating the development of specialist training and teaching tools harnessing eye tracking and AI to objectively assess interpretation performance.[56,61] Furthermore, there is growing interest in using eye tracking for automatic image segmentation and labelling to build training data sets for AI detection and diagnosis algorithms.[55]

Due to the relative novelty of DP, there are not many eye-tracking studies investigating image interpretation behaviour in this domain. Furthermore, due to the great size and complexity of the digital slide images and the essential image panning and zooming manipulation during the diagnostic reporting process, eye-tracking studies are technically very challenging. Early DP eye-tracking studies by Krupinski *et al.* demonstrated how less experienced pathologists exhibited increased time on task and made more visual fixations than more experienced pathologists.[52,62] However, pathologists in these studies viewed digital slides at fixed magnification levels that could not be zoomed or panned. Other studies allowed pathologists to navigate slides freely, but instead of tracking eye movement, tracked only the viewport as a proxy of slide visual coverage and to begin to investigate image navigation techniques including zoom and panning characteristics.[63,64] Only a limited number of studies have combined eye tracking and viewport tracking elements to comprehensively assess pathologist search and reading behaviour, but it is important to note that these studies employed custom-built digital slide viewers rather than commercially available slide viewers that are used for routine clinical reporting.[53,54]

In the eye-tracking substudy of this National Institute for Health and Care Research (NIHR) Digital Pathology Trial, we aimed to build on this limited knowledge base to further investigate the image search and reading behaviour in DP by analysing eye movements and digital slide image manipulation during the diagnostic reporting process, using a commercial, FDA-approved digital slide viewer. To explore this, we conducted a number of eye-tracking studies with different pathology cases to learn of the different reporting behaviours and to refine the DP eye-tracking protocol; including: (1) breast surgical resection specimens (pilot study),[65] (2) breast core needle biopsy specimens[50,66] and (3) GI specimens. Additionally, in order to conduct more thorough analysis with the eye-tracking data, we have developed a specialised software platform which utilises computer vision techniques to extract panning, zooming and tool use activity from recorded screen captures from the eye-tracking sessions.[67]

Methods

Breast surgical resection pilot study

This pilot included three pathologists from Nottingham University Hospitals NHS Trust. Participants reported on 10 surgically resected breast specimens chosen from the national histopathology database for UK breast pathology EQA scheme. GT diagnoses were agreed by an expert panel and only highly concordant cases were selected. All cases were invasive cancers and chosen by an experienced pathologist.

Participating pathologists were instructed to subdivide their reporting into five separate sections:

1. Overall diagnosis.
2. Epithelial proliferation or presence of in situ disease.
3. Lymphovascular invasion.
4. Measurements of lesion/invasive component.
5. Grading.

The above format was adapted from that used in the EQA schemes. Cases were reported verbally by each pathologist, with the answers recorded on a form by a research assistant.

Breast core needle biopsy study

Fourteen pathologists from Nottingham University Hospitals NHS Trust, University Hospitals of Derby and Burton NHS Trust and United Lincolnshire Hospitals NHS Trust were recruited. Participants reported 20 core needle biopsy breast specimens – one H&E slide per case. These specimens were obtained from the main Digital Pathology Trial. Samples were selected by an expert, to include a range of diagnoses and difficulties. The GT diagnoses were agreed by consensus after the slides had been read by the pathologists in the main study.

The participants were asked to report their findings on a standardised reporting form which was modified from the Royal College of Pathologist Guidelines for reporting non-operative diagnostic breast specimens and UK breast pathology EQA scheme core needle biopsy proforma.[68] The reporting form was divided into six sections to include the following domains:

Overall diagnosis:

1. B1 normal.
2. B2 benign.
3. B3 lesion of uncertain malignant potential – including epithelial proliferation with or without atypia.
4. B4 suspicious.
5. B5a malignant in situ – including nuclear grade and microinvasion.
6. B5b malignant invasive – including type and grade.

Gastrointestinal study

Nine pathologists from Nottingham University Hospitals NHS Trust and University Hospitals of Derby and Burton NHS Trust were recruited. Participants reported 20 single-slide H&E GI specimens. These specimens were obtained from the main Digital Pathology Trial. Samples were selected by an expert, to include a range of diagnoses and difficulties. The GT diagnoses were agreed by consensus after the slides had been read by the pathologists in the main study.

The participants were asked to report their findings on a standardised reporting form which was designed specifically for this study with the assistance of an expert GI pathologist. The form included information about the case, including specimen type (i.e. where the biopsy was taken from the GI tract) and included clinical details. The reporting form was divided into five sections to include the following domains:

Overall diagnosis:

1. Normal.
2. Benign.
3. Neoplastic malignant.
4. Inflammation.
5. Other.

Workstation and eye tracker set-up

A dedicated workstation was set up in each pathologist's normal reading environment with controlled ambient light to replicate ordinary reporting conditions. The digital slides were viewed on a 27-inch single-screen display with 2560 × 1440 pixels (Dell P2720D, Dell Computer Corporation, TX, USA), using the Philips IntelliSite DP slide viewer (Philips, Amsterdam, the Netherlands). This system allows the display of each image at a standard resolution and allows panning and zooming of up to 40× optical magnification and 100× digital magnification. The pathologists were provided with a standard mouse and keyboard for slide navigation and reporting.

Three remote, non-intrusive eye-tracking cameras (SmartEyePro, SmartEye AB, Gothenberg, Sweden) were mounted to the pathology monitor and passively recorded gaze and eye behavioural data at a sampling rate of 60 Hz as participants read the case set (*Figure 11*). Prior to each eye-tracking session, the eye trackers were adjusted to suit each participant's natural seated head position and calibrated using a 9-point gaze calibration to ensure accurate gaze tracking across the screen. A second PC was set up for eye tracker calibration and monitoring during the experiment to ensure that participants remained within the eye tracker's field of view for optimal eye tracking. A separate PC running the eye tracker was chosen such that the DP PC could dedicate its computing power on displaying the digital slides, and not competing with other demanding tasks, ensuring that the performance of the slide viewer was not compromised. During the eye-tracking experiment, the DP screen display was recorded at native resolution at a frame rate of 25 Hz. Participants' eye-tracking data, workstation screen capture, scene camera recording and mouse and keyboard data were all automatically recorded, compiled, synced in real time and timestamped (eyesDX software suite, eyesDx, IA, USA). The two PCs were time synced using network time protocol time server software.

FIGURE 11 Shows the DP eye-tracking set-up. The participating pathologist reviewed digitised pathology slides on the left monitor with eye-tracking cameras attached to non-intrusively record eye movements during the reporting process. The right monitor displayed a real-time eye-tracking feed for the research assistant to ensure that the quality of the eye movement data was maintained throughout the experiment.

Procedure

Some of the participants had no prior DP experience, and were provided with online training sessions, information and video tutorials on how to use the viewing software prior to data collection. Following consent, participants completed a brief demographic and experience survey. Participants were then given access to two practice cases to familiarise themselves with the set-up and reporting software before starting the test cases. Practice cases were not recorded, and this provided an opportunity to correctly adjust the remote eye-tracking cameras to ensure that the participants' natural reading position was within the eye tracker's field of view for optimal eye and head detection for accurate gaze data.

Slide navigation tracking software

During the reporting process, pathologists manipulate the slide images by panning and zooming, as in LM, to enable the assessment of tissue in greater detail to formulate a diagnosis. Due to this essential slide zooming and panning, the slide is dynamic for much of the reporting process, and hence the portion of the slide visible on-screen (termed the viewport) changes very frequently depending on the magnification level selected and the spatial location on the slide. For example, when zooming to a higher magnification, the viewport displays a smaller section of the slide image with greater detail, but spatial context is reduced (*Figure 12*).

Since eye trackers record eye movement (gaze) location in terms of screen co-ordinates, it is essential to also track the pathologists' slide navigation to determine exactly where on the slide the pathologist has viewed and what features (and at what magnification) the pathologist has fixated (*Figure 13*).

To capture the slide movement data, screen captures of the DP clinical workstation recorded during the eye-tracking sessions were analysed automatically using the developed software platform. Data were extracted from the screen capture video frame-by-frame, using computer vision techniques to identify and track landmark features on the user interface of the slide viewing software. This allowed the automatic identification of case name, zoom level and two-dimensional slide co-ordinates, relating to the centre of the viewport, for each timestamped frame of the recording. These data can then be combined to allow for the isolation of the viewable slide area from the raw DP slide image at each timestamp during the eye-tracking session. Then, the gaze data can be combined with the segmented slide image at each timestamp to reconstruct the overall gaze path/heatmap, including locating fixation points, and eye movements while the slide is in motion.

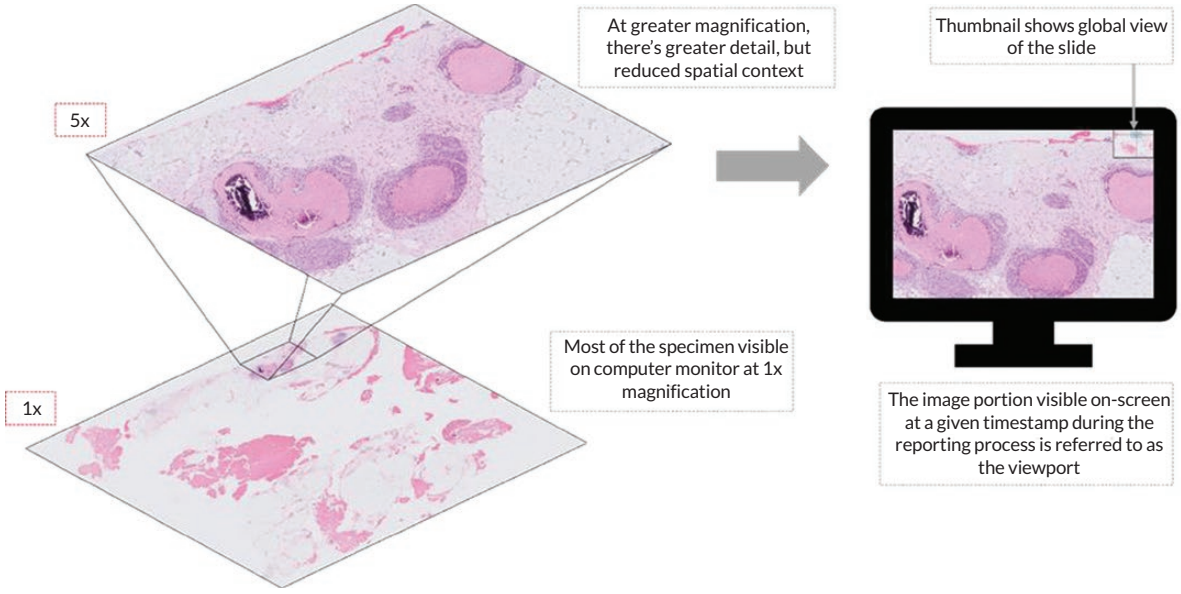

FIGURE 12 Shows a visualisation of how the on-screen content displayed to the pathologist can change dramatically due to panning and zooming behaviour during the reporting process.

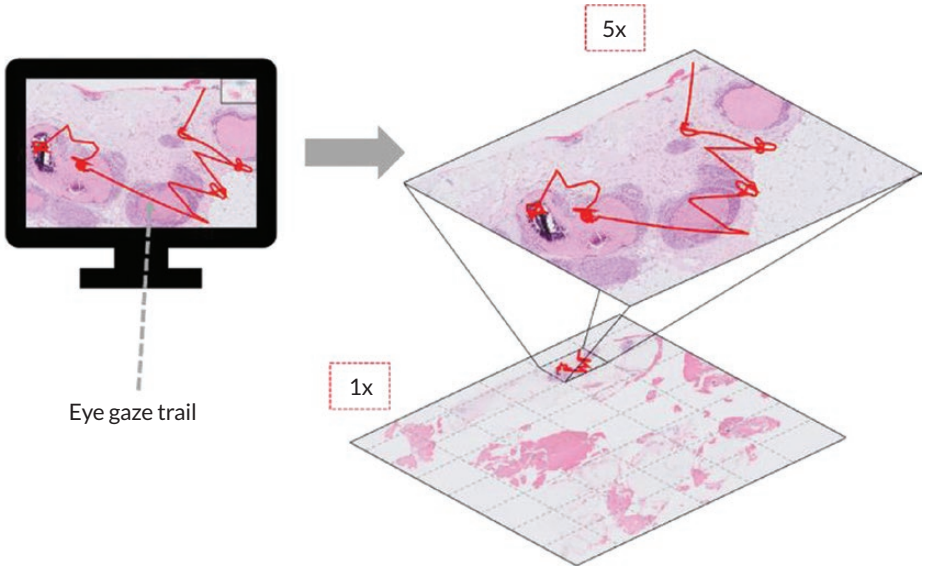

FIGURE 13 Schematic to show how gaze data captured in terms of screen co-ordinates need to be translated to the global slide co-ordinates for high-level assessment and analysis of digital slide image coverage and region of interest analysis.

Statistical analysis

In the current report, only data from the breast surgical resection pilot study and the first round of data from the breast core needle biopsy study (n = 5) are presented. Analyses including the additional data collected in the breast core needle biopsy study and the data collected in the GI study are currently underway at the time of submitting this report, and these results will be disseminated following report submission (see eye-tracking study dissemination plans for further details).

For both studies, the case reading time, fixations, magnification and diagnostic accuracy were assessed for each participant. For the breast surgical resection pilot, the most experienced pathologist (participant 1) was compared to the most inexperienced (participant 2). For the breast core needle biopsy study, repeated measures-analysis of variance was used to test for statistical significance, with post hoc pairwise t-tests using the Benjamini–Hochberg correction method. Additionally, in the breast core needle biopsy study, the relationship between experience and case reading time, fixation and magnification was investigated using mixed-effects linear regressions, with inverse transformations, wherever appropriate. Case ID was taken as a RE variable to account for within-slide correlations. The relationship between experience and diagnostic accuracy was investigated using mixed-effects logistic regression, again taking case ID as a RE variable. The threshold for statistical significance was set at α = 0.05.

Results

Breast surgical resection pilot study

Reading time

The three participants spent an average of 4 minutes reading each case. Pathologists with more experience had shorter reporting times. Participant 1, a senior breast pathology consultant, spent an average of 3.77 minutes per case, while participant 2, a trainee, spent 5.03 minutes (p = 0.002).

Fixations

The eye position data were analysed by examining the total number of fixations per case. The senior pathologist, participant 1, had an average of 55.30 fixations per case compared to the trainee pathologist, participant 2, who had an average of 178.90 fixations (p < 0.001).

Magnification

To examine differences in reading technique/strategy, the recorded data were analysed to compare the magnification technique employed by the senior and trainee pathologists (*Figure 14*). The magnification pattern for each case was grouped in 10-second increments up to 60 seconds. It was found that the senior pathologist utilised a higher magnification in the first 20 seconds of reporting, while the trainee pathologist used a higher magnification in the latter 40 seconds. This may signify that the trainee pathologist took a longer time to decide which areas to investigate further using the zoom function.

Diagnostic accuracy

Interestingly, although there was a wide range in the amount of time taken to read the slides, the diagnostic accuracy for all three participants is similar (*Table 25*).

Breast core needle biopsy study

The five participants included in the breast core needle biopsy study were ordered 1–5 in terms of increasing lengths of experience.

Reading time

A significant difference was identified between participants in terms of case durations ($p < 0.001$; *Figure 15*). Participant 1, who had the least experience in pathology, took on average 169.75 seconds, whereas participant 5 who had the most experience took on average 101.22 seconds. Post hoc pairwise t-tests showed that the less experienced participants 1 and 2, with an average of 2 years of experience, exhibited significantly longer reading times compared to the more experienced participants 3, 4 and 5 (mean years of experience: 25.67 years) ($p \leq 0.01$). There was a significant trend towards quicker reading times with increasing experience ($p < 0.001$). The changes were greatest for the first 10 years, which levelled off thereafter (*Figure 16*).

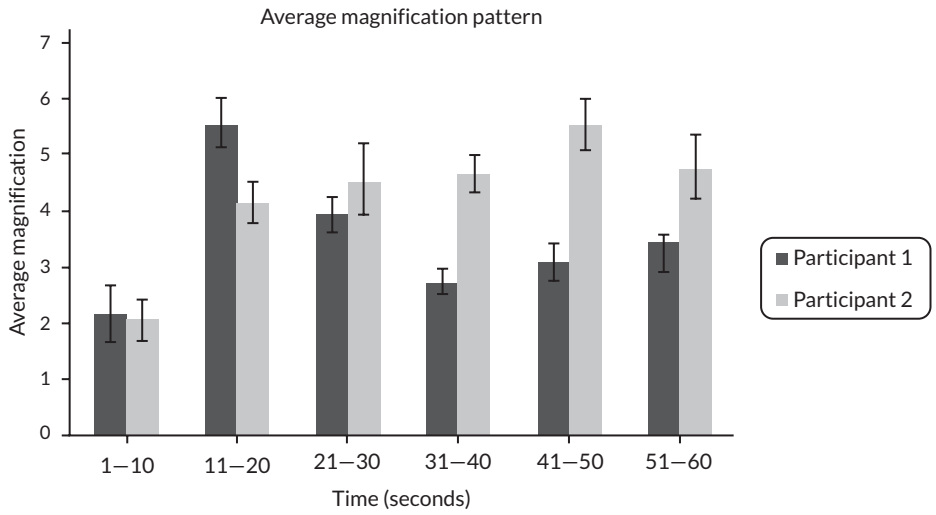

FIGURE 14 Average magnification pattern for participant 1 (senior pathologist) and participant 2 (trainee pathologist) over 10-second increments.

TABLE 25 Total number of incorrect diagnoses and grading for all participants across the 10 cases

	Number of incorrect cases		
	Participant 1	**Participant 2**	**Participant 3**
Diagnosis	1	1	2
Grading	3	3	1

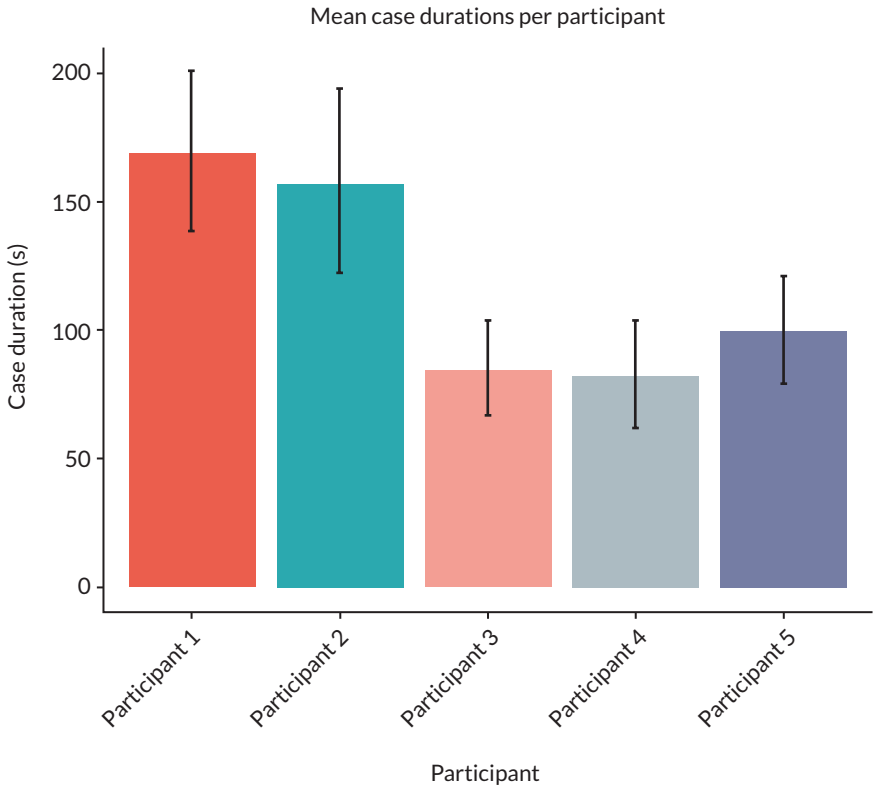

FIGURE 15 Comparison of mean case durations in seconds per participant. The black bars denote the 95% CIs of the means. Participants have been ordered in increasing lengths of experience. Figure reproduced with permission from Sudin *et al.*[50]

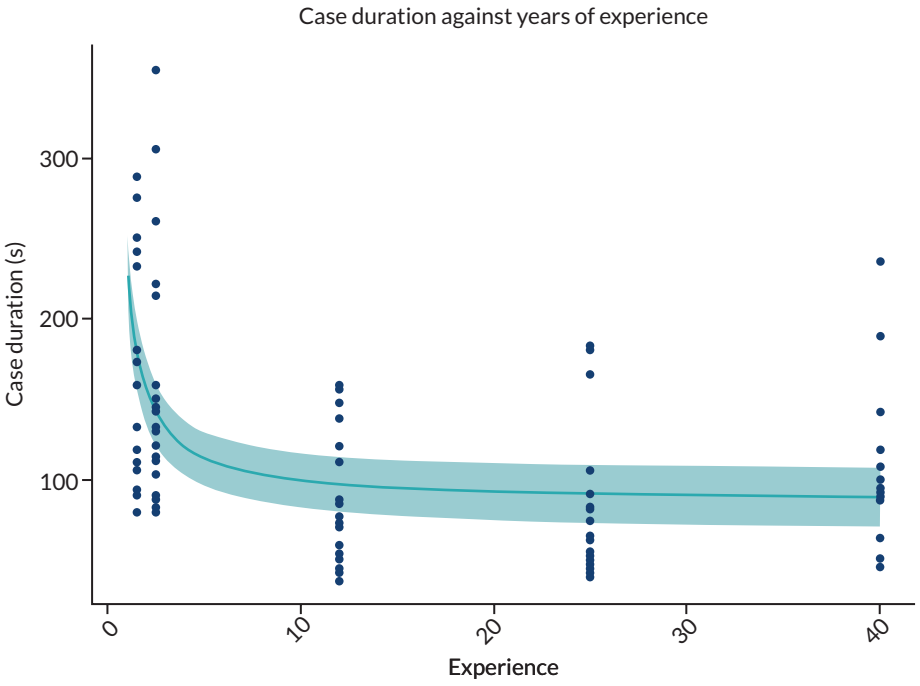

FIGURE 16 A graph of case duration against years of experience. The line denotes the regression line. The shaded area denotes the 95% CI of the regression line. Figure reproduced with permission from Sudin *et al.*[50]

Fixations

The least experienced participant (participant 1) had the greatest average number of fixations per case (mean: 65.5), and the most experienced participant (participant 5) had the least (mean: 33.1; *Figure 17*). Increased experience was significantly associated with a reduced number of fixations per case ($p < 0.001$; *Figure 18*).

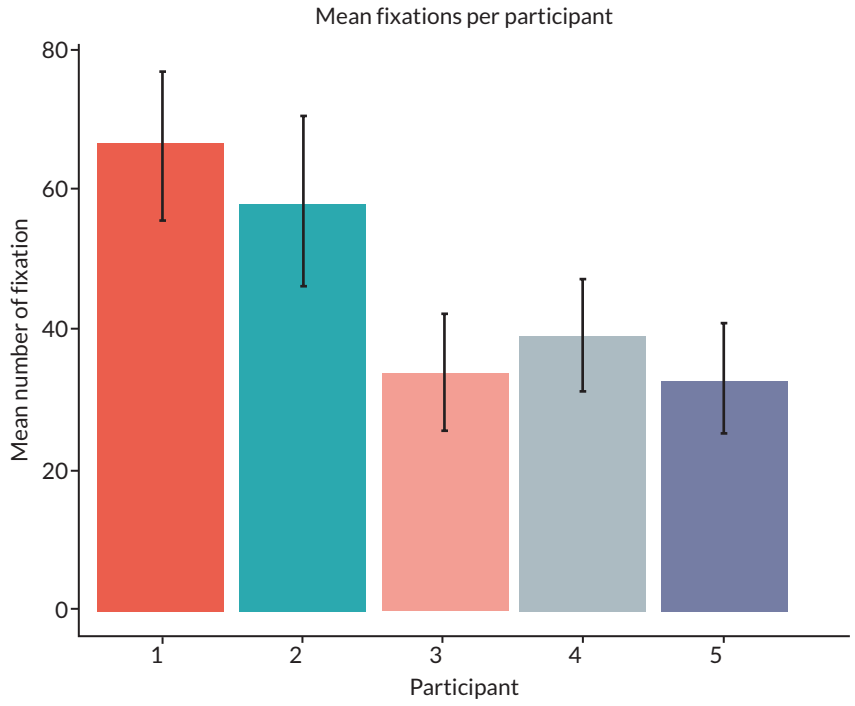

FIGURE 17 Comparison of the mean number of fixations per participant. The black bars denote the 95% CIs of the means. Figure reproduced with permission from Sudin *et al.*[50]

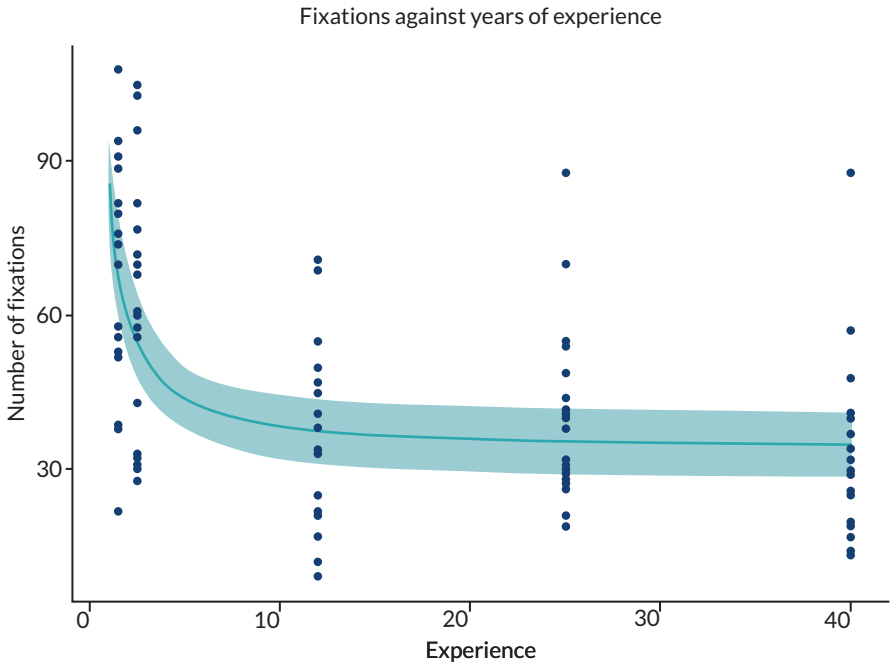

FIGURE 18 A graph of fixations against years of experience. The line denotes the regression line. The band denotes the 95% CI. Figure reproduced with permission from Sudin *et al.*[50]

Magnification

All pathologists exhibited a preference for 5× magnification for the majority of the examination session. The more experienced pathologists utilised the lower magnifications more than the less experienced pathologists, and conversely, the less experienced pathologists utilised the higher magnifications more than experienced pathologists (*Figure 19*). Greater experience was significantly associated with greater relative use of the lower magnifications: 0.1 × (p = 0.04) and 5 × (p = 0.008). In contrast, greater experience was significantly associated with lesser usage of higher magnifications: 5 × (p < 0.001) and 10 × (p = 0.002).

Additionally, we investigated how experience was related to 'zooming' and 'panning' behaviour, where 'zooming' was measured as the number of changes of magnification in a minute, and 'panning' was defined as the total proportion of time spent on any single magnification for longer than 2 seconds. Experience was associated with a near-significant decrease in the relative time spent panning in each case (p = 0.070; *Figure 20*). A significant association between experience and increased zooming behaviour was identified (p < 0.001; *Figure 21*).

Diagnostic accuracy

There was a significant correlation between greater experience and overall diagnostic accuracy of each participant (p = 0.033; *Figure 22*).

Discussion

From the studies and analyses conducted to this point as part of this NIHR Digital Pathology Trial, we have found significant differences in DP interpretation between pathologists of different experience levels. Pathologists with greater experience required less time to reach a diagnosis and exhibited fewer visual fixations when reporting compared to pathologists with less experience. Furthermore, greater experience was associated with differences in reading behaviour, specifically magnification use and slide exploration strategy, including 'zooming' and 'panning' behaviours.

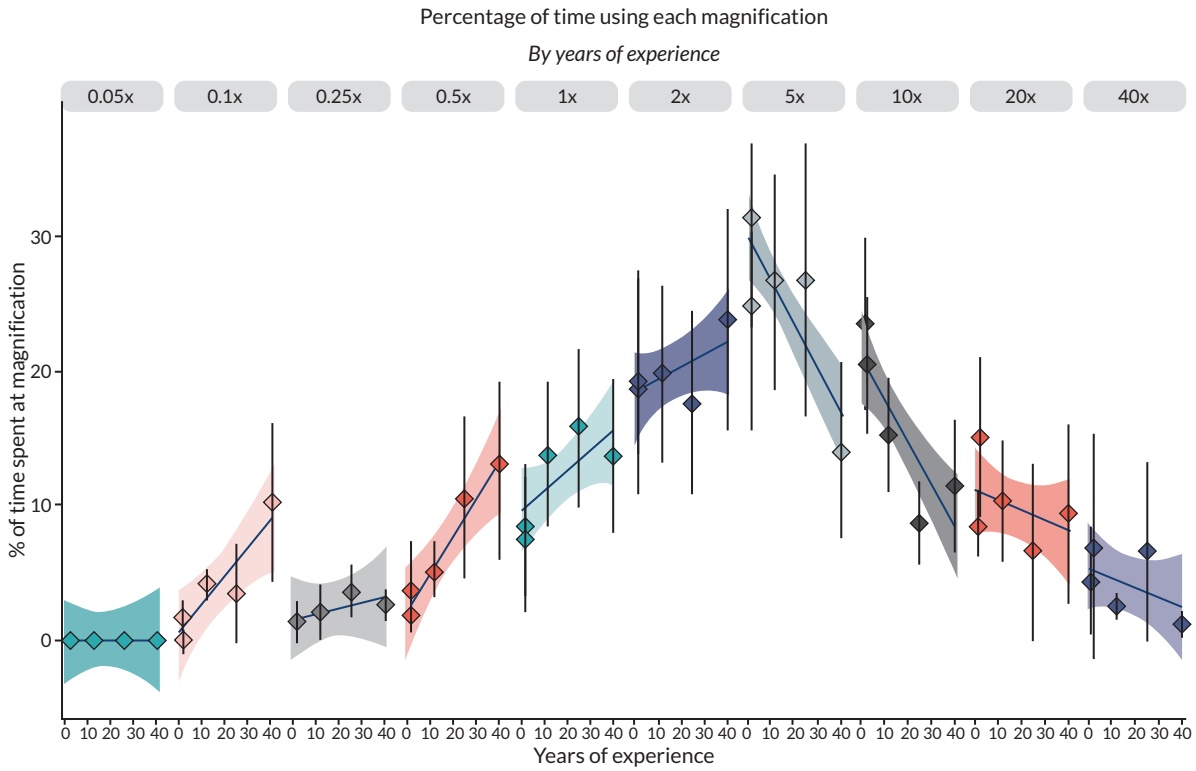

FIGURE 19 Graph showing the percentage of time spent at each magnification against years of experience. The points denote the mean, and the error bars denote a 95% CI. The blue line shows the regression line. The shadowing represents the SE of the regression line. Figure reproduced with permission from Sudin *et al.*[50]

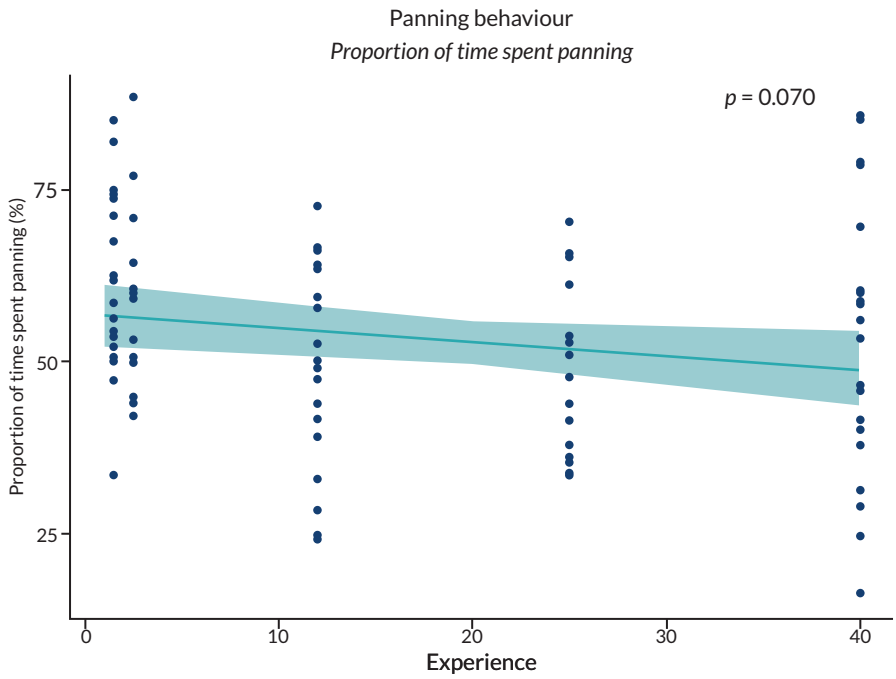

FIGURE 20 Relative time spent panning by years of experience. The line denotes the regression line, and the shaded band denotes the 95% CI of the regression line. Figure reproduced with permission from Sudin *et al.*[50]

Together, these findings suggest more efficient visual search patterns, better pattern recognition and greater visual processing capabilities of more experienced pathologists.

Magnification use

We identified that 5× zoom was the most frequently used magnification value for the interpretation. The time spent on other magnifications was approximately normally distributed around that value. We found that more experienced pathologists had tendencies towards lower magnifications, whereas less experienced pathologists exhibited tendencies towards higher magnification values. This is consistent with Mercan *et al.* and Brunyé *et al.*,[53] who observed an association between higher magnifications and errors arising from overinterpretations.[53,69] Our findings lend further support to the current and widely taught practice of utilising low power assessment of the slide to decide which areas of the slide are most important and merit high power assessment.

Additionally, we investigated the magnification use of our participants and quantified the amount of 'zooming' and 'panning'. A study by Mercan *et al.* found a significant association between 'panning' (called 'scanning' by the authors) and slower interpretation times compared to 'zooming' (called 'drilling' by the authors) in pathologists.[69] A similar effect was described in radiology by Drew *et al.*, who found that drilling strategies were associated with better performance on a variety of metrics.[60] Drew *et al.* postulated in their studies of radiologists that the better performance of drillers may be an effect of more experienced radiologists learning to use drilling strategies. This study provides evidence in support of this hypothesis. The significant association between experience and 'zooming' behaviours, alongside our confirmation that greater experience results in better diagnostic accuracy and reading speed, suggests this explanation as to the most likely mechanism of the observed effect.

Visual processing

Our studies allow for an indirect assessment of visual processing cognitive capabilities and are suggestive of improved visual processing with greater experience. The crudest of these are the findings of significant improvements in reading speed with greater experience, with greater accuracy. Moreover, we found that it took significantly fewer visual fixations to reach a diagnosis for experienced pathologists compared to inexperienced pathologists. In addition, greater zooming frequency among more experienced pathologists was indicative of greater visual processing speeds and better pattern recognition. It may be the case that experience develops the cognitive capabilities that are prerequisites for

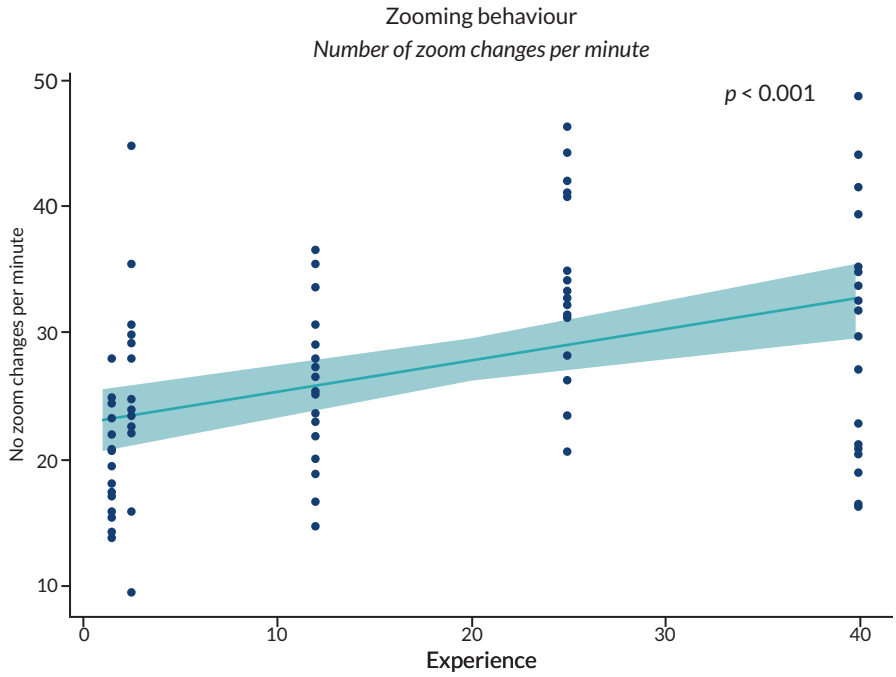

FIGURE 21 Number of zoom changes per minute by years of experience. The line denotes the regression line, and the shaded band denotes the 95% CI of the regression line. Figure reproduced with permission from Sudin *et al.*[50]

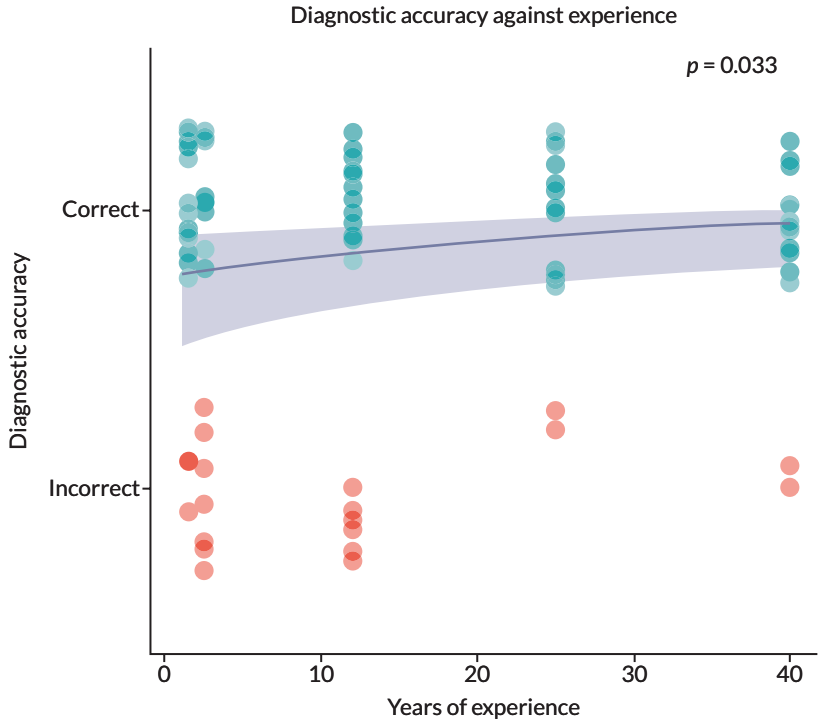

FIGURE 22 Graph showing the relationship between diagnostic accuracy and years of experience. Green dots represent correct diagnosis, and red dots represent incorrect diagnosis. The line denotes the logistic regression line. The shaded band denotes the 95% CI. The points are randomly scattered around the *y*-axis for the clarity of individual points. Figure reproduced with permission from Sudin *et al.*[50]

zooming behaviours. This could be an explanation for the tendency to dwell longer within a higher magnification value for less experienced pathologists who are yet to possess these cognitive capabilities.

Conclusions

Experienced pathologists made greater use of lower magnification compared to inexperienced trainees. Zooming behaviour (changing zoom values) was seen to be more prevalent among experienced pathologists, whereas panning behaviour (dwelling on any single magnification) was more prevalent among inexperienced pathologists. We found evidence that greater experience develops visual processing capabilities, which may be a prerequisite for zooming behaviours. Further research into these and other characteristics of expertise development in DP image interpretation, as well as increasing sample size and pooling participants from different centres in future studies, could demonstrate the use of these behaviours as markers of expertise in training programmes. As trainee pathologists progress through their education, eye-tracked test sets of DP images could be scheduled. Breaking down recorded interpretation behaviours alongside diagnostic performance and comparing these breakdowns throughout their training could help evidence their progression and expertise development.

Finally, as mentioned in the methods section, we have conducted further eye-tracking data collection, the results of which have not been detailed in this report since data analysis is currently underway. As part of this analysis, we have developed a software platform to track digital slide zooming and panning activity so that the eye gaze data can be accurately mapped to the digital slide image to enable analysis of image coverage and region of interest analysis and segmentation. The results of this work will be disseminated following report submission (see eye-tracking study dissemination plans for further details).

Chapter 7 Discussion

Interpretation and overall results

The main study has measured the assessment and reporting of 2024 cases by 16 consultant pathologists working across 6 sites in the UK and demonstrated extremely high levels of agreement. Overall figures from the study show CMC of 99.95% for intra-pathologist agreement between LM and DP readings, made on viewings separated by a minimum 6-week gap. The level of agreement shown between the two modalities exceeds the reference point initially set at the start of the study (98.3%), which established from previous studies.[15,22] The two modalities showed identical results when comparing individual pathologists reports with the consensus GT for both LM and DP (99.95 and 99.95, respectively). The overall inter-pathologist agreement is also identical in both LM and DP modalities (0.91 and 0.91, respectively).

These figures are similar to those seen in other studies which have conducted comparisons of similar case number but with differing methodologies and summarised in *Table 26*. However, this is the first study to also measure interobserver agreement on the same cases, thereby demonstrating pathologist interobserver performance is identical with DP and LM both in terms of inter-pathologist variation and comparing either modality to consensus GT. The study shows near identical results between the DP and LM platforms across all the specialty groups, as well as for breast and bowel cancer screening cases.

TABLE 26 Comparison of this study with other multisite validation studies previously published in the literature

Study ID	Tabata *et al.* 2017[1]	Mukhopadhyay *et al.* 2018[2]	Borowsky *et al.* 2020[70]	Babawale 2021[21]	This study
Number of participating sites	12	4	5	7	6
Number of cases	900	1992	2045	3001	2024
Total study readings	2140	15,925	15,031	3001	16,192
Sample enrichment with difficult cases	No	Yes	Yes	Not specified	Yes
Inclusion of cancer screening biopsies	Not specified	Not specified	Not specified	Not specified	Yes
Inclusion of immunofluorescence slides	No	No	No	No	Yes
Sample/slide selection	Up to five parts only. No IHC or special stains included	Selective slides from biopsies and resection	Selective slides from biopsies and resection	Not specified	All slides from biopsies. Selective slides if > 10 blocks resection
Number of DP/LM reading pairs	1070	7964	7423	3001	8100
Washout interval	> 2 weeks	Minimum 4 weeks	Minimum 31 days	No washout period	Minimum 6 weeks
Number of reading pathologists	9	16	19	22	16
Scanning magnification	20× and 40×	40×	20× and 40×	20×	40×
Samples randomised for reading modality	No	Yes	Yes	No	Yes
Reading pathologists' interpreted cases enrolled by different centres	No	Yes	No	No	Yes

TABLE 26 Comparison of this study with other multisite validation studies previously published in the literature (*continued*)

Study ID	Tabata *et al.* 2017[1]	Mukhopadhyay *et al.* 2018[2]	Borowsky *et al.* 2020[70]	Babawale 2021[21]	This study
LM and DP reading by the same pathologist	Yes	Yes	Yes	No	Yes
Adjudication by the reading pathologists	Yes	No	No	Not stated	No
Interobserver concordance measures	No	No	No	Yes	Yes
DP vs. LM clinical concordance	99.2%	95.1%	96.36%	97.1%	99.95%

Histopathology is an interpretive discipline and occasional discordance between reports issued on the same case is to be expected, even when re-reported by the same pathologist. This is more likely in cases known to contain lesions which are be challenging to interpret, where high intra- and interobserver variability has been previously reported with which this study was enriched. Clinically significant differences were observed in these cases and reflected in lower levels of agreement seen in the moderately difficult and difficult cases compared to routine cases *Table 9*. *Table 15* lists the most common diagnoses giving rise to these differences seen in breast, GI and skin groups, specialty areas which have been studied previously. As with the overall results, it is noticeable that the incidence of these differences is nearly identical in reports issued with DP and LM platforms. The previous studies have highlighted areas where DP may present problems for the reporting pathologist. These include recognition of bacteria, identification of amyloid and calcification, and a tendency to 'over-call' dysplasia or atypia.[15,22,71,72]

Examining further to see if there were trends present in these and other areas revealed nearly identical patterns across both DP and LM for the clinically significant differences recorded. For example, failure to recognise *H. pylori* in gastric biopsies was seen six times in LM and seven times in DP, and there were only single instances of *Giardia duodenalis* and *H. cytomegalovirus* respectively being missed, both in DP, whereas gastric amyloidosis was missed by two pathologists in both LM and DP reports. There were no errors recorded in breast due to failure to pick up calcification. Slight differences in breast were seen for B5a (in situ carcinoma) versus B5mi (microinvasive carcinoma) and in GI for low-grade versus high-grade dysplasia in adenomatous polyps. Further examination shows that the seven instances of B5a versus B5mi difference seen in DP showed a nearly equal division in 'over-' and 'under-calls', namely four instances where the invasion was not reported and three instances where the invasion was reported, contrary to the GT diagnosis. In GI, dysplasia grading was the second most common difference seen and occurred in 21 and 28 LM and DP reports, respectively, with both platforms showing greater differences of low-grade dysplasia against the GT of high-grade dysplasia than the reverse, which is the opposite to what would be seen if DP were leading to over-grading of dysplasia, but is an observation which is in keeping with the fact that high-grade dysplasia is much the less common diagnosis in practice. It is also worth pointing out that six of these differences occurred in one case, relating to a small focus of high-grade dysplasia in an adenomatous polyp, which is indicative of challenges around the correct identification of a small focus of diagnostic importance as opposed to a drift in overall grading due to altered perception of nuclear chromatin density, as has been implied in prior studies.[15]

Differences observed in the reports relate to the inherent differences in interpretation which accompany the assessment of histopathology slides, which one would expect to see in a series of cases re-examined for a second time. Careful analysis of these differences has failed to detect any trend or outlying results which would suggest the modality used contributed to the differences observed. Differences in areas which previously have been suggested as potential pitfalls were seen but were not noticeably more frequent in DP compared with LM.

It is important to note that there will be different opinions of what is considered clinically important in respect of patient management, partly due to variation in local practice protocols. The same group of clinicians advised throughout this

study and where similar errors re-occurred the same decision on arbitration was made in order to deliver consistent results regarding this point.

Presented with difficult cases, pathologists naturally used varying terminologies and arbitrators which are inevitably overridden in this didactic study design. Furthermore, pathologists, aware of the problem routinely, refer such cases to peer review from colleagues, which was not an option for the pathologists in this study.

In some cases, for example, recognising *H. pylori* or amyloid deposition, there is little doubt LM is superior to DP. As has been pointed out previously,[15] but the pathologist will know when they have confidently seen a region of interest to be able to make a diagnosis. Thus, the recognition of small objects such as bacteria, mitotic figures, intranuclear inclusions and similar subcellular objects may indeed be obvious on DP where they are present in reasonable numbers, but yet other cases, where either the image is not as good and/or regions are affected by artefacts, may still demand the superior discrimination and the ability to focus through the plane of section and examination with high-power oil immersion objectives to be fully confident the objects in question are, or are not, present in a case. These tasks are better served by LM. The advantages DP offer can still be fully exploited while retaining the undoubted superiority of LM has for these tasks. It seems likely that these superior attributes of LM may well account for the trend towards greater confidence in LM diagnoses than was seen in DP in this study. Therefore, laboratories need to ensure appropriate steps are in place to ensure pathologists working geographically separate from the slides have access to LM when it is needed. Either in the form of transport of slides to pathologist when its needed or review by a colleague with access to the slides would suffice.

This is the first study to demonstrate DP is equivalent to LM in cancer screening cases and renal biopsies The flexibility DP allows in the distribution of the workload is pivotal in both these areas where capacity demand and access to highly specialised services are currently important constraints of service delivery.[5] Given that the reporting of cancer screening cases is based on the same principles regardless of the tumour site, and indeed histopathology samples of any type, there is every reason to believe the results presented here should be translatable to other cancer screening samples such as uterine cervix and lung.

Renal biopsy samples pose significant challenges for DP as they require the fine resolution to recognise the presence of pathological changes within nephron, many of which relate to the presence of subcellular abnormalities that require both special and immunofluorescence stains for demonstration. The study demonstrates DP can deliver this capability satisfactorily. This is a significant development. Firstly, hitherto immunofluorescent studies have been non-permanent and are routinely discarded after a few days as the fluorescent staining is lost. DP provides permanent record of these slides for peer review by colleagues, presentation to clinicians and review in future whenever they should arise. Secondly, ability to report renal biopsies remote from the laboratory where the sample is located immediately offers opportunities to re-design renal pathology services to provide all centres with equitable access high standard specialist pathology services with round-the-clock cover. It is both difficult and expensive for every laboratory to replicate this at their own site, and furthermore prone to error and failure where either difficulty in appointing pathologist or unexpected absence places an unexpected burden on the incumbent pathologist(s). Therefore, DP should become an important tool to provide all laboratories with immediate access to a team of pathologists experienced and able to deliver this service around the clock. Healthcare providers need to plan for and embrace the opportunities that investment in DP offers to re-design and strengthen the service. Furthermore, the excellent agreement demonstrated in the renal biopsy cases gives confidence that DP should be equally successful in other specialty areas with similar requirements such as haematopathology and neuropathology.

This study is one of the largest and most detailed studies comparing DP and LM yet conducted. Previous studies of similar size shown in *Table 26* have been reviewed in a meta-analysis allowing data to be used to calculate sample size and margin of error for the current study. The enrichment with challenging cases, multiple readings for each case, pairwise comparison, independent arbitration and consensus GT diagnosis all strengthen the data provided by this study, delivering 24,288 report comparisons analysed through RE logistic regression. Some, but not all, of these factors have been included in previous studies – either with a limited sample size[73,74,75] or a single reading for each case[21] or powered specialty groups and consensus GT.[2]

In common with these previous studies, our results show excellent correlation between LM and DP. This study therefore provides definitive evidence that pathologists give equivalent results regardless of the modality used to assess the case.

There is a trend towards pathologist's being more confident in working with LM than DP. It is possible this relates to increased resolution and the ability to focus through the plane of section which LM offers and also most probably the familiarity of LM as a modality to report slides with.

The qualitative study shows views on DP are many and varied and influenced largely by the effectiveness with which the modality can be implemented into the existing laboratory workflow. High-quality integration with seamless connectivity between systems and rapid response to the viewing of slides are key to engaging with the laboratory staff and pathologist delivering the service. Conversely, there is often concern about how difficult this might be for staff all too familiar with the existing frailties of complex information technology systems. Introducing DP into the broader NHS laboratory environment may pose challenges, including the need for substantial infrastructure upgrades, staff training and overcoming resistance to change. Nevertheless, with careful planning and support, these barriers can be addressed, paving the way for new ways of working.

It is unlikely DP differs greatly from LM in speed of reporting slides. The data show a clear relationship with improved speed of reporting with DP with the increased experience of the modality over the course of the study. Health economic benefits of wider re-organisation of pathology services which DP permits were beyond the scope of the study, but the study conclusively demonstrated DP can be used effectively by all pathologists studied, who varied greatly in experience and were largely new to the modality. Therefore, transitioning to DP, given that adequate training and support are provided, should not be a barrier to its use in radical re-design of service delivery, if this is beneficial to the delivery of the service. Pathologists in training have used DP for some time to provide a convenient means of seeing rare and unusual cases, which can be of value in supporting experience derived from clinical practice. The wider adoption of DP could potentially expand this potential. However, it is important to reiterate that, in some cases, pathologists will need to revert to LM (examination under polarised light, where superior resolution is needed and to focus through the plane of a section); therefore, pathologists in training will continue to benefit from understanding of, and the ability to exploit, the full potential of LM, even if the majority of their work is done using DP.

Eye-tracking analysis clearly showed efficient examination technique accompanies experience and is associated with accurate reporting. This is the first time this approach has been used in DP and complements prior studies examining information on slide tracking provided in some DP systems.

Generalisability

The different specialty areas studied and the similarity of results across these areas are strong indicators these results should translate into other areas of pathology. DP works equally well on the commonly used staining techniques both with special stains, immunocytochemistry and immunofluorescence stains.

The inclusion of breast and bowel cancer screening samples shows the results for these samples do not differ from the other cases studied. Since the principles of cancer screening remain the same, there is every reason to believe these results will translate equally well to other cancer screening cases in uterine cervix and lung.

Limitations and further research

The study did not examine cytology samples. There is every reason to believe the results would translate to cytoblocks prepared from cytology aspirates, as these samples are analogous to small biopsies in many respects. The use of cytoblocks has increased in recent years, as these preparations are excellent for diagnosis and allow material for both immunocytochemistry and molecular analysis. The scanning of thin prep and conventional cytology smears is more likely to present difficulties with incomplete slide coverage and inability to focus through three-dimensional cell groups

unless scans are performed with a z stack facility. The use of DP for these more conventional cytology preparations remains an area for further research.

Oversized slides from large 'mega' blocks and frozen sections were not examined in the study, but the authors believe these results will translate into these preparations. Although there are important differences in the preparation and staining of these slides, there seems to be no reason why these should present any impediment to reproducing these slides as digital WSI. Indeed, the ability DP provides to allow multiple pathologists geographically separate from the laboratory offers many advantages to reporting of frozen sections. Due to the limitations imposed in a blinded crossover study of this type, it examined pathologists working in isolation from each other and their clinical colleagues, which does not reflect reality and may be the reason for some of the difference detected. However, if so, we observe these differences were evenly distributed across both platforms.

Lessons learnt

Great effort was made in this study to ensure pathologists were adequately trained in the use of DP prior to the study starting. Even with this training its apparent from the speed of reporting that pathologist's interaction with DP changed positively over the course of the study which it is no doubt related to experience using the system. Translating this to transitioning pathologists from LM to DP reinforces the need to ensure training is adequate, that this is essential to providing confidence to pathologists making the change. Naturally, individuals will differ in the rate of progress they make with this transition, but it is important to allow pathologists enough time to make this change and at all times ensure they can use LM when they require it to complete their task.

Overall conclusions

This study provides definitive data that pathologists deliver equivalent reports with DP as they would with LM. This finding is repeated in all areas studied, including cancer screening biopsies, and these findings should translate to other areas of reporting activity including those which require fine resolution. The flexibility that by DP offers opens up new opportunities for improving pathology services, addressing shortages of pathologists as well as offering greatly improved peer review of cases.

Additional information

CRediT contribution statement

David RJ Snead (https://orcid.org/0000-0002-0766-9650): Conceptualisation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – reviewing and editing.

Ayesha S Azam (https://orcid.org/0000-0003-2681-8153): Data curation, Investigation, Methodology, Project administration, Writing – original draft, Writing – reviewing and editing.

Jenny Thirlwall (https://orcid.org/0009-0006-5587-5180): Project administration.

Peter Kimani (https://orcid.org/0000-0001-8200-3173): Formal analysis, Writing – original draft, Writing – reviewing and editing.

Louise Hiller (https://orcid.org/0000-0001-8538-9163): Formal analysis, Supervision, Writing – original draft, Writing – reviewing and editing.

Adam Bickers (https://orcid.org/0009-0005-0153-9643): Investigation.

Clinton Boyd (https://orcid.org/0000-0002-3138-6220): Investigation.

David Boyle (https://orcid.org/0000-0003-3352-5092): Investigation.

David Clark (https://orcid.org/0000-0002-1575-8119): Investigation, Writing – original draft.

Ian Ellis (https://orcid.org/0000-0001-5292-8474): Investigation.

Kishore Gopalakrishnan (https://orcid.org/0000-0003-0459-6967): Investigation.

Mohammad Ilyas (https://orcid.org/0000-0001-7949-7504): Investigation.

Paul Kelly (https://orcid.org/0000-0002-4350-6998): Investigation.

Maurice Loughrey (https://orcid.org/0000-0001-8424-1765): Investigation.

Desley Neil (https://orcid.org/0000-0001-9800-6811): Investigation.

Emad Rakha (https://orcid.org/0000-0002-5009-5525): Investigation.

Ian SD Roberts (https://orcid.org/0000-0002-4885-7957): Investigation, Resources.

Shatrughan Sah (https://orcid.org/0009-0005-8299-8998): Investigation.

Maria Soares (https://orcid.org/0000-0001-6231-8987): Investigation, Resources, Writing – original draft.

YeeWah Tsang (https://orcid.org/0009-0000-3147-6712): Investigation, Resources, Writing – original draft, Writing – reviewing and editing.

Manuel Salto-Tellez (https://orcid.org/0000-0001-8586-282X): Investigation.

Helen Higgins (https://orcid.org/0000-0002-7095-4542): Project administration.

Donna Howe (https://orcid.org/0000-0002-8363-8127): Project administration.

Abigail Takyi (https://orcid.org/0000-0003-2696-6148): Investigation.

Yan Chen (https://orcid.org/0000-0003-3107-7898): Investigation.

Agnieszka Ignatowicz (https://orcid.org/0000-0002-5863-0828): Investigation, Writing – original draft.

Jason Madan (https://orcid.org/0000-0003-4316-1480): Investigation.

Henry Nwankwo (https://orcid.org/0000-0001-7401-1923): Investigation.

George Partridge (https://orcid.org/0000-0002-0832-0725): Investigation.

Janet Dunn (https://orcid.org/0000-0001-7313-4446): Formal analysis, Supervision, Writing – original draft, Writing – reviewing and editing.

Roger A'Hern: Supervision.

Imtiaz Ahmed: Investigation.

Eliot Chadwick: Resources.

Martin Collard: Data curation, Resources.

Sophie Cramp: Project administration.

Rahul Deb: Supervision.

Louise Dolan: Data curation, Resources.

Annemarie Donnelly: Data curation, Resources.

Hesham ElDaly: Investigation.

Emma Elliot: Data curation, Resources.

Helene Euston-Mellor: Data curation, Resources.

Harriet Evans: Investigation.

Khunsha Fatima: Investigation.

Alastair Gale: Investigation.

Sasha Gill: Data curation, Resources.

Catherine Goodway: Data curation, Resources.

Becky Haley: Project administration.

Emily Hero: Investigation.

Joseph Houghton: Investigation.

Rhian Hughes: Investigation.

Daniel Kearns: Resources.

Kirstie Lukhram: Data curation, Resources.

Stuart McIntosh: Investigation.

Carl McLaughlin: Software.

Gordon Moran: Investigation.

Chockalingham Muthiah: Software.

Jacobo Ortiz Fernandez-sordo: Investigation.

Rania Osman: Investigation.

Colin Purdie: Supervision.

Nasir Rajpoot: Funding acquisition.

Sinthuri Raveendran: Investigation.

Ben Storey: Investigation.

John Tippett: PPI.

Sophia Turner: PPI.

Evie Waddell: PPI.

Aileen Withington: PPI.

Patient data statement

This work uses data provided by patients and collected by the NHS as part of their care and support. Using patient data is vital to improve health and care for everyone. There is huge potential to make better use of information from people's patient records, to understand more about disease, develop new treatments, monitor safety and plan NHS services. Patient data should be kept safe and secure, to protect everyone's privacy, and it's important that there are safeguards to make sure that they are stored and used responsibly. Everyone should be able to find out about how patient data are used. #datasaveslives You can find out more about the background to this citation here: https://understandingpatientdata.org.uk/data-citation.

Data-sharing statement

Data will be made available to researchers whose full proposal for their use of the data has been approved by the Sponsor's Data Access Committee. Data will be provided after completion of a data sharing agreement. Contact Telephone: 02476 968582 E-mail: pathlake@uhcw.nhs.uk.

Ethics statement

This study was conducted in accordance with the World Medical Association Declaration of Helsinki and all relevant regulations, and adhered to the highest standards of research governance.

Health Research Authority (HRA) and Research Ethics Committee (REC) approval for the research study was issued on 29 August 2019, and. Before any site started to enrol samples into the study, confirmation of capacity was sought from the site's research and development (R&D) department.

Substantial amendments that required HRA and REC review were only implemented after the HRA and REC granted favourable opinion. For any amendment that affected the site's permission, the R&D department at each site confirmed that permission as ongoing.

Information governance statement

University of Warwick is committed to handling all personal information in line with the UK Data Protection Act (2018) and the General Data Protection Regulation (EU GDPR) 2016/679.

Under Data Protection legislation University of Warwick is the Data Processor; University Hospitals Coventry and Warwickshire is the Data Controller, and we process personal data in accordance with their instructions. You can find out more about how we handle personal data, including how to exercise your individual rights and the contact details for DHSC's Data Protection Officer here: www.uhcw.nhs.uk/privacy/

Disclosure of interests

Full disclosure of interests: Completed ICMJE forms for all authors, including all related interests, are available in the toolkit on the NIHR Journals Library report publication page at https://doi.org/10.3310/SPLK4325.

Primary conflicts of interest: David RJ Snead declares:

- Study funding from NIHR Health Technology Assessment (HTA) programme.
- Research funding from NIHR AI Awards.
- Paid consultancy Interview from Oliver Wyman.
- Support for attending Global Engage Digital Pathology Congress 2019, 2020, 2021 and 2022, from Global Engage.
- Patent applied for Iguana colon biopsy screening tool.
- Founder and Director of Histofy Ltd. Founded Histofy Ltd in 2021 this company develops computer algorithms for clinical DP use.
- Shareholder in Histofy Ltd.

Ayesha S Azam has declared no competing interests.

Jenny Thirlwall has declared no competing interests.

Peter Kimani declares:

- NIHR HTA grant payments to WCTU.

Louise Hiller declares:

- NIHR HTA grant payments to WCTU.

Adam Bickers has declared no competing interests.

Clinton Boyd has declared no competing interests.

David Boyle declares:

- Employee of Belfast Health and Social Care Trust, institution received payments from Digital Pathology study funded by NIHR.
- Consulting fees from Digital Pathology study funded by NIHR.

David Clark declares:

- NIHR HTA Grant funding payments made to Nottingham University Hospital NHS Trust.

Ian Ellis declares:

- Payment for Hong Kong IAP invited lectures. Payment for expert testimony for Medicolegal work.
- Payment for attending meeting and/or travel for invited lectures.
- Participation on Data Safety Monitoring Board or Advisory Board: Daiichi Sanyo HER2 Low advisory board.
- Leadership or fiduciary role in other board, society, committee or advocacy group: President Association of Breast Pathology – unpaid.
- Other financial or non-financial interests: Medical Director SourceLDPath.

Kishore Gopalakrishnan declares:

- NIHR funding for the duration of this project.

Mohammad Ilyas declares:

- Study NIHR grant as Co-Investigator (Co-I).
- Grant award (as Co-I) from MRC and Innovate UK. This was a grant to support biomarker studies using DP. It was totally unrelated and did not provide any data to the current study.
- Royalties pending for software licensed to TissueGnostics GmbH. This is software generated and licensed to TissueGnostics before the start of the study. It is totally unrelated to the study.
- Leadership or fiduciary role in other board, society, committee or advocacy group, paid or unpaid:
  - Previously standing member of the Medical Technology Advisory committee for NICE.
  - Previous Chair of the Pathology section for British Society of Gastroenterology.
  - Current Meetings secretary for Pathological Society of Great Britain.

Paul Kelly has declared no competing interests.

Maurice Loughrey declares:

- NCRI payments made to Institution.

Desley Neil declares:

- Liver transplant working group lead for donor steatosis assessment and Co-lead working group on preimplantation renal biopsy, for Banff Foundation for allograft pathology.
- Pathology lead for out-of-hours donor pathology for NHSBT.

Emad Rakha has declared no competing interests.

Ian SD Roberts declares:

- NIHR HTA Grant funding payments made Oxford University Hospitals NHS FT.
- Royalties from Elsevier for chapters in fourth edition of our *Diagnostic Pathology*: *Kidney Diseases*, and seventh edition of *Underwood's Pathology – A Clinical Approach*. Consulting fees from Novartis and Travere Therapeutics.
- Travel stipends from meeting organisers and professional societies for attending meetings of ASN, USCAP, BDIAP, KSN, ESN.
- Participation on Data Safety Monitoring Board or Advisory Board: Novartis C3G Advisory Board, 2023.
- Leadership or fiduciary role in other board, society, committee or advocacy group, paid or unpaid: President, British Division of the IAP.

Shatrughan Sah declares:

- Study NIHR HTA grant as Co-Investigator, payments made to UHCW NHS Trust.

Maria Soares has declared no competing interests.

YeeWah Tsang has declared no competing interests.

Manuel Salto-Tellez declares:

- NIHR HTA funding for Digital Pathology.
- Grants or contracts from any entity:
  - ImageDx CRC and ImageDx Lung: A centralised, AI-based solution for tissue-based delivery of personalised medicine testing for the NHS, application successful. Funder programme AI Award 2020.
  - Precision Medicine: Establishing a Precision Medicine Catapult Centre of Excellence in Northern Ireland – QUB Contribution, application successful.
  - PD-L1 in multiple solid tumours: A systematic approach to test validation and verification (Scoping Exercise), application successful.
  - Kelvin-2 – The High Performance Computing Centre in Northern Ireland (HPC-NI), application successful.
  - ACTIONED: The Actioned Consortium – integrated molecular soluTIons fOr diagNostics and Early Detection, application successful.
  - PathLEAD 104689: PathLEAD: Pathology Image Data Lake for Education, Analytics and Discovery, application successful.
- Consulting fees: Mindpeak and Sonrai Analytics. Payments personal and to the institution.
- Payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events: BMS, Roche, MSD, Sanofi and Incyte. Payments personal and to the institution.
- Support for attending meetings and/or travel: Associated with the lectures above.
- Participation on a Data Safety Monitoring Board or Advisory Board: Roche, Sanofi and Incyte. Payments personal and to the institution.

Helen Higgins has declared no competing interests.

Donna Howe declares:

- NIHR HTA grant payments to WCTU.

Abigail Takyi has declared no competing interests.

Yan Chen has declared no competing interests.

Agnieszka Ignatowicz declares:

- Grant – payments to institution. Rwanda912: Use of an electronic communications platform to improve pre-hospital transport of injured people – funded by NIHR RIGHT Programme.
- Grant – payments to institution. NIHR Global Health Group on Equitable Access to Quality Health Care for Injured People in Four Low- or Middle-Income Countries: Equi-injury – funded by NIHR Global Health Groups Programme.
- Grant – payments to institution. How to utilise the potential of Hospital at Home to deliver more acute non-COVID and COVID care outside of hospital – funded by NIHR Policy Research Programme.

Jason Madan declares:

- Five NIHR grants, funding made to institution.
- Two Bill and Melinda Gates Foundation grants, funding made to institution.

Henry Nwankwo declares:

- NIHR HTA grant payments to WCTU.

George Partridge has declared no competing interests.

Janet Dunn declares:

- NIHR HTA grant payments to WCTU.
- EME Strategy Advisory Committee 2019 and EME Funding Committee Members 2017–23.

Publications

Koh A, Roy D, Gale A, Mihai R, Atwal G, Ellis IO, *et al.* Understanding digital pathology performance: an eye tracking study. *Proc. SPIE Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment* 2020;**11316**:7–13. https://doi.org/10.1117/12.2550513

Azam AS, Miligy IM, Kimani PK, Maqbool H, Hewitt K, Rajpoot NM, Snead DRJ. Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis. *J Clin Pathol* 2021;**74**:448–55. https://doi.org/10.1136/jclinpath-2020-206764

Sudin E, Roy D, Kadi N, Triantafyllakis P, Atwal G, Gale A, *et al.* Eye tracking in digital pathology: identifying expert and novice patterns in visual search behaviour. *Proc. SPIE Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment* 2021;**11603**:253–62. https://doi.org/10.1117/12.2580959

Partridge GJW, Phillips P, Taib A, Chen Y. Challenges and solutions to processing and visualising eye tracking data from digital pathology studies. *Proc. SPIE Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment* 2023;**12467**:19–26. https://doi.org/10.1117/12.2652981

Poster presentations

Azam AS, Miligy IM, Kimani PKU, Maqbool H, Hewitt K, Rajpoot NM, Snead DRJ. Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis. *Liverpool Pathol* 2020 Sep15;74:448–55. https://doi.org/10.1136/jclinpath-2020-206764

Snead D, Tsang YW, Azam A, Thirlwall J, Kimani P, Graham S, Rajpoot N. *Multi-Site Validation of Digital Pathology for the Routine Reporting of Histopathology Samples*. ECDDP Congress, 14–17 June 2023.

References

1. Tabata K, Mori I, Sasaki T, Itoh T, Shiraishi T, Yoshimi N, *et al*. Whole-slide imaging at primary pathological diagnosis: validation of whole-slide imaging-based primary pathological diagnosis at twelve Japanese academic institutes. *Pathol Int* 2017;**67**:547–54. https://doi.org/10.1111/pin.12590
2. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, *et al*. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (Pivotal Study). *Am J Surg Pathol* 2017;**42**:39–52. https://doi.org/10.1097/PAS.0000000000000948
3. Duffy S, Vulkan D, Cuckle H, Parmar D, Sheikh S, Smith R, *et al*. Annual mammographic screening to reduce breast cancer mortality in women from age 40 years: long-term follow-up of the UK Age RCT. *Health Technol Assess* 2020;**24**:1–24. https://doi.org/10.3310/hta24550
4. Bainbridge S, Cake R, Meredith M, Furness P, Gordon B. *Testing Times to Come: An Evaluation of Pathology Capacity across the UK*. Cancer Research UK; 2016. URL: www.cancerresearchuk.org/sites/default/files/testing_times_to_come_nov_16_cruk.pdf
5. Rowlands GL. Histopathology workforce survey summary and reports (Reports 1 and 2 originally published in *The Bulletin* April 2018 edition). *Bull R Coll Pathol* 2018;**182**:78–86.
6. Harris G. Digitisation will transform the future of pathology. *Br J Healthcare Manag* 2020;**26**. https://doi.org/10.12968/bjhc.2020.0018
7. Pantanowitz L, Sharma A, Carter AB, Kurc T, Sussman A, Saltz J. Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *J Pathol Inform* 2018;**9**:40. https://doi.org/10.4103/jpi.jpi_69_18
8. Jahn SW, Plass M, Moinfar F. Digital pathology: advantages, limitations and emerging perspectives. *J Clin Med* 2020;**9**:3697. https://doi.org/10.3390/jcm9113697
9. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. *Histopathology* 2012;**61**:1–9. https://doi.org/10.1111/j.1365-2559.2011.03814.x
10. Retamero JA, Aneiros-Fernandez J, Del Moral RG. Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network. *Arch Pathol Lab Med* 2020;**144**:221–8. https://doi.org/10.5858/arpa.2018-0541-OA
11. Browning L, Colling R, Rakha E, Rajpoot N, Rittscher J, James JA, *et al*. Digital pathology and artificial intelligence will be key to supporting clinical and academic cellular pathology through COVID-19 and future crises: the PathLAKE consortium perspective. *J Clin Pathol* 2021;**74**:443–7. https://doi.org/10.1136/jclinpath-2020-206854
12. Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence – the third revolution in pathology. *Histopathology* 2019;**74**:372–6. https://doi.org/10.1111/his.13760
13. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol* 2019;**20**:e253–61. https://doi.org/10.1016/S1470-2045(19)30154-8
14. Burthem J, Brereton M, Ardern J, Hickman L, Seal L, Serrant A, *et al*. The use of digital 'virtual slides' in the quality assessment of haematological morphology: results of a pilot exercise involving UK NEQAS(H) participants. *Br J Haematol* 2005;**130**:293–6. https://doi.org/10.1111/j.1365-2141.2005.05597.x
15. Snead DR, Tsang YW, Meskiri A, Kimani PK, Crossman R, Rajpoot NM, *et al*. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* 2016;**68**:1063–72. https://doi.org/10.1111/his.12879

16. Al-Janabi S, Huisman A, Nap M, Clarijs R, van Diest PJ. Whole slide images as a platform for initial diagnostics in histopathology in a medium-sized routine laboratory. *J Clin Pathol* 2012;**65**:1107–11. https://doi.org/10.1136/jclinpath-2012-200878

17. Baidoshvili A, Bucur A, van Leeuwen J, van der Laak J, Kluin P, van Diest PJ. Evaluating the benefits of digital pathology implementation: time savings in laboratory logistics. *Histopathology* 2018;**73**:784–94. https://doi.org/10.1111/his.13691

18. Stathonikos N, Nguyen TQ, Spoto CP, Verdaasdonk MAM, van Diest PJ. Being fully digital: perspective of a Dutch academic pathology laboratory. *Histopathology* 2019;**75**:621–35. https://doi.org/10.1111/his.13953

19. Baidoshvili A. *How to Go Digital in Pathology*. LabPON Laboratorium Pathologie Oost-Nederland; 2016. URL: www.philips.com/c-dam/b2bhc/master/sites/pathology/resources/white-papers/labron-how-to-go-digital.pdf (accessed 20 June 2024).

20. Williams B, Hanby A, Millican-Slater R, Nijhawan A, Verghese E, Treanor D. Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study. *Histopathology* 2017;**72**:662–71. https://doi.org/doi: 10.1111/his.13403

21. Babawale M, Gunavardhan A, Walker J, Corfield T, Huey P, Savage A, *et al*. Verification and validation of digital pathology (whole slide imaging) for primary histopathological diagnosis: all Wales experience. *J Pathol Inform* 2021;**12**:4. https://doi.org/10.4103/jpi.jpi_55_20

22. Azam AS, Miligy IM, Kimani PKU, Maqbool H, Hewitt K, Rajpoot NM, Snead DRJ. Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis. *J Clin Pathol* 2021;**74**:448–55. https://doi.org/10.1136/jclinpath-2020-206764

23. National Institute for Health and Care Research. *Multi-Centred Validation of Digital Whole Slide Imaging for Routine Diagnosis*. National Institute for Health and Care Research; 2018. URL: https://fundingawards.nihr.ac.uk/award/17/84/07

24. International Standard Randomised Controlled Trial Number. *Is the Use of Digital Pathology in Routine Diagnosis Reliable and Safe in Comparison to Standard Microscopy?* 2018. https://doi.org/10.1186/ISRCTN14513591

25. Cross S, Furness P, Igali L, Snead D, Treanor D. *Best Practice Recommendations for Implementing Digital Pathology*. London: RCPath, Royal College of Pathologists; 2018.

26. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, *et al*.; College of American Pathologists Pathology and Laboratory Quality Center. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med* 2013;**137**:1710–22. https://doi.org/10.5858/arpa.2013-0093-CP.

27. Wood S, Scheipl F. *gamm4: Generalized Additive Mixed Models Using 'mgcv' and 'lme4'*. 2020.

28. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020. URL: www.R-project.org/ (accessed 20 June 2024).

29. Brooks ME, Kristensen K, Van Benthem KJ, Magnusson A, Berg CW, Nielsen A, *et al*. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R Journal* 2017;**9**:378–400.

30. Huang A. Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Stats Model* 2017;**17**:359–80. https://doi.org/10.1177/1471082x17697749

31. Saeed-Vafa D, Magliocco AM. Practical applications of digital pathology. *Cancer Control* 2015;**22**:137–41. https://doi.org/10.1177/107327481502200203

32. Griffin J, Treanor D. Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology* 2017;**70**:134–45. https://doi.org/10.1111/his.12993

33. Hanna MG, Ardon O, Reuter VE, Sirintrapun SJ, England C, Klimstra DS, Hameed MR. Integrating digital pathology into clinical practice. *Mod Pathol* 2022;**35**:152–64. https://doi.org/10.1038/s41379-021-00929-0

34. Cheng CL, Tan PH. Digital pathology in the diagnostic setting: beyond technology into best practice and service management. *J Clin Pathol* 2017;**70**:454–7. https://doi.org/10.1136/jclinpath-2016-204272

35. Eloy C, Vale J, Curado M, Polónia A, Campelos S, Caramelo A, *et al*. Digital pathology workflow implementation at IPATIMUP. *Diagnostics* 2021;**11**:2111. https://doi.org/10.3390/diagnostics11112111

36. Williams BJ, Treanor D. Practical guide to training and validation for primary diagnosis with digital pathology. *J Clin Pathol* 2020;**73**:418–22. https://doi.org/10.1136/jclinpath-2019-206319

37. Randell R, Ruddle RA, Treanor D. Barriers and facilitators to the introduction of digital pathology for diagnostic work. *Stud Health Technol Inform* 2015;**216**:443–7.

38. Tuthill EL, Maltby AE, DiClemente K, Pellowski JA. Longitudinal qualitative methods in health behavior and nursing research: assumptions, design, analysis and lessons learned. *Int J Qual Methods* 2020;**19**:1609406920965799. https://doi.org/10.1177/1609406920965799

39. Strauss A, Corbin J. *Basics of Qualitative Research*. Thousand Oaks, CA: SAGE Publications Ltd; 1990.

40. Greenhalgh T, Papoutsi C. Spreading and scaling up innovation and improvement. *BMJ* 2019;**365**:l2068. https://doi.org/10.1136/bmj.l2068

41. Shaw EC, Hanby AM, Wheeler K, Shaaban AM, Poller D, Barton S, *et al*. Observer agreement comparing the use of virtual slides with glass slides in the pathology review component of the POSH breast cancer cohort study. *J Clin Pathol* 2012;**65**:403–8. https://doi.org/10.1136/jclinpath-2011-200369

42. Murray E, Burns J, May C, Finch T, O'Donnell C, Wallace P, Mair F. Why is it difficult to implement e-health initiatives? A qualitative study. *Implement Sci* 2011;**6**:1. https://doi.org/10.1186/1748-5908-6-6

43. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, *et al*. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. *Implement Sci* 2017;**12**:113. https://doi.org/10.1186/s13012-017-0644-2

44. Kyratsis Y, Ahmad R, Holmes A. Technology adoption and implementation in organisations: comparative case studies of 12 English NHS Trusts. *BMJ Open* 2012;**2**:e000872. https://doi.org/10.1136/bmjopen-2012-000872

45. Shaw J, Shaw S, Wherton J, Hughes G, Greenhalgh T. Studying scale-up and spread as social practice: theoretical introduction and empirical case study. *J Med Internet Res* 2017;**19**:e244. https://doi.org/10.2196/jmir.7482

46. Murray E, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, *et al*. Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. *BMC Med* 2010;**8**:63. https://doi.org/10.1186/1741-7015-8-63

47. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance in information technology. *MIS Q* 1989;**13**:319–40.

48. Ho J, Ahlers SM, Stratman C, Aridor O, Pantanowitz L, Fine JL, *et al*. Can digital pathology result in cost savings? A financial projection for digital pathology implementation at a large integrated health care organization. *J Pathol Inform* 2014;**5**:33. https://doi.org/10.4103/2153-3539.139714

49. Buckley YM. *Generalised Linear Models*. In: Fox GA, Negrete-Yankelevich S, Sosa VJ, eds *Ecological Statistics Contemporary Theory and Application*. Oxford: Oxford University Press; 2014. 131–48.

50. Sudin, E. *et al*. Eye tracking in digital pathology: identifying expert and novice patterns in visual search behaviour. *Proc. SPIE* 2021;11603:116030Z. https://doi.org/10.1117/12.2580959

51. Pallua JD, Brunner A, Zelger B, Schirmer M, Haybaeck J. The future of pathology is digital. *Pathol Res Pract* 2020;**216**:153040. https://doi.org/10.1016/j.prp.2020.153040

52. 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. https://doi.org/10.1016/j.humpath.2006.08.024

53. Brunye TT, Mercan E, Weaver DL, Elmore JG. Accuracy is in the eyes of the pathologist: the visual interpretive process and diagnostic accuracy with digital whole slide images. *J Biomed Inform* 2017;**66**:171–9. https://doi.org/10.1016/j.jbi.2017.01.004

54. Drew T, Lavelle M, Kerr KF, Shucard H, Brunyé TT, Weaver DL, Elmore JG. More scanning, but not zooming, is associated with diagnostic accuracy in evaluating digital breast pathology slides. *J Vis* 2021;**21**:7. https://doi.org/10.1167/jov.21.11.7

55. Stember JN, Celik H, Krupinski E, Chang PD, Mutasa S, Wood BJ, *et al*. Eye tracking for deep learning segmentation using convolutional neural networks. *J Digit Imaging* 2019;**32**:597–604. https://doi.org/10.1007/s10278-019-00220-4

56. Mall S, Brennan PC, Mello-Thoms C. Can a machine learn from radiologists' visual search behaviour and their interpretation of mammograms–a deep-learning study. *J Digit Imaging* 2019;**32**:746–60. https://doi.org/10.1007/s10278-018-00174-z

57. Wu CC, Wolfe JM. Eye movements in medical image perception: a selective review of past, present and future. *Vision (Basel)* 2019;**3**:32. https://doi.org/10.3390/vision3020032

58. Waite S, Grigorian A, Alexander RG, Macknik SL, Carrasco M, Heeger DJ, Martinez-Conde S. Analysis of perceptual expertise in radiology – current knowledge and a new perspective. *Front Hum Neurosci* 2019;**13**:272. https://doi.org/10.3389/fnhum.2019.00213

59. Brunye TT, Drew T, Weaver DL, Elmore JG. A review of eye tracking for understanding and improving diagnostic interpretation. *Cogn Res Princ Implic* 2019;**4**:7. https://doi.org/10.1186/s41235-019-0159-2

60. Drew T, Vo ML, Olwal A, Jacobson F, Seltzer SE, Wolfe JM. Scanners and drillers: characterizing expert visual search through volumetric images. *J Vis* 2013;**13**:3. https://doi.org/10.1167/13.10.3

61. Mall S, Brennan P, Mello-Thoms CF. Fixated and not fixated regions of mammograms: a higher-order statistical analysis of visual search behavior. *Acad Radiol* 2017;**24**:442–55. https://doi.org/10.1016/j.acra.2016.11.020

62. Krupinski EA, Graham AR, Weinstein RS. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Hum Pathol* 2013;**44**:357–64. https://doi.org/10.1016/j.humpath.2012.05.024

63. Mello-Thoms C, Mello CAB, 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. https://doi.org/10.5858/arpa.2010-0697-OA

64. Molin J, Fjeld M, Mello-Thoms C, Lundstrom C. Slide navigation patterns among pathologists with long experience of digital review. *Histopathology* 2015;**67**:185–92. https://doi.org/10.1111/his.12629

65. Koh A, Roy D, Gale A, Mihai R, Atwal G, Ellis I, Snead D, *et al*. Understanding digital pathology performance: an eye tracking study. *Proc SPIE 11316 Med Imaging* 2020;**11316**:1–7. https://doi.org/10.1117/12.2550513

66. Sudin E, Searjeant M, Partridge G, Phillips P, Hiller L, Snead D, *et al*. Digital pathology: the effect of experience on visual search behavior. *J Med Imaging* 2022;**9**:035501. https://doi.org/10.1117/1.JMI.9.3.035501

67. Partridge GJW, Phillips P, Taib A, Chen Y. Challenges and solutions to processing and visualising eye tracking data from digital pathology studies. *Med Imaging* 2023;**12467**:19–26.

68. Lee A, Carder P, Deb R, Ellis IO, Howe M, Jenkins JA, Pinder SE. *Guidelines for Non-Operative Diagnostic Procedures and Reporting in Breast Cancer Screening*. London: Royal College of Pathologists Guidelines; 2021.

69. Mercan E, Shapiro LG, Brunye TT, Weaver DL, Elmore JG. Characterizing diagnostic search patterns in digital breast pathology: scanners and drillers. *J Digit Imaging* 2018;**31**:32–41. https://doi.org/10.1007/s10278-017-9990-5

70. Borowsky AD, Glassy EF, Wallace WD, Kallichanda NS, Behling CA, Miller DV, *et al*. Digital whole slide imaging compared with light microscopy for primary diagnosis in surgical pathology. *Arch Pathol Lab Med* 2020;**144**:1245–53. https://doi.org/10.5858/arpa.2019-0569-OA

71. Williams B, Hanby A, Millican-Slater R, Verghese E, Nijhawan A, Wilson I, *et al*. Digital pathology for primary diagnosis of screen-detected breast lesions – experimental data, validation and experience from four centres. *Histopathology* 2020;**76**:968–75. https://doi.org/10.1111/his.14079

72. Elmore JG, Longton GM, Pepe MS, Carney PA, Nelson HD, Allison KH, *et al*. A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis. *J Pathol Inform* 2017;**8**:12. https://doi.org/10.4103/2153-3539.201920

73. Vodovnik A, Aghdam MRF. Complete routine remote digital pathology services. *J Pathol Inform* 2018;9:36. https://doi.org/10.4103/jpi.jpi_34_18

74. Lee JJ, Jedrych J, Pantanowitz L, Ho J. Validation of digital pathology for primary histopathological diagnosis of routine, inflammatory dermatopathology cases. *Am J Dermatopathol* 2018;**40**:17–23. https://doi.org/10.1097/DAD.0000000000000888

75. Bauer TW, Schoenfield L, Slaw RJ, Yerian L, Sun Z, Henricks WH. Validation of whole slide imaging for primary diagnosis in surgical pathology. *Arch Pathol Lab Med* 2013;**137**:518–24. https://doi.org/10.5858/arpa.2011-0678-OA

Appendix 1 Health economics

Model specification

Two general specifications of the model were fit at each level of analysis and the results were reported. Both models are similar; however, model 2 differs from model 1 by the inclusion of interaction effects to explore the relationship between technology and specified sample characteristics.

Model 1

Let i denote each observation (level one) and j denote each histopathology sample (level two). y_{ij} is the response (dependent) variable, that is, time taken for diagnosis realised for observation i in histopathology sample j for $i = 1 \ldots 8$, $j = 1, \ldots, 2000$.

$$y_{ij} \sim Gamma(y_{ij}|\lambda_{ij}, a).$$

The probability distribution of the gamma model is described with a scale parameter a where $Gamma(y_{ij}|\lambda_{ij}, a) = \frac{1}{(a^{\lambda_{ij}} G\lambda_{ij})} y_{ij}^{\lambda_{ij}-1} \exp\left(-\left\{\frac{y_{ij}}{a}\right\}\right)$.

For $y_{ij}, \lambda_{ij}, a > 0$,

where λ_{ij} and a is the mean of the gamma distribution and the shape parameter, respectively.

The general form of the statistical model can be written as

$$y_{ij} = \lambda_{ij} + \varepsilon_{ij}.$$

The general form of the statistical model can be written as

$$\log(y_{ij}) = \beta_{0j} + \beta' x'_{ij} + e_{ij},$$

where y_{ij} is the gamma distributed response variable for the ith observation in the jth sample. x'_{ij} are matrix of fixed-effects predictors for that observation and β' are fixed-effect coefficients.

$$\beta_{0j} = \gamma_{00} + u_{0j},$$

where γ_{00} is the intercept for the jth sample and u_{0j} is the residual error for the jth sample.

Residual errors e_{ij} and u_{0j} are assumed to be independent and follow a gamma distribution.

Model 2 (interaction models)

Each pathologist reporting time was divided into 10 quantiles. For example, assuming a total reporting time of 20 months for pathologist X, observations in the first and second quantile represent reported dates within the first 2 months and between the second and fourth months, respectively. Observations in the 10th quantile represent reported dates falling within the 18th and 20th months.

Let i denote each observation (level one) and j denote each histopathology sample (level two). y_{ij} is the response (dependent) variable, that is, time taken for diagnosis realised for observation i in histopathology sample j for $i = 1 \ldots 8$, $j = 1, \ldots, 2000$.

Let T denote a variable indicating technology used, D denote a variable indicating quantiles of reported dates for each pathologist, S denote a variable indicating number of slides per observation and C denote a variable indicating case complexity (i.e. difficulty level).

The general form of the statistical model fit to explore the relationship between technology T and quantiles of reported dates D, between technology T and number of slides S, and between technology T difficulty level C can respectively be written as:

$$\log(y_{ij}) = \beta_{0j} + \beta_1 T_{ij} + \beta_2 D_{ij} + \beta_3 T_{ij} \times D_{ij} + \beta' x'_{ij} + e_{ij},$$

$$\log(y_{ij}) = \beta_{0j} + \beta_1 T_{ij} + \beta_2 S_{ij} + \beta_3 T_{ij} \times S_{ij} + \beta' x'_{ij} + e_{ij},$$

$$\log(y_{ij}) = \beta_{0j} + \beta_1 T_{ij} + \beta_2 C_{ij} + \beta_3 T_{ij} \times C_{ij} + \beta' x'_{ij} + e_{ij},$$

where y_{ij} is the gamma distributed response variable for the ith observation in the jth sample. x'_{ij} represent matrix of fixed-effects predictors for that observation and β' are fixed-effect coefficients.

$$\beta_{0j} = \gamma_{00} + u_{0j},$$

where γ_{00} is the intercept for the jth sample and u_{0j} is the residual error for the jth sample.

Residual errors e_{ij} and u_{0j} are assumed to be independent and follow a gamma distribution.

Appendix 2 Eye-tracking team dissemination plans

Overview

We are planning on disseminating further findings of the DP eye-tracking studies in the form of research article publication. Using the eye-tracking data collected towards the end of the DP trial, we anticipate that we will produce two academic papers:

1. A technical paper – formally writing up the methodology of the slide tracking software platform developed for detailed eye-tracking analysis.
2. A research paper – analysing the complete eye-tracking data set, utilising the slide tracking software for a more detailed analysis.

Objectives

The objective of the technical paper is to capture our approach in detail so it can be cited in the research paper. Furthermore, it will be useful to share methodologies and software developed in order to assist other research groups working in the field of DP image perception.

The objective of the research paper is to expand the current knowledge base of visual perception in the DP field. Findings could be useful for recommendations of effective reading strategies for trainee pathologists/pathologists new to DP as its uptake becomes more universal, as well as recommendations to DP reporting software vendors.

Audience

The audience of the research outputs will primarily be clinical researchers, pathologists and DP vendors. To effectively target these audiences, the outputs will be submitted to appropriate journals. The technical paper directed to a more technical journal (e.g. *Journal of Medical Imaging*) and the research paper will be submitted to a pathology journal for dissemination to pathologist audiences.

Timeline

Although progress has been made with the slide tracking software platform, refinement and further testing are required. We anticipate that a first draft of the technical paper can be completed by October 2023. The analysis for the research paper is currently underway, but it will require completion of the software for complete analysis. We anticipate that first draft of this paper will be completed in January 2024.

Resources

We have sufficient expertise for the dissemination activity within our team at the University of Nottingham and collaborators at Nottingham University Hospitals Foundation Trust. Additional funding will not be necessary.

EME
HSDR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library