The Lancet

Digital pathology for reporting histopathology samples, including cancer screening samples – definitive evidence from a multi-site study.

--Manuscript Draft--

Manuscript Number:	
Article Type:	Article
Keywords:	Whole slides image; digital pathology; validation; discordance; digital imaging; diagnosis
Corresponding Author:	David Robert John Snead, MB BS University Hospitals Coventry and Warwickshire NHS Trust UNITED KINGDOM
First Author:	David Robert John Snead, MB BS
Order of Authors:	David Robert John Snead, MB BS
	A S Azam
	YW Tsang
	S Sah
	K Gopalakrishnan
	C Boyd
	MB Loughrey
	P J Kelly
	D P Boyle
	M Salto-Tellez
	D Clark
	IO Ellis
	M Ilyas
	E Rakha
	A Bickers
	ISD Roberts
	M F Soares
	DAH Neil
	A Takyi
	S Raveendran
	E Hero
	H Evans
	R Osman
	K Fatima
	RW Hughes
	JA Dunn
	PK Kimani
	L Hiller
	J. Thirlwall

Manuscript Region of Origin:	UNITED KINGDOM
Abstract:	Introduction Digital pathology (DP) is the examination of digitised histopathology slides on computer workstations as opposed to brightfield and immunofluorescent light microscopy (LM). Deployment in routine practice requires demonstration that pathologists using DP provide equivalent reports in comparison to LM, the current standard of care. Purpose Multicentre comparison of DP with LM for reporting histopathology slides to measure intra and inter-observer variation on both modalities. Methods Sample size of 2000 cases (600 breast, 600 gastrointestinal (GI), 600 skin, 200 renal) was chosen to obtain precise estimates of percentage clinical management concordance (CMC), meaning identical diagnoses plus differences which do not affect patient management. Cases were examined by 4 pathologists (16 study pathologists across the 4 specialty groups), using LM and DP, with the order randomly assigned and 6 weeks between viewings. Random effects (RE) logistic regression models for estimating percentage CMC included crossed RE terms for case and pathologist. Findings were interpreted with reference to 98.3% CMC. Results 2024 cases (608 breast, 607 GI, 609 skin, 200 renal) were recruited, including 207 breast and 250 bowel cancer screening samples. Overall LM v DP comparisons, CMC levels were 99.95% (95%CI 99.90-99.97) for all groups and 98.96 (98.42, 99.32) for cancer screening samples. In specialty groups CMC for LM v DP showed: Breast 99.40% (99.06-99.62) overall and 96.27% (95%CI 94.63-97.43) for cancer screening samples; GI 99.96% (99.89-99.99) overall and 99.93% (95%CI 99.68-99.98) for bowel cancer screening samples; skin 99.99% (99.92-100.0); renal 99.99% (95%CI 99.57-100.0), Analysis of clinically significant differences revealed discrepancies in areas where inter-observer 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 target 98.3% providing compelling evidence that pathologist’s provide equivalent results for both routine and cancer screening samples irrespective of the modality used.

Digital pathology for reporting histopathology samples, including cancer screening samples – definitive evidence from a multi-site study.

A S Azam [1, 2], YW Tsang [1], J. Thirlwall [2], PK Kimani [2], S Sah [1], K Gopalakrishnan [1], C Boyd [3], MB Loughrey [3,4], P J Kelly [3], D P Boyle [3], M Salto-Tellez [4 ,5], D Clark [6], IO Ellis [6, 7], M Ilyas [6, 7], E Rakha [6, 7], A Bickers [8] , ISD Roberts [9] , M F Soares [9] , DAH Neil [10], A Takyi [1], S Raveendran [1], E Hero [1,11], H Evans [1,2], R Osman [1], K Fatima [2], RW Hughes [1], JA Dunn [2], L Hiller [2], DRJ Snead [1, 2].

[1] University Hospitals Coventry and Warwickshire NHS Trust, Coventry, [2] Warwick Medical School, University of Warwick, Coventry, [3] Belfast Health and Social Care Trust, Belfast, N Ireland [4] Queen's University, Belfast, [5] Institute for Cancer Research London , [6] Nottingham University Hospital NHS Trust, Nottingham UK, [7] University of Nottingham, Nottingham, [8] Northern Lincolnshire and Goole NHS Foundation Trust, Lincs UK, [9] Oxford University Hospitals NHS Foundation Trust, Oxford, [10] Birmingham NHSFT, Birmingham, [10] University Hospitals of Leicester NHS Trust, Leicester.

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

Abstract

Introduction: Digital pathology (DP) is the examination of digitised histopathology slides on computer workstations as opposed to brightfield and immunofluorescent light microscopy (LM). Deployment in routine practice requires demonstration that pathologists using DP provide equivalent reports in comparison to LM, the current standard of care.

Purpose: Multicentre comparison of DP with LM for reporting histopathology slides to measure intra and inter-observer variation on both modalities.

Methods: Sample size of 2000 cases (600 breast, 600 gastrointestinal (GI), 600 skin, 200 renal) was chosen to obtain precise estimates of percentage clinical management concordance (CMC), meaning identical diagnoses plus differences which do not affect patient management. Cases were examined by 4 pathologists (16 study pathologists across the 4 specialty groups), using LM and DP, with the order randomly assigned and 6 weeks between viewings. Random effects (RE) logistic regression models for estimating percentage CMC included crossed RE terms for case and pathologist. Findings were interpreted with reference to 98.3% CMC.

Results: 2024 cases (608 breast, 607 GI, 609 skin, 200 renal) were recruited, including 207 breast and 250 bowel cancer screening samples. Overall LM v DP comparisons, CMC rates were 99.95% (95%CI 99.90-99.97) for all groups and 98.96 (98.42-99.32) for cancer screening samples. In specialty groups CMC for LM v DP showed: Breast 99.40% (99.06-99.62) overall and 96.27% (94.63-97.43) for cancer screening samples; GI 99.96% (99.89-99.99) overall and 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 inter-observer 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 pathologist's provide equivalent results for both routine and cancer screening samples irrespective of the modality used.

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 around 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 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 means of examining pathology slides has been termed Digital pathology (DP) and has increased rapidly over the past decade.[5] DP allows remote viewing of slides, thereby allowing work to be moved easily between pathologists, either to assist flow, provide multi-disciplinary, 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] DP hitherto has been used for teaching and external quality assessment[12] but use in routine reporting of slides has only been delivered recently in 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 (less 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 were performed at a single institution, and without enrichment for challenging cases or samples from cancer screening programmes, leading to a lack of data supporting the use of digital pathology 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.

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

In this study, [24,25] we performed a multi-site comparison of breast, gastrointestinal (GI), skin and renal specialties 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 being intra-observer and inter-observer 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 College of American Pathologists. [26,27] A blinded crossover comparison compared pathologist's reports using LM and DP. Health Research Authority (National Health Service, UK) approved the study protocol and any subsequent amendments. The study protocol was published on the International Traditional Medicine Clinical Trial Registry.[25] The steering committee, including an independent chair, the chief investigator and patient representatives, provided study oversight .

Pathologists

Sixteen pathologists, all NHS consultants with 3-35 years experience worked in specialty 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. [26]

Sample selection

Prospective consecutive histopathology samples were recruited across the four sub-specialty areas including breast and bowel cancer screening biopsies. These were enriched with 20% cases considered either difficult or moderately difficult to report (see supplementary data). [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. The remaining groups 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 eosin (H&E), special, immunocytochemistry and immunofluorescence stains were included in the study except GI where only H&E stains were included.

Cases were excluded if:

- missing or damaged slides
- contained oversized slides
- faded slides or poor staining
- where a prior biopsy review was required for interpretation

Slides were scanned and viewed using proprietary equipment provided by Philips Eindhoven, Netherlands and 3D HISTECH Budapest, Hungary as detailed in the supplementary data.

Reporting of samples

Pathologist 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 UK NHS Bowel and Breast Cancer Screening programme and RCPath. minimum datasets 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 7-point Likert scale, from least confident to most confident.[26]

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 if the differences identified would more likely 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 by 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 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 intra-observer inter-modality clinical management concordance (CMC, identical diagnoses plus differences clinically insignificant differences) comparing pairs of LM and DP reports by the same pathologist, and inter-pathologist CMC across 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 is 40%-70% found in literature), and 75% for moderate cases (midpoint between routine and difficult).[23] Taking account for enrichment with difficult and moderately difficult cases the baseline intra-modality variability of the whole study sample was defined as 90%.
The study sample size was determined so that it was sufficient to analyse each specialty separately. Based on the precisions of intra-observer inter-modality

percentage CMC estimates, target recruitment was 2000 cases; 600 cases for each of breast, skin and GI specialties, and 200 cases for renal.

Four comparisons arising from four pathologists diagnosing 600 cases within the breast, skin and GI specialties 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 corresponds 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, for renal, the margin of error 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 intra-observer inter-modality 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.[28,29]

Additionally, using these models, ICC to estimate inter-observer agreement, first within LM and then within DP, was computed as

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

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 by using a RE generalised Poisson model with crossed RE terms for case and pathologist fitted using the "glmmTMB" package in R.[30]

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

Results

Characteristics of cases
A total of 2024 cases (62.8% female 37.2% male), enrolled between July 2019-July 2021, comprised 608 breast, 607 GI, 609 skin and 200 renal samples (Table 1 & Consort diagram fig 3). The four pathologists' reading reports on LM and DP resulted in 16,192 case readings and 8,096 comparisons in three possible combinations: LM vs DP, LM vs GT, DP vs GT, totalling 24,288 comparison combinations, excluding RD.

Primary outcome results
Reports' comparison data are summarised in Table 2. RE logistic regression model of the 8096 LM vs DP comparisons showed, over all 2024 cases, CMC between LM and DP was 99.95% (95% CI 99.90, 99.97) (Table 3). This primary endpoint result exceeds the pooled percentage CMC (98.3%) in a recent meta-analysis. [23] High CMC was also observed within the 4 specialty 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 CMC are very close so that one modality does not outperform the other in diagnosis accuracy (Table 3). Both modalities also have similar inter-observer agreements which, except for moderately difficult, difficult and breast screening cases, are very high with intra-class correlation (ICC) above 0.8 (Table 3).

Secondary outcomes
RE logistic regression models results for CC i.e. any difference regardless of clinical relevance, are given in supplementary data table 7. All LM-DP percentage CC (intra-observer 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.

Pathologists reported the highest confidence level in 88% of the diagnoses (Table 5). Within a modality, GI pathologists were most confident with their diagnoses closely followed by skin pathologists while renal pathologists were noticeably less confident compared to the other specialties' pathologists. Skin pathologists had approximately same level of confidence on LM and DP diagnoses whilst for the rest of the specialties and overall, confidence of DP diagnoses was slightly less than confidence of LM diagnoses. RE generalised model showed that, overall, lower confidence in DP diagnosis was borderline significant (rate ratio=0.92, 95%CI 0.85-1.00, p=0.053)(Table 6). 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 and supp data table 8). The renal differences, to be examined in a separate paper, are not discussed.

In all three specialties inter-pathologist differences appear similar in the comparisons, LM v GT and DP v GT and higher than intra-observer inter-modality differences LM v DP.

In breast slightly higher numbers of differences were seen in B5a v B5b microinvasion on DP (10) in comparison to LM (4). Three of the 10 DP differences the pathologist gave the same diagnosis on LM as they did for DP. In the 7 remaining cases 4 cases were reported as showing no invasion where the GT concluded invasion was present and 3 cases were reported as showing invasion where the GT concluded no invasion was present.

Slightly higher intra-observer inter-modality than inter-pathologist difference was seen in the in B2 v B3 (with atypia) (31) LM v DP as compared to either LM (20) or DP (19) to GT. The 31 intra-observer differences were equally divided between LM (15) and DP (16) in equal agreement with GT.

GI showed 31 instances where discrepancy between high grade and low grade dysplasia was recorded. Of these 21 LM and 27 DP diagnoses were different to GT. 14 LM & 19 DP diagnoses showing low grade dysplasia where GT was high grade as opposed to 7 LM and 9 DP showing high grade dysplasia and GT recorded low grade.

Discussion

This study has 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 8 supp data). However this is the first study to also measure inter-observer agreement on the same cases, demonstrating inter-observer performance is identical with DP and LM as measured by agreement to consensus GT. The study shows near identical results between the DP and LM platforms across all the specialty 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. This is more likely with difficult lesions with which this study was enriched.[16,18–22] 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 nearly identical 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.[13,23,31,32] Examining further for trends in these and other areas revealed nearly identical patterns across both DP and LM modalities. For example failure to recognise *H. pylori* in gastric biopsies was seen 6 times in LM and 7 times in DP, gastric amyloidosis was missed by two pathologists on both LM and DP reports. There were only single instances of *G. 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 areas where differences between reports are common, and further examination showed no consistent trend with the modality. In the GI group of cases 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. 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 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 sub-cellular objects may indeed be better on LM and it is possible this could account for the trend towards greater confidence in LM than DP seen. However the advantages DP offers can still be fully exploited whilst retaining the undoubted superiority that LM may have for some tasks. A timely reminder, if it were needed, to laboratories to ensure 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 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.[3] In breast cancer screening comparison between LM v DP for CMC was 96.27% which is very high but slightly below the reference of 98.3%. However the comparison to the GT for these samples shows slightly better agreement seen with DP (99.89) as opposed to LM (97.57), indicating, along with the lower inter-class correlation scores, these variances are more likely to be due to differences in interpretation of challenging biopsies than the modality. Reporting cancer screening cases is based on the same principles regardless of the tumour site, so there is every reason to believe the results presented here will translate to other cancer screening samples such as uterine cervix and lung. 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 DP is equivalent to LM for these samples, and should help healthcare providers embrace the opportunities DP offers to re-design and strengthen the service and give 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. In common with previous studies our results show excellent correlation between LM and DP including in cancer screening samples, providing definitive evidence that pathologists give equivalent results regardless of the modality used.

Study Group

First names	Surnames	Affiliation
S	McIntosh	Queen's University Belfast,
G	Moran	Nottingham University Hospital NHS Trust
J	Ortiz Fernandez-sordo	Nottingham University Hospital NHS Trust
NM	Rajpoot	Tissue Image Analytics Centre University of Warwick
B	Storey	Oxford University Hospitals NHS Foundation Trust
E	Chadwick	Nottingham University Hospital NHS Trust
I	Ahmed	University Hospitals Coventry and Warwickshire NHS Trust

Note all these authors wish to be indexed to this paper in PubMed.

Funding
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.

DRJS also reports funding through PathLAKE funded by Innovate UK through Industrial Strategy Fund reference 18181.

Ethics approval
The study protocol and any subsequent amendments were reviewed and approved by Health Regulation Authority (HRA) and Research Ethics Committee (REC). ISRCTN Number 14513591, IRAS Number 258799.

Declarations of interest
D Snead and N Rajpoot report they are co-founders, directors and shareholders of Histofy Ltd, a start-up company developing artificial intelligence algorithms for digital pathology.
D Snead has also worked in the past as a member of Philips computational pathology advisory board.
YW Tsang reports she is a shareholder in Histofy Ltd.
E Hero and H Evans report they working ad-hoc part time sessions for Histofy Ltd.

Authors 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 truth, analysis of clinical differences, manuscript preparation.
Warwick Clinical Trails Unit (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.

Data sharing

Will individual participant data be available (including data dictionaries)? Yes

What other documents will be available? Study protocol

With whom? PathLAKE data lake, University Hospitals Coventry and Warwickshire. https://www.pathlake.org email pathlake@uhcw.nhs.uk

By what mechanism will data be made available? Application to the PathLAKE access committee via the PathLAKE website or by email

Available from: the date of publication indefinitely.

Research in context panel

Evidence before this study -A review and meta analysis to analyse existing published studies on the diagnostic application of digital pathology (DP) to compile a comprehensive evaluation of the safety and reliability of DP for routine diagnosis (both primary and secondary) was conducted to Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) guidelines. [23] The review protocol was registered in the PROSPERO database (registration number CRD42019145977. This was based on a search of the following databases: PubMed, including Medline, EMBASE, Cochrane Library and Google Scholar between 2013 and August 2019. To identify any study being currently undertaken, a search of ClinicalTrials.gov (U.S. National Institutes of Health, Maryland) was performed. A detailed search strategy is available as online supplementary content (online supplementary appendix 1). To identify any other eligible articles a manual search was conducted via forward citation tracking and reference search of the included studies. Search terms used were digital pathology OR whole slide imaging AND light microscopy, OR validation OR comparison OR reporting. 994 records were retrieved 828 were screened using Quality Assessment of Diagnostic Accuracy Studies (QUADAS 2) (tool to evaluate the quality and risk of bias in each individual study) identifying 45 eligible studies.

Added value of this study – This is the first study to measure both intra- and inter-observer agreement on the same cases comparing LM and DP, thereby measuring differences between modalities against differences which occur anyway through inherent intra- and inter-observer variation, and the first study to examine cancer screening samples and renal biopsies with immunofluorescent stained sections.

Implications of this study – Pathologists deliver equivalent results regardless of modality. DP should be permitted for use on cancer screening samples (currently embargoed in England), and any other sample types currently examined on LM. This technology can help to address health inequalities wherever lack of access to pathologist expertise risks impacting on patient care. Service providers should note of the enormous potential DP offers to deliver results faster results, provide access to peer and expert review, and out of hours support. This is particularly so for complex samples such as renal and transplant biopsies.

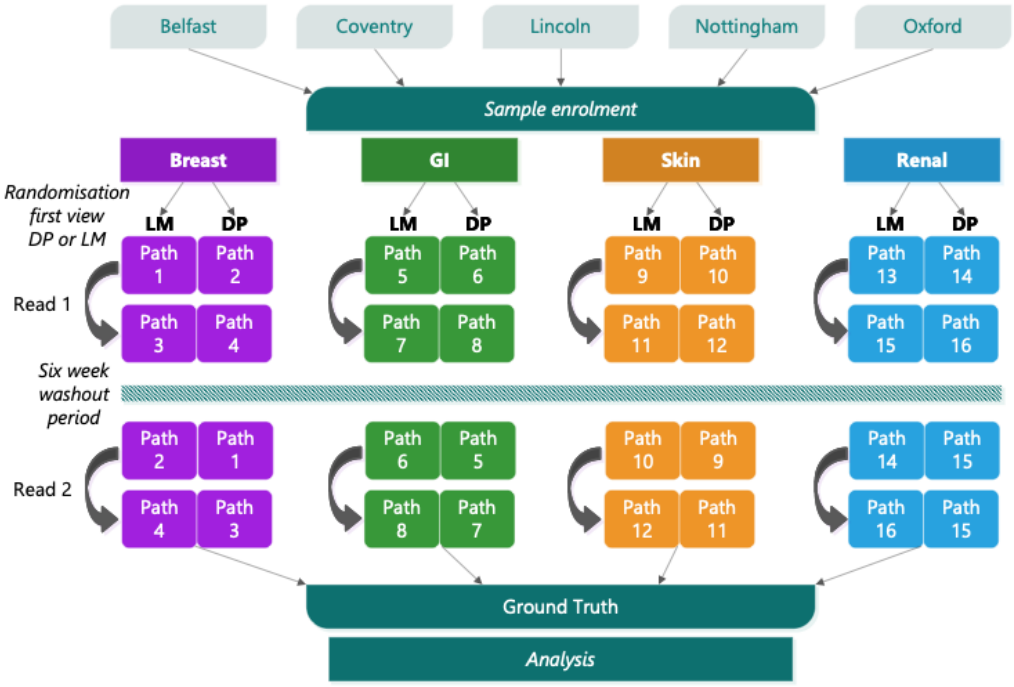

Figure 1: Study overview. Cases were recruited from participating sites in the four specialty groups, anonymised and enrolled into the study. In each group each case was examined by each pathologist twice using light microscopy (LM) and digital pathology (DP) respectively. The sequence of whether LM or DP was performed first was randomised and there was a six 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.

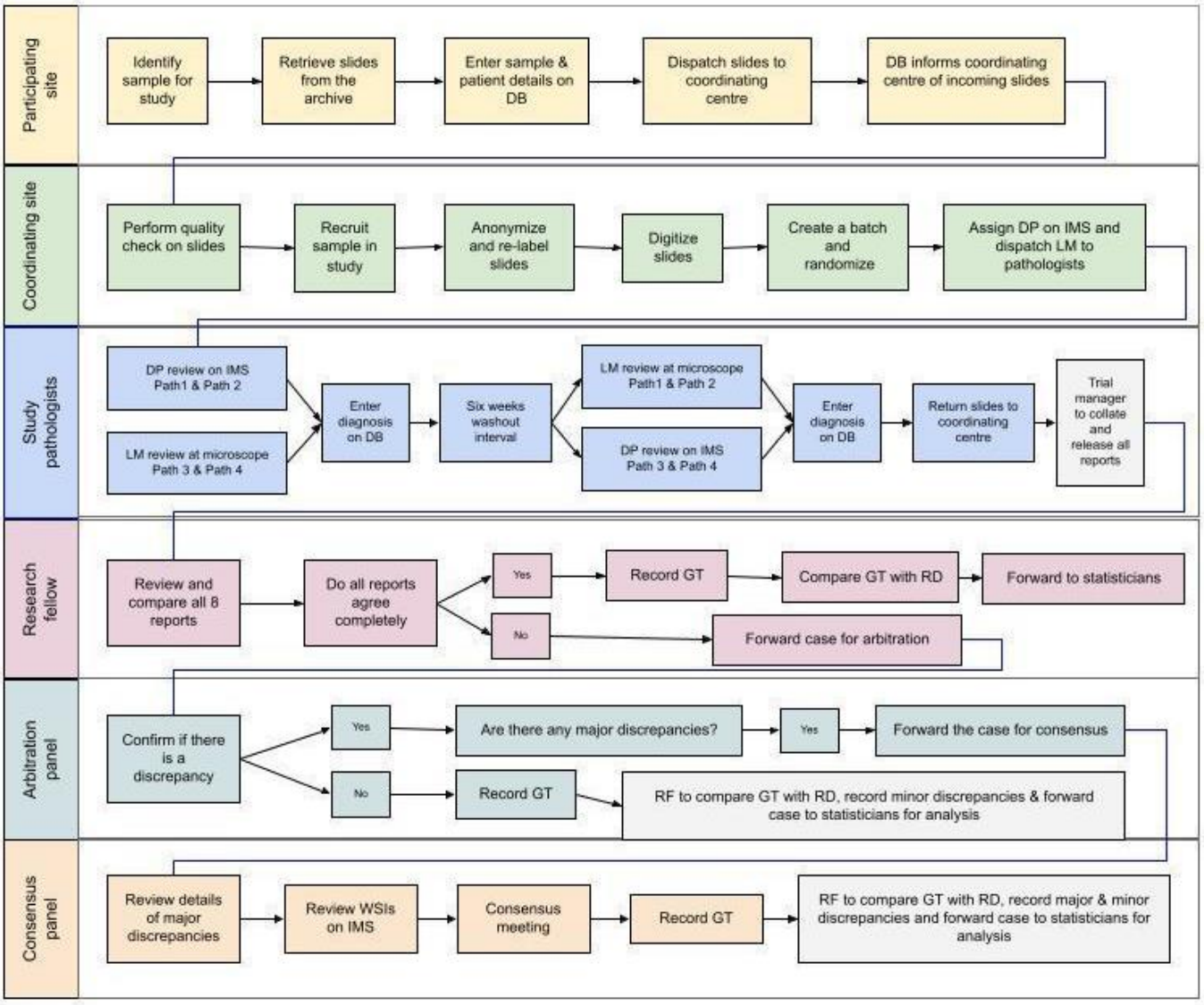

Figure 2: Overall study workflow, reports review, arbitration and consensus process.

Abbreviations: DB = database, DP = digital pathology, LM = light microscopy, GT = ground truth, RD = reference diagnosis

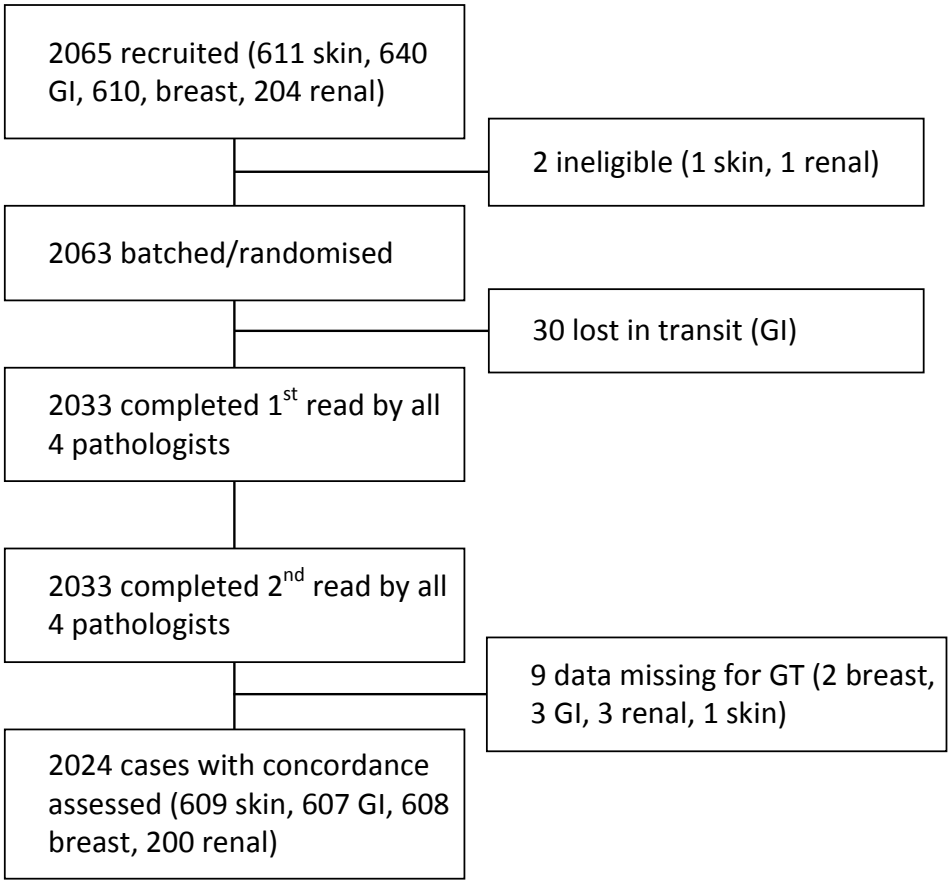

Figure 3: Consort diagram of cases entered into the study.

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 specialty 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 (%)	NA			NA	NA
Yes		207 (34.0)	250 (41.2)		
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)

Min = minimum; Max = maximum; LQ = Lower quartile; UQ = upper quartile

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)

Outcome	All cases (N=2024)	Breast (N=608)	GI (N=607)	Skin (N=609)	Renal (N=200)
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)
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 3: Summary of the clinical management concordance (CMC) analysis using RE logistic regression models

Cases included in the analysis	Percentage CMC (95% confidence interval)			Intra-class correlation coefficient (ICC)	
	Intra-observer LM v DP agreement	**LM v GT agreement**	**DP v 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 4 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 (see Table 2).

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)

Table 5: Comparison of diagnosis confidence data using RE generalised Poisson models

Data included	**Rate ratio (95% CI), p-value**
All the data (all pathologists and all specialties) (n=2024)†	0.92 (0.85-1.00), 0.053
Subgroup analysis by specialty	
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 specialties (n=1447)†	0.86 (0.76-0.98), 0.024
Moderate cases from all specialties (n=164)†	1.32 (1.00-1.75), 0.052
Difficult cases from all specialties (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 is the number of cases. Each case is reported by 4 pathologists on both LM and DP and so the number of rows for each case in the analysis is 8n. In the entire database, only five reports (out of 16,192 reports) had missing diagnosis confidence data.

Table: 6 Errors recorded in two or more instances in Breast, GI and Skin specialties.

Breast Difference type	**All**	**LM v GT**	**DP v GT**	**LM v DP**	**Screening cases**
Tumour type	56	37	37	39	13
B2 v B3	48	37	29	30	12
B2 v B3 with atypia	35	20	19	31	16
B2 v B1	26	15	18	19	15
B3 with atypia vs B5a	16	13	8	11	10
B5a v B5a mi	12	5	11	8	11
B3 with atypia v B3 no atypia	8	8	8	0	5
B5a v B5b	8	5	5	6	5
B3 with atypia vs B3	7	2	5	7	2
B4 v B5a	3	3	2	1	2
B2 v B4	2	2	1	1	1
B2 v 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 v GT**	**DP v GT**	**LM v DP**	**Screening cases**
HP v SSL	37	29	26	21	31
LGD v HGD	32	22	28	14	14
Tumour stage	13	10	9	6	1
Normal v HP	12	10	9	5	10
Missed H pylori	8	6	7	3	
TA v SSL	7	7	5	3	7
Normal v BA2	5	4	5	1	

TA v TA LGD	4		4	4	4
Inflammation NOS v IBD	4	4	3	1	1
Inflammation v LGD	3	3	3		
Quiescent v active colitis	3	3	3		
Inflammation v indefinite for dysplasia	3	2	2	2	
Gastritis v amyloidosis	2	2	2		
Normal v fundic polyp indefinite for dysplasia	2	2	2		
Quiescent v IBD NET	2	2	2		
Reactive v TA	2	2	2		
Reported incorrect case	2	2	2	1	2
TA v HP	2	2	2	2	
Tumour type	2	2	2		
Barretts v indefinite for dysplasia	2	1	2	1	
Normal v IEL	2		2	2	
TA vs polyp cancer	2		2		
Normal v non-specific inflammation	2	2	1	1	
Inflammation v IM	2	1	1	2	
Total	155	118	126	69	70
Skin Difference Type	**All**	**LM v GT**	**DP v GT**	**LM v DP**	
BCC with high risk component v BCC	18	9	11	16	
MM v naevus	11	8	6	8	
SCC margin involvement v no margin involvement	10	8	7	5	
BCC v SCC	6	3	6		
SCC v AK or IEC	6	5	5	2	

Breslow thickness	5	3	4	3	
Blue naevus v atypical naevus	5	4	3	5	
KA v SCC	5	5	3	2	
in-situ v invasive melanoma	5	4	2	4	
Melanoma margin involvement	4	3	4	1	
Adenoid cystic carcinoma v benign adnexal tumour	2	1	2	1	
DFSP v DF	2	1	2	1	
Herpes v alternative inflammatory lsesions	2	1	2	1	
Lichenoid keratosis v compound naevus	2	2	2	0	
Bowens disease v stasis	2	2	1	1	
Metastatic melanoma v benign node	2	2	1	1	
Viral wart v polyp	2	1	1	2	
Total	89	62	62	53	

Abbreviations:
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

v versus

Table 7: comparison of this study with other multi-site validation studies previously published in the literature

Study ID	Tabata et al. 2017[20]	Mukhopadhyay et al. 2017[19]	Borowsky et al.[22]	Babawale 2021[21]	This study
No of participating sites	12	4	5	7	6
No of cases	900	1,992	2,045	3,001	2,024
Total study readings	2,140	15,925	15,031	3001	16,192
Number of DP / LM reading pairs	1,070	7,964	7,509	3,001	8,096
Washout interval	>2weeks	Min 4 weeks	Min 31 days	No washout period	Min 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
Intra-observer concordance	Yes	Yes	Yes	No	Yes
Inter-observer concordance	No	No	No	Yes	Yes
DP vs LM clinical concordance	99.2%	95.1%	96.36%	97.1%	99.95%

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**(55): 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: 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). *Bulletin of the Royal College of Pathologists* 2018; **182**: 78-86.
4. Harris. Digitisation will transform the future of pathology. *Br J Health Care Manag* 2020; **26**.
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**(11).
7. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. *Histopathology* 2012; **61**(1): 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**(2): 221-8.
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**(7): 443-7.
10. Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence-the third revolution in pathology. *Histopathology* 2019; **74**(3): 372-6.
11. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol* 2019; **20**(5): e253-e61.
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**(2): 293-6.
13. Snead DR, Tsang YW, Meskiri A, et al. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* 2016; **68**(7): 1063-72.
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**(12): 1107-11.
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**(5): 784-94.
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**(5): 621-35.

17. Baidoshvili A. How to go digital in pathology. https://www.philips.com/c-dam/b2bhc/master/sites/pathology/resources/white-papers/labron-how-to-go-digital.pdf: LabPON Laboratorium Pathologie Oost-Nederland, 2016.
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**: 662-71.
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**(11): 547-54.
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**(10): 1245-53.
23. Azam AS, Miligy IM, Kimani PK, et al. Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis. *J Clin Pathol* 2020.
24. NIHR. Multi-centred validation of digital whole slide imaging for routine diagnosis. 2018.
25. ISRCTN. Is the use of digital pathology in routine diagnosis reliable and safe in comparison to standard microscopy? 2018; (ISRCTN14513591).
26. Cross S, Furness P, Igali L, Snead D, Treanor D. Best practice recomendations for implementing digtial pathology. Royal college of pathologists: RCPath, 2018.
27. 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**(12): 1710-22.
28. Wood S, Scheipl F. Estimate generalized additive mixed models via a version of function gamm() from 'mgcv', using 'lme4' for estimation. 2020-04-03 ed. gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'; 2020.
29. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. . 2020. https://www.R-project.org/.
30. Brooks ME, Kristensen K, van Benthem KJ, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal* 2017; **9**.
31. 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.
32. 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**(7): 968-75.