Virchows Archiv (2019) 474:269–287
https://doi.org/10.1007/s00428-018-02519-z

REVIEW AND PERSPECTIVES

The performance of digital microscopy for primary diagnosis in human pathology: a systematic review

Anna Luíza Damaceno Araújo[1] · Lady Paola Aristizábal Arboleda[1] · Natalia Rangel Palmier[1] · Jéssica Montenegro Fonsêca[1] · Mariana de Pauli Paglioni[1] · Wagner Gomes-Silva[1,2,3] · Ana Carolina Prado Ribeiro[1,2,4] · Thaís Bianca Brandão[2] · Luciana Estevam Simonato[4] · Paul M. Speight[5] · Felipe Paiva Fonseca[1,6] · Marcio Ajudarte Lopes[1] · Oslei Paes de Almeida[1] · Pablo Agustin Vargas[1] · Cristhian Camilo Madrid Troconis[7] · Alan Roger Santos-Silva[1]

Received: 9 August 2018 / Revised: 25 December 2018 / Accepted: 28 December 2018 / Published online: 26 January 2019

Abstract

Validation studies of whole slide imaging (WSI) systems produce evidence regarding digital microscopy (DM). This systematic review aimed to provide information about the performance of WSI devices by evaluating intraobserver agreement reported in previously published studies as the best evidence to elucidate whether DM is reliable for primary diagnostic purposes. In addition, this review delineates the reasons for the occurrence of discordant diagnoses. Scopus, MEDLINE/PubMed, and Embase were searched electronically. A total of 13 articles were included. The total sample of 2145 had a majority of 695 (32.4%) cases from dermatopathology, followed by 200 (9.3%) cases from gastrointestinal pathology. Intraobserver agreements showed an excellent concordance, with values ranging from 87% to 98.3% (κ coefficient range 0.8–0.98). Ten studies (77%) reported a total of 128 disagreements. The remaining three studies (23%) did not report the exact number and nature of disagreements. Borderline/challenging cases were the most frequently reported reason for disagreements (53.8%). Six authors reported limitations of the equipment and/or limited image resolution as reasons for the discordant diagnoses. Within these articles, the reported pitfalls were as follows: difficulties in the identification of eosinophilic granular bodies in brain biopsies; eosinophils and nucleated red blood cells; and mitotic figures, nuclear details, and chromatin patterns in neuropathology specimens. The lack of image clarity was reported to be associated with difficulties in the identification of microorganisms (e.g., *Candida albicans*, *Helicobacter pylori*, and *Giardia lamblia*). However, authors stated that the intraobserver variances do not derive from technical limitations of WSI. A lack of clinical information was reported by four authors as a source for disagreements. Two studies (15.4%) reported poor quality of the biopsies, specifically small size of the biopsy material or inadequate routine laboratory processes as reasons for disagreements. One author (7.7%) indicated the lack of immunohistochemistry and special stains as a source for discordance. Furthermore, nine studies (69.2%) did not consider the performance of the digital method—limitations of the equipment, insufficient magnification/limited image resolution—as reasons for disagreements. To summarize the pitfalls of digital pathology practice and better address the root cause of the diagnostic discordance, we suggest a

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00428-018-02519-z) contains supplementary material, which is available to authorized users.

✉ Alan Roger Santos-Silva
alan@unicamp.br

[1] Oral Diagnosis Department, Semiology and Oral Pathology Areas, Piracicaba Dental School, University of Campinas (UNICAMP), Av. Limeira, 901, Bairro Areião, Piracicaba, São Paulo 13414-903, Brazil

[2] Dental Oncology Service, Instituto do Câncer do Estado de São Paulo, Faculdade de Medicina da Universidade de São Paulo (FMSUP), São Paulo, Brazil

[3] Medical School of Nove de Julho University, São Paulo, Brazil

[4] Faculdade de Odontologia, Fernandópolis, Universidade Brasil, São Paulo, Brazil

[5] Unit of Oral and Maxillofacial Pathology, School of Clinical Dentistry, University of Sheffield, Sheffield, UK

[6] Clinic, Pathology and Odontological Surgery Department, Minas Gerais Dental School, Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil

[7] Dentistry Program, Health of Science Faculty, Corporación Universitaria Rafael Nuñez (CURN), Cartagena de Índias, Colombia

Categorization for Digital Pathology Discrepancies to be used in further validations studies. Among 99 discordances, only 37 (37.3%) had preferred diagnosis rendered by means of WSI. The risk of bias and applicability concerns were judged with the QUADAS-2. Two studies (15.4%) presented an unclear risk of bias in the sample selection domain and 2 (15.4%) presented a high risk of bias in the index test domain. Regarding applicability, all studies included were classified as a low concern in all domains. The included studies were optimally designed to validate WSI for general clinical use, providing evidence with confidence. In general, this systematic review showed a high concordance between diagnoses achieved by using WSI and conventional light microscope (CLM), summarizes difficulties related to specific findings of certain areas of pathology—including dermatopathology, pediatric pathology, neuropathology, and gastrointestinal pathology—and demonstrated that WSI can be used to render primary diagnoses in several subspecialties of human pathology.

Keywords Whole slide imaging · Intraobserver agreement · Systematic review

Introduction

Validation studies regarding the feasibility of whole slide imaging (WSI) systems have been conducted by pathology laboratories in a wide range of subspecialties to produce solid evidence and support the use of this technology for several applications, including primary diagnosis. The guideline statement of the College of American Pathologists Pathology and Laboratory Quality Center (CAP-PLQC) for WSI systems validation summarizes recommendations, suggestions, and expert consensus opinion about the methodology of validation studies in an effort to standardize the process. This guideline encompasses the need to include a sample set of at least 60 cases for one application and to establish a diagnostic concordance between digital and glass slides for the same observer—intraobserver variability—with a minimum washout period of 2 weeks between views [1]. Surprisingly, the recommendations do not suggest a consecutive or random selection of the cases or a need to blind evaluators, but they do highlight that the viewing can be random or non-random.

Validation studies are cross-sectional studies by definition, and their designs have many methodological variations, which should be considered when evidence is assembled [2]. All these variations lead to skewed estimates about the test accuracy. The most important variation concerns how the sample was selected, included, and analyzed [3]. Some aspects regarding configuration, the purpose of the test, and the risks that prevent the test from serving its purposes may have been considered in validation studies, since performance may be influenced by analysis bias; reproducibility; washout period; response time; and size, scope, and suitability of certain types of specimens. Besides that, the learning curve and performance problems may be related to the method or to the pathologists [2]. Apparently, the order of analyses—digital or conventional—does not affect the interpretation in this context [3].

The most common biases in diagnostic studies are verification bias/detection bias/work-up bias (when the reference standard is not applied in all sample), incorporation bias (when the index test and reference standard are not independent, which leads to overestimation of the sensitivity and specificity of the test), and inspection bias (when the tests are not blinded). The methodological characteristics should be individually evaluated by domain, which represents the way that the study was conducted [4].

The most common problems identified in the design of previously published validation studies are the case selection—samples selected have a narrow range of subspecialty specimens or known malignant diagnoses—and the comparisons of the study results with a "gold standard"/consensus diagnosis/expert diagnosis instead of establishing the concordance by assessing the intraobserver agreement [5].

The FDA recently approved a WSI system for primary diagnosis purposes [6] and, even though this statement highlighted some assurance about the safety and feasibility of the digital system, only one device was tested and approved. Regardless of this achievement, individual validation studies conducted by each laboratory and customized for each service and WSI system used are still necessary and will provide the best evidence to attest the feasibility of digital pathology, especially if based on CAP-PLQC guidelines.

Given the absence of a broader collective agreement on the use of WSI in a human pathology context, it is necessary to assemble evidence regarding the performance of digital microscopy in order to establish whether this technology can be used to provide a primary diagnosis. Therefore, this systematic review tested the diagnostic performances of WSI in human pathology. In addition, this review provided access to the main reasons for disagreement occurrences.

Materials and methods

The present systematic review was conducted following the guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [7] and was registered with the PROSPERO database under the protocol CRD42018085593. The review question defined was: "Is digital microscopy performance as reliable for use in clinical practice and routine surgical pathology for diagnostic purposes as conventional microscopy?" The best evidence to answer this question is from intraobserver agreement [1].

Definition of eligibility criteria

The eligibility criteria (Table 1) were elaborated based on two important recommendations and one suggestion established by CAP-PLQC guideline [1]: the validation process should include a sample set of at least 60 cases for one application, the validation study should establish diagnostic concordance between digital and glass slides for the same observer (i.e., intraobserver variability), and a washout period of at least 2 weeks should occur between viewing digital and glass slides.

Literature review

Recognizing the need to check if there are similar systematic reviews registered, executed, in progress or published with the same theme, the primary researcher (ALA) conducted a previous literature review. A systematic review with a similar proposal registered with the PROSPERO in 2015 was in progress, entitled: "The diagnostic accuracy of digital microscopy: a systematic review"; it was under the protocol CRD42015017859. Two published systematic reviews were found: "A systematic analysis of discordant diagnoses in digital pathology compared with light microscopy" [8] and "The Diagnostic Concordance of Whole Slide Imaging and Light Microscopy: A Systematic Review" [9]. Based on these findings, the research team decided to proceed with the present systematic review, since the methodology of the present review focused on studies supported by the CAP-PLQC guidelines [1]. These well-designed studies can provide much more reliable evidence about the utilization of WSI systems performance to provide a primary diagnosis in human pathology than the previously published systematic reviews.

Table 1 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Validation cross-sectional study	Articles published in foreign language;
At least 60 cases[a]	Articles about telepathology, cytopathology or immunohistochemistry;
Intraobserver agreement	Sample with a known malignant diagnose;[b]
The concordance percentage	Articles with lack of information about how the sample was analyzed;
or kappa index should be reported[a]	
At least 2 weeks[a] washout period	Studies which the primary goal was not to examine diagnostic concordance between WSI and CLM;
	Studies which aimed to establish the intraobserver agreement but instead: used two different samples; in which each pathologist only performed diagnosis by one method; in which whole slide imaging diagnosis was compared to a consensus panel or original diagnosis (it is not intraobserver agreement)[b]

[a] CAP-PLQC Guidelines for WSI systems validation (Pantanowitz et al., 2013)

[b] Cornish et al. (2012)

Search strategy

An electronic search was carried out in these databases: Scopus (Elsevier, Amsterdam, the Netherlands), MEDLINE (Medline Industries, Mundelein, Illinois) by PubMed platform (National Center for Biotechnology Information, US National Library of Medicine, Bethesda, Maryland) and Embase (Elsevier, Amsterdam, the Netherlands). Scopus was the first database used (due to its interdisciplinary basis and article indexing capabilities) in order to align the keywords. The search strategy used was the following: [ALL (validation) AND ALL ("whole slide image")]. In sequence, the search was reproduced in the other databases. As result, 599 articles from Scopus, 132 from Embase, and 115 from PubMed were retrieved. A manual search was conducted in order to identify any eligible articles that may not have been retrieved by the search strategy, but none were compatible with the eligibility criteria.

Article screening and eligibility evaluation

Two reviewers (ALDA and ARSS) independently conducted the screening of articles by reading the title and abstract and excluding articles that clearly did not fulfill the eligibility criteria. The assessment of eligibility was guided by a flow diagram drawn on phase 2 of the quality assessment. The two reviewers proceeded to read the full text of the articles, screened them to identify the eligible articles; all primary reasons for exclusions were registered for the composition of the article selection flow chart. Rayyan QCRI was used as the reference manager to perform the screening of the articles, exclusion of duplicates, and registration of a primary reason for exclusion [10].

Extraction of qualitative and quantitative data and quality assessment

The data extraction was conducted by the primary researcher (ALDA) and guided by a tailored extraction data form (Appendix 1) originally suggested by The Cochrane Collaboration [11]. The tailored tool has 5 sections: general information, eligibility, interventions participants and sample, methods, the risk of bias assessment, applicability and outcomes. The section of "risk of bias assessment" and "applicability" was added based on the tailored QUADAS-2 (University of Bristol, Bristol, England), a tool designed to assess the quality of primary diagnostic accuracy studies. Specific guidance for each signaling question was produced and some signaling questions—which did not apply to the review—were removed (Appendix 2). Qualitative and quantitative data were tabulated and processed in Microsoft Excel®. The studies identified in this review were highly heterogeneous with regard to equipment utilized, magnification, the number of pathologists involved, specimen type (subspecialty), washout time, and mainly how the sample was analyzed. These variations in study design represent limitations

and did not justify meta-analysis but only allowed a narrative synthesis of the findings from the included studies.

Results

PRISMA flowchart

The search strategy identified a total of 846 records through database searching. After duplicates were removed, 681 records were screened; among these, 48 articles were selected to be assessed for eligibility. A total of 13 articles [12–24] were included and 35 articles were excluded based on eligibility criteria. The composition of the article selection flow is shown in Fig. 1.

One article (2.1%) [25] was excluded for being published in French, 1 (2.1%) [26] for having insufficient sample size, 1 (2.1%) [27] for having a sample with a known malignant diagnosis, and 11 studies (22.2%) [28–38] for presenting only abstracts (gray literature). Two studies (4.1%) [39, 40] were excluded because the main objective was not to examine diagnostic concordance between WSI and conventional light microscope (CLM). Four studies (8.3%) [8, 41–43] were excluded because they utilized insufficient washout time between the analyses.

The most important eligibility criteria establish that the intraobserver agreement should be the preferred measure to assess the performance of digital microscopy, according to CAP-PLQC guidelines [1]. Thirteen studies (27.1%) did not fit that criteria and were excluded for the following reasons: in six studies (12.5%) [44–49], the pathologists only assessed WSI and the concordance was reached by comparing WSI diagnosis with the original glass slide diagnosis; in four studies (8.3%) [50–53], the WSI diagnosis was compared to a consensus panel diagnosis; in one study (2.1%) [54], two groups of students only assessed WSI and the other only assessed glass slides; in two studies (4.1%) [55, 56], the sample analyzed was not the same in both methods. Two studies (4.1%) [57, 58] did not report either intraobserver concordance percentage or kappa value. Disagreements among the reviewers at the screening and assessment of eligibility were confronted and resolved by consensus.

Methodological characteristics of the studies

Publication dates ranged from 2010 to 2017. Only six articles (46.1%) [18–21, 23, 24] mentioned the use of CAP-PLQC guidelines, but methodologies of all the included studies were according to these guidelines. The included studies used scanners from eight different manufacturers. The most commonly used scanner was Scan Scope (Aperio, Vista, CA), which was reported in eight studies (61.5%) [13–16, 18, 19, 21, 23] (Table 2).

The aims of the studies were highly variable: five (38.4%) [13–15, 17, 21] aimed to test the feasibility of digital methods, two (15.4%) [21, 24] aimed to determine the utility of CAP-PLQC guidelines [1], two (15.4%) [20, 23] intend to assess primary digital pathology reporting, one (7.7%) [22] proposed to determine the accuracy of WSI interpretation, one (7.7%) [12] proposed to investigate whether conventional microscopy of skin tumors can be replaced by virtual microscopy, one (7.7%) [19] proposed to evaluate whether diagnosis from WSI is inferior to diagnosis of glass slides, and one (7.7%) [16] aimed to evaluate the use of WSI for diagnosis of placental tissue and pediatric biopsies.

The most relevant methodological characteristics of the included studies are shown in Table 3. Full information about the methodological characteristics of the studies included in this systematic review is available as supplementary material (Supplementary Table 1). Included studies performed

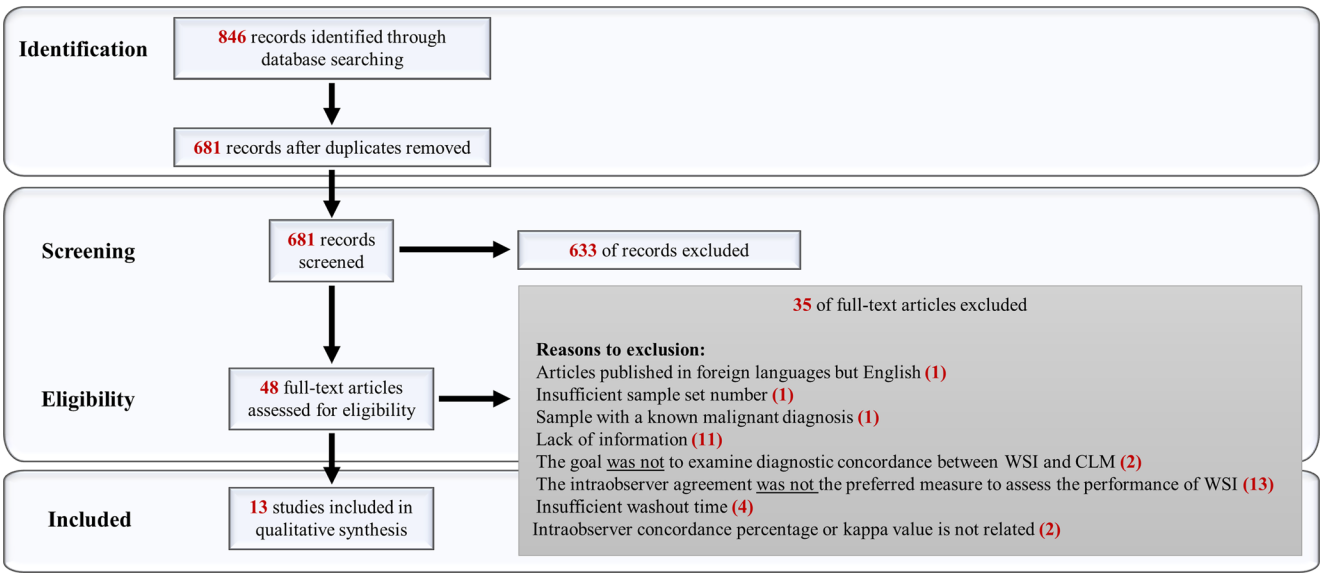


Fig. 1 Flow diagram of literature search adapted from PRISMA (Moher et al. 2009)

Table 2 Technical characteristics of the equipment used in included studies

Author/year	WSI system specifications
Al-Janabi et al. (2012)a	Scanner: ScanScope XT (Aperio Technologies Inc., Vista, CA, USA), 20x;
	Monitor settings/resolution: Samsung 245B (Samsung, Seoul, South Korea) displays of 24″ (resolution of 19,203 × 1200 pixels).
Al-Janabi et al. (2012)b	Scanner: ScanScope XT (Aperio Technologies Inc., Vista, CA, USA), 20x;
	Monitor settings/resolution: not mentioned.
Al-Janabi et al. (2012)c	Scanner: ScanScope XT (Aperio Technologies Inc., Vista, CA, USA);
	Monitor settings/resolution: 24-in displays (Samsung, Seoul, South Korea) with 1920 × 1200 pixels.
Al-Janabi et al. (2013)	Scanner: ScanScope XT (Aperio Technologies Inc., Vista, CA, USA), 20x;
	Monitor settings/resolution: not mentioned.
Al-Janabi et al. (2014)	Scanner: not mentioned
	Magnification: 20x
	Monitor settings/resolution: not mentioned.
Arnold et al. (2015)	Scanner: Aperio Model XT (Aperio Technologies Inc., Vista, CA, USA), 20x or 40x;
	Monitor settings/resolution: Dell monitors (Dell Corporation, Austin, TX, USA) with 1280–31,024-pixel.
Kent et al. (2017)	Scanner: AT2 Image Scope (Aperio Technologies Inc., Vista, CA, USA), 20x;
	Monitor settings/resolution: not mentioned
Loughrey et al. (2015)	Scanner: Hamamatsu Nanozoomer (Hamamatsu, United Kingdom), 40x;
	Monitor settings/resolution: not mentioned
Nielsen et al. (2010)	Scanner: Mirax Scan (Carl Zeiss MicroImaging, Göttingen, Germany), 20x;
	Monitor resolution: not mentioned
Pekmezci et al. (2016)	Scanner: ScanScope XT (Aperio Technologies Inc., Vista, CA, USA), 40x;
	Monitor resolution: not mentioned
Saco et al. (2017)	Scanner: Ventana iScan HT (Ventana Medical Systems, Tucson, AZ, USA) 400x;
	Monitor settings/resolution: 30 Coronis fusion MDC4130 monitor 4 Megapixels (Barco Electronic Systems, Barcelona, Spain)
Tabata et al. (2017)	Scanners, magnifications and monitor resolutions:
	IntelliSite Ultra Fast Scanner (Phillips Health, Amsterdam, Netherlands), 40x, 0.25 mm/pixel;
	Aperio AT2 Scanner (Leica Biosystems, San Diego, CA, USA), 20x, 0.5 mm/pixel;
	NanoZoomer 2.0-HT C9600-13 (Hamamatsu photonics, Hamamatsu, Shizuoka, Japan), 20x, 0.46 mm/pixel;
	NanoZoomer 2.0-RS C10730-13 (Hamamatsu photonics), 20x, 0.46 mm/pixel;
	NanoZoomer 2.0-RS C10730-13 (Hamamatsu photonics), 40 (0.23 mm/pixel) VS800 (Olympus Corporation, Tokyo, Japan), 40x, 0.185 mm/pixel;
	FINO (CLARO, Hirosaki, Aomori, Japan), 40x, 0.25 mm/pixel.
Thrall et al. (2015)	Scanner: iScan Coreo Au, 20x.
	Monitor resolution: 1280 3 1084 pixels

validations in the following areas: dermatopathology, neuropathology, gastrointestinal, genitourinary, breast, liver, and pediatric pathology. Surgical pathology specimens of pediatric pathology were gastrointestinal, heart, liver, lung, neuropathology, placenta, rectal suction, skin, and tonsil. Subsets also included endocrine, head and neck, hematopoietic organ, hepatobiliary-pancreatic organ, soft tissue, bone, hematopathology, medical kidney, and transplant biopsies.

These 13 papers included a total sample of 2145 glass slides and corresponding digital slides, in which the majority of 695 (32.4%) were from dermatopathology, followed by 200 (9.3%) from gastrointestinal pathology. The mean number of samples within the included studies was 165. Four studies included cases from various pathology subspecialties.

The samples were analyzed in two different ways: (1) pathologists assessed the cases with one modality and—after a washout period—they reassessed the cases with the other modality; (2) when WSI diagnoses were compared to original glass slides diagnoses, the cases were addressed to the original pathologist, providing a satisfactory washout period and maintaining the intraobserver agreement as the preferred measure. In one study (7.7%) [24], the cases were first evaluated half as glass slides and half as digital images and reviewed with the other modality after washout. The washout period between views within the included studies ranged from 2 weeks to 12 months.

Three studies (23%) [12, 23, 24] reported set training and eight (61.5%) stated that the pathologists had previous

Table 3 Methodological characteristics of the included studies

Author/ year	Subspecialty/ specimens (*n*)	Intraobserver agreement	Disagreements and PD	Disagreement reason	Conclusion of the study
Al-Janabi et al. (2012)a	Dermatopathology (*n* = 100)	94% (95% CI = 0.87–0.97)	WSI: 1 CLM: 5	Borderline cases	Primary histopathological diagnosis of skin biopsies and resections can be done digitally using WSI.
Al-Janabi et al. (2012)b	Gastrointestinal (*n* = 100)	95% (95% CI = 0.89–0.98)	WSI: 3 CLM: 2	Identification of microorganisms[c]	Histopathological diagnosis of routine gastrointestinal biopsies and resections can be done well on WSIs acquired using today's scanning technology.
Al-Janabi et al. (2012)c	Breast pathology (*n* = 100)	93% (95% CI = 86–97)	WSI:4 CLM: 0	Borderline cases	This study demonstrates that upfront histopathological diagnosis of breast biopsies and resections can reliably be done on digital slide image.
Al-Janabi et al. (2013)	Pediatric pathology (*n* = 80)	90% (95% CI = 0.84–0.96)	WSI: 1 CLM: 9	Identification of microorganisms[c]	Histopathological diagnosis of biopsies and resections can generally be done well on WSI acquired using today's scanning technology. 20× magnification was not optimal for exploring placental tissue.
Al-Janabi et al. (2014)	Genitourinary (*n* = 100)	87% (95% CI = 0.80–0.94)	WSI: 6 CLM: 7	LCI; absent of multidisciplinary discussion; lack of routine; limited image resolution; suboptimal navigation tools;	Primary diagnostics of urinary tract specimens can be reliably done on WSI.
Arnold et al. (2015)	Pediatric pathology (*n* = 473)	98.3%	WSI: 0 CLM: 1	Misidentification of the EGB (in brain biopsies), eosinophils and NRBC.	This study demonstrates that specimens representing the spectrum of pediatric surgical pathology practice can be reviewed using WSI.
Kent et al. (2017)[a]	Dermatopathology (*n* = 499)	94%	NML: 6 ML: 6 IC: 2 PD not mentioned	The inherent subjectivity of dysplasia in NML; challenging lesions with subjective nuances in ML; LCI (as photographs) in IC.	Diagnosis from WSI was found to be noninferior compared with diagnosis from traditional microscopy.
Loughrey et al. (2015)[a]	Gastrointestinal (*n* = 100)	95%	WSI: 4 CLM:10	Borderline cases	The study provides further evidence to support validation of digital slide viewing as an alternative to light microscopy for primary reporting in the setting of gastrointestinal pathology.
Nielsen et al. (2010)[a]	Dermatopathology (*n* = 96)	$\kappa = 0.93$	–	Diagnostic interpretation; complexity of the cases (actinic keratosis); LCI; poor quality of the biopsies/inadequate routine laboratory processes; lack of experience of the pathologists;	It is feasible to make histologic diagnosis on the skin tumor types represented in this study using virtual microscopy.
Pekmezci et al. (2016)	Neuropathology (*n* = 97)	Path 1: 94.9% Path 2: 88%	Path 1: 5 Path 2: 10 PD not mentioned.	Misidentification of mitotic figures; loss of nuclear details and distortion of the chromatin pattern; LCI; non-utilization of special stains;	An all-encompassing conclusion about the utility of WSI for diagnostic purposes may not be available. We recommend independent validation for each subspecialty of pathology to identify subspecialty specific concerns, so they can be properly addressed.
Saco et al. (2017)[a]	Liver (*n* = 100)	Path 1: 96.6% $k = 0.9$ (95% CI: 0.9–1) Path 2: 90.3% $k = 0.9$ (95% CI: 0.8–0.9)	–	Small size of the material or difficulty of the case.	WSI can be safely used for primary histological diagnosis of liver biopsies, including native and transplantation specimens.
Tabata et al. (2017)	GI tract, female genital and genitourinary organ, breast, endocrine, head and neck, skin, soft tissue and bone, hematopoietic, and	96% (95% CI = 94.2–96.8)	Discrepant cases: WSI: 1 CLM: 8 Minor discrepancies:	–	The results of this study demonstrated that WSI had good performance and usefulness for primary diagnosis.

Table 3 (continued)

Author/year	Subspecialty/specimens (*n*)	Intraobserver agreement	Disagreements and PD	Disagreement reason	Conclusion of the study
	hepatobiliary-pancreatic organ, ($n = 100$)		WSI: 17 CLM: 20		
Thrall et al. (2015)	Phase 1: hematopathology, neuropathology, medical kidney, and transplant biopsies; Phase 2: neoplasia (lymphomas, neuropathology tumors, melanomas, and soft tissue neoplasms), liver and gastroesophageal. (2 sets of 100 cases)	79%	–	Difficulty in seeing microorganisms[c]; lack of image clarity at magnification above 320; challenging cases; individual interpretation; insufficient attention to the critical foci; uncarefully analysis by pathologists; limited experience of the pathologists; WSI is disorienting and difficult to comprehensively analyze.	The results were felt to validate the use of WSI for the intended applications in our multiinstitucional laboratory system.

[a] Interobserver agreement were reported additional to intraobserver agreement

[b] Clinical information was provided

[c] *Candida albicans*, *Helicobacter pylori*, and *Giardia lamblia*

GI gastrointestinal, *PD* preferred diagnosis, *WSI* whole-slide imaging, *CLM* conventional light microscope, *LCI* lack of clinical information, *EBG* eosinophilic granular bodies, *NRBC* nucleated red blood cells, *NML* nonmelanocytic lesions, *ML* melanocytic lesions, *IC* inflammatory conditions

experience with WSI systems. One study (7.7%) [20] did not include a trained pathologist in the validation process but claimed that the pathologist was familiar with the method. Set training or previous experience was not mentioned in one study (7.7%) [18].

Only one study (7.7%) [13] measured the scan time of slides (stating that they took on average of 2.5 min) and only one (7.7%) [24] measured the diagnosis time (median time for glass slides was 132 s and 210 s for WSI). Two studies (15.4%) [16, 17] considered WSI more time-consuming than CLM, although no formal timings had been performed. A consensus diagnosis was mentioned to be used in three included studies (23%) [19, 20, 23].

Intraobserver concordance

Within the included studies, one (7.7%) [12] did not report the percentage of concordance but reported an almost perfect kappa index of 0.93. Two other studies (15.4%) [21, 22] reported the concordance percentage for each pathologist, instead of an overall concordance percentage. For these reasons, these three studies are not graphically represented on Fig. 2; however, they are detailed in Table 3. The majority of the intraobserver agreements reported showed an excellent concordance, with values ranging from 87% to 98.3% (κ coefficient range 0.8–0.98). Only one study (7.7%) [24] showed a lower concordance of 79%. All values of the intraobserver agreement are shown in Table 3. Interobserver agreements were reported additionally to the intraobserver agreement in four studies (30.7%) [12, 19, 20, 22].

Reasons for disagreements

Within these 13 included studies, ten (77%) reported a total of 128 disagreements [13–21, 23]. The other three studies (23%) [12, 22, 24] lacked the exact number and nature of the disagreement. We provide an overview of the reasons for disagreements—i.e., pitfalls—according to the subspecialties of pathology (Fig. 3). Among all the reasons that might explain the occurrence of disagreements, the most frequent were borderline, difficult, or challenging cases, which were reported in seven articles (53.8%) [12, 13, 15, 19, 20, 22, 24]. Along with the subjective nuances of non-melanocytic lesions with dysplasia and melanocytic lesions (widely considered to be challenging cases), the study by Kent et al. also reported the lack of clinical information in inflammatory lesions as reasons for disagreements [19]. Nielsen reported the challenging diagnosis—specifically referring to actinic keratosis—as reasons for discordances, as well as poor quality of the biopsies, the lack of clinical information, and inexperience of the pathologists.

Six authors reported limitations of the equipment and/or limited image resolution as reasons for disagreements. Within these, one study (7.7%) [18] indicated pitfalls regarding the identification of eosinophilic granular bodies in brain biopsies, eosinophils, and nucleated red blood cells (which demonstrate refractile eosinophilic cytoplasm). Another study (7.7%) [21] reported

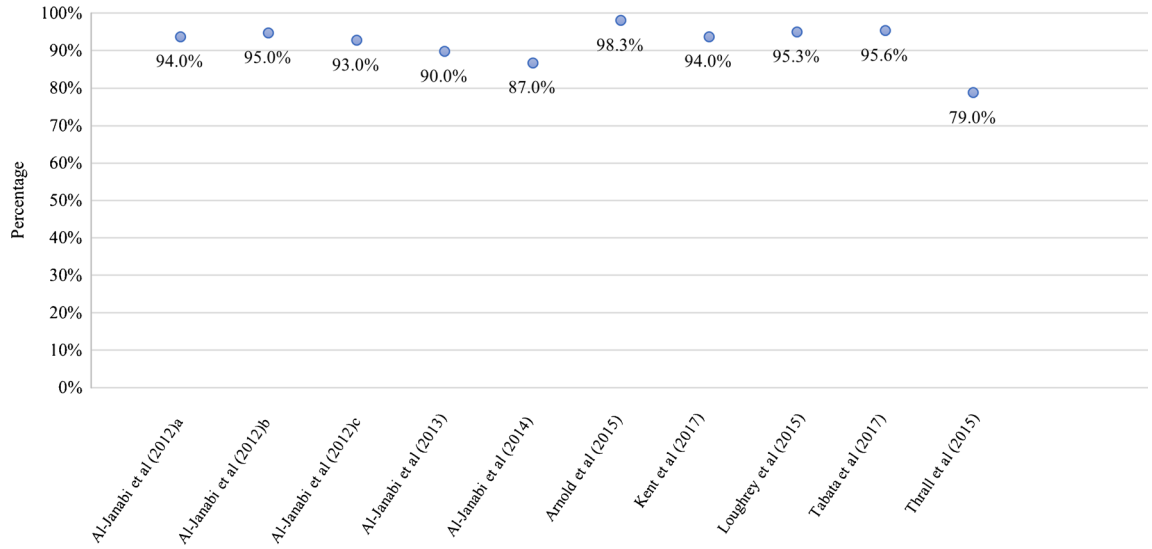


Fig. 2 Graphic presentation of intraobserver agreement of included studies

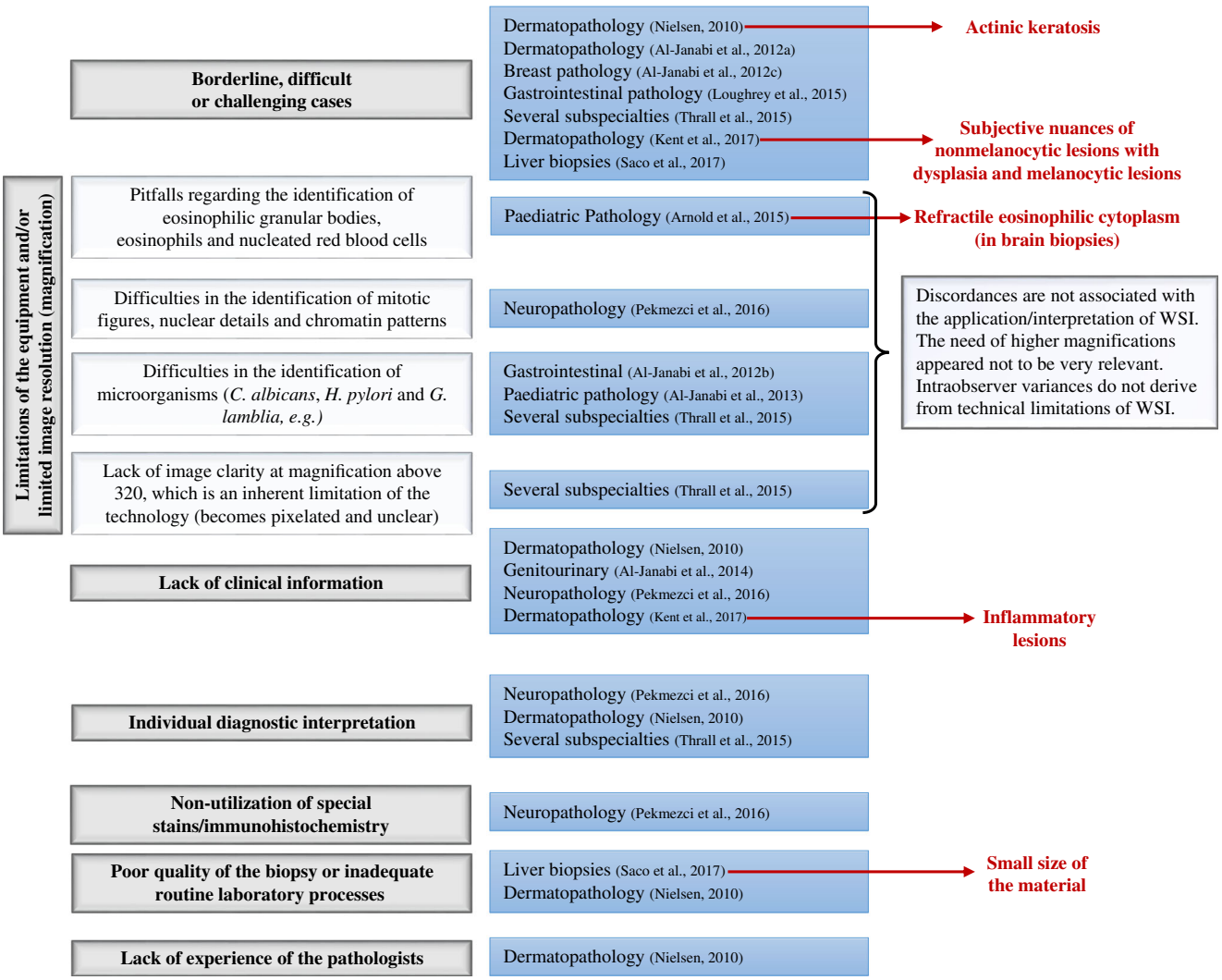


Fig. 3 Overview of the reasons for disagreements (pitfall) according to subspecialties of pathology

difficulties in the identification of mitotic figures, nuclear details, and chromatin patterns in neuropathology specimens. Three articles (23%) [14, 16, 24] reported difficulties in the identification of microorganisms (e.g., *Candida albicans*, *Helicobacter pylori*, and *Giardia lamblia*). Thrall et al. also reported a limitation of the technology related to lack of image clarity at a magnification above 320×: the image becomes pixelated and unclear [24]. However, authors stated that the intraobserver variances do not derive from technical limitations of WSI.

The lack of clinical information was reported by four authors [12, 17, 19, 21] as a source for disagreements.

Two studies (15.4%) reported poor quality of the biopsy, specifically the small size of the material [22] or inadequate routine laboratory processes [12] as reasons for disagreements.

Another reason cited was the utilization of suboptimal navigation tools reported by two authors (15.4%) [16, 17]. One author (7.7%) [23] remarked upon the difficulty to determine whether the discordance depends on disagreement between the methods or intraobserver disagreement of pathological diagnosis; it is possible the author intended to refer to the variations on the interpretations of pathological diagnosis, so intraobserver disagreement should not be used in this context. One author (7.7%) indicated the lack of immunohistochemistry special stains as a source for discordance [21].

Furthermore, nine studies (69.2%) [12–16, 19, 21, 22, 24] did not consider the performance of the digital method—i.e., limitations of the equipment, insufficient magnification/limited image resolution—as reasons for disagreements.

Eight studies (61.5%) [13–18, 20, 23] provided a preferred diagnosis when disagreements occurred. These preferred diagnoses were reached upon reviewing the discordant cases and choosing the most correct diagnosis. Among 99 disagreements, only 37 (37.3%) had preferred diagnoses rendered by means of WSI.

Categorization for digital pathology discrepancies

To summarize pitfalls of digital pathology practice and better address the root cause of the discordances, we developed a Categorization for digital Pathology Discrepancies, which can be used to report reasons for disagreements in further validation studies. This categorization can help to establish if there are valid concerns about the performance of the digital method (Table 4). We based this categorization on data retrieved from this systematic review and from the previously published systematic reviews [9, 59].

Quality assessment (risk of bias)

The results of the quality assessment are shown in Table 5 and Fig. 4. In 13 included articles, two (15.4%) [13, 14] presented an unclear risk of bias in the sample selection domain due to selection criteria of the sample remained unclarified (e.g., if it was randomized or consecutive). One study [21] excluded several lesions (pituitary adenomas, degenerated diseases or other reactive lesions, metastatic carcinomas and melanomas, vascular malformations, and other benign or descriptive diagnoses such as meningoceles, dermoid cysts, or focal cortical dysplasia) not relevant to the study and also excluded cases for which the slides were not available for WSI scanning. These exclusions were acceptable and do not indicate bias. Two studies (15.4%) [21, 24] presented a high risk of bias in the index test due to the absence of specification of a threshold. The term "threshold" is related to the parameters used to classify the diagnoses—e.g., if they were concordant, slightly discordant, or discordant. The risk of bias was considered low in 100% of the other domains in the remaining included studies. Regarding applicability, all studies included were classified as a low concern in all domains.

Table 4 Categorization for digital pathology discrepancies

Category	Description
A	Borderline, difficult or challenging cases/Individual diagnostic interpretation 1. Subjective nuances as lesions with dysplasia 2. Invasion areas missed 3. Inflammatory reaction patterns/architecture
B	Limitations of the equipment and/or limited image resolution (magnification) 1. Difficult in the identification of specific cells (inflammatory cell, e.g.) or nuclear details/features 2. Difficult in the identification of mucin or amyloid 3. Difficulties in the identification of microorganisms 4. Pixelated image/lack of image clarity
C	Lack of clinical information
D	Absent of special stains/immunohistochemistry
E	Poor quality of the biopsy or inadequate routine laboratory processes 1. Badly positioned sections, chatter artifact, tissue folds and bubbles formed during coverslipping 2. Small size of the material

Discussion

Validation studies have been improved over time and the recommendations of CAP-PLQC guidelines are particularly important in this aspect, since the standardization of study designs provides validations with homogeneous methodology [1]. The main purpose of systematic reviews is to minimize the chance of type I (systematic) error, by eliminating studies with high risk of bias. Therefore, exclusion of studies with highly discrepant methodologies allowed the comparison of only well-designed studies and the reaching of solid, reliable conclusions. The way the sample is analyzed should encompass the index test and the reference standard with timing between analyses of paired samples (glass slide and correspondent digital slides). The analyses must be blinded, and the sample flow should encompass the analysis of all glass slides by CLM and, after the washout, the analysis of all correspondent digital slides.

Table 5 QUADAS-2

No.	Author	RISK OF BIAS				APPLICABILITY CONCERNS		
		Sample selection	Index test	Reference standard	Flow and timing	Sample selection	Index test	Reference standard
1	Al-Janabi et al (2012)a	?	☺	☺	☺	☺	☺	☺
2	Al-Janabi et al (2012)b	?	☺	☺	☺	☺	☺	☺
3	Al-Janabi et al (2012)c	☺	☺	☺	☺	☺	☺	☺
4	Al-Janabi et al (2013)	☺	☺	☺	☺	☺	☺	☺
5	Al-Janabi et al (2014)	☺	☺	☺	☺	☺	☺	☺
6	Arnold et al (2015)	☺	☺	☺	☺	☺	☺	☺
7	Kent et al (2017)	☺	☺	☺	☺	☺	☺	☺
8	Loughrey et al (2015)	☺	☺	☺	☺	☺	☺	☺
9	Nielsen et al (2010)	☺	☺	☺	☺	☺	☺	☺
10	Pekmezci et al (2016)	☺	😐	☺	☺	☺	☺	☺
11	Saco et al (2017)	☺	☺	☺	☺	☺	☺	☺
12	Tabata et al (2017)	☺	☺	☺	☺	☺	☺	☺
13	Thrall et al (2015)	☺	😐	☺	☺	☺	☺	☺

Studies with a known malignant diagnosis, which may lead to a false high performance, and studies that compared WSI diagnosis with original or consensus diagnosis were excluded. These issues represent the most common problems in validation studies [60] and generate selection bias [4]. The use of the index test alone and the comparison with a consensus panel refers to a concept of accuracy, which is not a recommended design for this particular purpose. Three articles included in this systematic review mentioned a consensus diagnosis in two different, yet justifiable situations: to include in the sample only cases appropriate for the intended purpose [19] and to reach a preferred diagnosis in discordant cases [20, 23]. The importance of reaching a preferred diagnosis lies in the possibility of identifying the pitfalls and missing details of the pathology, which are determinants in some cases [1].

Among included studies, one (7.7%) [22] proposed to determine the accuracy of WSI interpretation but presented intraobserver agreement instead. The accuracy is defined as concordance between the result of the method tested and the diagnosis established by a consensus or gold standard, while the intraobserver agreement is basically the percentage of concordance between diagnoses reached by an observer when assessing two diagnostic modalities [1]. The outcome of this study was not aligned with the aim but was found to provide appropriate data, which allowed the correct interpretation of the results. Another study [12] proposed to evaluate if the diagnosis can be replaced by virtual microscopy and, for this purpose, the accuracy, sensitivity, specificity, and positive/negative predictive values were measured. The accuracy, in this context, was defined as the addition of the percentage level of concordance and minor discordance, which is not the best concept definition. The diagnostic performance was intended to be calculated by means of sensitivity and specificity. However, sensitivity and specificity are used to calculate the reliability of the method and indicate the consistency of the results as the test is repeated, not the performance of the test. Fortunately, this study also provided the percentage of concordance (intraobserver agreement) between WSI and CLM diagnosis. It is very important to correctly delineate the study design

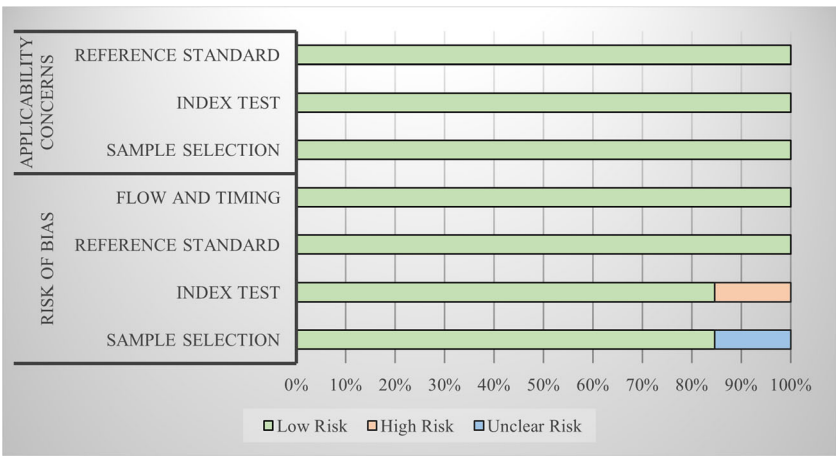


Fig. 4 Graphic presentation for QUADAS-2 results for included studies

according to the aim. These sources of inconsistency generate divergent measures and provide conflicting and unreliable data.

Validated pathology areas included dermatopathology, neuropathology, gastrointestinal, genitourinary, breast, liver, and pediatric pathology. Surgical pathology specimens of pediatric pathology were gastrointestinal, heart, liver, lung, neuropathology, placenta, rectal suction, skin, and tonsil. Subsets also included endocrine, head and neck, hematopoietic organ, hepatobiliary-pancreatic organ, soft tissue, bone, hematopathology, medical kidney, and transplant biopsies. However, Saco et al. considered, in 2016, that the areas of hematopathology, endocrine pathology, soft tissue, and bone had not been fully studied [61]. Tabata et al., in 2017 [23], included soft tissue specimens and bone pathology in the sample, but it is not possible to know how representative these specimens were, and a more targeted and specific validation is recommended. Saco et al. had also pointed out the need for validations in the head and neck area because there was only one study in this subspecialty. Fortuitously, our research group recently published a validation in oral pathology [62], adding original evidence of high performance of WSI in this unexplored area. This study was not added to this systematic review because it was published after the search.

The washout time is highly variable in the literature, and there is no consensus of what period is most appropriated to avoid recall bias; either an inferior or an overextended washout may produce bias due to the sample flow. A small period of washout may cause memorization bias in the test, and a long washout may allow diagnostic criteria to change over time [12]. Surprisingly, this systematic review found that the study with the lowest intraobserver agreement has been conducted with one of the shortest washout periods: 3 weeks [24]. This study also stated that intraobserver variations do not derive from the technical limitations of WSI.

The inclusion of trained pathologists encompasses one of the recommendations of CAP-PLQC and appears to provide better concordance rates and minor diagnosis time [1]. One included study reported the lack of experience of the pathologists as a reason for disagreement [12]. However, the study methodology reported the inclusion of four trained pathologists. In addition to increasing the intraobserver disagreement, an inexperienced pathologist also increases the time needed for diagnosis. Most pathologists are convinced that the digital method is more time-consuming. However, the learning curve [39, 43] and the utilization of suboptimal tools for navigation [16, 17] are likely explanations for this increasing time and may also be related to the lack of confidence of the pathologist in the WSI manipulation [63].

Although no formal assessment of timing has been conducted, two included studies [16, 17] stated that the utilization of suboptimal navigation tools, such as a computer mouse, is not adequate to explore the glass slide and may increase digital diagnosis time.

The scan time may also be influenced by the file size, which is dependent on the magnification of scanning [64] and represents an extra step in the diagnostic process. This is one of the chief barriers to digital pathology acceptance, even greater than the time required to render a diagnosis [8]. However, scan time is also highly variable and depends on the type of scanner used and its throughput capacity. It is therefore very difficult to include scan time as a part of the validation studies since it does not provide a reproducible parameter. This may explain the absence of timing in most validation studies. These issues should be considered an integral part of the digital methodology and not a disadvantage.

A higher intraobserver agreement is related to the high quality of digital slides and a better workflow provided by WSI systems [65], which appears to be easier to navigate compared to handling glass slides [66]. Some studies stated that digital microscopy provides the best definition of histologic images and confers the best method for the identification of microscopic structures [67]. Intraobserver agreement values of the included studies support the high performance of the digital method, and even the study with a lower intraobserver agreement [24] dismissed the technical limitations of WSI as a reason for the occurrence of discordances. However, it is important to be able to recognize when an overestimation of the test's performance occurs. Validation studies have incorporation bias since index tests and reference standards are not independent. In addition, intraobserver variability also increases when comparing the same glass slide over time. Interobserver variability can also be increased in difficult cases. This fact supports the cross-analysis of intraobserver and interobserver variability [24]. However, CAP-PLQC advocated that it is important for validation purposes to have one pathologist reproducing the same diagnosis with both modalities—i.e., intraobserver agreement—and the main objective is to accomplish a higher concordance rate [1]. The interobserver agreement should not be used to evaluate the performance of the test because this introduces bias due to the individual diagnostic interpretations of each pathologist [68].

The secondary objective of this review was to identify the reported reasons for disagreements and to determine the cause of these occurrences, which is also stated by CAP-PLQC as an important outcome [1]. In this systematic review, the most commonly reported reason for diagnostic discordance were borderline cases. The difficulty caused by borderline cases is inherent in the diagnostic process and can occur in CLM as well [20]. Sometimes, there is a need for higher magnifications to visualize subtle details, which could be present in difficult cases [64]. The subjectivity of some diagnoses, such as the interpretation of dysplasia [19], often indicates a greater complexity and also correlates directly to the individual interpretation and experience of the pathologists. This systematic review identified a higher frequency of borderline and challenging cases in dermatopathology validation studies.

Seven authors reported the limitations of the equipment and/or the limited image resolution as pitfalls. Among these, one study [18] indicated pitfalls regarding the identification of eosinophilic granular bodies, eosinophils, and nucleated red blood cells in a neuropathology specimen of a pediatric validation study, but the

authors did not consider this fact as a failure of the WSI method. Pekmezci et al. reported difficulties in the identification of mitotic figures, nuclear details, and chromatin patterns in a neuropathology validation study [21]. Also, difficulties in the identification of microorganisms were reported in three studies [14, 16, 24], but the need for higher magnifications appeared to be of little relevance in these studies. Thrall et al. stated that the lack of image clarity was a limitation of the technology but dismissed this fact as a reason for the intraobserver variances [24]. The impairment in recognizing eosinophilic granular bodies, eosinophils, mitotic figures or nuclear details and chromatin pattern, as well as some microorganisms—such as *Candida albicans*, *Helicobacter pylori*, and *Giardia lamblia*—points to a limitation of the scanner and occur more frequently in some subspecialties of pathology (neuropathology, gastrointestinal pathology, and pediatric pathology within a neuropathology specimen). These pitfalls highlight the need for more advanced scanners, which should certainly be improved with the advent of technological improvement. Therein lies the need for regulation of these devices, which should be standardized and improved. It is important to emphasize that difficulties in the identification of microorganisms were pointed to as reasons for disagreements, but higher magnifications were not considered to be very relevant by the authors [14, 16].

The lack of clinical information supplied with cases in both analyses represents an absence of reproducibility [1], increases the difficulty in the diagnostic process, and may lead to wrong diagnoses. Four included studies [12, 17, 19, 21] did not provide clinical data for the analyses. Nielsen reported that this absence could make it more difficult to render the diagnoses, which may add an element of error [12], while Al-Janabi et al. indicated that the provision of clinical data may decrease these errors [17]. According to Kent et al., the lack of clinical information leads to disagreement in a sample of inflammatory lesions. However, this author also reported a high level of concordance with inflammatory lesions that had a mixed infiltrate with eosinophils [19]. One included study [23] did not mention if clinical data was provided and did not correlate the absence of this information with the occurrence of discordant diagnoses. Fortunately, the majority of validation studies recognized the need to correlate the histopathological and the clinical information to provide a correct diagnosis, either through glass or digital slides.

The selection and inclusion of the cases should, ideally, be consecutive or random. However, this selection strategy may not provide a sample with the most relevant diagnosis and a broad range of tissue sources. A stratified uniform sampling is more appropriate to select the cases; it gives smaller error estimation and may be useful to do measurements and estimates using cases grouped into strata [69]. Unfortunately, none of the included studies followed this methodology. Additionally, two studies included in this systematic review [13, 14] did not clarify how the samples were retrieved. An inappropriate exclusion of cases may result in overly optimistic estimates of diagnostic accuracy [4]. One included study reported exclusions [21], but it was acceptable and coherent with the proposal of the study. The prespecification of the test threshold is important so there is no bias in interpreting the results; this could otherwise lead to an overoptimistic estimate of the test performance [70]. Two included studies [21, 24] did not mention the threshold previously, but one [24] mentioned a deliberately low threshold setting to maximize the identification of discordances.

In general, this systematic review showed a high concordance between diagnoses achieved by using WSI and CLM. The included studies were optimally designed to validate WSI for general clinical use, providing evidence with confidence and—most importantly—it is possible to confirm that this technology can be used to render primary diagnoses in several subspecialties of human pathology. The reported difficulties related to specific findings of certain areas of pathology reinforce the need for validation studies in some areas not fully studied, such as hematopathology, endocrine, and bone and soft-tissue pathology.

Authors' contributions All authors had substantial contributions to the conception, draft and design of this work, (Anna Luíza Damaceno Araújo, Natália Rangel Palmier, Cristhian Camilo Troconis, Paul M. Speight, Oslei Paes de Almeida, Marcio Ajudarte Lopes and Alan Roger Santos-Silva), as well as participation of the acquisition (Lady Paola Aristizábal Arboleda, Natália Rangel Palmier, Jéssica Montenegro Fonsêca, Mariana de Pauli Paglioni, Ana Carolina Prado Ribeiro, Pablo Agustin Vargas, Luciana Estevam Simonato and Wagner Gomes-Silva), analysis (Anna Luíza Damaceno Araújo, Felipe Paiva Fonseca and Lady Paola Aristizábal Arboleda), and interpretation (Anna Luíza Damaceno Araújo, Cristhian Camilo Troconis, Thaís Bianca Brandão and Alan Roger Santos-Silva) of data for the work. The final version of this work was reviewed and approved for publication by all parts included. Authors Anna Luíza Damaceno Araújo and Alan Roger Santos-Silva takes full responsibility for the work as a whole, including the study design, access to data and the decision to submit and publish the manuscript.

Funding information Financial support was received from the Coordination for the Improvement of Higher Education Personnel (CAPES/PROEX, Brazil), the National Council for Scientific and Technological Development (CNPq, Brazil) and the grants from São Paulo Research Foundation (FAPESP, Brazil) process number: 2009/53839-2.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical responsibilities of author section All authors had substantial contributions to the conception, draft and design of this work, as well as participation of the acquisition, analysis and interpretation of data for the work. The final version of this work was approved for publication by all parts included. If there is a need, all authors agreed to be accountable for any aspects of the work and we ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The authors also state that the material is original, has not been published elsewhere, and is being submitted only to the Virchows Archiv.

Appendix 1

DATA COLLECTION FORM

Review title or ID

Study ID *(surname of first author and year first full report of study was published e.g. Smith 2001)*

Report IDs of other reports of this study *(e.g. duplicate publications, follow-up studies)*

1 GENERAL INFORMATION

Date form completed	
Name of person extracting data	
Report title	
Reference details *(author, year)*	
DOI/PUI/another ID of the publication	
Publication type *(full report, abstract, letter)*	

2 ELIGIBILITY

Study Characteristics	**Review Inclusion Criteria**	**Yes/ No / Unclear**	**Values**	**Location in text**
Type of study	Cross sectional		-	
Sample set	At least 60 cases			
Types of intervention	Reference standards - conventional light microscopy (CLM) Index test– whole slide imaging system (WSI)		-	
Washout time	> 2 weeks			
Types of outcome measures	Was the intraobserver agreement the preferred measurement? (each observer assessing all sample by both methods – digital and conventional – with an appropriated wash out period between the analyses).			
	The percentage of concordance was reported?			
	Kappa index was reported?			
Decision:				
Reason for exclusion	Articles published in foreign languages but English. Insufficient sample set number. Sample with a known malignant diagnosis Studies with lack of information (mainly about how the sample was analyzed); Studies which the primary goal was not to examine diagnostic concordance between WSI and CLM; The intraobserver agreement of the methods is the preferred measure to assess the performance of digital microscopy. Intraobserver concordance percentage or kappa value is not mentioned.			
Notes:				

DO NOT PROCEED IF STUDY EXCLUDED FROM REVIEW

3 INTERVENTIONS, PARTICIPANTS AND SAMPLE

	Description	**Location in text**
WSI system utilized and magnification of scanner:		
Computer settings/ monitor resolution:		
Pathologist number:		
Sample set quantity (n):		

4 METHODS

	Descriptions as stated in report/paper	**Location in text**
Type of study:		
The study was based on some stated Guideline?		
Aim of study:		
Pathologists were previous trained?		
How sample was analyzed?		
Was there any information available along with the cases?		
Scan time or diagnosis time were measured?		
Washout time:		
Notes:		

5 RISK OF BIAS ASSESSMENT

Domain			Location in text
1. Sample selection **Describe methods of sample selection:**			
Was the sample selection consecutive or random? *Yes/No/Unclear*		*Risk of bias* *High (if at least one was reached as 'no' or 'unclear')* *Low (all reached as "yes"*	
A known malignant sample was avoided? *Yes/No/Unclear*			
Did the study avoid inappropriate exclusions? *Yes/No/Unclear*			
Could the selection of patients have introduced bias? *RISK: LOW/HIGH/UNCLEAR*			
Notes:			
2. Index test **Describe the index test and how is conducted and interpreted:**			
Were the index test results interpreted without knowledge of the results of the reference standard? *Yes/No/Unclear*		*Risk of bias* *High (if at least one was reached as 'no' or 'unclear')* *Low (all reached as "yes"*	
If a threshold (classification of the agreement) was used, was it pre-specified? *Yes/No/Unclear*			
Could the conduct or interpretation of the index test have introduced bias? *RISK: LOW/HIGH/UNCLEAR*			
Notes:			
3. Reference Standard **Describe the reference standard and how it was conducted and interpreted:**			
Is the reference standard likely to correctly classify the target conditions (diagnosis)? *Yes/No/Unclear*		*Risk of bias* *High (if at least one was reached as 'no' or 'unclear')* *Low (all reached as "yes"*	
Was the reference standard results interpreted without knowledge of the results of the index test? Yes/No/Unclear			
Could the reference standard, its conduct, or its interpretation have introduced bias? *RISK: LOW/HIGH/UNCLEAR*			
Notes:			
4. Flow and timing **Describe the time interval and any interventions between index test(s) and reference standard:**			
Could the patient flow have introduced bias? *RISK: LOW/HIGH/UNCLEAR*			
Notes:			

6 APPLICABILITY

Domain		Location in text
1. Sample selection **Describe included cases (specimen type, subspecialty, biopsy location)?**		
Is there a concern that the included cases do not match the review question? *CONCERN: LOW/HIGH/UNCLEAR*		
Notes:		
2. Index test **Is there concern that the index tests, its conduct, or interpretation differs from the review question?** *CONCERN: LOW/HIGH/UNCLEAR*		
Notes:		
3. Reference Standard **Is there concern that the reference standard, its conduct, or interpretation does not match with the review question?** *CONCERN: LOW/HIGH/UNCLEAR*		
Notes:		

7 OUTCOMES

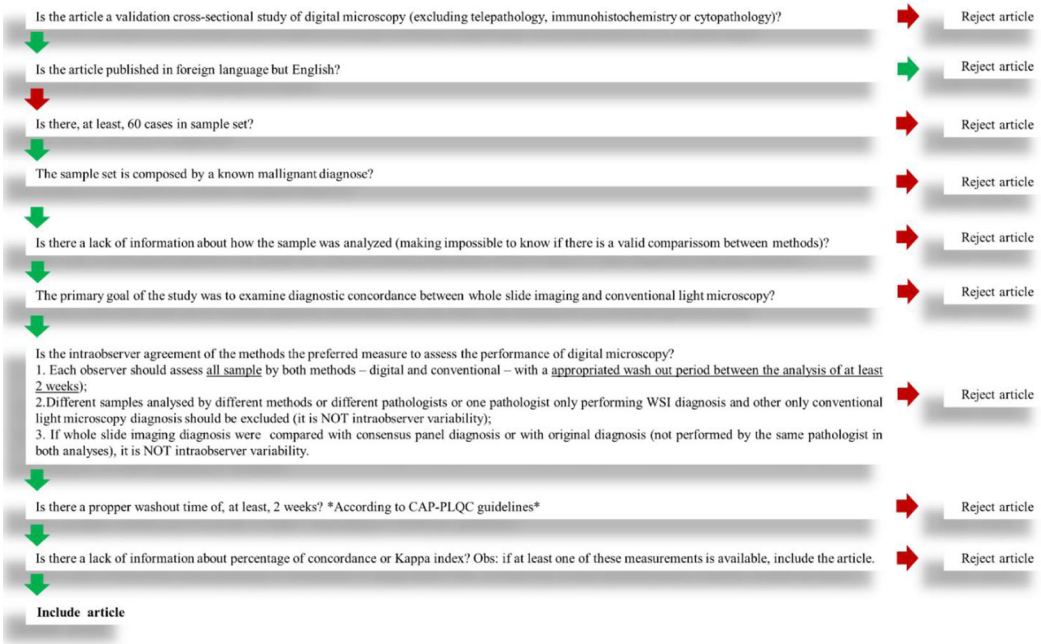
	Description as stated in report/paper	Location in text
Intraobserver agreement Concordance (%)		
Discordance (%)		
Kappa index		
Main reason for disagreement:		
In disagreement, what diagnosis was preferred?		
Discordant cases were reviewed?		
Conclusion of the study:		
Notes:		

Appendix 2

TAILORED QUADAS-2

Phase 1. State the review question: The review question we intend to elucidate is: "Is digital microscopy performance reliable for use in clinical practice and routine surgical pathology for diagnostic purposes as conventional microscopy?". For this purpose, we evaluate previous studies, which compared digital microscopy (index test) with conventional microscopy (reference standard), in several pathology areas (target conditions) for diagnostic purposes (intended use). To assess the performance of whole-slide imaging systems, we focus on intra-observer agreement (preferred measurement stated by CAP-PLQC guidelines). The sample set should include at least 60 cases and the pathologists involved on validation studies should perform an evaluation of all cases by two methods (conventional and digital) with a wash out period superior of 2 weeks. All these parameters obey the CAP-PLQC guidelines. Because the performance of the index test may depend on where it will be used in the diagnostic pathway, we should reinforce the need of a blind and independent analyses by both methods. Pathologists must assess either glass slides or correspondent whole slide images with a proper washout period (> 2 weeks). If WSI diagnosis were compared with original diagnosis (by glass slide), it is important that the digital slide be assessed by the same pathologist who made the original report (ensuring that the measure is intra-observer agreement).

Phase 2. Draw a flow diagram for the primary study:



Phase 3. Risk of bias and applicability judgments

Instructions:
Risk of bias (could be answered as: yes, no or unclear) - If all signaling questions for a domain are answered "yes" then the risk of bias can be judged "low." If any signaling issue is answered "no," this signals the potential for bias. The "unclear" category should only be used when insufficient data is reported to allow for judgment.

Applicability (could be answered as: low, high or unclear) - The applicability sections are structured similarly to the polarization sections, but do not include signaling issues. The review authors should record the information on which the applicability judgment is made and then assess their concerns that the study does not match the review question. The "unclear" category should be used only when insufficient data are reported to allow judgment.

DOMAIN 1: PATIENT SELECTION
A. Risk of Bias
Describe methods of sample selection:

Was a consecutive or random sample enrolled? Yes/No/Unclear
A study should, ideally, include selected samples consecutively or randomly - otherwise, it has the potential to bias. If the sample includes both (consecutively/randomly and non-consecutively/non-randomly), the risk of bias may be considered "low" if the percentage of non-consecutively/non-randomly cases was less than 10% of the total number of cases. If the selection of the samples was not clear, this signaling question must be rated as "unclear"

A known malignant sample was avoided? Yes/No/Unclear
A known malignant sample may lead to a super estimation of diagnostic accuracy (Cornish et al, 2012).

Did the study avoid inappropriate exclusions? Yes/No/Unclear
Inappropriate exclusion may result in over-optimistic estimates of diagnostic accuracy. If the study excluded > 10% the sample with or without specific motives, exclusions must be considered inadequate. This limit was determined pragmatically.

Could the selection of patients have introduced bias? RISK: LOW/HIGH/UNCLEAR

B. Concerns regarding applicability
Describe included cases (specimen type, subspecialty, biopsy location)?

Is there a concern that the included cases do not match the review question?
CONCERN: LOW/HIGH/UNCLEAR

DOMAIN 2: INDEX TEST(S)
A. Risk of Bias
Describe the index test and how it was conducted and interpreted:

Were the index test results interpreted without knowledge of the results of the reference standard? Yes/No/Unclear
Interpretation of the results of the index tests can be influenced by the knowledge of the standard reference results (Whiting et al, 2004). The bias potential is related to the subjectivity of the test and the order of the test. Studies needs do clearly report blindness to answer this question with 'yes'.

If a threshold (classification of the agreement) was used, was it pre-specified? Yes/No/Unclear
For this question to be answered with 'yes', the study needs to mention which type of threshold was used and clearly indicate that it was specified prior to the start of the study. Selecting the test threshold to optimize sensitivity and/or specificity may lead to over-optimistic estimates of test performance, which is likely to be poorer in an independent sample of patients in whom the same threshold is used (Leeflang et al, 2008).
Could the conduct or interpretation of the index test have introduced bias? RISK: LOW/HIGH/UNCLEAR

B. Concerns regarding applicability
Is there a concern that the index tests, its conduct, or interpretation differs from the review question?
Variations in test technology, execution, or interpretation may affect estimates of its diagnostic accuracy. If index tests methods vary from those specified in the review question there may be concerns regarding applicability.
CONCERN: LOW/HIGH/UNCLEAR

DOMAIN 3: REFERENCE STANDARD
A. Risk of Bias
Describe the reference standard and how it was conducted and interpreted:

Is the reference standard likely to correctly classify the target condition (diagnosis)? Yes/No/Unclear
Estimates of test accuracy are based on the assumption that the reference standard is 100% sensitive and specific disagreements between the reference standard and index test are assumed to result from incorrect classification by the index test (Biesheuvel, Irwig and Bossuyt, 2007; van Rijkom and Verdonschot, 1995).

Where the reference standard results interpreted without knowledge of the results of the index test? Yes/No/Unclear

Could the reference standard, its conduct, or its interpretation have introduced bias?
Potential for bias is related to the potential influence of prior knowledge on the interpretation of the reference standard (Whiting et al, 2004).
RISK: LOW/HIGH/UNCLEAR

B. Concerns regarding applicability
Is there concern that the reference standard, its conduct, or interpretation does not match with the review question?
CONCERN: LOW/HIGH/UNCLEAR

DOMAIN 4: FLOW AND TIMING
A. Risk of Bias
Describe the time interval and any interventions between index test(s) and reference standard:

Could the sample flow have introduced bias? RISK: LOW /HIGH/UNCLEAR

References
1. *Biesheuvel C, Irwig L, Bossuyt P. Observed differences in diagnostic test accuracy between patient subgroups: is it real or due to reference standard misclassification? Clin Chem 2007; 53(10):1725-1729.*
2. Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clinical Chemistry 2008; 54(4):729-737.
3. *van Rijkom HM, Verdonschot EH. Factors involved in validity measurements of diagnostic tests for approximal caries--a meta-analysis. Caries Research 1995; 29(5):364-70.*
4. *Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004; 140(3):189-202.*

References

1. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, Beckwith BA, Evans AJ, Lal A, Parwani AV, College of American Pathologists Pathology and Laboratory Quality Center (2013) 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 137:1710–1722. https://doi.org/10.5858/arpa.2013-0093-CP
2. Pantanowitz L, Evans A, Pfeifer J, Collins LC, Valenstein PN, Kaplan KJ, Wilbur DC, Colgan TJ (2011) Review of the current state of whole slide imaging in pathology. J Pathol Inform 2:36. https://doi.org/10.4103/2153-3539.83746
3. Koch LH, Lampros JN, Delong LK, Chen SC, Woosley JT, Hood AF (2009) Randomized comparison of virtual microscopy and traditional glass microscopy in diagnostic accuracy among dermatology and pathology residents. Hum Pathol 40:662–667. https://doi.org/10.1016/j.humpath.2008.10.009
4. Whiting P, Harbord R, Kleijnen J (2005) No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 5:19. https://doi.org/10.1186/1471-2288-5-19
5. Cornish TC, Swapp RE, Kaplan KJ (2012) Whole-slide imaging: routine pathologic diagnosis. Adv Anat Pathol 19:152–159. https://doi.org/10.1097/PAP.0b013e318253459e
6. Food and Drug Administration (2017) FDA allows marketing of first whole slide imaging system for digital pathology. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm552742.htm. Accessed 16 Mar 2017
7. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 6:e1000097. https://doi.org/10.1371/journal.pmed.1000097
8. Williams BJ, Hanby A, Millican-Slater R, Nijhawan A, Verghese E, Treanor D (2018) Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study on digital pathology validation and training. Histopathology 72:662–671. https://doi.org/10.1111/his.13403
9. Goacher E, Randell R, Williams B, Treanor D (2017) The diagnostic concordance of whole slide imaging and light microscopy: a systematic review. Arch Pathol Lab Med 141:151–161. https://doi.org/10.5858/arpa.2016-0025-RA
10. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A (2016) Rayyan—a web and mobile app for systematic reviews. Syst Rev 5:210. https://doi.org/10.1186/s13643-016-0384-4
11. Cochrane Effective Practice and Organisation of Care (EPOC) (2017). EPOC Resources for review authors. https://epoc.cochrane.org/resources/epoc-resources-review-authors
12. Nielsen PS, Lindebjerg J, Rasmussen J, Starklint H, Waldstrøm M, Nielsen B (2010) Virtual microscopy: an evaluation of its validity and diagnostic performance in routine histologic diagnosis of skin tumors. Hum Pathol 41:1770–1776. https://doi.org/10.1016/j.humpath.2010.05.015
13. Al-Janabi S, Huisman A, Vink A et al (2012) Whole slide images for primary diagnostics in dermatopathology: a feasibility study. J Clin Pathol 65:152–158. https://doi.org/10.1136/jclinpath-2011-200277
14. Al-Janabi S, Huisman A, Vink A et al (2012) Whole slide images for primary diagnostics of gastrointestinal tract pathology: a feasibility study. Hum Pathol 43:702–707. https://doi.org/10.1016/j.humpath.2011.06.017
15. Al-Janabi S, Huisman A, Willems SM, Van Diest PJ (2012) Digital slide images for primary diagnostics in breast pathology: a feasibility study. Hum Pathol 43:2318–2325. https://doi.org/10.1016/j.humpath.2012.03.027
16. Al-Janabi S, Huisman A, Nikkels PGJ et al (2013) Whole slide images for primary diagnostics of paediatric pathology specimens: a feasibility study. J Clin Pathol 66:218–223. https://doi.org/10.1136/jclinpath-2012-201104
17. Al-Janabi S, Huisman A, Jonges GN et al (2014) Whole slide images for primary diagnostics of urinary system pathology: a feasibility study. J Ren Inj Prev 3:91–96. https://doi.org/10.12861/jrip.2014.26
18. Arnold MA, Chenever E, Baker PB, Boué DR, Fung B, Hammond S, Hendrickson BW, Kahwash SB, Pierson CR, Prasad V, Nicol KK, Barr T (2015) The College of American Pathologists Guidelines for whole slide imaging validation are feasible for pediatric pathology: a pediatric pathology practice experience. Pediatr Dev Pathol 18:109–116. https://doi.org/10.2350/14-07-1523-OA.1
19. Kent MN, Olsen TG, Feeser TA, Tesno KC, Moad JC, Conroy MP, Kendrick MJ, Stephenson SR, Murchland MR, Khan AU, Peacock EA, Brumfiel A, Bottomley MA (2017) Diagnostic accuracy of virtual pathology vs traditional microscopy in a large dermatopathology study. JAMA Dermatol 153:1285–1291. https://doi.org/10.1001/jamadermatol.2017.3284
20. Loughrey MB, Kelly PJ, Houghton OP, Coleman HG, Houghton JP, Carson A, Salto-Tellez M, Hamilton PW (2015) Digital slide viewing for primary reporting in gastrointestinal pathology: a validation study. Virchows Arch 467:137–144. https://doi.org/10.1007/s00428-015-1780-1
21. Pekmezci M, Uysal SP, Orhan Y et al (2016) Pitfalls in the use of whole slide imaging for the diagnosis of central nervous system tumors: a pilot study in surgical neuropathology. J Pathol Inform 7:25. https://doi.org/10.4103/2153-3539.181769
22. Saco A, Diaz A, Hernandez M, Martinez D, Montironi C, Castillo P, Rakislova N, del Pino M, Martinez A, Ordi J (2017) Validation of whole-slide imaging in the primary diagnosis of liver biopsies in a university hospital. Dig Liver Dis 49:1240–1246. https://doi.org/10.1016/j.dld.2017.07.002
23. Tabata K, Mori I, Sasaki T, Itoh T, Shiraishi T, Yoshimi N, Maeda I, Harada O, Taniyama K, Taniyama D, Watanabe M, Mikami Y, Sato S, Kashima Y, Fujimura S, Fukuoka J (2017) Whole-slide imaging at primary pathological diagnosis: validation of whole-slide imaging-based primary pathological diagnosis at twelve Japanese academic institutes. Pathol Int 67:547–554. https://doi.org/10.1111/pin.12590
24. Thrall MJ, Wimmer JL, Schwartz MR (2015) Validation of multiple whole slide imaging scanners based on the guideline from the College of American Pathologists pathology and laboratory quality center. Arch Pathol Lab Med 139:656–664. https://doi.org/10.5858/arpa.2014-0073-OA
25. Camparo P, Ramirez A, Claude V et al (2009) Whole slide imaging in daily routine examination in a pathologic department: Experience of a military hospital network in Paris. Rev Fr Lab 38:49–55 RFL-01-2008-38-408-1773-035x-101019-200812623
26. Wang M, Liu S, Xie C et al (2015) Making primary diagnosis on liver allograft biopsies with whole slide images - a validation study. Am J Clin Pathol 144:A168. https://doi.org/10.1093/ajcp/144.suppl2.168
27. Gage JC, Joste N, Ronnett BM, Stoler M, Hunt WC, Schiffman M, Wheeler CM (2013) A comparison of cervical histopathology variability using whole slide digitized images versus glass slides: experience with a statewide registry. Hum Pathol 44:2542–2548. https://doi.org/10.1016/j.humpath.2013.06.015
28. Zeitouni J, Jorda M, Reyes C, Nadji M (2012) Validation of whole slide imaging for the first line diagnosis of prostate biopsies. Lab Invest 92:519A–520A. https://doi.org/10.1038/labinvest.2012.24

29. Gerhard R, Honorio A, Gentili A et al (2014) Primary histopathological diagnosis using whole slide imaging (WSI): a validation study. Lab Invest 94:399A. https://doi.org/10.1038/labinvest.2014.28
30. Goodman S, Kandil D, Khan (2014) A Diagnosis of breast needle core biopsies using whole slide imaging. Lab Invest 94:399A. https://doi.org/10.1038/labinvest.2014.28
31. Parimi V, Borys A, Zhou Y et al (2016) Validation of whole frozen section slide image diagnosis in surgical pathology. Lab Invest 96: 399A–400A. https://doi.org/10.1038/labinvest.2016.15
32. Bradshaw S, Driman D, Dupre M et al (2013) Inter- and intra-observer agreement in diagnosing dysplasia in Barrett's esophagus: comparison of routine glass slide vs. digital image examination. Lab Investig 93:471–489. https://doi.org/10.1038/labinvest.2013.36
33. Sturm B, Fleskens S, Bot F et al (2013) Larynx virtual microscopy validation study. Virchows Arch 463:109–352. https://doi.org/10.1007/s00428-013-1444-y
34. Sturm B, Mooi W, Creytens D et al (2017) Validation of diagnosing melanocytic lesions on whole slide images- does z-stack scanning improve diagnostic accuracy? Virchows Arch 471:S15. https://doi.org/10.1007/s00428-013-1444-y
35. Maleeff BE (2014) Validation of a digital pathology whole slide imaging system. Microsc Microanal 20:1410–1411. https://doi.org/10.1017/S1431927614008782
36. Eccher A, Calio A, Colombari R et al (2015) Validation of digital whole slide imaging according to the College of American Pathologists Guidelines in the evaluation of pre-implant kidney biopsies. Lab Investig 95:499A. https://doi.org/10.1038/labinvest.2015.25
37. Hoffmann J, McGinnis L, Mafnas CT et al (2016) Validation of digital whole slide imaging system for intraoperative breast sentinel lymph node touch prep analysis: a single institution experience. Lab Investig 96:391–402. https://doi.org/10.1177/20101058110200S101
38. Wilson I, Treanor D, Williams B (2017) Belfast Pathology 2017. 10th joint meeting of the British division of the international academy of pathology and the pathological Society of Great Britain & Ireland, 20-23 June 2017. J Pathol 243:S1–S41. https://doi.org/10.1002/path.4984
39. Randell R, Ruddle RA, Mello-Thoms C, Thomas RG, Quirke P, Treanor D (2013) Virtual reality microscope versus conventional microscope regarding time to diagnosis: An experimental study. Histopathology 62:351–358. https://doi.org/10.1111/j.1365-2559.2012.04323.x
40. Lee JJ, Jedrych J, Pantanowitz L (2017) Validation of digital pathology for primary histopathological diagnosis of routine, inflammatory dermatopathology cases 0:1–7. https://doi.org/10.1097/DAD.0000000000000888
41. Rodriguez-Urrego PA, Cronin AM, Al-Ahmadie HA et al (2011) Interobserver and intraobserver reproducibility in digital and routine microscopic assessment of prostate needle biopsies. Hum Pathol 42:68–74. https://doi.org/10.1016/j.humpath.2010.07.001
42. Jara-Lazaro AR, Tan PH (2012) Comparing digital and optical microscopy diagnoses of breast and prostate core biopsies. Pathology 44:46–48. https://doi.org/10.1097/PAT.0b013e32834e4254
43. Krishnamurthy S, Mathews K, McClure S, Murray M, Gilcrease M, Albarracin C, Spinosa J, Chang B, Ho J, Holt J, Cohen A, Giri D, Garg K, Bassett RL Jr, Liang K (2013) Multi-institutional comparison of whole slide digital imaging and optical microscopy for interpretation of hematoxylin-eosin-stained breast tissue sections. Arch Pathol Lab Med 137:1733–1739. https://doi.org/10.5858/arpa.2012-0437-OA
44. Campbell W, Lele S, West W et al (2012) Diagnoses rendered by whole slide imaging (WSI) alone are accurate for use in a general surgical pathology practice. Lab Investig 92:494–509. https://doi.org/10.1038/labinvest.2012.23
45. Campbell WS, Lele SM, West WW, Lazenby AJ, Smith LM, Hinrichs SH (2012) Concordance between whole-slide imaging and light microscopy for routine surgical pathology. Hum Pathol 43:1739–1744. https://doi.org/10.1016/j.humpath.2011.12.023
46. Campbell WS, Hinrichs SH, Lele SM, Baker JJ, Lazenby AJ, Talmon GA, Smith LM, West WW (2014) Whole slide imaging diagnostic concordance with light microscopy for breast needle biopsies. Hum Pathol 45:1713–1721. https://doi.org/10.1016/j.humpath.2014.04.007
47. Brunelli M, Beccari S, Colombari R et al (2014) 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 9(Suppl 1):S12. https://doi.org/10.1186/1746-1596-9-S1-S12
48. Ordi J, Castillo P, Saco A, del Pino M, Ordi O, Rodríguez-Carunchio L, Ramírez J (2015) Validation of whole slide imaging in the primary diagnosis of gynaecological pathology in a university hospital. J Clin Pathol 68:33–39. https://doi.org/10.1136/jclinpath-2014-202524
49. Snead DRJ, Tsang YW, Meskiri A, Kimani PK, Crossman R, Rajpoot NM, Blessing E, Chen K, Gopalakrishnan K, Matthews P, Momtahan N, Read-Jones S, Sah S, Simmons E, Sinha B, Suortamo S, Yeo Y, el Daly H, Cree IA (2016) Validation of digital pathology imaging for primary histopathological diagnosis. Histopathology 68:1063–1072. https://doi.org/10.1111/his.12879
50. Fónyad L, Krenács T, Nagy P, Zalatnai A, Csomor J, Sápi Z, Pápay J, Schönléber J, Diczházi C, Molnár B (2012) Validation of diagnostic accuracy using digital slides in routine histopathology. Diagn Pathol 7:35. https://doi.org/10.1186/1746-1596-7-35
51. Shah KK, Lehman JS, Gibson LE, Lohse CM, Comfere NI, Wieland CN (2016) Validation of diagnostic accuracy with whole-slide imaging compared with glass slide review in dermatopathology. J Am Acad Dermatol 75:1229–1237. https://doi.org/10.1016/j.jaad.2016.08.024
52. Elmore J, Longton G, Pepe M et al (2017) A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis. J Pathol Inform 8:12. https://doi.org/10.4103/2153-3539.201920
53. Mills AM, Gradecki SE, Horton BJ et al (2018) Diagnostic Efficiency in Digital Pathology: A Comparison of Optical Versus Digital Assessment in 510 Surgical Pathology Cases. Am J Surg Pathol. 42(1):53–59. https://doi.org/10.1097/PAS.0000000000000930
54. Foad AFA (2017) Comparing the use of virtual and conventional light microscopy in practical sessions: virtual reality in Tabuk University. J Taibah Univ Med Sci 12:183–186. https://doi.org/10.1016/j.jtumed.2016.10.015
55. Bauer TW, Schoenfield L, Slaw RJ, Yerian L, Sun Z, Henricks WH (2013) Validation of whole slide imaging for primary diagnosis in surgical pathology. Arch Pathol Lab Med 137:518–524. https://doi.org/10.5858/arpa.2011-0678-OA
56. Buck T, Dilorio R, Havrilla L, O'Neill D (2014) Validation of a whole slide imaging system for primary diagnosis in surgical pathology: a community hospital experience. J Pathol Inform 5:43. https://doi.org/10.4103/2153-3539.145731
57. Bauer TW, Slaw RJ (2014) Validating whole-slide imaging for consultation diagnoses in surgical pathology. Arch Pathol Lab Med 138:1459–1465. https://doi.org/10.5858/arpa.2013-0541-OA
58. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, Cathro HP, Cheng L, Cooper K, Dickey GE, Gill RM, Heaton RP Jr, Kerstens R, Lindberg GM, Malhotra RK, Mandell JW, Manlucu ED, Mills AM, Mills SE, Moskaluk CA, Nelis M, Patil DT, Przybycin CG, Reynolds JP, Rubin BP, Saboorian MH, Salicru M, Samols MA, Sturgis CD, Turner KO, Wick MR, Yoon JY, Zhao P, Taylor CR (2017) Whole slide imaging versus microscopy for primary diagnosis in surgical pathology. Am J Surg Pathol 42:1. https://doi.org/10.1097/PAS.0000000000000948
59. Williams BJ, DaCosta P, Goacher E, Treanor D (2017) A systematic analysis of discordant diagnoses in digital pathology compared with

light microscopy. Arch Pathol Lab Med 141:1712–1718. https://doi.org/10.5858/arpa.2016-0494-OA
60. Cornish TC, Swapp RE, Kaplan KJ (2012) Whole-slide Imaging. Adv Anat Pathol 19:152–159. https://doi.org/10.1097/PAP.0b013e318253459e
61. Saco A, Ramírez J, Rakislova N, Mira A, Ordi J (2016) Validation of whole-slide imaging for Histolopathogical diagnosis: current state. Pathobiology 83:89–98. https://doi.org/10.1159/000442823
62. Araújo ALD, Amaral-Silva GK, Fonseca FP, Palmier NR, Lopes MA, Speight PM, de Almeida OP, Vargas PA, Santos-Silva AR (2018) Validation of digital microscopy in the histopathological diagnoses of oral diseases. Virchows Arch 473:321–327. https://doi.org/10.1007/s00428-018-2382-5
63. Sanders DSA, Grabsch H, Harrison R, Bateman A, Going J, Goldin R, Mapstone N, Novelli M, Walker MM, Jankowski J, on behalf of the AspECT trial management group and trial principal investigators (2012) Comparing virtual with conventional microscopy for the consensus diagnosis of Barrett's neoplasia in the AspECT Barrett's chemoprevention trial pathology audit. Histopathology 61:795–800. https://doi.org/10.1111/j.1365-2559.2012.04288.x
64. Romero Lauro G, Cable W, Lesniak A, Tseytlin E, McHugh J, Parwani A, Pantanowitz L (2013) Digital pathology consultations - a new era in digital imaging, challenges and practical applications. J Digit Imaging 26:668–677. https://doi.org/10.1007/s10278-013-9572-0
65. Boyce BF (2015) Whole slide imaging: uses and limitations for surgical pathology and teaching. Biotech Histochem 90:321–330. https://doi.org/10.3109/10520295.2015.1033463
66. Vodovnik A (2016) Diagnostic time in digital pathology: a comparative study on 400 cases. J Pathol Inform 7:4. https://doi.org/10.4103/2153-3539.175377
67. Fernandes C, Bonan R, Bonan P et al (2018) Dental Students' Perceptions and Performance in Use of Conventional and Virtual Microscopy in Oral Pathology. J Dent Educ 82:883–890. https://doi.org/10.21815/JDE.018.084
68. Fallon MA, Wilbur DC, Prasad M (2010) Ovarian frozen section diagnosis: use of whole-slide imaging shows excellent correlation between virtual slide and original interpretations in a large series of cases. Arch Pathol Lab Med 134:1020–1023. https://doi.org/10.1043/2009-0320-OA.1
69. Särndal C-E (2003) Stratified sampling. In: Model Assisted Survey Sampling. Springer, pp 100–109
70. Leeflang MMG, Moons KGM, Reitsma JB, Zwinderman AH (2008) Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin Chem 54:729–737. https://doi.org/10.1373/clinchem.2007.096032