Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis

Ayesha S Azam 1,2, Islam M Miligy 3, Peter K-U Kimani 4, Heeba Maqbool 1, Katherine Hewitt 1, Nasir M Rajpoot 4, David R J Sneed 1

ABSTRACT

Background Digital pathology (DP) has the potential to fundamentally change the way that histopathology is practised, by streamlining the workflow, increasing efficiency, improving diagnostic accuracy and facilitating the platform for implementation of artificial intelligence–based computer-assisted diagnostics. Although the barriers to wider adoption of DP have been multifactorial, limited evidence of reliability has been a significant contributor. A meta-analysis to demonstrate the combined accuracy and reliability of DP is still lacking in the literature.

Objectives We aimed to review the published literature on the diagnostic use of DP and to synthesise a statistically pooled evidence on safety and reliability of DP for routine diagnosis (primary and secondary) in the context of validation process.

Methods A comprehensive literature search was conducted through PubMed, Medline, EMBASE, Cochrane Library and Google Scholar for studies published between 2013 and August 2019. The search protocol identified all studies comparing DP with light microscopy (LM) reporting for diagnostic purposes, predominantly including H&E-stained slides. Random-effects meta-analysis was used to pool evidence from the studies.

Results Twenty-five studies were deemed eligible to be included in the review which examined a total of 10 410 histology samples (average sample size 176). For overall concordance (clinical concordance), the agreement percentage was 98.3% (95% CI 97.4 to 98.9) across 24 studies. A total of 546 major discordances were reported across 25 studies. Over half (57%) of these were related to assessment of nuclear atypia, grading of dysplasia and malignancy. These were followed by challenging diagnoses (26%) and identification of small objects (16%).

Conclusion The results of this meta-analysis indicate equivalent performance of DP in comparison with LM for routine diagnosis. Furthermore, the results provide valuable information concerning the areas of diagnostic discrepancy which may warrant particular attention in the transition to DP.

INTRODUCTION

Digital pathology (DP), frequently referred to as whole slide imaging (WSI), is a rapidly emerging technology. 1 2 It has the potential to fundamentally change the way that histopathology is practised by streamlining the workflow, 3 increasing efficiency, 4 improving diagnostic accuracy and facilitating the platform for the implementation of artificial intelligence–based computer-assisted diagnostics. 5 Inevitably, these advantages will dictate the adoption of this technology which is very likely to gather pace rapidly in the next few years.

Currently, only a small number of pathology laboratories in the UK and elsewhere are using DP for their routine sign-out purposes. 6 For more widespread adoption of DP in clinical laboratories, evidence of safety and reliability is needed, ideally in the form of adequately powered multi-site validation studies, demonstrating equivalent performance of DP compared with the existing gold standard of light microscopy (LM). 7

A number of validation studies have been published to date, but most are small single-site studies and there is considerable variation in the study designs. Previous systematic narrative reviews have summarised the qualitative evidence on the diagnostic reliability of WSI. 8–10 A systematic review of 38 validation studies between 1999 and 2015 reported an overall diagnostic concordance ranging from 63% to 100%, with a weighted mean of 92.4%. 3 That review recognised the limitation of small sample size (mean number of cases 140) with variable study design and case types. A subsequent study based on the systematic review of the discordant diagnoses reported 333 discordances (4%) among 8069 comparisons of digital and LM diagnoses. A significant proportion of those discordances were concerning the diagnosis of dysplasia (32%). 9 Araújo et al conducted a systematic review of studies from 2010 to 2017 reporting intra-observer agreement ranging from 87% to 98.3% (k coefficient range 0.8–0.98). 10 That review was again based on a small selected series of 13 studies with a mean sample size of 165. Studies comparing the digital diagnosis with either consensus diagnosis or original diagnosis by a different pathologist were not included in this review.

Since the publication of these reviews, larger validation studies have been performed including studies supporting regulatory approval and development of current recommendations by national organisations, providing more guidance and practical advice on the validation process. 11,12

A meta-analysis to demonstrate the combined accuracy and reliability of DP is still lacking. In the hierarchy of evidence-based healthcare, systematic reviews and meta-analysis using statistical methods to combine the results of individual studies can provide a precise estimate of the effect of size with considerably increased statistical power, placing them at the apex of the evidence pyramid.
The aim of this systematic review and meta-analysis was to review the quantitative evidence across the validation studies, synthesise statistical data and to summarise the evidence on safety and reliability of DP.

**MATERIALS AND METHODS**

**Review protocol and registration**

This review was conducted in accordance with the guidelines by the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA). The review protocol was registered with the PROSPERO database (registration number CRD42019145977: Centre for Reviews and Dissemination, University of York, England), the international prospective register of systematic reviews.

**Problem statement**

To review the existing literature on the diagnostic (primary and secondary) use of DP and its comparison with LM in the context of the validation process.

- Uncover the strength of the concordance evidence between DP and LM, in order to highlight the usefulness of transition to DP.
- To analyse recent evidence on safety and reliability of DP by identification of discordant diagnosis on the digital platform.

**Definitions**

**Concordance and discordance**

Diagnostic concordance was defined as ‘degree of agreement between digital reading and the LM reading for the same sample’. Conversely, any difference or variance between the digital and LM report would reflect discordance. Intra-observer concordance was the preferred method of evaluation, where possible, as per CAP and RCPath guidelines.

Minor and major discordance

Minor discordance reflects a difference between two reports which is clinically insignificant and would not affect patient management decisions. However, a major discordance leads to or can lead to a difference in clinical decision for patient management.

**Overall concordance**

For this review, the overall clinical concordance was also recorded, which included concordance as well as minor clinically insignificant discordances.

**Search strategy**

A literature search was conducted by the primary researcher (ASA) through the key electronic databases: PubMed platform (National Center for Biotechnology Information, U.S. National Library of Medicine, Maryland) including Medline (Medline Industries, Illinois, USA), EMBASE (Elsevier, Amsterdam, The Netherlands), Cochrane Library (London, England) and Google Scholar (Google, California) 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).

In order to identify any more potentially eligible articles not captured through the aforementioned search, a manual search was conducted via forward citation tracking and reference search of the included studies.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Inclusion and exclusion criteria for screening of literature search results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inclusion criteria</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Digital pathology (DP) and light microscopy comparison/validation studies</td>
<td>Studies involving other uses of DP</td>
</tr>
<tr>
<td>For diagnostic purpose (primary or secondary)</td>
<td>including education, research, molecular or image analysis</td>
</tr>
<tr>
<td>Primarily using H&amp;E slides</td>
<td>Predominantly involving immunocytochemistry, special stains, fluorescence or frozen section slides</td>
</tr>
<tr>
<td></td>
<td>Cytopathology, autopsy, neuropathology</td>
</tr>
<tr>
<td></td>
<td>Not using whole slide images (telepathology, robotic or static microscopy)</td>
</tr>
</tbody>
</table>

**Article screening and eligibility evaluation**

Using Rayyan Qatar Computing Research Institute (QCRI), all results were screened against a predefined eligibility criterion (table 1) including full abstract and study title by two independent reviewers (ASA and HM).

Studies were included provided they met the full inclusion criteria: studies comparing DP with LM reporting for diagnostic purposes, predominantly including H&E-stained slides. Studies were excluded if they explored applications of DP other than diagnosis, predominantly involving ancillary studies or other sub-specialist areas.

The screened results were displayed under one of the following categories: ‘included’, ‘excluded’ and ‘maybe’. The reasons for exclusion were also recorded.

Any disagreements highlighted by Rayyan QCRI between the two reviewers were resolved by discussion. Articles in the ‘maybe’ category were further reviewed by a third reviewer (DRJS). Full texts of all potentially eligible articles were retrieved and reviewed in detail for further evaluation.

**Data extraction**

A comprehensive data extraction protocol was developed based on the Cochrane Effective Practice and Organization of Care template. In addition to the generic domains adapted from the Cochrane—good practice data collection, domains relevant to this review were added. A tailored data record form was designed in Microsoft Excel (V.16.38).

Data extraction was conducted by ASA and supported by other reviewers (DRJS, HM, IMM, KH, NMR). In case of discrepancy between two reviewers, a consensus was reached by discussion. For each included article, the following data items were extracted: study information, participants, interventions, sample, study method, outcome variables and quality assessment (box 1). Corresponding authors of the included studies were contacted to request any further details, where required.

**Quality assessment (QUADAS 2—tailored)**

To assess the quality and risk of bias in individual studies, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS 2) tool was used. QUADAS 2 was tailored according to this review’s protocol and some of the signalling questions not applicable to the validation studies were excluded. For each of the signalling questions, clear and precise instructions were produced. For each QUADAS domain, the risk of bias was assessed as either ‘low’ or ‘high’ based on the answers to the signalling questions. If insufficient information was provided in the study, the risk of bias was assessed as ‘unclear’.
Box 1 Data collection form with details of domains recorded

General Information
Study ID, study title, year of publication, country of origin, single/multicentre, retrospective/prospective/crossover and journal reference.

Participants
Number of pathologists involved, consultant or trainee level or both, and training provided to pathologists.

Interventions
Whole slide imaging scanning system manufacturer/model, scanning magnification, resolution of viewing monitor and calibration.

Samples
Sample size, total number of slides, total number of readings, number of readings per case, total number of reading pairs, stains evaluated (H&E, Immunohistochemistry (IHC), special), single/multiple specialties included, specialties involved, case selection criteria and type of specimen; biopsies/resections/both.

Methodology
Sample enrolment, randomisation, anonymisation, degree of blinding, length of washout time, intra-observer/inter-observer, provision of clinical details, analysis strategy.

Results
Number of concordant cases, concordance percentage, number of discordant cases, discordance percentage, minor discordance, major discordance, concordance+minor discordance, 95% CI/ Kappa index, details of discordance, preferred modality, reasons for discordance and discordance category.

The quality assessment of all included studies was performed by two reviewers. A review specific tailored QUADAS-2 tool is available as online supplementary content (online supplementary appendix 2).

Statistical analysis
For each study, we recorded the number of DP–LM comparisons and the number of comparisons where DP and LM diagnoses agreed (number of agreements). We considered two definitions for agreement (concordance): no difference in clinical management (overall concordance) and complete agreement. In three studies,18–20 DP–LM comparison for each case was made by different samples, analysis and results.23 The two phases of the same research paper were recorded separately into the total number of included studies. The 25 research papers included in the systematic review were based on the evaluation of total 19 468 LM versus digital comparisons. The quantitative meta-analysis included 24 of these studies.

RESULTS
PRISMA flowchart
Figure 1 shows the PRISMA flowchart summarising the results of the review process. The initial systematic search of the literature yielded 994 records in total. After removing the duplicate results, abstracts of 828 records were screened for eligibility.

Eligibility screening identified 45 research studies for full-text review of these articles of which 25 were deemed eligible for inclusion in the review. These studies included a total of 10 410 histology samples with an average sample size of 176. The third reviewer was needed to reach consensus in 8/45 articles. One publication incorporated two distinct study phases with different samples, analysis and results.23 The two phases of the same research paper were recorded separately into the total number of included studies. The 25 research papers included in the systematic review were based on the evaluation of total 19 468 LM versus digital comparisons. The quantitative meta-analysis included 24 of these studies.

Study demographics
Twenty-three studies had been conducted at single centres (92%) and only two (8%) were multi-centre validations. The majority of studies (14; 56%) were from America, 10 (40%) from Europe and 1 (4%) was from Asia. The majority of these studies (56%, n=14) were published from 2015 onward.

Study characteristics: samples, participants, training, washout time and equipment
Sample size (figure 2) varied from 6024 to 301725 cases with an average of 176 cases. The majority (15; 60%) examined 200 samples or less with only 3 studies examining 1000 or more cases. The largest sample size reported in validation literature was 3017 to date.24 Single specialty samples were selected in 48% (n=12) of the studies, whereas more than one specialty samples were included in 52% (n=13) studies (figure 3). None of the studies stated the inclusion of cancer screening samples.

The number of pathologists who participated in reporting study samples ranged from 1 to 57 with 15 (60%) studies involving...
fewer than 5 pathologists. In one study, 57 pathologists participated in reporting study samples (n=200) with 28 of them being residents/fellows. The remaining 23 studies included experienced consultant pathologists while one study did not document the experience level of the participating pathologists.

The participating pathologists were provided training for using the WSI system before reporting the study samples in 13 studies (52%). No training was provided in four studies because the participants had previous experience with using DP for reporting, teaching or tumour boards. Eight studies did not state whether training was provided or not.

Clinical information was provided to the reporting pathologists with each case in the majority of the studies (80%, n=20). Five studies did not state this.

The length of washout time between DP and LM reporting was variable in the included studies and ranged from 2 weeks to more than a year (figure 4). Four studies (16%) did not use any washout time between the two readings due to the live-validation approach.

The whole slide image scanning devices used in the included studies comprised seven different scanner manufacturers, Aperio (Leica biosystems) being the most commonly used (n=14, 56%). Nine studies had performed slide scanning at ×40 magnification, eight at ×20, six used a combination of ×20 and ×40, and one study used a combination of ×40 and ×60 depending on the sample requirement. Viewing monitor resolution was not commented on in 10 studies. The details of the scanning systems and scan magnification are shown in table 2.

Only five studies performed a prior sample size calculation by statistical methods. The sample size for three studies was based on non-inferiority tests, but with different non-inferiority margins and percentage agreements between DP and LM. Snead et al collected the baseline multi-disciplinary team meeting review data to calculate overall intra-observer and inter-observer concordance on ‘LM’ and reached a sample size of just over 3000 cases.

A summary of main characteristics of included studies is available online as online supplementary material (online supplementary appendix 3).

Diagnostic concordance and discordance

The overall concordance was assessed by measuring complete agreement along with clinically insignificant variations between digital and LM reports. Individual studies percentage agreement across 24 studies included in the meta-analysis ranged from 92.3% to 100%, with the majority (23/24) having percentage agreement above 95% and three studies having 100% agreement (figure 5). The pooled percentage agreement for overall concordance was 98.3% (95% CI 97.4 to 98.9) across 24 studies. One study used the kappa coefficient (k=0.81) and did not state the concordance percentage. The pooled percentage agreement for complete concordance was 92% (95% CI 87.2 to 95.1) (figure 6).

The studies were heterogeneous (I²=90%, p<0.0001). The inter-modality (ie, digital and light microscopy) readings were performed by the same pathologist in 23 (92%) studies. In one study, two-thirds of cases were examined by different pathologists and one-third by the same pathologist. In one study, the two modalities were examined by different pathologists. A total of 546 major discordances were reported across 21 studies. Three studies reported only minor differences in results. One study reported the kappa coefficient and did not state the discordance percentage.

Gastrointestinal tract and gynaecological pathology were two most commonly reported specialties among the discordant cases, followed by skin, breast, genitourinary and renal pathology.

Categorisation of diagnostic discrepancies

Out of 546 major discordances reported across the 25 studies, details of diagnosis and preferred modality were provided for 158 instances. In order to identify and summarise areas of difficulty in the diagnostic performance of WSI, we categorised all the major discordances into four main groups based on the underlying diagnostic discrepancy. Over half (57%) of the reported discordant diagnoses were related to the assessment of nuclear atypia, grading of dysplasia and malignancy (group A). These
were followed by challenging diagnoses (26% in group C) and identification of small objects (16% in group B). Each category was further sub-classified into three to seven sub-classifications to capture the nature of discordance. This categorisation was based on the findings of this review and previously reported reviews.9 10

Within group A, a total of 68 discordant instances concerning the grading of epithelial dysplasia and nuclear atypia were recorded. The preferred modality for these instances was LM, the DP under-called (ie, lower grade of dysplasia compared with LM) in 30 of 48 (62.5%) and over-called (higher grade of dysplasia) in 18 of 48 (37.5%).

Table 3 shows each category with nature of discordance and organs/sites involved as well as the percentage of all reported major discordances. Figure 7 shows the distribution of four groups of discordances across the specialties involved.

The remaining underlying reasons for disagreement were stated as follows: difficult case requiring consultation, textural quality of amyloid hard to appreciate on a digital display, lack of clinical information and non-availability of ancillary stains.

Risk of bias and applicability

The results of quality assessment for risk of bias and applicability in individual studies are displayed in table 4. Across the four domains (case selection, index test, reference standard, flow and timing), the percentage of studies with low risk of bias ranged from 75% (20/25) to 92% (23/25), the percentage of studies with a high risk of bias ranged from 8% to 25%. An unclear risk of bias was found in 12% of the studies. Regarding applicability, all studies showed low concern across case selection domain. Applicability concern was classed as high across the index test domain in one study.

DISCUSSION

This is the first meta-analysis of DP studies and largest systematic review to date, based on 25 studies covering 10 412 samples and 19 468 glass versus digital comparisons, including from two recent multi-centre validation studies.31 32

These studies demonstrated percentage agreement of 98.3% (95% CI 97.4 to 98.9) for overall concordance and 92% (95% CI 87.2 to 95.1) for complete concordance. This high level of agreement across multiple studies provides strong evidence that DP is a viable alternative to LM for routine practice and, given the multiple additional advantages it offers, can be expected to

---

**Table 2** Various scanning systems and scanning magnifications used in the included studies

<table>
<thead>
<tr>
<th>Scanner manufacturers</th>
<th>Number of studies</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ventana (Roche Diagnostics)</td>
<td>5</td>
<td>Campbell et al,56, Ordi et al,57, Reyes et al,58, Saco et al,59, Thrall et al,60</td>
</tr>
<tr>
<td>Nanozoomer (Hamamatsu)</td>
<td>4</td>
<td>Houghton et al,61, Loughrey et al,62, Villa et al,63, Tabata et al</td>
</tr>
<tr>
<td>Ultra-fast scanner (Philips Intellisite Pathology system)</td>
<td>2</td>
<td>Mukhopadhyay et al,64, Araújo et al</td>
</tr>
<tr>
<td>Omnyx VL120 (GE Healthcare)</td>
<td>2</td>
<td>Lee et al,65, Sneed et al,66</td>
</tr>
<tr>
<td>Mikroskan vs800 (Olympus Corporation)</td>
<td>1</td>
<td>Tabata et al</td>
</tr>
<tr>
<td>FINO (CLARO, Hirosaki)</td>
<td>1</td>
<td>Tabata et al</td>
</tr>
</tbody>
</table>

**Figure 5** Forest plot representing percentage agreement for overall concordance across 24 studies with the number of comparisons, participating pathologists and digital pathology training.

**Figure 6** Forest plot representing percentage agreement for complete concordance across 24 studies.
replace the LM as the main tool for diagnostic histopathology. The studies examined samples from multiple different tissue types suggesting the results are fully representative of the breadth of diagnostic material encountered in routine practice. However, it is noticeable that no studies included ophthalmic samples and too few studies examined renal or paediatric samples to enable meaningful conclusions to be drawn. None of the studies stated which samples, if any, were generated by cancer screening programmes, which has been an area of concern in the UK National Health Service cancer screening programmes (Public Health England personal communication). However, the inclusion of breast, gynaecological and gastrointestinal samples in a large proportion of these studies suggest the results should be relevant to these sample types.

The 546 discordances were analysed to determine the nature of discrepancy and its relevance to patient management. The majority (57%) of clinically significant discordances were related to grading of dysplasia, atypia and malignancy in various tissue types. Grading dysplasia is an important feature which relies largely on subjective assessments, and which is a common tissue types. Grading dysplasia is an important feature which relates diagnostic areas already known to be contentious and discrepancies all relate diagnostic areas already known to be contentious and where differences of opinion are to be expected irrespective of the modality used to examine the slides.

Some of the diagnostic discrepancies are however likely to be related to DP. These include assessment of objects requiring an appreciation of textural quality, such as deposits of amyloid, mucin and Weddellite calcifications. In these instances, the inability to focus through the ‘z’ plane means the observer using DP is unable to detect the changes in texture normally visible in LM. Second, the recognition of small objects, less than 5 µm in size, such as bacteria has been identified as area of difficulty on DP. Reproduction of these objects in DP systems is clearly inferior to LM because of a combination of loss of detail in the image acquisition, restrictions in the scanning objective and the reliance of a single focal plane. Finally, any object requiring examination under polarised light cannot be adequately catered for in the DP systems currently available. Awareness and knowledge of these issues is essential if pathologists are to understand the limitations of DP and be able to adapt to its use in their own practice safely.

Across this meta-analysis, there are potentially important variations in study design including sample selection criteria, equipment used and length of washout time that could influence the results. Where possible, these potential sources of bias have been assessed using the QUADAS 2 tool, which demonstrated a low risk of bias for the majority of studies. As expected with an emerging new technology, this meta-analysis has captured results from studies using...
differing technologies. Despite improvements in image quality, the more recent equipment can provide there was no difference in discrepancy rates over the course of these studies. This most likely indicates that even the earlier DP studies used equipment capable of delivering a diagnostic tool equivalent to LM.

Although this review provides strong and consistent evidence of the equivalent performance of DP in comparison with LM, it also highlights certain areas where the evidence is still weak. First, the majority of the studies provide no evidence base for the evidence base further. This most likely indicates that even the earlier DP studies used equipment capable of delivering a diagnostic tool equivalent to LM.

Although this review provides strong and consistent evidence of the equivalent performance of DP in comparison with LM, it also highlights certain areas where the evidence is still weak. First, the majority of the studies provide no evidence base for the evidence base further. This most likely indicates that even the earlier DP studies used equipment capable of delivering a diagnostic tool equivalent to LM.

### Table 4 QUADAS 2—assessment of risk of bias and applicability concerns across 25 studies

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Case selection</th>
<th>Index test</th>
<th>Reference standard</th>
<th>Flow and timing</th>
<th>Case selection</th>
<th>Index test</th>
<th>Reference standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al-Janabi et al10</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Bauer et al10</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Campbell et al18</td>
<td>Unclear</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Al-Janabi et al21</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Brunelli et al26</td>
<td>Unclear</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Houghton et al30</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Reyes et al30</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Bauer and Slaw3</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Ordi et al22</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Bucks et al(1)23</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Bucks et al(2)23</td>
<td>High</td>
<td>Unclear</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Arnold et al24</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>Low</td>
<td>Unclar</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Loughrey et al25</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Unclar</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Thrall et al26</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Shah et al26</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Snead et al25</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Saco et al29</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Kent et al26</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Tabata et al31</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Araújo et al26</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Lee et al32</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Villa et al33</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Mukhopadhyay et al32</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Williams et al36</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Unclar</td>
<td>Low</td>
</tr>
<tr>
<td>Hanna et al44</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
</tbody>
</table>

QUADAS2, Quality Assessment of Diagnostic Accuracy Studies.

CONCLUSION

Although the barriers to wider adoption of DP have been multifactorial, limited evidence of reliability has been a significant contributor. The COVID-19 pandemic acted as a catalyst to change working practice across the health sector. The flexibility provided by digitising the workflow is fundamental to these changes, and DP is a key step to enabling this to happen in cellular pathology. The demand for widespread adoption of DP is growing further as a result.

The results of this meta-analysis represent significant evidence to indicate the equivalent performance of DP for routine diagnosis. Furthermore, the categorisation of diagnostic discordances highlights a number of potential limitations, where alternative solutions may be needed. The findings also indicate how the enrichment of sample selection in future studies may improve the evidence base further.

Handling editor Runjan Chetty.

Twitter Ayesha S Azam @AyeshaSAzam and Katherine Hewitt @kezia1501

Acknowledgements We acknowledge the following individuals: Dr Louise Hillier (Warwick Medical School, University of Warwick) for her valuable advice on the search protocol, data collection and results analysis; Dr Lesley Anderson (Centre for Public Health, Queens University Belfast) for valuable advice on the search strategy and review protocol; Dr M Nimir for assistance with data acquisition. We thank the following individuals for their responses: Professor Paul J van Diest, Dr Darren Treanor, Professor Muhammad Illyas, Dr Bethany Williams, Dr Anne M Mills and Dr Matteo Brunelli.

Contributors Conception, design, protocol and manuscript writing: ASA and DRJS. Data acquisition and article review: ASA, IMM, HM, KH, NMR and DRJS. Data
REFERENCES


34. van der Wei MI, Duits LC, Ten Kate FF, et al. Sus365 digital microscopy is a valid alternative to conventional microscopy for diagnosing Barrett’s esophagus. Gastroenterology 2015;148:A282.


