The Diagnostic Concordance of Whole Slide Imaging and Light Microscopy

A Systematic Review

Edward Goacher, BSc; Rebecca Randell, PhD; Bethany Williams, MBBS; Darren Treanor, MB, BSc, PhD, FRCPATH

• Context.—Light microscopy (LM) is considered the reference standard for diagnosis in pathology. Whole slide imaging (WSI) generates digital images of cellular and tissue samples and offers multiple advantages compared with LM. Currently, WSI is not widely used for primary diagnosis. The lack of evidence regarding concordance between diagnoses rendered by WSI and LM is a significant barrier to both regulatory approval and uptake.

Objective.—To examine the published literature on the concordance of pathologic diagnoses rendered by WSI compared with those rendered by LM.

Data Sources.—We conducted a systematic review of studies assessing the concordance of pathologic diagnoses rendered by WSI and LM. Studies were identified following a systematic search of Medline (Medline Industries, Mundelein, Illinois), Medline in progress (Medline Industries), EMBASE (Elsevier, Amsterdam, the Netherlands), and the Cochrane Library (Wiley, London, England), between 1999 and March 2015.

Conclusions.—Thirty-eight studies were included in the review. The mean diagnostic concordance of WSI and LM, weighted by the number of cases per study, was 92.4%. The weighted mean κ coefficient between WSI and LM was 0.75, signifying substantial agreement. Of the 30 studies quoting percentage concordance, 18 (60%) showed a concordance of 90% or greater, of which 10 (33%) showed a concordance of 95% or greater. This review found evidence to support a high level of diagnostic concordance. However, there were few studies, many were small, and they varied in quality, suggesting that further validation studies are still needed.

Arch Pathol Lab Med. 2017;141:151–161; doi: 10.5858/arpa.2016-0025-RA

Whole Slide Imaging

Whole slide imaging was first developed in the 1990s, generating fully colored digital images of entire glass slides at resolutions of less than 0.5 μm/pixel, comparable to a light microscope. Whole slide imaging uses a high-resolution camera, coupled with 1 or more high-quality microscopy objectives to capture images of adjacent areas from glass slides, either as tiles or stripes. Specialized software then combines these individual images to generate a single whole slide image. Whole slide images can be viewed and analyzed digitally on a computer screen.

In comparison to LM, WSI offers numerous advantages. Several pathologists in different locations are able to independently view and assess a slide simultaneously, and viewing of whole slide pathology images. Digital pathology evolved as a practical technology in the 1980s with the development of “telepathology” technology. Initially, telepathology existed in 2 forms: static-image telepathology and dynamic robotic telepathology. Static-image telepathology involved the transmission of preselected regions of microcopy images. Dynamic robotic telepathology enabled real-time image alterations by granting a remote user control of the microscope. Since these 2 initial telepathology technologies, more than 12 types of telepathology systems have evolved. Whole slide imaging is the latest technology used in digital pathology.

WHOLE SLIDE IMAGING

Traditionally, light microscopy (LM) has been the reference method used in anatomic pathology in the diagnosis of many diseases. However, advances in digital imaging hardware and software have led to the development of whole slide imaging (WSI) devices. These devices allow the digital capture, analysis, storage, sharing, and rendering by WSI and LM. Studies were identified following studies assessing the concordance of pathologic diagnoses compared with those rendered by LM.
individual pathologists can examine multiple slides, allowing side-by-side comparisons of different magnifications of the same case. Slides can be annotated and subjected to standardized-image analysis software. The whole slide images generated are stored and shared virtually, decreasing the time taken to render second opinions and preventing slide degradation and physical damage.

At present, WSI is routinely used in both undergraduate and postgraduate education and research. Despite increasing use of WSI in Europe and North America in secondary diagnosis, its use in primary clinical diagnosis remains limited. The authors are aware of a handful of projects worldwide in which WSI is used routinely in primary diagnosis (Sweden; Toronto, Ontario, Canada). Barriers to its implementation include low acceptability of digital pathology among pathologists and the costs associated with implementation. However, another current, significant barrier is the lack of evidence that validates the diagnostic concordance between WSI and LM. Although certain WSI devices have been permitted for use in primary diagnosis in the European Union and Canada, the approval to use WSI in diagnosis from regulatory bodies, such as the US Food and Drug Administration (FDA), has not yet been established. Whole slide imaging systems are currently categorized as a class III (highest risk) medical device by the FDA, meaning they need premarket approval before the FDA permits their sale for clinical use; currently, no devices have been granted that approval. However, the Digital Pathology Association (Indianapolis, Indiana) has recently started to encourage vendors to submit de novo applications for WSI devices to be considered as class II device. Furthermore, pathologists cite the lack of evidence regarding the equivalence of WSI with LM as a reason for nonadoption.

To date, few studies that have compared the diagnostic concordance of WSI and traditional LM. In 2012, Lindsköld et al undertook a systematic review of these studies for a health technology assessment. They reported the diagnostic intraobserver agreement (variation in diagnoses between the same individual) to range from 61% to 100% and a Cohen κ coefficient range of 0.55 to 0.81. Interobserver diagnostic (variation in diagnoses among different users) agreement ranged from 70% to 100% with a Cohen κ coefficient ranging from 0.28 to 0.42. The study concluded that diagnostic disagreements were associated with differences of minor clinical importance but that the quality of the evidence was low. The wide eligibility criteria used focused only on study design, language, and date of publication. No consideration was given to study factors, such as case type, slide type, and study participants. The subsequent heterogeneity displayed among the included studies may have contributed to the large range of diagnostic agreement found.

Addressing the need to validate WSI, the College of American Pathologists (CAP) produced 12 guideline statements in 2013 for studies wishing to validate the diagnostic concordance of WSI and LM. Their guidelines included specific recommendations for validation studies; at least 60 routine cases per application, training in WSI for participants, and, for intraobserver studies, a washout period of at least 2 weeks between viewing the slide sets in each condition. Development of the CAP guidelines; advances in WSI scanners, software, and hardware; and an increased push toward digital technology in health care indicate a need for an updated, systematic review of the diagnostic concordance of WSI and LM.

The primary aim of this systematic review was to examine the published literature on the concordance of pathologic diagnoses rendered by WSI compared with those rendered by LM. Secondary outcome measures, including time to diagnosis and diagnostic confidence, were examined where possible.

**MATERIALS AND METHODS**

The review was registered with PROSPERO database (registration number: CRD4215017859; Centre for Reviews and Dissemination, University of York, Heslington, York, England), the international prospective register of systematic reviews. The review protocol can be accessed online. 12

**Search Strategy**

An electronic search was carried out on the databases: Medline (Medline Industries, Mundelein, Illinois), EMBASE (Elsevier, Amsterdam, the Netherlands), and the Cochrane Library (Wiley, London, England) between 1999 and March 2015. A search of clinicaltrials.gov (US National Institutes of Health, Bethesda, Maryland) was performed to identify any ongoing studies. The ProQuest (Ann Arbor, Michigan) and the Health Management Information Consortium (London School of Hygiene & Tropical Medicine, London, England) databases were also searched to identify any relevant grey literature in an attempt to minimize study bias. Included studies underwent manual reference searching and citation tracking through PubMed (National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland) and Google Scholar (Google Inc, Mountain View, California). Corresponding authors (n = 33) of included studies (n = 33; 92%) were contacted, where possible, to identify any subsequent or ongoing research. A detailed breakdown of the search strategy is provided in the supplemental digital content available at www.archivesofpathology.org in the January 2017 table of contents.

**Article Screening**

Two reviewers (E.G. and D.T.) independently subjected the abstracts of articles to the screening algorithm shown in Figure 1. In cases of disagreement, a third independent reviewer was consulted. Full texts of all articles that fulfilled the initial screening algorithm were retrieved and reviewed.

**Data Extraction**

A standardized data-extraction protocol was applied to all included studies. The protocol was developed from the Cochrane Effective Practice and Organisation of Care template. Data were extracted from studies by the primary researcher (E.G.) and the extracted data was reviewed independently by a second reviewer (D.T.). For studies reporting results as κ statistics, the Landis and Koch classification was used to interpret κ values: no agreement to slight agreement (<0.20), fair agreement (0.21–0.40), moderate agreement (0.41–0.60), substantial agreement (0.61–0.80), and excellent agreement (>0.81). Studies were classified according to their organ system and study design (determined as shown in Figure 2).

**Quality Assessment**

The methodological quality of all included studies was assessed by 2 independent reviewers (E.G. and D.T.) using the updated quality assessment of studies of diagnostic accuracy included in the systematic reviews (QUADAS-2, University of Bristol, Bristol, England) tool, as recommended by the Cochrane Collaboration. Two signaling questions were omitted because they were not relevant to WSI, and an additional 3 signaling questions were added to the tool. The modified QUADAS-2 tool used is shown in Table 1—“Modified Version of the QUADAS-2 Quality Assessment Tool”—with the additional signaling questions marked. Whole slide imaging was considered the index test, and LM was the reference standard. Studies that did not provide participants with the corresponding clinical information for cases, that involved

---

**Whole Slide Imaging Systematic Review—Goacher et al**
Figure 1. Article-screening algorithm. Two reviewers independently screened 1155 abstracts using this screening algorithm. The number of studies rejected at each stage is shown parenthetically. Of the 56 abstracts that met the screening algorithm, 10 were presentation abstracts only. Full texts of the remaining 46 studies were retrieved and reviewed. Abbreviation: H&E, hematoxylin–eosin.

Figure 2. Study design classification based on diagnostic comparisons. A, Retrospective retrieval and review using whole slide imaging (WSI). B, Retrospective retrieval and review using light microscopy (LM). C, Prospective, comparative review of WSI and LM. D, Prospective, comparative review using WSI. E, Prospective, comparative review using LM. Studies that performed more than one of the above comparisons were classified as crossover studies. Comparisons C and D were combined for this review and termed a prospective comparative review.
users not trained in WSI, or that used alternate hardware, such as tablets (Apple, Cupertino, California), were considered to have both a high risk of bias and a high applicability concern for both the index test and the reference standard. The CAP guidelines recommend a 2-week minimum washout period between slide views.23 Therefore, a 2-week minimum interval between the index test and the reference standard was considered an appropriate interval between the index test and reference standard for the flow and timing domain.

Quantitative Synthesis

The studies identified in this review demonstrated high levels of heterogeneity in organ systems, study designs; WSI scanners, hardware, and software; index test conditions; and outcome measures. Therefore, statistical meta-analysis was not justified.27 A narrative review of the studies is provided.

RESULTS

In total, 1155 studies were identified. Of those, 1127 (98%) were sourced from electronic databases, 12 (1%) were from a grey literature search, 5 (<1%) were from citation tracking, and 11 (<1%) from manual reference searching. Two additional studies were obtained from contacted authors. Of the 33 authors contacted, 11 (33%) responded. Of the 1155 studies, 56 (5%) were identified as potentially relevant after an initial abstract screen, and the full text of those articles was sought. The number of studies excluded at each stage is shown in Figure 3. Ten (18%) of the included studies were presentation abstracts only and were subsequently excluded. Of the remaining 46 studies, 10 (22%) were excluded after review of the full text. Two included articles (6%) were each felt to incorporate 2 distinctly different studies.28,29 Outcomes for each of those studies were recorded separately. In total, 38 studies were included in the review.6,7,18,28–60 The study selection process is shown in Figure 3. Intereviewer agreement for article screening was excellent, with a Cohen’s κ coefficient of 0.90 and a 95% CI of 0.84–0.95.

Study Characteristics

The 38 included studies consisted of 6 crossover studies (16%), 19 prospective comparative reviews (50%), and 13 retrospective retrieval and review studies (34%). The mean (SD) number of cases within the included studies was 140 (160). Sixteen studies (42%) used participants trained in using WSI systems. Washout periods between comparisons ranged from none to more than 12 months. Eight WSI scanner manufacturers were represented in the studies, with Aperio (Aperio, Vista, California) scanners used in the majority of the studies (n = 23; 61%). Interobserver agreement was measured in 6 studies (16%), whereas 32 studies (84%) measured intraobserver agreement. The most commonly studied individual organ system was the gastrointestinal system (n = 7; 18%). Ten studies (26%) were a mix of 2 or more distinct organ systems. A detailed breakdown of individual study characteristics can be found in the supplemental digital content.
Quality Assessment

A tabulated display of the quality-assessment results for individual studies is shown in Table 2. Graphic depictions of the quality-assessment results by assessment domain for risk of bias and applicability concerns are shown in Figure 4, A and B.

Risk of Bias

Across the 4 domains (patient selection, index test, reference standard, and flow and timing), the percentage of studies with a high risk of bias ranged from 11% (n = 4) to 16% (n = 6) (Figure 4, A). The percentage of studies with a low risk of bias ranged from 32% (n = 12) to 74% (n = 28). The index test domain showed the highest risk of bias, with 16% of studies (n = 6) having a high risk. For the same domain, an unclear risk of bias was found in 53% of the studies (n = 20). The flow and timing domain had the lowest risk of bias (74% [28 cases] in the low-risk category).

Applicability

The patient-selection domain caused the least concern regarding applicability, with 100% of studies (n = 38) being classified as low concern (Figure 4, B). Greatest concern was for applicability of the index test domain with 18% (n = 7) of studies being classified as high concern and only 32% (n = 12) classified as low concern. Applicability of studies in the reference standard domain was reasonable (61% [n = 23] low concern).

Diagnostic Concordance

Diagnostic concordance was reported as the percentage of concordance (n = 25; 66%), κ agreement (n = 8; 21%), or both (n = 5; 13%). The diagnostic intraobserver reported concordance ranged from 63% to 100%, (κ coefficient range, 0.48–0.87). The diagnostic interobserver reported concordance ranged from 84% to 100%. A single interobserver κ coefficient value of 0.91 was reported. To obtain an idea of overall concordance corrected for study size, the cited concordance was adjusted to account for the number of cases per study. Across all studies, the mean percentage of diagnostic concordance was 92.4%, and the mean κ agreement was 0.75 (substantial agreement). Concordance across retrospective retrieval and review studies, prospective comparative review studies, and crossover studies was 92.9%, 92.4%, and 91.2%, respectively. Crossover studies and retrospective retrieval and review studies showed excellent agreement for calculated mean (SD) κ coefficients (0.87 [0] and 0.81 [0], respectively). Prospective comparative review studies showed substantial agreement (0.68 [0.16]).

Of the 30 studies (79%) that provided percentage of diagnostic concordance measurements, 18 (60%) reported a concordance of 90% or greater, and 10 of these (56%) showed a concordance rate of 95% or greater. Six studies
(20%) reported a concordance of less than 85%. The weighted mean percentage diagnostic concordance of study LM diagnosis with original LM diagnosis was 93.4% across the 10 studies that measured it. Whole slide imaging and LM concordance across the same 10 studies was 90.9%. Graphic representations of the percentage of diagnostic concordance against study-design factors are shown in Figure 5, A through F.

The percentage of concordance range (PCR) among the 10 studies with a mixed case load ranged from 75% to 97%.* The PCR for studies on the gastrointestinal system ranged from 70.0% to 99%.* Table 3 displays the PCRs for each organ system included.

### Table 3. Percentage of Concordance Range (PCR) by Organ System

<table>
<thead>
<tr>
<th>Organ System</th>
<th>Frequency, No. (%)</th>
<th>PCR, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case mix</td>
<td>10 (26)</td>
<td>75.0–97.0</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>7 (18)</td>
<td>70.0–99.0</td>
</tr>
<tr>
<td>Breast</td>
<td>6 (16)</td>
<td>91.0–99.0</td>
</tr>
<tr>
<td>Skin</td>
<td>5 (13)</td>
<td>76.0–100</td>
</tr>
<tr>
<td>Genitourinary</td>
<td>4 (11)</td>
<td>63.0–95.0</td>
</tr>
<tr>
<td>Gynecological</td>
<td>2 (5)</td>
<td>94.0</td>
</tr>
<tr>
<td>Pediatric/placental</td>
<td>2 (5)</td>
<td>90.0–98.3</td>
</tr>
<tr>
<td>Prostate</td>
<td>2 (5)</td>
<td>76.0</td>
</tr>
</tbody>
</table>

*References 7, 29, 36–39, 41, 44, 46, 60.

### Time to Diagnosis

Out of the 4 studies that reported time to diagnosis, the 3 (75%) that compared WSI and LM times to diagnosis all found a longer time to diagnosis using WSI.38,43,45,58 Gui et al* found an average WSI review time of 91.9 seconds (95% CI, 65.3–118.5 seconds) compared with an average LM review time of 57.1 seconds (95% CI, 55.9–72.3 seconds). Valez et al* compared the viewing time per slide of 2 WSI modalities and 23 seconds for LM. Jen et al* reported the

---

### Table 2. QUADAS-2 Assessment of Individual Studies

<table>
<thead>
<tr>
<th>Source, y</th>
<th>Risk of Bias</th>
<th>Applicability Concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al Habeeb et al,30 2012</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Al-Janabi et al,18 2014</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Al-Janabi et al,31 2012</td>
<td>Low</td>
<td>Unclear</td>
</tr>
<tr>
<td>Al-Janabi et al,32 2012</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Al-Janabi et al,33 2012</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Arnold et al,35 2015</td>
<td>Low</td>
<td>Unclear</td>
</tr>
<tr>
<td>Bauer et al,36 2013</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Bauer and Slaw,37 2014</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Brunelli et al,38 2014</td>
<td>Unclear</td>
<td>Unclear</td>
</tr>
<tr>
<td>Buck et al,39 2014 (validation phase)</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Buck et al,40 2014 (parallel study phase)</td>
<td>High</td>
<td>Unclear</td>
</tr>
<tr>
<td>Campbell et al,28 2014 (crossover)</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Campbell et al,29 2014 (prospective comparative review)</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Chargari et al,41 2011</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Fonyadi et al,42 2012</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Gage et al,43 2013</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Gilbertson et al,44 2006</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Gui et al,45 2012</td>
<td>Low</td>
<td>Unclear</td>
</tr>
<tr>
<td>Houghton et al,46 2014</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Ho et al,47 2006</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Jen et al,48 2013</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Jukic et al,49 2011</td>
<td>Low</td>
<td>Unclear</td>
</tr>
<tr>
<td>Krishnamurthy et al,50 2013</td>
<td>Low</td>
<td>Unclear</td>
</tr>
<tr>
<td>Mooney et al,51 2011</td>
<td>Unclear</td>
<td>Unclear</td>
</tr>
<tr>
<td>Nielsen et al,52 2010</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Ordri et al,53 2015</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Oztok et al,54 2012</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Reyes et al,55 2014</td>
<td>Low</td>
<td>Unclear</td>
</tr>
<tr>
<td>Rios et al,56 2010</td>
<td>Low</td>
<td>Unclear</td>
</tr>
<tr>
<td>Rodriguez-Urrego et al,57 2011</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Sanders et al,58 2012</td>
<td>Unclear</td>
<td>High</td>
</tr>
<tr>
<td>Shaw et al,59 2012</td>
<td>Unclear</td>
<td>Unclear</td>
</tr>
<tr>
<td>Van der Post et al,60 2013</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Velez et al,61 2008</td>
<td>Low</td>
<td>Unclear</td>
</tr>
<tr>
<td>Wendum et al,62 2009</td>
<td>High</td>
<td>Unclear</td>
</tr>
<tr>
<td>Willbur et al,63 2009</td>
<td>High</td>
<td>Unclear</td>
</tr>
</tbody>
</table>

a Definitions: low, low risk; high, high risk; unclear, unclear risk.

b The tabulated display of QUADAS-2 assessment results by individual study. Results of QUADAS-2 assessment cannot be quantified; however, this table demonstrates specific strengths and weaknesses in the quality of the individual studies. Interestingly, more-recent studies appear to have a higher-quality index. Data were derived from the QUADAS-2 resource page.25,69
average time spent examining digital slides to be 1.4 times greater than spent on glass slides ($P < .03$).

**Diagnostic Confidence**

Of the 2 studies to report diagnostic confidence (5%), only Jukic et al$^{46}$ compared WSI diagnostic confidence to LM diagnostic confidence.$^{41}$ Mean diagnostic confidence in WSI and LM was calculated at 4.0 of 5.0 (95% CI, 2.9–5.1) and 4.1 of 5.0 (95% CI, 4.0–4.2), respectively.

**DISCUSSION**

Systematic reviews form the cornerstone of evidence-based medicine, at the top of the study-design hierarchy along with meta-analyses.$^{61}$ Although there have been many studies published on the diagnostic concordance of WSI and LM, there has been no systematic assimilation of those studies, apart from the Lindsköld$^{20}$ health technology assessment. This review was structured according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines.$^{62}$ The review was conducted by a team experienced in conducting systematic reviews, pathology, and digital-pathology research. The primary researcher (E.G.) undertook a formal, systematic-review training course before conducting the review.

This systematic review identified 1155 studies, 36 of which (3%) were included in the review. Two of the included studies seemed to incorporate multiple studies and were each viewed as 2 separate studies for the purposes of this review, resulting in a total of 38 studies included in the review. For these 38 studies included, the mean diagnostic concordance between diagnoses rendered by WSI and those rendered by LM was 92.4%. The mean $k$ agreement was
Figure 5. Graphs showing relationship among reported concordance and study design factors. A, A scatter plot of the total number of cases per study compared with the percentage of diagnostic concordance. B, A distribution dot plot of study design against the percentage of diagnostic concordance. C, A distribution dot plot showing whether study participants were trained in using whole slide imaging (WSI) compared with the percentage of diagnostic concordance. D, A distribution dot plot of organ system assessed compared with the percentage of diagnostic concordance. E, A distribution dot plot comparing study date to the percentage of diagnostic concordance. F, A distribution dot plot comparing the length of the washout period between views of slides to the percentage of diagnostic concordance. All graphs were created using StataIC 13 (Stata Statistical Software, Release 13, 2013. StataCorp, College Station, Texas). Abbreviation: NA, not applicable.
0.75 (substantial agreement). A mean diagnostic concordance of 93.4% was found among the 10 studies (26%) that compared prospective and retrospective diagnosis using LM. Few studies reported time to diagnosis and diagnostic confidence (n = 4 [11%] and n = 2 [5%], respectively). Where measured, the time to diagnosis was increased, and diagnostic confidence was less using WSI compared with LM.

These results complement the 2012 Lindsköld health technology assessment review, particularly for the intraobserver agreement ranges found. The PCR for interobserver ranges found in this review was less than the corresponding range reported in the Lindsköld review (84%–100% and 70%–100%, respectively), which may be due to the stricter eligibility criteria used in this review, reducing the level of heterogeneity between included studies. The Lindsköld review used the quantifiable GRADE (Grading of Recommendations Assessment, Development and Evaluation) system to assess the quality of studies. Unlike the GRADE system, the QUADAS-2 tool does not provide quantitative measures; therefore, we cannot determine whether the studies included in this review were of higher, equal, or lower quality. The QUADAS-2 assessment is, however, a more in-depth assessment of study quality and has subsequently been recommended by the Cochrane Collaboration.

For WSI to be validated for use in routine diagnostic work, diagnostic concordance does not need to be 100%. It does, however, need to be established as being noninferior to LM. Unfortunately, few studies have investigated the intraobserver concordance between LM diagnoses. This would suggest that the most appropriate study design for the validation of WSI is a crossover study. Such a study facilitates the direct comparison of intraobserver concordance for LM and intraobserver diagnostic concordance of WSI and LM. However, an insignificant difference between the concordance means does not necessarily imply equivalence. An adequate study size, determined by a noninferiority power calculation, is required to demonstrate the diagnoses rendered by WSI are not inferior to those rendered by LM in concordance.

The secondary outcomes measured in this review complement the findings in existing literature. A slower time to diagnosis has been reported in WSI. The few studies that reported the time to diagnosis all showed an increased time to diagnosis when using WSI compared with LM. An increased time to diagnosis highlights the inefficiency associated with WSI at present, which is a particular concern in financially pressured health care systems. Increased time to diagnosis and a reduced diagnostic confidence are likely to reduce the acceptability of WSI among pathologists, already a barrier to WSI implementation. However, the development of image-analysis software and improvement in workflow with digital systems may potentially decrease time to diagnosis in the future. Increased pathologist experience with WSI devices is also likely to decrease time to diagnosis and increase diagnostic confidence.

The quality and size of studies appeared to be rising over time. This could be related to the increasing guidance being published by authorities, such as the CAP and the FDA, and to the increasing adoption and understanding of WSI. Because of the strict eligibility criteria used in this review, which included only studies that used hematoxylin–eosin–stained human tissue, patient selection was 100% applicable. The risk of bias and applicability concerns of the index test were affected by studies that did not train participants in using WSI and by studies that did not provide participants with the corresponding clinical information. Figure 5, C, shows the effect of training in WSI on diagnostic concordance. In some cases, less-applicable technology was also used for the index test. For example, Brunelli et al used an iPad to view digital slides. This resulted in greater concern about its applicability because, in routine diagnoses in a hospital setting, a workstation with a specialized high-resolution monitor would most likely be used.

Seventy-four percent of studies (n = 28) had a low risk of bias in the flow and timing domain. The main variable affecting this domain was the washout period used between slide views. This review considered a washout period longer than 2 weeks to be appropriate, based on the recommendations in the 2013 CAP guidelines. The 2-week washout period is intended to minimize recall bias between slide views. In 2013, there was a shortage of evidence in this area. However, Campbell et al have since found recall bias between slide views to occur in washout periods of up to 4 weeks.

Because of the limited timescale of this review, its scope included published articles only. The vendors of WSI systems have conducted their own validation studies for use in submission to regulatory authorities, such as the FDA or Conformité Européenne. To date, vendors have neither published nor publically released these internal data. However, follow-up work is planned that will include vendor data, where available. Twenty-nine percent of studies (n = 11) included in this review used fewer than 60 cases. This review endeavored to minimize the effect of such small studies by calculating means weighted according to the number of cases per study. Future reviews may wish to include a minimum number of cases as part of the eligibility criteria. Alternatively, the number of cases could be used in the quality assessment of included studies. In addition, future reviews may also wish to examine whether concordance was affected by differing definitions of concordance and differing case complexity among the included studies. By including only published journal articles, this review did not take into account the potential for publication bias. However, the grey literature search performed was intended to minimize the effect of such publication bias.

This review provides limited evidence of diagnostic concordance between WSI and LM. However, this finding is predominantly based on small studies that displayed heterogeneous study designs, participants, and case type. Larger study designs would provide greater confidence in the measured outcomes. Bauer et al conducted an a priori power calculation to estimate the number of cases needed to test the hypothesis that diagnoses rendered with WSI are not inferior to those rendered with LM. They found that 450 cases (225 glass and 225 whole slide images) were needed to establish a WSI noninferiority of 4% at a significance level of .05. This number of cases is noticeably different from the minimum number of 60 cases recommended by the 2013 CAP guidelines. Interestingly, only 1 of the 11 studies (9%) that had fewer than 60 cases showed a concordance percentage of 90% or greater. However, at present, it is not clear whether such sample-size calculations should determine the number of cases, the number of slides, or the number of pathologists.
There is a need for further work into the effect of the duration of the washout period on recall bias. In a 2015 study, Campbell et al. found that, after 2 weeks—the CAP recommended washout period—pathologists were capable of recalling 40% of cases previously seen. Even after 4 weeks, pathologists were able to recall up to 31% of cases previously seen. The CAP guidelines acknowledged at the time of publication, the lack of studies comparing the effect of the duration of the washout period on outcomes. The FDA recommends that sufficient time be allowed between intraoperator reviews of the same imaging to reduce recall bias, but it does not provide a quantifiable definition of sufficient.

Although this review provides evidence to support the diagnostic concordance of WSI and LM in routine diagnoses, it also highlights the significant heterogeneity among validation study designs. This is unfortunate because, if the included studies had been of sufficient size, quality, and homogeneity, they would have enabled us to perform a meta-analysis on more than 5000 cases, significantly enhancing the quality of evidence available to pathologists and regulators about the validation of WSI. One method of reducing heterogeneity among future validation studies is for such studies to consult and adhere to the 2013 CAP guidelines.

At present, however, there is a lack of available evidence to validate the use of WSI in routine primary diagnosis. Regulators, industry, health care providers, and the academic community are all interested in the digitizing of pathology, so future validation studies are inevitable, and many are currently ongoing. By demonstrating the types of study designs available, this review may help in the design of future validation studies. In addition, this review highlights weaknesses present in previous validation studies, which future studies could avoid.

Advice on modifying the QUADAS-2 tool was provided by Penny Whiting. We thank the following individuals for their responses: W. Scott Campbell; Liron Pantanowitz; László Fonyad; Peter Furness; Anna R. Graham; Joseph P. Houghton; Bela Molnar; Ellen Mooney; Emily Shaw; Thomas P. Buck; and Carolina Reyes. Karen Lee assisted in the organization of the project.

References

38. Brunelli M, Beccari S, Colombari R, et al. iPathology cockpit diagnostic station: validation according to College of American Pathologists Pathology and
Laboratory Quality Center recommendation at the Hospital Trust and University of Verona. Diagn Pathol. 2014;9(suppl 1):S12.


