Review Article

Diagnostic Concordance Between Whole Slide Imaging and Conventional Light Microscopy in Cytopathology: A Systematic Review

Ilaria Girolami, MD [1]; Liron Pantanowitz, MD [2]; Stefano Marletta, MD[1]; Matteo Brunelli, MD[1]; Claudia Mescoli, MD[3]; Alice Parisi, MD[1]; Valeria Barresi, MD[1]; Anil Parwani, MD[4]; Desley Neil, MD[5]; Aldo Scarpa, MD[1]; Esther Diana Rossi, MD [6]; and Albino Eccher, MD [1]

Many studies have examined the diagnostic concordance of whole slide imaging (WSI) and light microscopy (LM) for surgical pathology. In cytopathology, WSI use has been more limited, mainly because of technical issues. The aim of this study was to review the literature and determine the overall diagnostic concordance of WSI and LM in cytopathology. A systematic search of PubMed, Scopus, and the Cochrane Library was performed, with data extracted from the included articles. A quality assessment of studies was performed with a modified Quality Assessment of Diagnostic Accuracy Studies 2 tool. The primary outcome was concordance for the diagnoses rendered by WSI and LM as shown by the concordance rate with the original diagnosis, intra-observer and interobserver concordance with the κ coefficient, or a percentage. Secondary outcomes included the time taken to reach a diagnosis and the quality and perception of WSI. A descriptive survey was provided. Among 1867 publications, a total of 19 studies (1%) were included. Overall, the concordance between WSI and the original diagnosis was 84.1%, the intra-observer concordance between WSI and LM was 92.5% with a κ coefficient of 0.66, and the interobserver κ coefficient was 0.69. The time to reach a diagnosis was longer with WSI in all studies. The quality of WSI was good, but diagnostic confidence and cytologist preference were higher for LM. In conclusion, the concordance of WSI with LM is acceptable and in line with systematic reviews in surgical pathology. However, the time required for scanning and technical issues represent barriers to complete adoption. It is foreseeable that technical advances and rigorous validation study design will help to improve the diagnostic concordance of WSI with LM in cytopathology. ***Cancer Cytopathol* 2020;128:17-28**.

KEY WORDS: agreement; cytology; cytopathology; diagnostic concordance; review; whole slide imaging.

INTRODUCTION

Digital pathology became popular with the application of telepathology in the 1980s, at which time the 2 main technical systems available used either static or robotic digital images. Static imaging involves the transmission of a single microscopic field of view (eg, a still microphotograph) acquired with a digital camera mounted on a microscope. Robotic imaging allows the end user to remotely control a microscope and offers full access to navigate the entire slide. Whole slide imaging (WSI) is newer technology that allows glass slides

Corresponding author: Albino Eccher, MD, Department of Pathology and Diagnostics, University and Hospital Trust of Verona, Verona, Italy, Piazzale A. Stefani 1, 37126 Verona, Italy; albino.eccher@aovr.veneto.it

[1]Department of Pathology and Diagnostics, University and Hospital Trust of Verona, Verona, Italy; [2]Department of Pathology, UPMC Shadyside Hospital, University of Pittsburgh, Pittsburgh, Pennsylvania; [3]Surgical Pathology and Cytopathology Unit, Department of Medicine, University and Hospital Trust of Padua, Padua, Italy; [4]Department of Pathology, Ohio State University, Columbus, Ohio; [5]Department of Histopathology, University Hospital Birmingham, National Health Service Foundation Trust, Birmingham, United Kingdom; [6]Division of Anatomic Pathology and Histology, Catholic University of Sacred Heart, Agostino Gemelli School of Medicine, Rome, Italy

The first 2 authors contributed equally to this article.

Additional supporting information may be found in the online version of this article.

Received: August 18, 2019; **Revised:** September 21, 2019; **Accepted:** September 23, 2019

Published online October 10, 2019 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cncy.22195, wileyonlinelibrary.com

to be digitized (scanned) to generate large whole slide images (also known as virtual slides or e-slides). Whole slide images acquired with a 20× scan typically have a resolution of 0.25 μm/pixel, which is comparable to examining a glass slide with light microscopy (LM). This kind of digital image allows the user to navigate the slide, zoom in and out, and annotate areas of the image that are of particular interest.[1]

The advantages of WSI over LM include easy portability and sharing of digital slides, the possibility of simultaneous access to slides by multiple users, side-by-side comparisons of slides on a monitor, the use of image analysis, and easier archiving. As a result, WSI has become popular in academic settings for second-opinion diagnoses (teleconsultation), educational purposes, and research activity.[2] More recently, the adoption of WSI for primary diagnosis has become a reality in some countries despite some barriers to implementation, such as cost, pathologist resistance, and regulatory issues. To facilitate clinical adoption, the College of American Pathologists (CAP) published a formal guideline on validating WSI for primary diagnosis.[3] The CAP is revising this guideline and intends to release an update.[4]

Most efforts and considerations to date have been concerned mainly with the use of WSI for surgical pathology. Glass slides containing formalin-fixed, paraffin-embedded histological sections that are stained with hematoxylin-eosin, special stains, or immunohistochemistry are typically easier to digitize because the tissue material to scan has a relatively uniform thickness of 3 to 5 μm with a flat topography. Exceptions do occur with occasional tissue folds or other artifacts (eg, air bubbles). For WSI scanners, image quality and focusing are dependent on the devices' optics (eg, the objective numerical aperture), digital camera (eg, sensors), and scanning along the vertical axis (the z-axis). Most scanners use algorithms to which only 1 level of the z-axis is acceptable. However, in cytopathology, where glass slides may contain material with a variable smear thickness or 3-dimensional cell groups, Z-stacking (scanning with multiple focal planes) is preferred.[1] With current scanners, Z-stacking comes at the cost of increasing scan time and digital file size, with the latter sometimes several gigabytes per image. These impediments are some of the main reasons for the reduced implementation of WSI in cytopathology.[5,6]

Not surprisingly, the literature on the use of WSI in cytopathology for primary diagnosis is limited. In a review performed a decade ago, it was noted that WSI for cytopathology had only limited applications such as proficiency testing for cytologists.[7] Since then, WSI technology has progressed, and some of the barriers (eg, economics) have improved. This has accordingly resulted in an increase in the number of publications regarding the validation of WSI for cytopathology diagnostic use. The aim of this review is to systematically examine the published literature in which cytopathology diagnoses rendered by WSI are compared with those made with LM.

MATERIALS AND METHODS

Framing the Review Question

We intended to structure our work according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines.[8] The primary aim of the study was to evaluate the diagnostic concordance of WSI (digital modality) and LM (traditional glass slide) diagnoses in cytopathology. Secondary aims included additional elements related to making a digital diagnosis, such as the time required to render a diagnosis, the cytologist's ease with and perception of using digital slides, and the types of pathology settings of users more prone to using digital slides. Studies encountered were likely to follow a crossover study design in which "multiple cases with multiple readers" were used for validation purposes. In these studies, enrolled cases were mostly already assessed by glass slide LM and then were subsequently reassessed by WSI, and diagnoses made by both modalities were compared with each other or the reference (so-called original "ground truth") diagnosis rendered by LM. Taking into account the fact that in the field of cytopathology the diagnostician can involve a pathologist and/or a cytotechnologist, we chose to combine both types of cytologists as reading diagnosticians. Cytopathology studies performed entirely for quality-assurance reasons were also included.

Search Strategy

A search strategy was built according to a modified Population, Intervention/Index Test, Comparison/Comparator Test, and Outcome (PICO) model. The population term was restricted to human studies and excluded studies concerning microorganisms and veterinary pathology. Because the primary aim was to compare the use of WSI with traditional LM for cytopathology diagnosis, we decided that the index test terms used must be restricted

to WSI technology. We accordingly excluded other technologies such as the transmission of static images, robotic microscopy, video streaming, smartphones, and software solutions that did not involve WSI.[9] Consequently, the index test term was represented by free text referring to WSI, but at the same time, more general terminology such as *digital pathology* and *telepathology* was also used. For the comparator term, free text referring to light, traditional, or conventional microscopy was used. Because the primary aim was to evaluate the concordance between WSI and LM for cytopathology diagnosis, studies reporting the use of automated screening systems, image analysis, or other automation tools run before human examination of slides were excluded. Outcome terms were represented by any measure of diagnostic agreement. Inclusive measures likely to be reported in retrieved studies were concordance or agreement rates, κ statistics, and any other measure of diagnostic concordance. Terms defining the setting of cytology (eg, *smear* and *touch preparation*) were also added to the search strategy (see the supporting information).

Article Screening

The PubMed, Scopus, and Cochrane Library electronic databases were searched with no language restrictions up until May 29, 2019. Another search of ClinicalTrials.gov was also performed to identify any ongoing studies. Two investigators (I.G. and S.M.) independently screened article titles and abstracts with the aid of the Rayyan QCRI reference manager web application.[10] After screening, any studies with disagreement were resolved by consensus. Full texts of the articles that fulfilled the initial screening criteria were acquired and reviewed for subsequent inclusion against the eligibility criteria.

Data Extraction

Data were extracted from studies by 2 investigators (I.G. and S.M.), and the extracted data were reviewed independently by the senior researcher (A.E.). A standardized form for extraction and presentation was used. The data extracted were as follows: number of cases, number of slides, type of pathology/organ system, type of cytological specimen, staining, type of scanner used, presence and number of Z-stacked planes, number of diagnosticians, washout period for readings (ie, the time between digital and LM reads), presence of training in WSI use, measure of the primary outcome, and measure of secondary outcomes if present. For studies reporting results as κ statistics, the interpretation of values followed the Landis and Koch classification[11]: no agreement to slight agreement (0.00-0.20), fair agreement (0.21-0.40), moderate agreement (0.41-0.60), substantial agreement (0.61-0.80), and excellent agreement (≥0.81).

Quality Assessment

The methodological quality of included studies was assessed by 2 independent reviewers (I.G. and S.M.) using a modified version of the Quality Assessment of Diagnostic Accuracy Studies tool (Quality Assessment of Diagnostic Accuracy Studies 2 [QUADAS-2]).[12] Two signaling questions were removed from the tool because they were not relevant for WSI: one in the patient selection domain and the other in the index test domain. The index test was WSI, and the reference test was LM. We added 4 additional questions to address specific issues of WSI and cytopathology diagnosis: 1) for the index test domain, we looked for training of participants in the use of the index test because an absence of training could have hampered diagnostic performance; 2) we looked for a clear declaration of scanning modality; 3) we noted if single or multiple Z-stacking was used because it could have the potential to influence the quality of the image and thus the rendered diagnosis; and 4) for both the index test and reference test domains, we searched for whether clinical details, in the form of a brief clinical history or demographic data, were provided to participants before the reading of slides. The modified version of QUADAS-2 can be found in Supporting Table 1. The studies that did not provide clinical details of cases to participants, that did not show the presence of training in the use of the index test, and that did not provide insight into the technology features were considered to have a high risk of bias for these domains. For the question about a washout period in the flow and timing domain, we followed the CAP guideline publication for WSI validation,[3] which recommends that a minimum washout period of 2 weeks be used.

Synthesis and Reporting

Because the search retrieved a broad heterogeneity of studies in terms of the study design, types of scanners used, index test conditions, outcome measures reported, and varied diagnostic settings of each study, no quantitative statistical meta-analysis was possible. Therefore, a descriptive synthesis of these studies is provided.

RESULTS

A total of 1867 articles were identified after the removal of duplicates. Of these, 52 (3%) were identified as potentially relevant after the initial abstract screening, and the full text was retrieved. After the full text was read, 33 articles (63%) were excluded. Reasons for exclusion were as follows: a lack of any outcome measures or comparison with glass slides in 13 (39%), no use of WSI in 8 (24%), no cytological cases in 5 (15%), and other miscellaneous reasons in 7 (21%). A detailed flow diagram of the screening and exclusion of all articles is shown in Figure 1.

Study Characteristics

The 19 publications that were included consisted of 12 retrospective studies (63%), where rereads of archival cases were compared with the original diagnosis; 6 prospective studies (32%); and 1 article (5%) with a comparison of the digital cytological diagnosis rendered in a teleconference with the final diagnosis from the surgical pathology specimen. The number of cases per study ranged from 4 to 1005 (median, 22; mean, 93). The number of slides per study ranged from 4 to 1005 (median, 30; mean, 113). Eleven studies (58%) dealt with cervical cytology (Papanicolaou tests), and 2 of these studies also incorporated a minor fraction of other nongynecological cytology cases. Four publications (21%) dealt only with nongynecological cytology cases, and 1 of these included pediatric patients. The stain most frequently used was Papanicolaou (13 studies [68%]). Eight scanner providers were represented in the studies; they included Leica/Aperio (n = 7 [37%]), Hamamatsu (n = 4 [21%]), and Roche/Ventana (n = 3 [16%]). Z-stacking information was present in 13 studies (68%), with a single Z-plane used in 9 of the 13 studies (69%) and multiple layers ranging from 3 to 21 Z-stacks used in the other 4 studies (31%). A washout period was not used in 7 studies (37%), was not stated in 4 studies (21%), and ranged between 2 days and 9 months in the remaining studies.

Quality Assessment

A summary of the quality assessment for single studies is graphically displayed according to single domains in Figure 2 (for the results of single studies, see Supporting Table 2). The domain with a higher proportion of studies with a high risk of bias was the flow and timing domain, where the washout period and the inclusion of all cases were the main red flags for the quota of publications with a high risk of bias (n = 5 [26%]). As for the index test, 10 studies (53%) were judged to be at low risk of bias because the main criteria for assessing this domain were fulfilled, and 4 (21%) were at high risk of bias. As for the patient selection domain, 4 studies (21%) were judged to be at high risk of bias, with no random or consecutive selection of cases or with inappropriate exclusion of cases, whereas 9 articles (47%) did not report information that was clear enough for judgment. As expected, the domain with the lowest proportion of studies with a high or unclear risk of bias was the reference test domain, with 17 studies (89%) showing a low risk of bias. Overall, judgment on the applicability of the study to the review question showed high concern in 3 studies (16%) regarding the index test and in 1 study (5%) regarding the reference standard.

Diagnostic Concordance

Diagnostic concordance was reported as a percentage of concordance, a κ coefficient, or both. Intra-observer concordance between the 2 modalities was reported as a percentage (n = 4 [21%]),[14-17] a κ coefficient (n = 1 [5%]),[18] or both (n = 2 [11%]).[19,20] Twelve studies (63%)[14,17,21-30] reported a concordance rate for WSI with the original diagnosis, and 2 of these studies also reported other concordance measures.[14,17] One study reported the κ coefficient of WSI with the original diagnosis.[31] One study compared the diagnosis made via WSI with the final diagnosis based on the definitive surgical pathology diagnoses.[32] Interobserver concordance was reported as a κ coefficient in 5 studies (26%): 3 reported values for both WSI and LM,[15,17,18] 1 reported values only for WSI,[20] and 1 reported values only for LM.[19] The intra-observer concordance ranged from 77.5% to 100%, and the κ coefficient ranged from 0.44 to 0.93. The interobserver concordance with virtual slides varied, with κ ranging from 0.57 to 0.82. For studies comparing the WSI diagnosis with the original sign-out diagnosis on LM, the concordance ranged from 14% to 100%. To obtain an idea of the overall concordance with a correction for study size, the concordance percentages and κ coefficients reported were adjusted for the number of cases evaluated per study on the basis of the reported values or the mean value. Across the studies, the mean percentage of intra-observer concordance between WSI and LM was 92.5%, and the

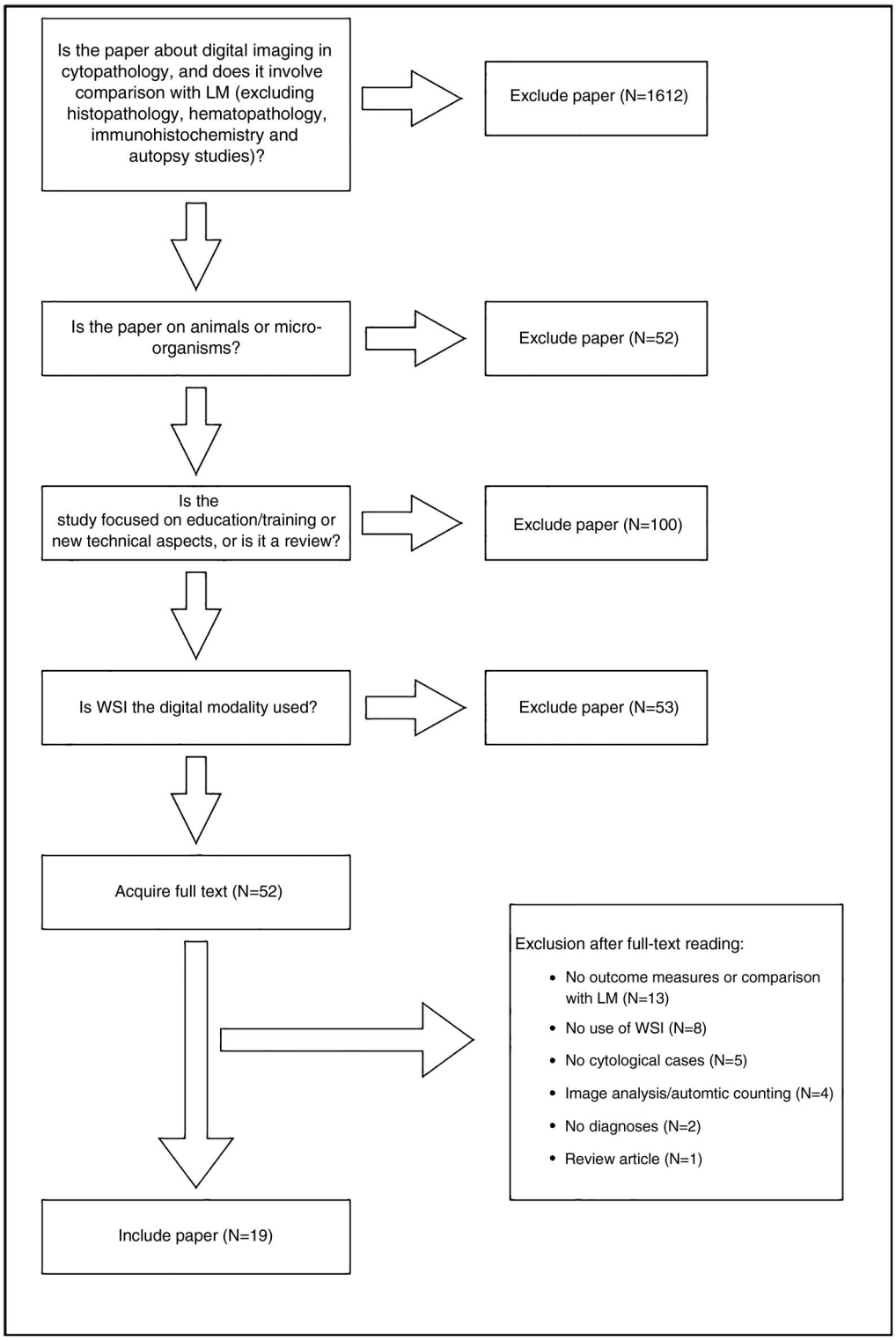

Figure 1. Flow diagram of the study selection process. LM indicates light microscopy; WSI, whole slide imaging.

κ coefficient was 0.66, whereas the mean κ coefficient for interobserver concordance with WSI was 0.69. The mean percentage of diagnostic concordance with the original reference diagnosis was 84.1%. Some studies[15,17-19] also reported a κ coefficient for interobserver concordance with LM, which ranged from 0.67 to 0.94 with an overall

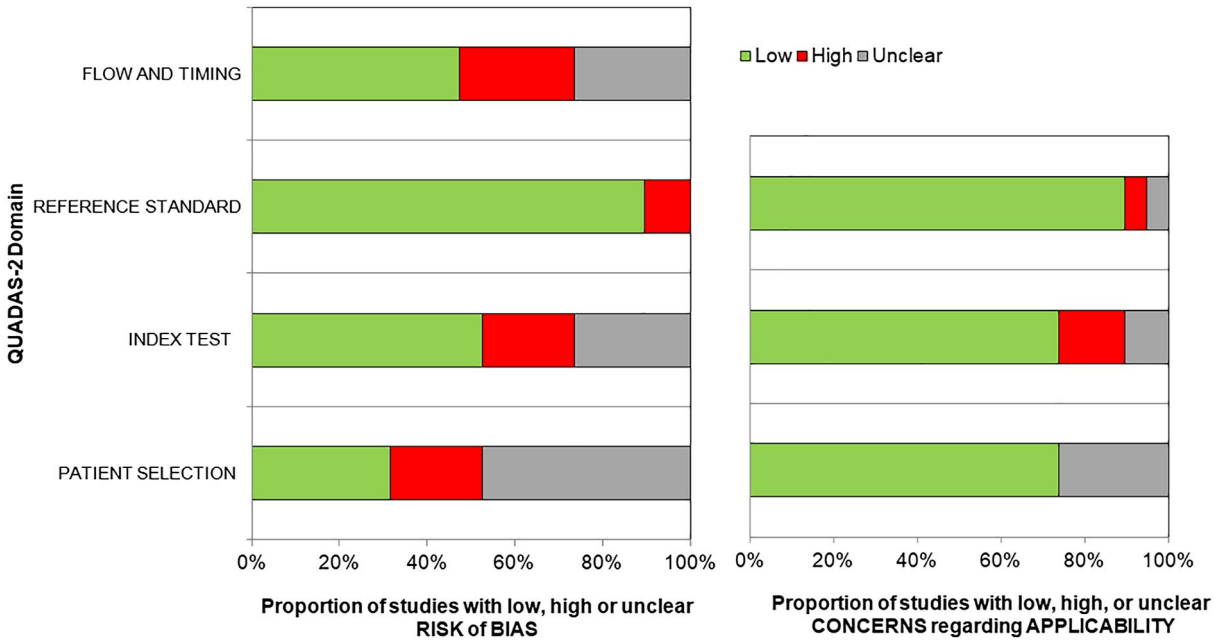

Figure 2. Graphic display of the QUADAS-2 quality assessment. Templates for the graphs are derived from the QUADAS-2 resource page.[13] QUADAS-2 indicates Quality Assessment of Diagnostic Accuracy Studies 2.

mean value of 0.78. One study provided measures of κ coefficients separately for adequacy and the final diagnostic category, with the κ value for intra-observer agreement on adequacy reported to be higher than that for the diagnostic category (0.86-1.00 vs 0.75-0.93).[18] As for studies dealing only with cervical specimens, all but 3 used liquid-based cytology (LBC) specimens, and in this clearly identifiable subgroup, the overall concordance of the WSI diagnosis with the original reference diagnosis was 88.2%. A representative example of a WSI digital slide of an LBC cervical specimen is shown in Figure 3. Notably, for the subgroup of nongynecological, non-LBC studies, the overall concordance with the original diagnosis was 81.4%, and the intra-observer and interobserver κ coefficients were 0.61 and 0.60, respectively, which were slightly lower than the overall values of the entire study population.

Time to Diagnosis

Eleven studies (58%) reported the time needed to screen a digital slide or to render a diagnosis, and 7 of these compared measurements with the time spent on LM using glass slides. The mean time to a diagnosis ranged from 2 minutes 2 seconds to 40 minutes.[14,16-18,21,22,24,27,28,30,31] The time needed to render a diagnosis appeared to be longer with WSI than LM, and this remained true for both LBC and non-LBC studies.

Additional Outcomes

Six studies explored other issues: the perception of quality of WSI slides[14-16,26-28] and the end user's confidence in rendering a diagnosis.[16] In one study, the investigators quantified the quality of WSI images in comparison with glass slides and microphotographs, the ease of navigation, and the speed and accessibility of images on a scale of 1 to 4. In that study, virtual slides ranged from fair to good; however, it had both 2- and 3-dimensional slides.[14] In another study with a similar 1 to 4 scale, the WSI slides were scored 3 or 4 by all participants.[28] One study that included mostly non-LBC specimens found that participants judged WSI slides to be of poor quality in less than 10% of cases, mainly because of the presence of bubbles in the preparation, hypocellularity, air-drying artifacts, and images with a blurry, suboptimal focus.[27] In one study with only LBC cases, there was a survey with a Likert-like scale for rating WSI and LM. In that study, although the main strengths of WSI were reported to be the ease of navigation and switching between slides, a relative majority of respondents still preferred using LM with glass slides for diagnostic work.[15] LM was perceived to be

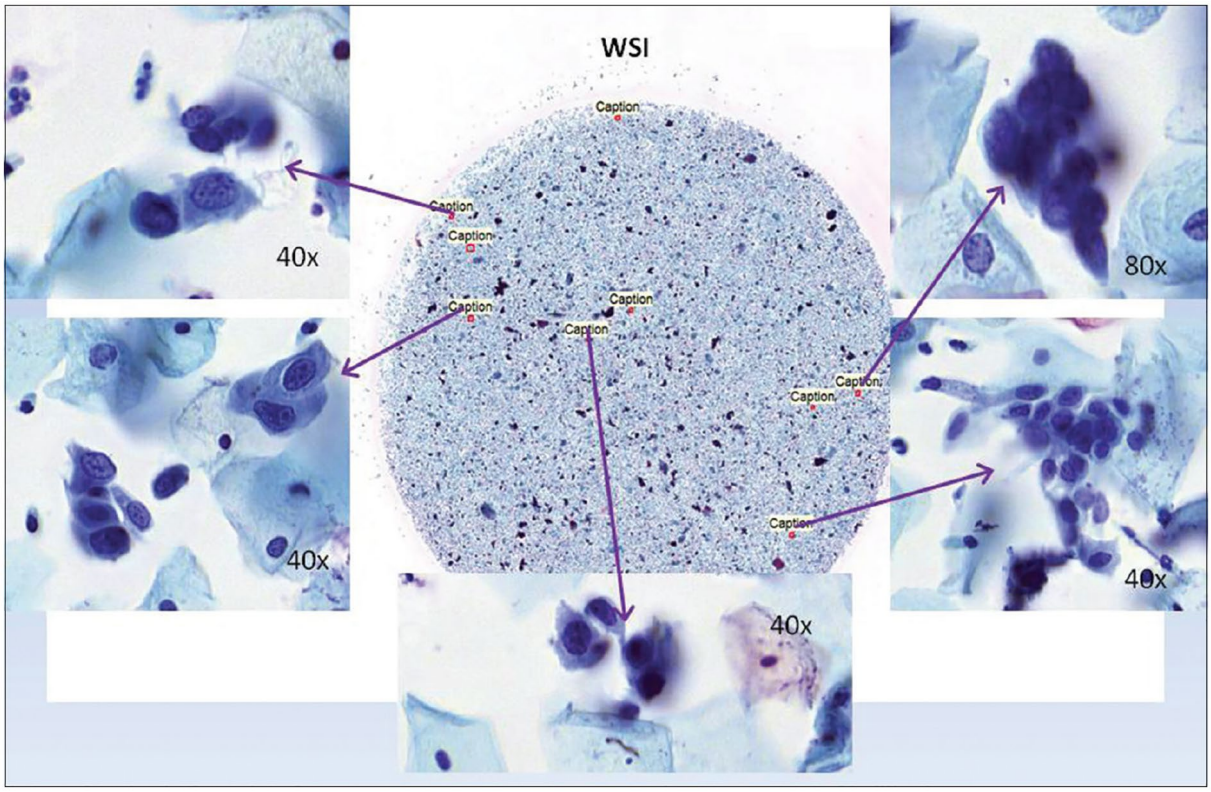

Figure 3. A representative example of a WSI digital slide of a liquid-based cytology cervical specimen with cytological detail. The case comes from Bongaerts et al.[19] WSI indicates whole slide imaging.

superior by 72% of the participants in one study[26] and to be superior with statistical significance on a 10-point rating scale by 75% of the participants in another.[16] Only the study by Hanna et al[16] reported separate evaluations of confidence and perceived quality of WSI for LBC and non-LBC cases, with WSI judged to be of inferior quality by 50% of the participants for LBC cases and with no difference in quality found for non-LBC cases; confidence in diagnosis was lower with WSI in both cases, however, with different degrees of statistical significance.

A summary of included studies is shown in Table 1 (see Supporting Table 3 for complete information).

DISCUSSION

WSI has represented a disruptive technology in the field of pathology. The potential to replace traditional LM with WSI has driven many studies to explore the concordance between these 2 modalities for making a diagnosis. Systematic reviews on the concordance of WSI and LM have shown that the overall diagnostic concordance is higher than 90%, sometimes with an excellent κ coefficient.[33,34] Some of these validation studies were designed specifically to satisfy CAP guidelines.[34] However, cytopathology slides differ from those used in surgical pathology. Cytology smears may cover the entire glass slide surface, may have areas of variable thickness, typically contain 3-dimensional groups of cells, and may have obscuring material (eg, blood, mucus, inflammatory cells, or ultrasound gel), and this makes conventional smears more difficult to digitize. As a result, focusing on cytology material warrants scanning at multiple planes (ie, Z-stacking). LBC preparations, which uniformly distribute and concentrate cells in a reduced area of the glass slide, can partly help to overcome this limitation,[6] and this is reflected in the slightly better performance of WSI in the LBC subgroup of studies. These barriers have limited the widespread adoption of WSI in cytopathology for routine use in clinical practice. Consequently, studies of comparisons regarding the diagnostic performance of WSI in cytology have been limited largely to academic institutions and to research or training purposes.[5,7]

Across studies that directly compared WSI and LM diagnoses, the overall percentage of concordance with the original reference diagnosis was 84.1%, the mean

TABLE 1. Summary of Included Studies

Source	No. of Cases	Pathology Type	Scanner; Z-Stack; No. of Planes	Main Outcome	Other Outcomes
Arnold 2015[23]	21	Mixed pediatric	Aperio model XT; single plane	Concordance rate of WSI with original diagnosis, 74.3%	NP
Bongaerts 2018[19]	1005	Cervical specimens (thin layer, SurePath)	3DHistech 250 flash scanner; single plane	Intra-observer κ for WSI-LM, 0.67 (CI, 0.60-0.74); overall, 95.3% (CI, 93%-96.9%)	Interobserver κ for LM, 0.75 (CI, 0.68-0.80); overall, 97.8% (CI, 96%-99%)
Dee 2007[14]	5	Cervical specimens (LBC, NOS)	MicroBrightField; Z-stack, NOS	Intra-observer concordance for WSI-LM, 92%; concordance rate with original diagnosis, 96% for LM and 94% for WSI	Time needed to examine slides, 8-11 min for WSI and 5-6 min for LM; perception of quality, ease, and speed >2 on a scale of 1-4
Della Mea 2006[31]	4	Cervical specimens, NOS	Institutionally built system	κ for WSI with reference diagnosis, 0.60	Time needed to examine slides, >12 min
Donnelly 2013[15]	192	Cervical specimens (LBC, SurePath)	iScan Coreo Au scanner (Ventana); 3 focal planes	Interobserver κ, 0.94 for LM (CI, 0.89-0.99) and 0.82 fir WSI (CI, 0.78-0.86); intra-observer concordance for WSI-LM, 89%-97%	Participants' perception of WSI use highly variable
Evered & Dudding 2011[24]	20	Cervical specimens (LBC, SurePath)	Hamamatsu NanoZoomer HT; 5 and 21 focal planes	Concordance with reference diagnosis, 70% for WSI and 74% for LM	Time to diagnosis, 18 min for WSI and 8 min for LM
Gerhard 2013[20]	202	Thyroid FNAC	Hamamatsu NanoZoomer 2.0HT	Intra-observer concordance for WSI-LM, 77.5%; κ, 0.54 (CI, 0.44-0.64)	Interobserver concordance with WSI, 80.2%; κ, 0.57 (CI, 0.46-0.67)
Gould & Saikali 2012[25]	30	CNS smears	Hamamatsu NanoZoomer 2.0HT	Concordance with original diagnosis, 29/30	NP
Hang 2015[26]	10	5 cervical, 5 mixed nongynecological	Leica SCN400; single plane	Concordance of WSI with original diagnosis, 28.5%-93.7%	WSI quality perceived to be inferior to LM quality
Hanna 2017[16]	30	10 cervical (LBC, ThinPrep), 20 mixed	Aperio ScanScope XT; single plane	Intra-observer concordance for WSI-LM, 90%-100%	Time on slide, 2 min for LM and 3.5 for min WSI ($P < .001$); lower quality of WSI on a 1-10 scale; no significant differences in confidence
House 2013[27]	22	Mixed	Aperio ScanScope XT; single plane	Concordance with original diagnosis, 92.5% for LM and 85.9% for WSI	Longer time for WSI (3.0-5.6 min) than LM (2.0-3.1 min); more than 90% of WSI slides of good quality
Marchevsky 2006[28]	20	Cervical specimens	Aperio ScanScope; single plane	Number of correct diagnoses per participant with 2 grading systems, 20%-100% (range)	Time to diagnosis, 4-40 min; perception of quality on a 1-4 scale
Mukherjee 2015[17]	12	Thyroid FNAC	iScan Coreo Au scanner (Ventana); 7, 5, and 3 focal planes	Intra-observer concordance for WSI-LM, 62%-100%; κ for interobserver concordance with WSI, 0.69-0.77 at different planes; κ for LM, 0.77; concordance of both LM and WSI with original diagnosis, 83%-86%	Screening time significantly lower for 2 of 3 participants with LM
Ross 2018[29]	56	Mixed	Aperio ScanScope XT; >1 for some cases, NOS	Concordance of WSI with original reference diagnosis, 40%-99% (range)	NP
Slodkowska 2009[32]	28	Thoracic	Aperio ScanScope SC	Concordance of diagnosis with final diagnosis on gross specimen, 28/28	NP
Steinberg & Ali 2001[30]	10	Cervical specimens (LBC, AutoCyte)	BLISS system, NOS	Concordance with original diagnosis, 98.3%	Scanning time (58-min average for 20%-30% of slide area); time per case diagnosis, 4.1 min
Stewart 2007[21]	30	Cervical specimens (LBC, ThinPrep)	Aperio T3 ScanScope; line scan; single plane	Concordance with original diagnosis, 93.3%-100%	Time-to-screen range, 74.6-189.3 min

TABLE 1. *Continued*

Source	No. of Cases	Pathology Type	Scanner; Z-Stack; No. of Planes	Main Outcome	Other Outcomes
Wright 2013[22]	11	Cervical specimens (LBC, 6 SurePath and 5 ThinPrep)	iScan Coreo Au 3.0 (Ventana); 1 and 7 planes	Concordance with original diagnosis, 14%-100% for WSI and 57%-100% for LM	Time to diagnosis significantly longer for WSI
Yao 2018[18]	60	Mixed	Hamamatsu NanoZoomer 9600-12; single Z-stack	Intra-observer κ for adequacy, 0.86-1.00; intra-observer κ for diagnosis, 0.75-0.93; interobserver κ for adequacy, 0.74 for WSI and 0.74 for LM; interobserver κ for diagnosis, 0.70 for WSI and 0.67 for LM	Average time to diagnosis, 113 s for LM and 122 s WSI

Abbreviations: CI, confidence interval; CNS, central nervous system; FNAC, fine-needle aspiration cytology; LBC, liquid-based cytology; LM, light microscopy; NOS, not otherwise specified; NP, not present in the study; WSI, whole slide imaging.

intra-observer percentage concordance for WSI was 92.5%, and the mean intra-observer κ coefficient was 0.66. The importance of intra-observer concordance for preventing inter-reader variation for validation purposes was stressed in the CAP guideline.[3] It would be interesting to compare the intra-observer concordance achieved with LM with that achieved with WSI, but unfortunately, intra-observer concordance only for LM is rarely reported. Only Bongaerts et al[19] reported an overall intra-observer concordance between LM diagnoses (97.8%) and between WSI and LM diagnoses (95.3%), with slightly overlapping confidence intervals. Such a comparison would permit us to demonstrate the noninferiority of WSI to LM for diagnostic concordance, as reported in previous validation studies for WSI in surgical pathology[35] and in recent systematic reviews on the topic.[33,34] At the same time, across studies that reported interobserver concordance with the κ coefficient,[15,17-20] the mean κ coefficient was 0.69 for WSI and 0.78 for LM, which were both in the range of substantial agreement. Interobserver variability is likely to be influenced by factors other than just the viewing modality, such as the expertise of the diagnosticians, the types and difficulty of the cases, and previous training in using the digital modality. The κ coefficient values that we found are in line with those reported by Goacher et al[33] in their review of WSI concordance for surgical pathology. However, the intra-observer κ coefficient of 0.66 from our analysis is lower than that found by Araujo et al[34] in their recent review of surgical pathology cases. However, Araujo et al included only studies designed according to CAP guidelines published after 2012, whereas our review comprises studies of different designs starting from 2001 with relevant heterogeneity; this is similar to what is seen in Goacher et al's work. Moreover, this finding could also reflect the difficulty of focusing related to cytology. Studies involving liquid-based Papanicolaou tests included the largest proportion of cases evaluated. In these particular studies, the overall percentage of concordance with the original diagnosis was 88.2%, which was slightly higher than the overall concordance of 84.1% across all studies. On the other hand, when only the subgroup of non-LBC studies was considered, the overall percentage of concordance with the original diagnosis decreased to 81.4%, and the intra-observer and interobserver κ coefficients decreased to 0.61 and 0.60, respectively, which were slightly lower than the values for the entire population of studies. As

mentioned previously, LBC is more amenable to WSI than the scanning of direct smears. The reasons for the worse performance of WSI in nongynecological, non-LBC cases may reside in the variable thickness of direct smears and the presence of obscuring material and artifacts. Cytopathology specimens in this group varied and included thyroid fine-needle aspiration[17,20] and central nervous system smears[25] as well as fine-needle aspiration samples of other sites and body fluid preparations.[18,23,26,29]

The secondary outcome explored in our review was the time it takes by diagnosticians to make a diagnosis. Such information was recorded in 11 studies (58%).[14,16-18,21,22,24,27,28,30,31] In all of these studies, the time for slide screening and diagnosis rendering was longer (1-2 times more) with WSI than LM. Three studies found a statistically significant difference in time spent, with a clear advantage from using LM,[16,17,22] and this was found in both LBC and non-LBC studies. When diagnosticians in one study were divided according to their expertise with cytology and WSI use, the authors found that cytotechnologists were the fastest with both modalities and that residents were the slowest with both modalities.[27] Making a diagnosis with WSI has also been reported to take longer in surgical pathology.[36-38] Furthermore, in contrast to surgical pathology, the time for a diagnosis is of particular importance in Papanicolaou test screening programs, where a high diagnostic workload is present. Scanning times were not reported in the majority of studies, and when they were reported, they appeared to be very long. However, these studies were performed with older technology.[30,31]

Another outcome that we investigated was the perceived quality of WSI slides and the confidence of the cytologist when making a diagnosis. These topics were documented in less than one-third of the included studies, and when they were reported, a nonstatistical evaluation was used. Nonetheless, they indicate that glass slides are better with respect to perceived quality and confidence in the diagnosis. This is not surprising because most of the studies used WSI scanning with only a single plane of focus. Notably, similar considerations were also found in a study with Z-stacking.[15] Even if Z-stacking can help to achieve better quality digital images and increase confidence in the diagnosis, a deeper comparison is limited by the fact that among the studies using more than a single plane, only 1 dealt with non-LBC specimens.[17] Training may also have a bearing when one is evaluating the perception of WSI slides, as suggested by related studies involving surgical pathology.[33] It is hypothesized that training with and exposure to WSI use will increase diagnostic confidence and decrease the time for a diagnosis. In 9 of the articles (47%) included in this review, the participants did receive training or basic instruction in the usage of WSI, and in 4 of these 9 articles, there also was reporting of the perceived quality of WSI images, which varied from 3 to 4 on a scale of 0 to 4[14,28] to significantly lower on a 10-point scale in another.[16] In the other 2 studies reporting the perceived quality of WSI images, the participants were not trained, and the quality of the slides was reported as lower in one study[26] and higher in the other one.[27]

In general, the studies included for review were heterogeneous in terms of the types of cytopathology cases investigated, the types of participants, and the number of cases evaluated. According to the CAP guidelines on validation, at least 60 cases are necessary for clinical validation of WSI. We found that only 4 of the included studies reached this volume.[15,18-20] Arnold et al[23] declared that for cytological specimens, it was impossible to reach this required number because of technical difficulties with image acquisition and quality, and they thus underlined again the difference between cytology and histology slides. For the washout period, this was not documented in certain studies where a comparison was made only with the original diagnosis. This point represented a main variable for the risk of bias and applicability in the flow and timing domain of QUADAS-2. In the patient selection domain, the most important parameter found was the nonrandom or nonconsecutive selection of cases. Unfortunately, in a large proportion of the publications, there was no explanation of the criteria used for the selection of cases.

In conclusion, this review provides limited evidence on the diagnostic concordance of WSI and LM in cytopathology, which appears to be in the range of substantial agreement when it is assessed as a κ coefficient for both intra-observer and interobserver agreement with WSI. A slightly better performance of WSI is achieved in LBC cases because this kind of specimen is less prone to the technical difficulties of conventional smears (eg, a more uniform distribution of material in a monolayer). However, this is based only on a few retrospective studies that documented heterogeneous characteristics with respect to case selection, user training, and outcome

measurements. These data show that the time required to make a diagnosis appears to be longer with WSI than LM, and this represents a major barrier to routine use in practice that could have a negative impact on Papanicolaou test screening programs with a high diagnostic workload. In the near future, technical advances related to WSI scanner speeds, Z-stacking, user interfaces (eg, image galleries), and the application of artificial intelligence will help to overcome some of these barriers. We anticipate that future studies using scanners with better technological capabilities and investigations with a more focused study design will provide stronger evidence when they compare WSI and LM for the purpose of rendering cytological diagnoses.

FUNDING SUPPORT

No specific funding was disclosed.

CONFLICT OF INTEREST DISCLOSURES

Liron Pantanowitz reports personal fees from Hamamatsu and Leica outside the submitted work. The other authors made no disclosures.

REFERENCES

1. Park S, Pantanowitz L, Parwani AV. Digital imaging in pathology. *Clin Lab Med*. 2012;32:557-584. doi:10.1016/j.cll.2012.07.006
2. Pantanowitz L, Wiley CA, Demetris A, et al. Experience with multimodality telepathology at the University of Pittsburgh Medical Center. *J Pathol Inform*. 2012;3:45. doi:10.4103/2153-3539.104907
3. Pantanowitz L, Sinard JH, Henricks WH, et al. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med*. 2013;137:1710-1722. doi:10.5858/arpa.2013-0093-CP
4. College of American Pathologists. Validating Whole Slide Imaging for Diagnostic Purposes in Pathology. Published 2019. Accessed July 28, 2019. https://www.cap.org/protocols-and-guidelines/cap-guidelines/current-cap-guidelines/validating-whole-slide-imaging-for-diagnostic-purposes-in-pathology
5. Khalbuss WE, Pantanowitz L, Parwani AV. Digital imaging in cytopathology. *Patholog Res Int*. 2011;2011:264683. doi:10.4061/2011/264683
6. Capitanio A, Dina RE, Treanor D. Digital cytology: a short review of technical and methodological approaches and applications. *Cytopathology*. 2018;29:317-325. doi:10.1111/cyt.12554
7. Pantanowitz L, Hornish M, Goulart R. The impact of digital imaging in the field of cytopathology. *Cytojournal*. 2009;6:6. doi:10.4103/1742-6413.48606
8. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. *PLoS Med*. 2009;6:e1000097. doi:10.1371/journal.pmed.1000097
9. Groen R, Abe K, Yoon HS, et al. Application of microscope-based scanning software (Panoptiq) for the interpretation of cervicovaginal cytology specimens. *Cancer Cytopathol*. 2017;125:918-925. doi:10.1002/cncy.21921
10. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5:210. doi:10.1186/s13643-016-0384-4
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
12. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529-536. doi:10.7326/0003-4819-155-8-201110180-00009
13. University of Bristol. QUADAS. Accessed July 19, 2019. http://www.bristol.ac.uk/population-health-sciences/projects/quadas/resources/
14. Dee FR, Donnelly A, Radio S, Leaven T, Zaleski MS, Kreiter C. Utility of 2-D and 3-D virtual microscopy in cervical cytology education and testing. *Acta Cytol*. 2007;51:523-529. doi:10.1159/000325788
15. Donnelly AD, Mukherjee MS, Lyden ER, et al. Optimal z-axis scanning parameters for gynecologic cytology specimens. *J Pathol Inform*. 2013;4:38. doi:10.4103/2153-3539.124015
16. Hanna MG, Monaco SE, Cuda J, Xing J, Ahmed I, Pantanowitz L. Comparison of glass slides and various digital-slide modalities for cytopathology screening and interpretation. *Cancer Cytopathol*. 2017;125:701-709. doi:10.1002/cncy.21880
17. Mukherjee MS, Donnelly AD, Lyden ER, et al. Investigation of scanning parameters for thyroid fine needle aspiration cytology specimens: a pilot study. *J Pathol Inform*. 2015;6:43. doi:10.4103/2153-3539.161610
18. Yao K, Shen R, Parwani A, Li Z. Comprehensive study of telecytology using robotic digital microscope and single Z-stack digital scan for fine-needle aspiration–rapid on-site evaluation. *J Pathol Inform*. 2018;9:49. doi:10.4103/jpi.jpi_75_18
19. Bongaerts O, Clevers C, Debets M, et al. Conventional microscopical versus digital whole-slide imaging-based diagnosis of thin-layer cervical specimens: a validation study. *J Pathol Inform*. 2018;9:29. doi:10.4103/jpi.jpi_28_18
20. Gerhard R, Teixeira S, Gaspar da Rocha A, Schmitt F. Thyroid fine-needle aspiration cytology: is there a place to virtual cytology? *Diagn Cytopathol*. 2013;41:793-798. doi:10.1002/dc.22958
21. Stewart J, Miyazaki K, Bevans-Wilkins K, Ye C, Kurtycz DFI, Selvaggi SM. Virtual microscopy for cytology proficiency testing: are we there yet? *Cancer*. 2007;111:203-209. doi:10.1002/cncr.22766
22. Wright AM, Smith D, Dhurandhar B, et al. Digital slide imaging in cervicovaginal cytology: a pilot study. *Arch Pathol Lab Med*. 2013;137:618-624. doi:10.5858/arpa.2012-0430-OA
23. Arnold MA, Chenever E, Baker PB, et al. The College of American Pathologists guidelines for whole slide imaging validation are feasible for pediatric pathology: a pediatric pathology practice experience. *Pediatr Dev Pathol*. 2015;18:109-116. doi:10.2350/14-07-1523-OA.1
24. Evered A, Dudding N. Accuracy and perceptions of virtual microscopy compared with glass slide microscopy in cervical cytology. *Cytopathology*. 2011;22:82-87. doi:10.1111/j.1365-2303.2010.00758.x
25. Gould PV, Saikali S. A comparison of digitized frozen section and smear preparations for intraoperative neurotelepathology. *Anal Cell Pathol (Amst)*. 2012;35:85-91. doi:10.3233/ACP-2011-0026
26. Hang JF, Liang WY, Hsu CY, Lai CR. Integrating a web-based whole-slide imaging system and online questionnaires in a national cytopathology peer comparison educational program in Taiwan. *Acta Cytol*. 2015;59:278-283. doi:10.1159/000430901
27. House JC, Henderson-Jackson EB, Johnson JO, et al. Diagnostic digital cytopathology: are we ready yet? *J Pathol Inform*. 2013;4:28. doi:10.4103/2153-3539.120727
28. Marchevsky AM, Khurana R, Thomas P, Scharre K, Farias P, Bose S. The use of virtual microscopy for proficiency testing in gynecologic cytopathology: a feasibility study using ScanScope. *Arch Pathol Lab Med*. 2006;130:349-355. doi:10.1043/1543-2165(2006)130[349:TUOVMF]2.0.CO;2

29. Ross J, Greaves J, Earls P, Shulruf B, Van Es SL. Digital vs traditional: are diagnostic accuracy rates similar for glass slides vs whole slide images in a non-gynaecological external quality assurance setting? *Cytopathology*. 2018;29:326-334. doi:10.1111/cyt.12552
30. Steinberg DM, Ali SZ. Application of virtual microscopy in clinical cytopathology. *Diagn Cytopathol*. 2001;25:389-396. doi:10.1002/dc.10021
31. Della Mea V, Demichelis F, Viel F, Dalla Palma P, Beltrami CA. User attitudes in analyzing digital slides in a quality control test bed: a preliminary study. *Comput Methods Programs Biomed*. 2006;82:177-186. doi:10.1016/j.cmpb.2006.02.011
32. Slodkowska J, Pankowski J, Siemiatkowska K, Chyczewski L. Use of the virtual slide and the dynamic real-time telepathology systems for a consultation and the frozen section intra-operative diagnosis in thoracic/pulmonary pathology. *Folia Histochem Cytobiol*. 2009;47:679-684. doi:10.2478/v10042-010-0009-z
33. Goacher E, Randell R, Williams B, Treanor D. The diagnostic concordance of whole slide imaging and light microscopy: a systematic review. *Arch Pathol Lab Med*. 2017;141:151-161. doi:10.5858/arpa.2016-0025-RA
34. Araujo ALD, Arboleda LPA, Palmier NR, et al. The performance of digital microscopy for primary diagnosis in human pathology: a systematic review. *Virchows Arch*. 2019;474:269-287. doi:10.1007/s00428-018-02519-z
35. Bauer TW, Schoenfield L, Slaw RJ, Yerian L, Sun Z, Henricks WH. Validation of whole slide imaging for primary diagnosis in surgical pathology. *Arch Pathol Lab Med*. 2013;137:518-524. doi:10.5858/arpa.2011-0678-OA
36. Randell R, Ruddle RA, Thomas RG, Mello-Thoms C, Treanor D. Diagnosis of major cancer resection specimens with virtual slides: impact of a novel digital pathology workstation. *Hum Pathol*. 2014;45:2101-2106. doi:10.1016/j.humpath.2014.06.017
37. Jen KY, Olson JL, Brodsky S, Zhou XJ, Nadasdy T, Laszik ZG. Reliability of whole slide images as a diagnostic modality for renal allograft biopsies. *Hum Pathol*. 2013;44:888-894. doi:10.1016/j.humpath.2012.08.015
38. Velez N, Jukic D, Ho J. Evaluation of 2 whole-slide imaging applications in dermatopathology. *Hum Pathol*. 2008;39:1341-1349. doi:10.1016/j.humpath.2008.01.006