Histopathology

Histopathology 2024 DOI: 10.1111/his.15129

Digital pathology for reporting histopathology samples, including cancer screening samples – definitive evidence from a multisite study

Ayesha S Azam,[1,2] Yee-Wah Tsang,[1] Jenny Thirlwall,[2] Peter K Kimani,[2] Shatrughan Sah,[1] Kishore Gopalakrishnan,[1] Clinton Boyd,[3] Maurice B Loughrey,[3,4] Paul J Kelly,[3] David P Boyle,[3] Manuel Salto-Tellez,[4,5] David Clark,[6] Ian O Ellis,[6,7] Mohammad Ilyas,[6,7] Emad Rakha,[6,7] Adam Bickers,[8] Ian S D Roberts,[9] Maria F Soares,[9] Desley A H Neil,[10] Abi Takyi,[1] Sinthuri Raveendran,[1] Emily Hero,[1,11] Harriet Evans,[1,2] Rania Osman,[1] Khunsha Fatima,[2] Rhian W Hughes,[1] Stuart A McIntosh,[4] Gordon W Moran,[7] Jacobo Ortiz-Fernandez-Sordo,[7] Nasir M Rajpoot,[12] Ben Storey,[9] Imtiaz Ahmed,[1] Janet A Dunn,[2] Louise Hiller[2] & David R J Snead[1,2,12]

[1]*University Hospitals Coventry and Warwickshire NHS Trust,* [2]*Warwick Medical School, University of Warwick, Coventry,* [3]*Belfast Health and Social Care Trust,* [4]*Queen's University, Belfast,* [5]*Institute for Cancer Research, London,* [6]*Nottingham University Hospital NHS Trust,* [7]*University of Nottingham, Nottingham,* [8]*Northern Lincolnshire and Goole NHS Foundation TrustScunthorpe,* [9]*Oxford University Hospitals NHS Foundation Trust, Oxford,* [10]*Birmingham NHS Foundation Trust, Birmingham,* [11]*University Hospitals of Leicester NHS Trust, Leicester and* [12]*Computer Science Department, University of Warwick, Coventry, UK*

Date of submission 31 July 2023
Accepted for publication 13 December 2023

Azam A S, Tsang Y-W, Thirlwall J, Kimani P K, Sah S, Gopalakrishnan K, Boyd C, Loughrey M B, Kelly P J, Boyle D P, Salto-Tellez M, Clark D, Ellis I O, Ilyas M, Rakha E, Bickers A, Roberts I S D, Soares M F, Neil D A H, Takyi A, Raveendran S, Hero E, Evans H, Osman R, Fatima K, Hughes R W, McIntosh S A, Moran G W, Ortiz-Fernandez-Sordo J, Rajpoot N M, Storey B, Ahmed I, Dunn J A, Hiller L & Snead D R J (2024) *Histopathology*. https://doi.org/10.1111/his.15129

Digital pathology for reporting histopathology samples, including cancer screening samples – definitive evidence from a multisite study

Aims: To conduct a definitive multicentre comparison of digital pathology (DP) with light microscopy (LM) for reporting histopathology slides including breast and bowel cancer screening samples.
Methods: A total of 2024 cases (608 breast, 607 GI, 609 skin, 200 renal) were studied, including 207 breast and 250 bowel cancer screening samples. Cases were examined by four pathologists (16 study pathologists across the four speciality groups), using both LM and DP, with the order randomly assigned and 6 weeks between viewings. Reports were compared for clinical management concordance (CMC), meaning identical diagnoses plus differences which do not affect patient management. Percentage CMCs were computed using logistic regression models with crossed random-effects terms for case and pathologist. The obtained percentage CMCs were referenced to 98.3% calculated from previous studies.
Results: For all cases LM versus DP comparisons showed the CMC rates were 99.95% [95% confidence interval (CI) = 99.90–99.97] and 98.96 (95% CI = 98.42–99.32) for cancer screening samples. In speciality groups CMC for LM versus DP showed: breast 99.40% (99.06–99.62) overall and 96.27% (94.63–97.43) for cancer screening samples; [gastrointestinal (GI) = 99.96% (99.89–99.99)] overall and

Address for correspondence: D Snead, Pathology Department, UHCW NHS Trust, Coventry CV2 2DX, UK. e-mail: david.snead@uhcw.nhs.uk

99.93% (99.68–99.98) for bowel cancer screening samples; skin 99.99% (99.92–100.0); renal 99.99% (99.57–100.0). Analysis of clinically significant differences revealed discrepancies in areas where interobserver variability is known to be high, in reads performed with both modalities and without apparent trends to either.

Conclusions: Comparing LM and DP CMC, overall rates exceed the reference 98.3%, providing compelling evidence that pathologists provide equivalent results for both routine and cancer screening samples irrespective of the modality used.

Keywords: diagnosis, digital imaging, digital pathology, discordance, validation, whole slides image

Introduction

Histopathology is the light microscopic (LM) examination of tissue sections and is an integral component of many patient pathways. Increasing workload remains a global problem for laboratories due to advances regarding early detection of cancer, improved life expectancy, expanding cancer screening programmes, molecular tests and allied ancillary tests.[1–3] In this context, the most efficient use of a limited cellular pathology workforce is vital to maintain standard of care and patient safety.[4]

Capturing histopathology slides at high resolution and stitching these digital images together enables pathology slides to be recreated on computer 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 increased rapidly during the past decade, aided by high-throughput automated slide scanners requiring minimal input from laboratory technicians that fit seamlessly into the laboratory workflow.[5] DP allows remote viewing of slides, thereby allowing work to be moved easily between pathologists, either to assist flow, provide for multidisciplinary review, expert out-of-hours review or review of previous slides, or where patients move between sites for treatment.[6–8] DP thereby provides almost limitless flexibility in the management of this workload: a factor exploited by many laboratories in response to the COVID-19 pandemic.[9] DP also enables analysis of pixel data contained in the images to be exploited to develop aids to improve diagnosis.[10,11] Hitherto, DP has been used for teaching and external quality assessment,[12] but use in routine reporting of slides has only been delivered recently in a small number of laboratories.[13–18]

Novel technologies require definitive evidence of comparable accuracy with the existing standard. Multiple studies have assessed comparison of LM to DP, most looking at small numbers of cases (fewer than 1000); there have been few large-scale studies aimed at providing evidence for clinical adoption.[13,19–22] A recent meta-analysis demonstrated high concordance rates between the digital and glass readings in these studies.[23] However, the majority (92%) of those studies was performed at a single institution without enrichment for challenging cases or samples from cancer screening programmes, leading to a lack of data supporting the use of DP in this setting and preventing wider adoption. Additionally, to date, few studies have evaluated the accuracy of DP for samples from medical renal biopsies with immunofluorescence slides, a speciality comprising highly complex and low-volume samples where DP may prove to have important benefits in providing improved access to specialist expertise.[24,25]

Examining histopathology slides depends upon interpretation of histological features in light of the clinical setting, and is subject both to inter- and intraobserver variation. The studies comparing DP to 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,[26,27] we performed a multisite comparison of breast, gastrointestinal (GI), skin and renal specialities with consultant pathologists experienced in reporting these samples, comprising routine biopsies, cancer screening samples and resections, as well as cases known to contain challenging lesions. The primary outcomes were intra- and interobserver agreement for pathologists' diagnoses using DP as opposed to LM.

Methods

STUDY DESIGN

The study design was developed incorporating principles published by the Royal College of Pathologists

(RCPath) and the College of American Pathologists.[28,29] A blinded crossover comparison compared pathologists' reports using LM and DP (Figure 1). The Health Research Authority (National Health Service, London, UK) approved the study protocol and any subsequent amendments. The study protocol was published in the International Traditional Medicine Clinical Trial Registry.[27] The steering committee, including an independent chair, the chief investigator and patient representatives, provided study oversight.

ETHICAL APPROVAL AND CONSENT

The study protocol and any subsequent amendments were reviewed and approved by the Health Regulation Authority (HRA) and Research Ethics Committee (REC), ISRCTN number 14513591, IRAS number 258799, 2018. Samples recruited from Oxford (renal) had generic consent for research. Consent was not sought for the remaining cases.

PATHOLOGISTS

Sixteen pathologists, all National Health Service (NHS) consultants with 3–35 years' experience worked in speciality areas of their normal practice. All completed training on the study DP image management system. Eleven pathologists not using DP for routine practice completed DP training following the Royal College of Pathologists' best practice recommendations.[28]

SAMPLE SELECTION

The sample pathway is summarised in Figure 2. Prospective consecutive histopathology samples, enrolled between July 2019 and July 2021, were recruited throughout the four subspeciality areas, including breast and bowel cancer screening biopsies. These were enriched with 20% cases considered either difficult or moderately difficult to report (see Supporting information, Table S1).[23] Renal biopsy samples, all deemed difficult due to the nature of these biopsies, comprised a consecutive series of native and transplant biopsies prospectively recruited from one centre (Oxford). All the other speciality group cases were recruited equally from the departments of the study pathologists.

The glass slides were retrieved along with the corresponding reports. The original report was the reference diagnosis (RD). All slides were included for biopsies. For some large (> 10 blocks) breast and GI resection samples, submitting pathologists selected representative slides sufficient to provide the report. All the available stains, including haematoxylin and

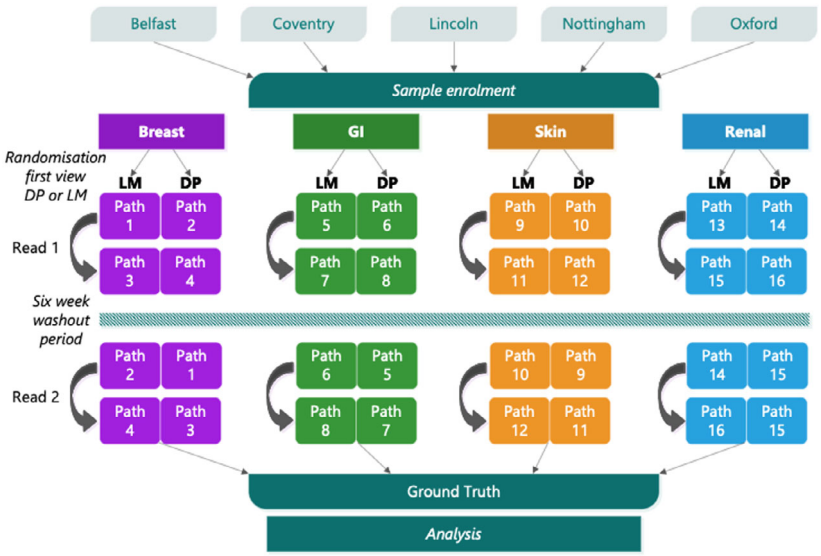

Figure 1. Study overview. Cases were recruited from participating sites in the four speciality groups anonymised and enrolled into the study. In each group each case was examined twice by each pathologist using light microscopy (LM) and digital pathology (DP), respectively. The sequence of whether LM or DP was performed first was randomised and there was a six-week gap between readings. On completion of the eight reads all clinically significant differences were reviewed in consensus meetings, held by the reporting pathologists, to agree the ground truth diagnosis.

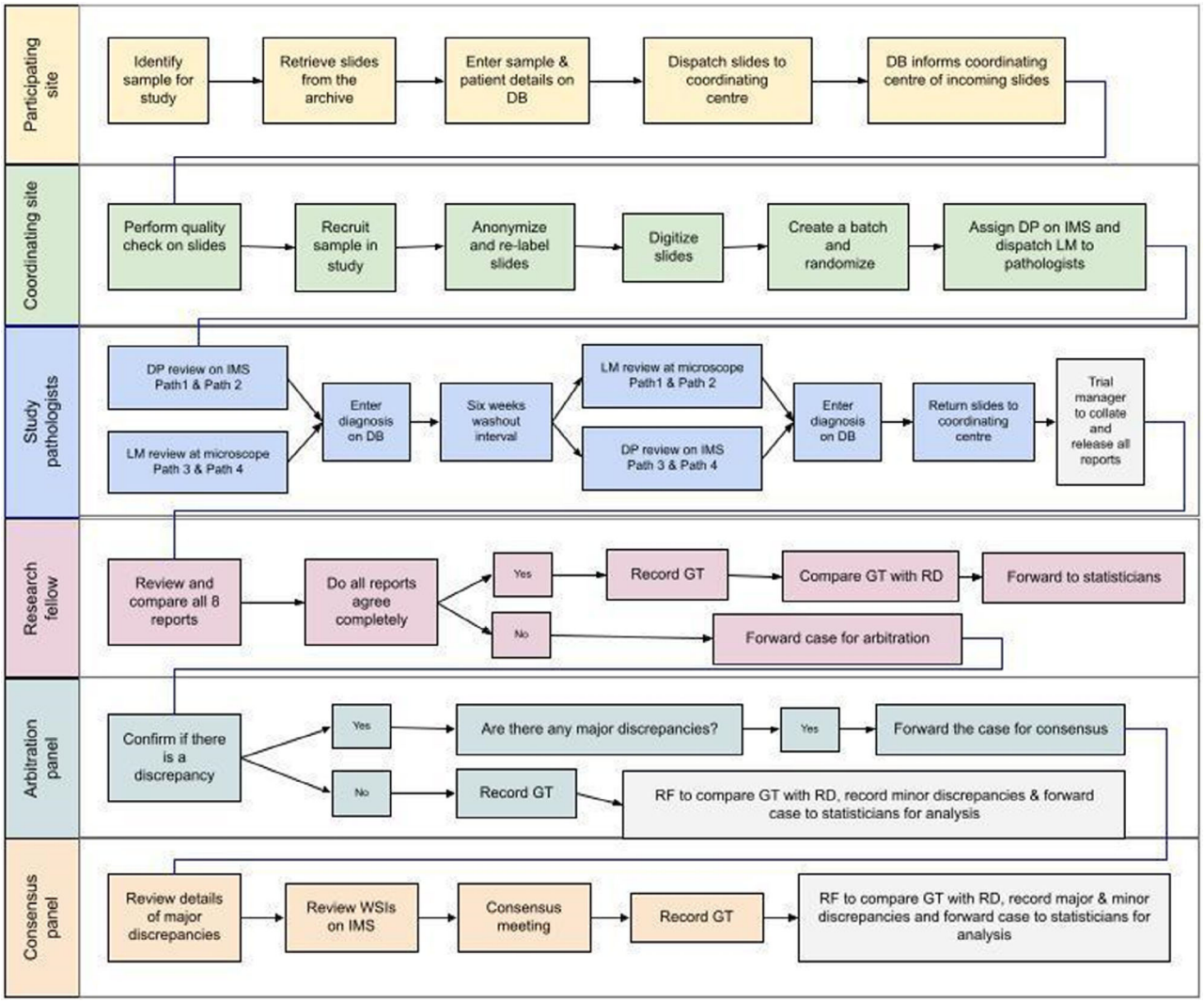

Figure 2. Overall study workflow, reports review, arbitration and consensus process. DB, database; DP, digital pathology; GT, ground truth; LM, light microscopy; RD, reference diagnosis.

eosin (H&E), special, immunocytochemistry and immunofluorescence stains, were included into the study with the exception of GI, where only H&E stains were included. Pen marks were cleaned from the slides and overhanging or badly marked coverslips were replaced, otherwise no additional preparation of slides was performed prior to scanning. Specifically, no attempts were made to correct for imperfections in section quality.

Cases were excluded if:

- there were missing or damaged slides;
- contained oversized slides;
- a prior biopsy review was required for interpretation.

The skin, GI and breast slides were scanned with Philips IntelliSite Pathology Solution (Philips, Eindhoven, the Netherlands) using a single Philips Ultra-Fast Scanner (UFS 1.8, IVD-CE), with automated focal point selection and tissue detection. Cases were viewed using the Philips Image Management System (IMS version 3.3.1; Philips). Once digitised at equivalent ×40 magnification, 0.25 μm per pixel, the WSIs were stored locally at the UHCW Coventry (network connection:1GB/s bandwidth) in two HP DL380 iron servers with a net 24 TB storage capacity. WSI were checked by laboratory technicians at low power to detect obvious errors in focusing or tissue detection, and rescanned if required. All participating sites were

provided with internet-enabled viewing access to images on the firewall-enabled server via a secure network (SSL) connection.

All renal cases were scanned at Oxford University Hospital NHS Trust using a dual-function (bright-field and fluorescent) 3DHISTECH PANNORAMIC SCAN II (3DHistech, Budapest, Hungary) at equivalent ×40 magnification, 0.25 μm per pixel. Brightfield scans used automated focal point selection. Single-channel (fluorescein isothiocyanate) immunofluorescence slides were captured at a single layer using five focus points set by the laboratory technician. WSI were stored on a secure cloud-based server provided by the vendor accessible to all renal pathologists via image the viewing system 3DHISTECH CaseCentre, version 2.9.

Pathologists used standard IVD-CE marked HP workstations (Z4) (comprising a dual-core @3GHz CPU; Microsoft Windows Server version 2012 R2 SP1, RAM 3GB with upgraded graphics cards) and Philips 27″ display monitors (resolution 1920 × 1200; brightness $> 300\ cd/m^2$; contrast 1000:1).

REPORTING OF SAMPLES

Pathologists reported each study sample twice: once using DP and once using LM. The order was randomised, and there was a minimum 6-week gap between viewings. Clinical and macroscopic details were accessed on the study database. LM was conducted using the microscopes used for routine diagnostic work and DP using the workstations provided. Where possible, reporting proformas were used. Reporting followed the UK NHS Bowel and Breast Cancer Screening programme and RCPath minimum data sets requirements.

The annotations and measurement tools available on the DP systems were permitted, but hidden from fellow pathologists. Pathologists recorded their diagnostic confidence for each report on a seven-point Likert scale, from least to most confident.[28]

REPORTS COMPARISON, ARBITRATION AND CONSENSUS PROCESS

The reports were compared by study reviewers blinded to modality, participating site and pathologist. Any variations between reports were forwarded for arbitration. Two pathologists, not involved in reporting of the cases, decided whether the differences identified would more probably have resulted in differences in management (clinically significant) or not (clinically insignificant). In uncertain cases, this decision was referred to a consulting clinician.

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

Consensus ground truth

Where there was one or more clinically significant difference, the WSI (glass slides were available on request) and all the reports (study and reference reports) were reviewed by the study pathologists reporting the case and a consensus ground truth (GT) was agreed.

Outcomes

The primary endpoints of the study were intraobserver intermodality clinical management concordance (CMC, identical diagnoses plus differences clinically insignificant differences) comparing pairs of LM and DP reports by the same pathologist, and interpathologist CMC among the four DP and LM diagnoses, respectively, and the GT.

The secondary outcome measures included: repetition of these comparisons in terms of complete concordance (CC), pathologists' diagnosis confidence separately rated for their LM and DP diagnoses.

Sample size

Percentage CMC for routine and difficult-to-diagnose cases were assumed to be, respectively, 98.8%[13] and 55% (based on the range of 40–70% found in the literature), and 75% for moderate cases (midpoint between routine and difficult).[23] Taking account of enrichment with difficult and moderately difficult cases, the baseline intramodality variability of the whole study sample was defined as 90%.

The study sample size was determined so that it was sufficient to analyse each speciality separately. Based on the precisions of intraobserver intermodality percentage CMC estimates, target recruitment was 2000 cases; 600 cases for each of breast, skin and GI specialities and 200 cases for renal.

Four comparisons arising from four pathologists diagnosing 600 cases within the breast, skin and GI specialities resulted in a total of 2400 LM:DP comparisons. An overall ICC was estimated at 0.8. Hence, the design effect is 1 + ICC (comparisons per case-1) = 3.4. Consequently, 2400 LM:DP comparisons correspond to 705 independent comparisons. This allows a margin of error of 2.2%, so precision is high while analysing breast, skin and GI specimens separately. Due to smaller sample size, the margin of error for renal is 3.1%.

STATISTICAL ANALYSIS

Random-effects (RE) logistic regression models, with crossed RE terms for case and pathologist, were used to estimate both the primary endpoint of intraobserver intermodality percentage CMC (between a pathologist's LM and DP pair of reports) and the secondary endpoints of CMC between a pathologist's LM and GT and between a pathologist's DP and GT. The 'gamm4' package in R statistical program was used.[30,31]

Additionally, using these models, ICC to estimate interobserver agreement, first within LM and then within DP, was computed as:

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

where σ^2_{path} and σ^2_{case} are the RE estimates for pathologist and case, respectively; 500 bootstrap samples were used to compute ICC 95% confidence intervals (CIs). CC data were analysed using the same approach.

LM and DP diagnosis confidence data were compared using a RE generalised Poisson model with crossed RE terms for case and were pathologist-fitted using the 'glmmTMB' package in R.[32]

Subgroup analyses were defined by speciality, screening/non-screening and difficulty level.

Results

CHARACTERISTICS OF CASES

A total of 2024 cases (62.8% female 37.2% male) comprising 608 breast, 607 GI, 609 skin and 200 renal samples (Table 1 and consort diagram Figure 3) were recruited, producing 7750 slides. In total, 766 slides required rescanning, the majority for out-of-focus regions or missing fatty tissue fragments. The four pathologists' reading reports on LM and DP resulted in 16 192 case readings and 8096 comparisons in three possible combinations: LM versus DP, LM versus GT and DP versus GT, totalling 24 288 comparison combinations, excluding RD.

PRIMARY OUTCOME RESULTS

The reports' comparison data are summarised in Table 2. An RE logistic regression model of the 8096 LM versus DP comparisons showed, over all 2024 cases, that CMC between LM and DP was 99.95% (95% CI = 99.90, 99.97; Table 3). This primary

Table 1. Characteristics of patients and cases

Characteristic	All cases ($N = 2024$)	Breast ($n = 608$)	GI ($n = 607$)	Skin ($n = 609$)	Renal ($n = 200$)
Difficulty level, *n* (%)					
Routine	1447 (71.5)	486 (79.9)	477 (78.6)	484 (79.5)	All cases in the speciality difficult
Moderate	164 (8.1)	54 (8.9)	53 (8.7)	57 (9.4)	
Difficult	413 (20.4)	68 (11.2)	77 (12.7)	68 (11.2)	
Screening cases, *n* (%)					
Yes	NA	207 (34.0)	250 (41.2)	NA	NA
No		401 (66.0)	357 (58.8)		
Age of patients (years)					
Min–Max	0–96	18–94	0–89	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)
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)

LQ, lower quartile; Min, minimum; Max, maximum; UQ, upper quartile; NA, not applicable; SD, standard deviation.

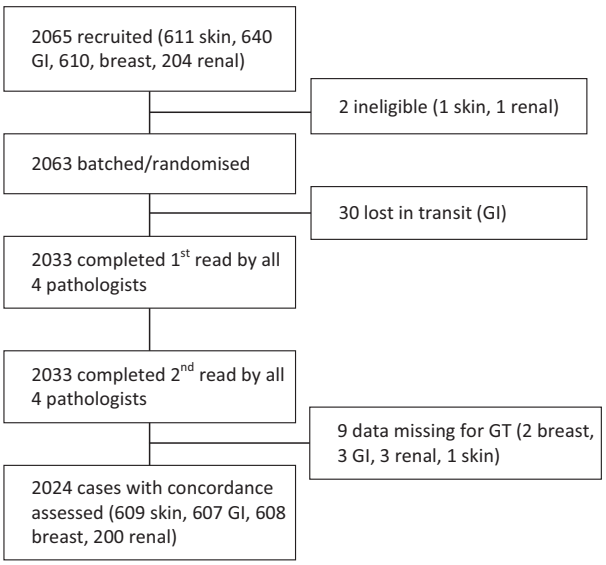

Figure 3. Consort diagram of cases entered into the study.

endpoint result exceeds the pooled percentage CMC (98.3%) in a recent meta-analysis.[23] High CMC was also observed within the four speciality areas [breast: 99.40% (95% CI = 99.06–99.62); GI = 99.96% (95% CI = 99.89–99.99); skin 99.99% (95% CI = 99.92–100); renal 99.99% (95% CI = 99.57–100)], within the difficulty levels [routine cases 99.98% (95% CI = 99.94, 99.99); moderate cases 95.34% (95% CI = 93.09, 96.89); difficult cases 99.84% (95% CI = 99.62, 99.93)] and for screening cases [breast 96.27% (95% CI = 94.63, 97.43); GI = 99.93% (95% CI = 99.68, 99.98); combined breast and GI = 98.96% (95% CI = 98.42, 99.32)].

Respective LM–GT and DP–GT percentage CMCs are very close, so that one modality does not outperform the other in diagnosis accuracy (Table 3). Both modalities also have similar interobserver agreements which, except for moderately difficult, difficult and breast screening cases, are very high, with intraclass correlation (ICC) above 0.8 (Table 3).

SECONDARY OUTCOMES

Summary comparison and RE logistic regression models results for CC, i.e. any difference regardless of clinical relevance, are shown in Supporting information, Tables S2 and S3. All LM–DP percentage CC (intraobserver agreements) are above 88%. Overall, and in subgroup analyses, respective LM–GT and DP–GT percentage CC are close, so that one modality does not outperform the other. Agreement between modalities appeared similar over the longitudinal course of the study, as shown by agreement levels in the various batches of cases (see Supporting information, Tables S4–S6).

Pathologists reported the highest confidence level in 88% of the diagnoses (Table 4). Within a modality, GI pathologists were the most confident with their diagnoses, closely followed by skin pathologists, while renal pathologists were noticeably less confident compared to the other specialities' pathologists. Skin pathologists had approximately the same level of confidence on LM and DP diagnoses, while for the rest of the specialities and overall, the confidence of DP diagnoses was slightly less than the generalised model showing that, overall, lower confidence in DP diagnosis was borderline significant (rate ratio = 0.92, 95% CI = 0.85–1.00, $P = 0.053$; Table 5). Lower confidence with DP diagnoses was significant for the routine cases (rate ratio = 0.86, 95% CI = 0.76–0.98, $P = 0.024$).

CLINICALLY IMPORTANT DIFFERENCES

Clinically important differences were grouped into common themes (Table 6). The renal differences, to be examined in a separate paper, are not discussed. In all three specialities, interpathologist differences appear similar in the comparisons: LM versus GT and DP versus GT and higher than intraobserver intermodality differences LM versus DP.

In breast, slightly higher numbers of differences were seen in B5a versus B5b microinvasion on DP (10) in comparison to LM (four). In three of the 10 DP differences the pathologist gave the same diagnosis on LM as they did for DP. In the seven remaining cases four cases were reported as showing no invasion where the GT concluded that invasion was present, and three cases were reported as showing invasion where the GT concluded that no invasion was present.

A slightly higher intraobserver intermodality than interpathologist difference was seen in the B2 versus B3 (with atypia) (31) LM versus DP compared to either LM (20) or DP (19) to GT. The 31 intraobserver differences were equally divided between LM (15) and DP (16), in equal agreement with GT.

GI showed 31 instances where discrepancy between high- and low-grade dysplasia was recorded. Of these, 21 LM and 27 DP diagnoses were different to GT. Fourteen LM and 19 DP diagnoses showed low-grade dysplasia where GT was high-grade, as opposed to seven LM and nine DP showing high-grade dysplasia and GT recorded low-grade.

Table 2. 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)
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)
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)

Table 2. (*Continued*)

Outcome	All cases ($N = 2024$)	Breast ($n = 608$)	GI ($n = 607$)	Skin ($n = 609$)	Renal ($n = 200$)
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)

LM, light microscopy; GT, ground truth; DP, digital pathology.

Discussion

This study measured the assessment and reporting of 2024 cases by consultant pathologists working at six sites in the United Kingdom, and demonstrated extremely high levels of agreement (99.95% agreement) between DP and LM readings. The level of agreement between the two platforms is identical to that of either platform with the consensus GT.

These figures are similar to those seen in other studies (Table 7), some of which used different DP systems, indicating that the results are likely to translate to laboratories using other equipment. Randomisation of the platform used for first view, and a washout period of 6 weeks, a period longer than those used in similar studies,[13,20,22] were used to reduce 'recall bias'. Recall bias does not affect interpathologist concordance, and this is the first study, to our knowledge, to measure interobserver agreement on the same cases, demonstrating that interobserver performance is identical to DP and LM, as measured by agreement to consensus GT. The study shows near-identical results between the DP and LM platforms among all the speciality groups, as well as for cancer screening cases in breast and GI groups.

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 with an identical clinical context. This is more likely with difficult lesions, with which this study was enriched.[33–35] Clinically significant differences were observed in these cases and reflected in lower levels of agreement seen. Table 6 lists the most common themes giving rise to differences in breast, GI and skin groups. It is noticeable that the incidence of these differences is similar in reports issued with DP and LM platforms.

Previous studies have highlighted areas where DP may present difficulties. These include recognition of bacteria, identification of amyloid and calcification and a tendency to 'over-call' dysplasia or atypia.[13,23,36,37] Examining these and other areas revealed no apparent trends patterns across the DP and LM modalities. For example, failure to recognise *Helicobacter pylori* in gastric biopsies was seen six times in LM and seven times in DP; gastric amyloidosis was missed by two pathologists on both LM and DP reports. There were only single instances of *Giardia duodenalis* and *H. cytomegalovirus*, respectively, being missed, both in DP. There were no errors recorded in breast due to failure to pick up calcification.

Where slight differences between LM and DP were seen, for example in breast B5a (*in-situ* carcinoma) versus B5mi (microinvasive carcinoma) and in GI grading dysplasia in adenomatous polyps; these were in areas where differences between reports are common, and further examination showed no consistent trend with either modality. Regarding dysplasia grading, the second most common difference seen in the GI series, this difference occurred in 21 and 28 LM and DP reports, respectively. However, DP, in common with LM, showed greater differences of low-grade dysplasia against the GT of high-grade dysplasia (i.e. undercalling the dysplasia grade) than the reverse, which is the opposite to what would be seen if DP were indeed leading to overcalling of dysplasia grade. Therefore, we can find no evidence that the platform used has any bearing on these differences.

It is important to note that challenging cases are recognised as such by pathologists at the time of reporting and reflected in lower confidence levels and varying terminologies in the reports, and that arbitrators can have different opinions of what is considered

Table 3. Summary of the clinical management concordance (CMC) analysis using random effects (RE) logistic regression models

Cases included in the analysis	Percentage CMC (95% confidence interval)			Intra-class correlation coefficient (ICC)	
	Intra-observer LM versus DP agreement	LM versus GT agreement	DP versus GT agreement	LM Inter-observer agreement	DP inter-observer agreement
Primary analysis					
All cases ($n = 2024$)[†]	99.95 (99.90, 99.97)[‡]	99.95 (99.91, 99.97)	99.95 (99.91, 99.97)	0.91 (0.89, 0.92)	0.91 (0.89, 0.93)
Subgroup analysis by specialty					
Breast ($n = 608$)[†]	99.40 (99.06, 99.62)	99.76 (99.54, 99.87)	99.88 (99.73, 99.95)	0.83 (0.60, 0.89)	0.88 (0.77, 0.91)
GI ($n = 607$)[†]	99.96 (99.89, 99.99)	99.92 (99.80, 99.97)	99.89 (99.74, 99.95)	0.90 (0.83, 0.93)	0.89 (0.77, 0.93)
Skin ($n = 609$)[†]	99.99 (99.92, 100.0)	99.99 (99.93, 100.0)	99.98 (99.91, 100.0)	0.94 (0.92, 0.95)	0.93 (0.92, 0.95)
Renal ($n = 200$)[†]	99.99 (99.57, 100.0)	100 (99.24, 100.00)	99.18 (97.84, 99.69)	*	*
Subgroup analysis by difficulty level					
Routine ($n = 1447$)[†]	99.98 (99.94, 99.99)	99.98 (99.94, 99.99)	99.98 (99.94, 99.99)	0.93 (0.91, 0.94)	0.93 (0.91, 0.94)
Moderate ($n = 164$)[†]	95.34 (93.09, 96.89)	93.91 (90.95, 95.94)	94.24 (91.41, 96.17)	0.53 (0.36, 0.78)	0.53 (0.36, 0.76)
Difficult excluding renal ($n = 213$)[†]	96.78 (94.27, 98.22)	97.78 (96.11, 98.74)	98.40 (97.14, 99.11)	0.42 (0.13, 0.53)	0.62 (0.24, 0.77)
Difficult including renal ($n = 413$)[†]	99.84 (99.62, 99.93)	97.63 (96.02, 98.60)	97.68 (96.00, 98.67)	0.33 (0.14, 0.90)	0.33 (0.17, 0.91)
Subgroup analysis of the screening cases					
Breast ($n = 207$)[†]	96.27 (94.63, 97.43)	97.57 (96.18, 98.47)	98.23 (97.03, 98.94)	0.53 (0.33, 0.87)	0.59 (0.35, 0.88)
GI ($n = 250$)[†]	99.93 (99.68, 99.98)	99.97 (99.78, 100.0)	99.98 (99.83, 100.0)	0.93 (0.89, 0.96)	0.94 (0.90, 0.96)
Breast and GI ($n = 457$)[†]	98.96 (98.42, 99.32)	99.87 (99.68, 99.95)	99.89 (99.71, 99.96)	0.88 (0.67, 0.92)	0.89 (0.73, 0.93)

[†]n is the number of cases. Each case is reported by four pathologists and so the number of comparisons in the analysis is $4n$.
[‡]Primary objective intra-observer inter-modality clinical management concordance.
*These ICC's could not be estimated reliably because there were only few cases where there was discordance between LM and GT reports and between DP and GT reports (Table 2).

a clinically important difference based on variation in local practice. Pathologists in practice are aware of these challenges and routinely refer such cases to peer review from colleagues.

Pathologists know when they have confidently seen a region of interest to be able to make a diagnosis. The recognition of (and absence of) bacteria and similar subcellular objects may indeed be better on LM. It is possible that this could account for the trend towards greater confidence in LM than DP seen in this study, although further work is needed to fully understand the reason for increased confidence scores with LM. Irrespective of this finding, the advantages DP offers can still be fully exploited while retaining the undoubted superiority that LM may have for some tasks: a timely reminder, if it were needed, to

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

LM, light microscopy; DP, digital pathology; GI, gastrointestinal.

laboratories to ensure that support exists for pathologists working geographically separate from the slides; the slides may need to be examined by LM before the case is reported. Either transport of slides to pathologist when needed or review by a colleague with access to the slides would suffice.

This is the first study, to our knowledge, to demonstrate that DP is equivalent to LM in cancer screening cases and renal biopsies. The flexibility that 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.[3] In breast cancer screening, comparison between LM versus DP for CMC was 96.27%, which is very high, but slightly below the reference of 98.3%. However, comparison to the GT for these samples shows slightly better agreement seen with DP (99.89) as opposed to LM (97.57) indicating, together with the lower ICC scores, that these variances are more likely to be due to differences in the interpretation of challenging biopsies than the modality. The reporting of cancer screening cases for other sites, such as uterine cervix and lung, is based upon similar principles (i.e. assessment of features of atypia and invasive carcinoma on H&E sections), so there is every reason to believe that the results presented here will translate to these sites. Renal biopsy cases require both fine optical resolution and access to immunofluorescence studies. The data for these samples is being published in greater detail in a separate paper but, overall, this study demonstrates that DP is equivalent to LM for these samples, and should help healthcare providers to embrace the opportunities DP offers to redesign and strengthen the service and provide confidence that DP should be equally successful in other speciality 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. In common with previous studies our results show excellent correlation between LM and DP, including in cancer

Table 5. Comparison of diagnosis confidence data using random effects (RE) generalised Poisson models

Data included	Rate ratio (95% CI), *P*-value
All the data (all pathologists and all specialities) ($n = 2024$)[†]	0.92 (0.85–1.00), 0.053
Subgroup analysis by speciality	
Breast cases ($n = 608$)[†]	0.90 (0.78–1.02), 0.108
GI cases ($n = 607$)[†]	0.87 (0.71–1.07), 0.189
Skin cases ($n = 609$)[†]	1.04 (0.86–1.25), 0.701
Renal cases ($n = 200$)[†]	0.91 (0.79–1.05), 0.208
Subgroup analysis by difficulty level	
Routine cases from all special-ties ($n = 1447$)[†]	0.86 (0.76–0.98), 0.024
Moderate cases from all specialities ($n = 164$)[†]	1.32 (1.00–1.75), 0.052
Difficult cases from all specialities ($n = 413$)[†]	0.92 (0.82–1.02), 0.124
Difficult cases excluding renal cases ($n = 213$)[†]	0.92 (0.78–1.09), 0.357
Subgroup analysis of screening cases	
Combined breast and GI screening cases ($n = 457$)[†]	0.87 (0.70–1.07), 0.176
Breast screening cases ($n = 207$)[†]	0.84 (0.67–1.05), 0.119
GI screening cases ($n = 250$)[†]	1.00 (0.60–1.66), 0.994

[†]n = the number of cases. Each case is reported by four pathologists on both light microscopy (LM) and digital pathology (DP), so the number of rows for each case in the analysis is $8n$. In the entire database, only five reports (of 16 192 reports) had missing diagnosis confidence data. GI, gastrointestinal.

screening samples, providing definitive evidence that pathologists give equivalent results regardless of the modality used.

Acknowledgements

This study was funded by the National Institute of Health Research, UK through the Health Technology Assessment Programme 17/84/07 reference 126020. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. D.R.J.S. also reports funding through PathLAKE funded by Innovate UK through Industrial Strategy Fund reference 18181.

Author contributions

ASA: conducted a literature review and managed the study including sample recruitment and supervision of recruitment, establishment of digital reporting across sites, training of pathologists, competence assessment, arbitration of differences, consensus meetings and recording of GT results, preparation of manuscript YWT, SAS, KG, CB, MBL, PJK, DOPB, DC, IOE, MI, ER, AB, ISDR, MFS, DAHN: recruitment and reporting samples, consensus meetings and establishment of GT, comments on clinical differences detected. SR, EH, HE, RO, KF, RWH: reviewed reports to detect differences MST, AT, YWT: arbitrated differences detected into clinically significant or not significant categories. SM, GM, JOF, BS, EC, IA: arbitrated differences detected where it was unclear to the arbitrating pathologists if they were clinically significant or not. NMR: conceived the study and contributed to study protocol. JAD, PKK, LH: designed study protocol and constructed the statistical analysis plan, managed data collection and database, analysed the results manuscript preparation. JT: managed the study, collated progress reports, managed the database, recorded results. JAD, LH, DRJS, JT: reported to the steering committee. DRJS: chief investigator conceived the study, designed protocol, recruitment and reporting of samples, consensus meeting to establish ground

Table 6. Errors recorded in two or more instances in breast, gastrointestinal (GI) and skin specialties

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

Table 6. (*Continued*)

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

AK, actinic keratosis; B1–B5 NHS Breast Screening Programme pathology category classification 1–5 (a, *in situ*, b, invasive, mi, micro-invasive carcinoma); BA2 Barrett's metaplasia; BCC, basal cell carcinoma; DCIS, ductal carcinoma *in situ*; DF, dermatofibroma; DFSP, dermatofibrosarcoma protuberans; HGP, high-grade dysplasia; HP, hyperplastic polyp; IBD, inflammatory bowel disease; IEC, intra-epidermal carcinoma; IEL, intra-epithelial lymphocytosis; IM, intestinal metaplasia; KA, keratoacanthoma; LGP, low-grade dysplasia; MM, malignant melanoma; NET, neuroendocrine tumour; SCC, squamous cell carcinoma; SSL, sessile serrated lesion; TA, tubular adenoma; LM, light microscopy; DP, digital pathology, GT, gastrointestinal.

Table 7. Comparison of this study with other multisite validation studies previously published in the literature

Study ID	Tabata *et al.*[20]	Mukhopadhyay *et al.*[19]	Borowsky[22]	Babawale[21]	This study
No. of participating sites	12	4	5	7	6
No. of cases	900	1992	2045	3001	2024
Total study readings	2140	15,925	15,031	3001	16,192
Number of DP/LM reading pairs	1070	7964	7509	3001	8096
Washout interval	> 2 weeks	Minimum 4 weeks	Minimum 31 days	No washout period	Minimum 6 weeks
Sample enrichment with difficult cases	No	Yes	Yes	Not specified	Yes
No of reading pathologists	9	16	19	22	16
Samples randomised for reading modality	No	Yes	Yes	No	Yes
Intraobserver concordance	Yes	Yes	Yes	No	Yes
Interobserver concordance	No	No	No	Yes	Yes
DP versus LM clinical concordance	99.2%	95.1%	96.36%	97.1%	99.95%

LM, light microscopy; DP, digital pathology.

truth, analysis of clinical differences, manuscript preparation. JAD, PKK, & LH curated all results and conducted the statistical analysis independently from the rest of the authors. All authors reviewed, edited and agreed the final version of the manuscript.

Conflicts of interest

D.R.J.S. and N.M.R. report that they are co-founders, directors and shareholders of Histofy Ltd, a start-up company developing artificial intelligence algorithms for digital pathology. D.R.J.S. has also worked in the past as a member of Philips computational pathology advisory board and has received an honorarium from Oliver Wyman, New York, USA. Y.W.T. reports that she is a shareholder in Histofy Ltd. E.H. and H.E. report working *ad-hoc* part-time sessions for Histofy Ltd.

Data availability statement

Data from this study is stored in the PathLAKE data lake and are available for further research. Applications for access to the data should be made to the PathLAKE Access Committee via the PathLAKE website.

References

1. Duffy S, Vulkan D, Cuckle H *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.
2. Bainbridge S, Cake R, Meredith M, Furness P, Gordon B. *Testing times to come? an evaluation of pathology capcity across the UK.* London, Cancer Research UK, 2016.
3. 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.
4. Harris G. Digitisation will transform the future of pathology. *Br. J. Health Care Manag.* 2020; **26**; 1–4. https://doi.org/10.12968/bjhc.2020.0018.
5. 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.
6. Jahn SW, Plass M, Moinfar F. Digital pathology: advantages, limitations and emerging perspectives. *J. Clin. Med.* 2020; **9**; 3697–3714. https://doi.org/10.3390/jcm9113697.
7. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. *Histopathology* 2012; **61**; 1–9.
8. 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**; 221–228.
9. Browning L, Colling R, Rakha E *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**; 443–447.

10. Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence-the third revolution in pathology. *Histopathology* 2019; **74**; 372–376.
11. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol.* 2019; **20**; e253–e261.
12. Burthem J, Brereton M, Ardern J *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**; 293–296.
13. Snead DR, Tsang YW, Meskiri A *et al*. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* 2016; **68**; 1063–1072.
14. 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**; 1107–1111.
15. 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**; 784–794.
16. Stathonikos N, Nguyen TQ, Spoto CP, Verdaasdonk MAM, van Diest PJ. Being fully digital: perspective of a Dutch Academic Pathology Laboratory. *Histopathology* 2019; **75**; 621–635.
17. Baidoshvili A. How to go digital in pathology (LabPON Laboratorium Pathologie Oost-Nederland). 2016 https://www.philips.com/c-dam/b2bhc/master/sites/pathology/resources/white-papers/labron-how-to-go-digital.pdf.
18. 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**; 662–671.
19. Mukhopadhyay S, Feldman MD, Abels E *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.
20. Tabata K, Mori I, Sasaki T *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**; 547–554.
21. Babawale M, Gunavardhan A, Walker J *et al*. Verification and validation of digital pathology (whole slide imaging) for primary histopathological diagnosis: all Wales experience. *J. Pathol. Inform.* 2021; **12**; 4.
22. Borowsky AD, Glassy EF, Wallace WD *et al*. Digital whole slide imaging compared with light microscopy for primary diagnosis in surgical pathology. *Arch. Pathol. Lab. Med.* 2020; **144**; 1245–1253.
23. Azam AS, Miligy IM, Kimani PKU *et al*. Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis. *J. Clin. Pathol.* 2020; **74**; 448–455.
24. Barisoni L, Troost JP, Nast C *et al*. Reproducibility of the NEPTUNE descriptor-based scoring system on whole-slide images and histologic and ultrastructural digital images. *Mod. Pathol.* 2016; **29**; 671–684.
25. L'Imperio V, Brambilla V, Cazzaniga G, Ferrario F, Nebuloni M, Pagni F Digital pathology for the routine diagnosis of renal diseases: a standard model. *J. Nephrol.* 2021; **34**; 681–688.
26. NIHR. Multi-centred validation of digital whole slide imaging for routine diagnosis. 2018.
27. ISRCTN. Is the use of digital pathology in routine diagnosis reliable and safe in comparison to standard microscopy? 2018.
28. Cross S, Furness P, Igali L, Snead D, Treanor D. *Best practice recomendations for implementing digtial pathology*. London, RCPath, Royal College of Pathologists, 2018.
29. Pantanowitz L, Sinard JH, Henricks WH *et al*. 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**; 1710–1722.
30. Wood S, Scheipl F. gamm4: generalized additive mixed models using 'mgcv' and 'lme4'. 2020.
31. Team RC. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2020. https://www.R-project.org/.
32. Brooks ME, Kristensen K, van Bentham KJ *et al*. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J.* 2017; **9**; 378–400.
33. Onega T, Barnhill RL, Piepkorn MW *et al*. Accuracy of digital pathologic analysis vs traditional microscopy in the interpretation of melanocytic lesions. *JAMA Dermatol.* 2018; **154**; 1159–1166.
34. Rakha EA, Aleskandarani M, Toss MS *et al*. Breast cancer histologic grading using digital microscopy: concordance and outcome association. *J. Clin. Pathol.* 2018; **71**; 680–686.
35. Kent MN, Olsen TG, Feeser TA *et al*. Diagnostic accuracy of virtual pathology vs traditional microscopy in a large dermatopathology study. *JAMA Dermatol.* 2017; **153**; 1285–1291.
36. Elmore JG, Longton GM, Pepe MS *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.
37. Williams B, Hanby A, Millican-Slater R *et al*. Digital pathology for primary diagnosis of screen-detected breast lesions—experimental data, validation and experience from four centres. *Histopathology* 2020; **76**; 968–975.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1.