Rapid Validation of Whole-Slide Imaging for Primary Histopathology Diagnosis

A Roadmap for the SARS-CoV-2 Pandemic Era

Megan I. Samuelson, MD,[1] *Stephanie J. Chen, MD,*[1] *Sarag A. Boukhar, MBChB,*[1] *Eric M. Schnieders,*[2] *Mackenzie L. Walhof,*[2] *Andrew M. Bellizzi, MD,*[1,*] *Robert A. Robinson, MD, PhD,*[1] *and Anand Rajan KD, MBBS*[1,*]

From the [1]Department of Pathology, University of Iowa Hospitals and Clinics, University of Iowa, Iowa City, IA, USA; and [2]Roy J. and Lucille A. Carver College of Medicine, University of Iowa, Iowa City, IA, USA.

Key Words: Whole-slide imaging; WSI validation; Digital pathology; Case management software; Primary digital diagnosis; SARS-CoV-2

Am J Clin Pathol 2021;XX:1–11

DOI: 10.1093/AJCP/AQAA280

ABSTRACT

Objectives: *The ongoing global severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic necessitates adaptations in the practice of surgical pathology at scale. Primary diagnosis by whole-slide imaging (WSI) is a key component that would aid departments in providing uninterrupted histopathology diagnosis and maintaining revenue streams from disruption. We sought to perform rapid validation of the use of WSI in primary diagnosis meeting recommendations of the College of American Pathologists guidelines.*

Methods: *Glass slides from clinically reported cases from 5 participating pathologists with a preset washout period were digitally scanned and reviewed in settings identical to typical reporting. Cases were classified as concordant or with minor or major disagreement with the original diagnosis. Randomized subsampling was performed, and mean concordance rates were calculated.*

Results: *In total, 171 cases were included and distributed equally among participants. For the group as a whole, the mean concordance rate in sampled cases (n = 90) was 83.6% counting all discrepancies and 94.6% counting only major disagreements. The mean pathologist concordance rate in sampled cases (n = 18) ranged from 90.49% to 97%.*

Conclusions: *We describe a novel double-blinded method for rapid validation of WSI for primary diagnosis. Our findings highlight the occurrence of a range of diagnostic reproducibility when deploying digital methods.*

Key Points

- We describe a novel method of whole-slide imaging (WSI) validation that incorporates subsampling for concordance measurement from a larger set of test cases to achieve double-blinding and reduction of bias.
- The method can be implemented rapidly, achieving WSI validation for primary diagnosis within days. This is foundational for further pandemic-related mitigative measures such as remote sign-out.
- Based on our findings, we suggest that validation studies are better conducted with a prespecified range of concordance in mind rather than a single fixed target figure.

Digital pathology and whole-slide imaging (WSI) are versatile tools that fulfill many roles in pathology teaching, clinical conferencing, slide archival, and research. Limited but focused routine implementations in telepathology and frozen-section interpretation have been carried at many institutions in the United States. Unlike radiology, which has incorporated digital imaging into routine practice for decades, pathology has moved much more slowly toward digitization. In 2017, the US Food and Drug Administration (FDA) approved the first WSI system for primary diagnosis in surgical pathology.[1] This occurred after a long process primarily because digital microscopy with WSI scanners was classified as a closed end-to-end class III device by the FDA and had to follow a premarket approval pathway requiring a clinical trial.[2] Subsequent devices, however, are considered class II, with a different FDA approval pathway.[3]

Until recently, primary digital diagnosis and consultation in histopathology remained predominantly a value-added service[4] rather than a new value proposition.

Prior regulatory barriers—both perceived and real—have significantly and negatively affected the digital performance of clinical services by surgical pathology departments and practices, especially when carried out remotely.[5,6] The FDA-approved Phillips and Leica implementations offered standardized pathology revenue-generation workflows incorporating digital methods but in the form of end-to-end closed systems. This in turn affects acquisition and deployment of the components of a digital-focused infrastructure, which carries significant up-front monetary and personnel costs. Digital pathology has been mired in the resulting cul-de-sac for close to a decade[6,7] in the United States.

The global severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic[8] has adversely affected functioning of pathology departments in various ways.[9] It has become apparent that hospital and health systems' services are better provided by remaining open and functional, to the maximum extent possible, even under pandemic conditions.[10] The Clinical Laboratories Improvement Amendments (CLIA) regulations require cytology and histopathology to be practiced in laboratory locations licensed for the purpose. This necessitates exposure of essential components of the clinical workforce—laboratory staff engaged in routine histopathology work, consulting and attending anatomic pathologists, residents, and fellows—to increased risk on a daily basis for an as-yet unforeseen length of time. Recognizing the potential for digital pathology to mitigate this risk, on March 26, 2020, the US government announced "enforcement discretion" of the CLIA regulations, amounting to a temporary relaxation of the rules prohibiting digital remote sign-out of cases.[11] On April 24, 2020, the FDA joined by declaring that it "does not intend to object to" to previous FDA-cleared restrictions on digital pathology devices and their marketing for remote use.[12] With large-scale noninferiority trials showing clear equivalency in diagnostic performance between glass and digital methods,[2,13] the College of American Pathologists (CAP) (March 26, 2020) issued a "COVID-19 Remote Sign-Out Guidance" that stated pathologists "may use a non-FDA approved system as long as it has been properly validated" for primary diagnosis. These brisk changes have greatly spurred interest in deployment of digital pathology for consultation and primary diagnosis as a medium- and long-term pandemic mitigative measure, even if not for the benefits that digital pathology brings **Table 1**.[4] As part of digital mitigative measures to be put into place, validation of the use of WSI for primary and consultative diagnosis would need to be completed by laboratories.

Validation of WSI for primary diagnosis ensures clinical precision and accuracy in unique, individual settings and provides a way for laboratories to demonstrate equivalence (or noninferiority) with the use of glass slides in diagnosis. In 2013, after extensive data review, validation guidelines applicable to WSI systems for diagnostic purposes were published by CAP.[15] The CAP guidelines offer 12 recommendations gathered from the review of 23 published studies. The recommendations chart a pathway for laboratories to perform a comprehensive end-to-end assessment of diagnosis using digital means while providing flexibility in the specific mechanisms of implementation.

Implementation of digital pathology is a multistep and continuous process[16] involving histotechnologists, imaging technologists, and information technology expertise, of which a validation study would be only the first step. The Department of Pathology at the University of Iowa Hospitals and Clinics is a midsized academic anatomic pathology program (36 faculty, 20 residents, 7 fellows) that had previously committed to a process of stepwise integration of digital capabilities into quantitative image analysis, tumor boards and teaching conference presentations, social media dissemination, case archiving, outside slide retention, and allied applications. The department acquired a P1000 whole-slide scanner in mid-2018 that was in use for teaching and research. No case or image management software was provided with the instrument or in use. As part of this effort, pathologists acquired experience in the basic functions of the WSI slide viewer program. We sought to perform validation of digital pathology for primary diagnosis in this context. We had 2 goals: (1) to perform rapid, robust validation of the scanner and associated digital infrastructure (computer hardware, software, file handling protocols) and (2) to move through the development life cycle[17] of digital diagnosis so that recommendations, problems, and barriers could be identified and iteratively addressed. We evolved and implemented specific methods that sought to fulfill recommendations in the CAP guidelines in an evidence-based manner. We present data that show successful validation utilizing well over 150 cases and involving 5 pathologists.

Materials and Methods

Case Inclusion and Scanning

We retrieved routine H&E slides from 180 surgical pathology cases from the files of the Department of Pathology at the University of Iowa Hospitals and Clinics. Cases with diagnoses rendered by each of the 5 participating study pathologists in the preceding

DOI: 10.1093/ajcp/aqaa280

Table 1

Distribution of Digital Cases (n = 171) by Small and Large Case Type and Anatomic Site by Subspecialty

Site	Case Type		
	Small	Large	Total
Head and neck			
Hard palate	1		1
Larynx	2	1	3
Mandible	3		3
Maxilla	2		2
Nose	1		1
Oral cavity	6	5	11
Parathyroid	2		2
Thyroid		2	2
Tonsil		1	1
Total	17	9	26
Skin			
Abdomen	1		1
Arm	2		2
Back	2		2
Ear	2		2
Hand	1		1
Melanoma		1	1
Neck	3		3
Nose	1		1
Penis	2		2
Right thumb	1		1
Scalp	2		2
Other sites	5		5
Umbilicus	1		1
Total	23	1	24
Genitourinary			
Kidney	3	3	6
Prostate	2	8	10
Ureter	1		1
Urethra	1		1
Urinary bladder	5	1	6
Vas deferens	6		6
Vulva		3	3
Total	18	15	33
Gastrointestinal			
Appendix	3		3
Colon	8	1	9
Esophagus		1	1
Gallbladder	1		1
Ileum		2	2
Liver	2		2
Multiple biopsies	1	5	6
Peritoneum		1	1
Rectum	1		1
Small intestine	1		1
Stomach	4		4
Total	21	10	31
Breast			
Axilla	1		1
Breast	11	5	16
Capsule	2		2
Total	14	5	19
Thoracic			
Bronchus	1		1
Heart	1		1
Lung	2	1	3
Total	4	1	5
Gynecologic			
Cervix	10		10
Endometrium	4		4
Endometrium + cervix	1		1
Fallopian tubes	1	1	2
Fallopian tubes + ovaries		1	1
Uterus	1	1	2
Uterus + fallopian tubes		2	2
Uterus + fallopian tubes + ovaries		5	5
Uterus + fallopian tubes + ovary		1	1
Vagina	1		1
Vulva	3	1	4
Total	21	12	33
Grand total	118	53	171

6-month to 2-week window were included (n = 36) such that 2 weeks or more had lapsed from the glass slide–based sign-out. Slides spanned the subspecialty disciplines of gastrointestinal (GI), gynecologic, head and neck, breast, genitourinary, and dermatologic pathology. Of the 36, two-thirds (n = 24) were "small" cases, including diagnostic biopsies and small resection specimens (eg, cholecystectomies), and the remainder (n = 12) were larger resections or multipart biopsies with a high number of slides per case. Within this framework, individual case selection was random and carried out such that the proportion of small and large cases would be roughly similar among pathologists and distributed evenly across subspecialties. During case search and retrieval, accession numbers for which slides were not on file or were comprised solely of frozen section slides were excluded. For each case, in line with CAP recommendations, frozen sections, special stains, and immunohistochemistry slides were removed by the adjudicating pathologist, and only those necessary and enough to arrive at the initial reported diagnosis were retained. Once assembled, slides were digitally scanned with the 20× objective (0.24µm/pixel resolution) by a P1000 Pannoramic scanner (3DHistech) by an imaging technologist. Glass slide markings for key findings and annotations (eg, lymph node status and count) were either removed at the scanning preview stage or after scanning, with the slide export function creating a second digital slide with only the manually selected area of interest. Quality control was performed by multiple personnel who examined digital slides for blurred areas, artifacts arising from scanner focusing errors, and tissue exclusion due to improper scanner thresholding (Figure 1, supplemental data; all supplemental materials can be

DOI: 10.1093/ajcp/aqaa280

found at *American Journal of Clinical Pathology* online). Slides with scanning errors were subject to rescan. Owing to the lack of a digital case manager, WSI slides were manually placed in subfolders named with the corresponding accession number and made available to study pathologists from a secure departmental network location. Brief clinical histories (ranging from 1 or 2 lines to a page) and gross descriptions including part labels and slide designations were automatically generated from the laboratory information system EPIC AP Beaker (EPIC), checked for formatting, and provided to the participants in the form of mock-up pathology requisition sheets. This process was carried out to closely replicate the original sign-out environment and to avoid logging into the pathology laboratory information system or electronic medical record to view details necessary for slide interpretation. The process is depicted in Figure 1.

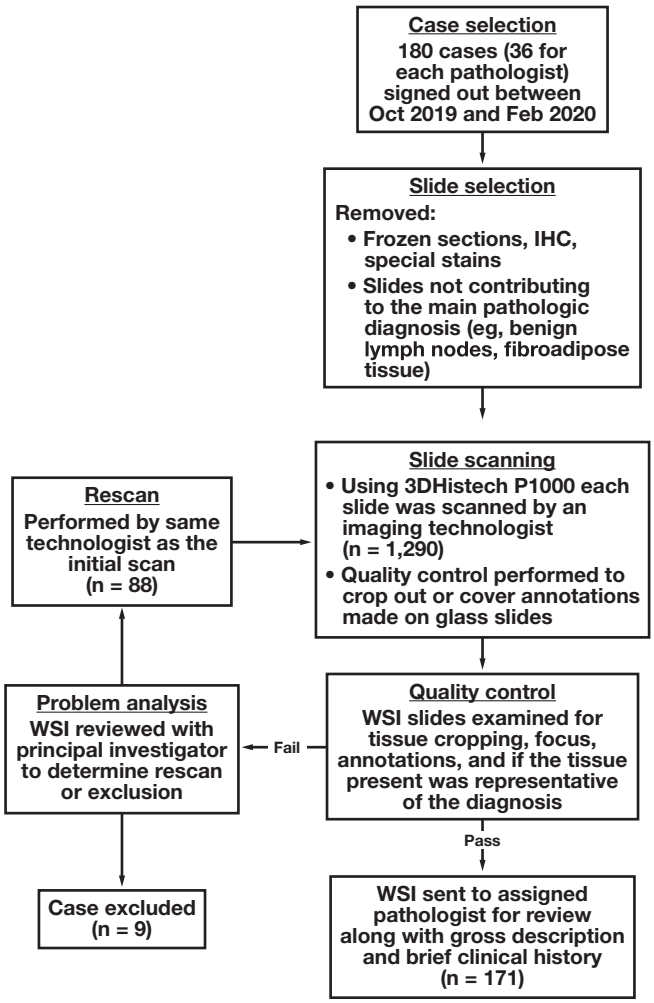

Figure 1 Case selection and exclusion and scanning quality control process used in the study. IHC, immunohistochemistry; WSI, whole-slide imaging.

Interpretation of Whole-Slide Images

Digital slides were viewed using CaseViewer 2.3.0 (3DHistech). The study pathologists recorded diagnoses, including major pertinent histopathologic findings such as the presence of lymphovascular or perineural invasion, tumor distances from inked margins, and the presence and size of metastatic tumor deposits.

Adjudication and Scoring

Standardized methods for case arbitration and discrepancy assessment in WSI have been described in the literature.[18] Given the smaller volume of cases to be analyzed, we adopted single-investigator classification and subject matter expert consultation as a method of adjudication. Post–digital review debriefing with the study pathologists was performed to inform participants of the review outcomes. Cases were classified as concordant or discordant, exhibiting major or minor discrepancy with the original diagnosis. A major discrepancy was defined as a change in diagnosis that would potentially affect either patient therapy or management after biopsy or surgery.[19,20] Intraobserver agreement was calculated for each study pathologist counting both major and minor disagreements and counting major disagreements alone. Although concordance was assessed on a per-part basis, the full case was considered discrepant even if any part was classified as being so. In addition, disagreement rates were calculated for the full set of cases for each pathologist and for the whole group.

Randomization

Case sampling was implemented to achieve blinding and randomization. Overall, 1,000 random samples (n = 18) with replacement were drawn for each pathologist from their total examined cases, and mean percentage of agreement and 95% CIs were computed. Likewise, 1,000 random samples (n = 90) were drawn from the full data set to compute mean percentage of agreement and 95% CIs for the whole sample set (see supplemental data for scripts used).

Validation Threshold Selection

As adopted by CAP, concordance (synonymous with intraobserver agreement) was the targeted metric, and the clinically reported diagnosis was set as the standard. CAP approves the medical director arriving at an acceptable intraobserver agreement rate in conjunction with the published data. In comparing the study sample set (n = 90) with available CAP data in which mean glass vs digital concordance rates range from 75% (n = 20 cases) to 91% (n = 200

DOI: 10.1093/ajcp/aqaa280

cases), the upper end was selected to be 91%. Although familiar with many routine functions, the study pathologists were not previously trained formally in the use of WSI, and the included cases encompassed those with diagnosis rendered well over the CAP-recommended washout period of 2 weeks; both situations could potentially have effects on intraobserver concordance. The CAP data review found that pathologists who were trained in using WSI showed greater concordance that those who were not (89% vs 84%; ie, a difference of 5%) based on a study of dermatopathology specimens.[21] Regarding the rate of disagreement with a gold standard diagnosis, manual (glass) slide review in a large scale multicenter double-blind study of 2,045 cases[13] was found to exhibit a 3.20% rate of disagreement. Taking the above 2 factors into account, a target concordance rate for the present validation protocol was prespecified as 81% to 91% counting major discrepancies alone. The χ^2 test was used to compare proportions of cases. Statistical analysis was performed in IBM SPSS version 26, and RStudio version 1.2.1335 running R 64-bit version 3.6.3.

Results

The profiles of the cases scanned are shown in Table 1 and **Figure 2**. Of the 180 initial cases, 9 were excluded from digital conversion. The main reason for exclusion was a high rate of scanning failure at first attempt within the case. Three breast resection cases were removed before scanning because they had high numbers of slides per case, and scanning failures occurred most frequently with breast tissue sections. A total of 88 glass slides were rescanned, with the highest number of rescans occurring in breast cases (**Table 2**, Figure 1). The WSI file sizes ranged from approximately 500 MB to 2.5 GB. The average number of slides per case was roughly equal among pathologists (6-7.6), excepting one (pathologist 3, 11.1 slides per case) who had a greater proportion of breast resections. The number of cases reviewed per pathologist ranged from 32 to 37. Following case adjudication, cases were classified as being concordant or with minor or major disagreement with the original conventional light microscopic ("glass") diagnosis. **Table 3** lists all disagreements adjudicated in the study. There were disagreements in grading of breast carcinoma (n = 5), prostate Gleason grade and score (n = 2), and oral dysplasia (n = 2). Other notable ones included misidentification of small diagnostic features in GI biopsies (n = 2) and dysplasia recognition in GI biopsies (n = 2). Counting all cases, the base concordance rate including major and minor disagreements for each pathologist ranged from 71.8% to 96.9% **Table 4**. The concordance including only major disagreements ranged from 93.7% to 96.9%. The mean pathologist concordance rate in sampled cases (n = 18) ranged from 72.2% to 94.18% **Table 5**. Mean concordance excluding minor disagreements ranged from 90.49% to 97% (Table 5). Assessment of the concordance rate for the group in sampled cases (n = 90) showed a mean of 83.62% counting all discrepancies and 94.72% counting only major disagreements **Figure 3**. No significant associations were identified between the occurrence of discrepancies (major or minor) and study pathologists (χ^2 test, $P = .19$) or subspecialty (χ^2 test, $P = .14$). No significant associations were identified between the occurrence of major discrepancies and pathologists (χ^2 test, $P = .91$), case type (large vs small; χ^2 test, $P = .06$), or subspecialty ($\chi 2$ test, $P = .31$). A small but significant increase in major disagreements was found in large cases compared with small cases ($\chi 2$ test, $P = .04$; see Figure 4, supplemental data).

Discussion

The Department of Pathology at the University of Iowa Hospitals and Clinics is staffed by 11 general surgical pathologists, each of whom is specialized in various subdisciplines of anatomic pathology, with an annual surgical pathology case load of approximately 50,000 cases. Three GI pathologists exclusively sign out GI pathology cases, although all 3 also participate in a portion of the general surgical sign-out. The cases included in the study (n = 171) roughly reflect the balance and proportion encountered in routine practice. These included a component of repeated case types, as the same type of surgical cases were included in sample sets reviewed by different pathologists (see Table 1). In our experience, as a partially subspecialized service, careful consideration needed to be given to subspecialty case inclusion because each type of case included had to be representative of the material that the group typically encounters as a whole. The mean concordance rate counting only major disagreements was 94.7%, which was above the prespecified threshold and comparable to the 95% concordance found by the CAP data review, thereby passing validation. The validation process not only satisfies CAP recommendations for digital diagnosis but also functions as a demonstration for first-time adopters of the viability, practicality, and role that an incoming digital system could play in routine practice.[1]

There has been debate about whether validation would require review of glass slides or digital slides first.[22] A recent rapid validation was performed with neuropathology cases (n = 30) but without assessment of intraobserver agreement (concordance).[23] Consequently,

DOI: 10.1093/ajcp/aqaa280

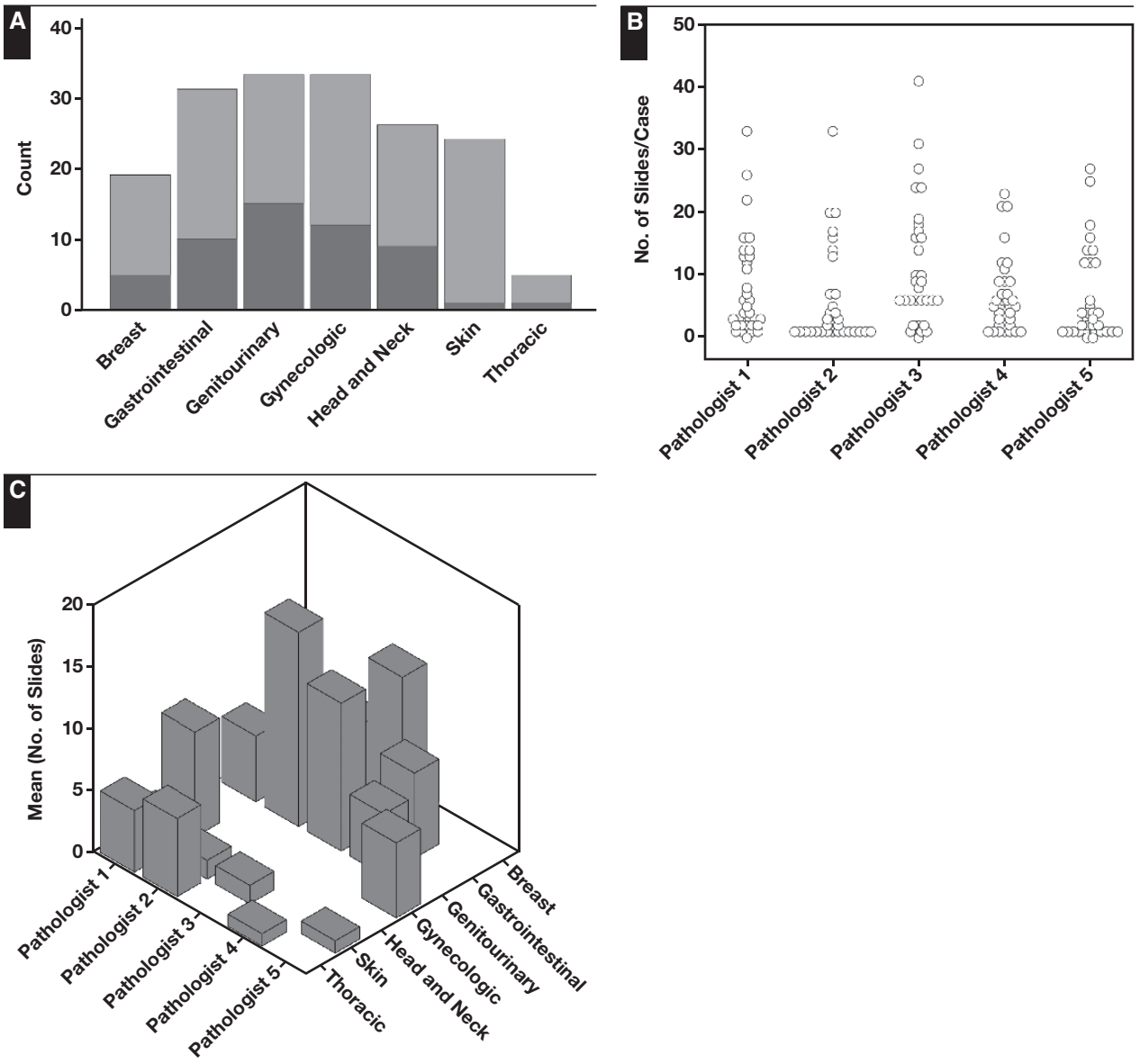

Figure 2 Distribution of digital cases (n = 171) by small (light gray) and large (dark gray) types (**A**) and numbers of slides per case (**B**) and by subspecialties (**C**) across study pathologists.

Table 2
Distribution of Cases and Slides Scanned per Participant

Pathologist	Cases	Slides	Slides Rescanned	Case Exclusions	Average Slides/Case
1	35	265	4	1	7.6
2	33	221	32	1	6.7
3	32	355	36	4	11.1
4	34	205	11	3	6
5	37	244	5	—	6.59
Total	171	1,290	88	9	7.5

it did not satisfy CAP recommendations. The 2013 CAP data review concluded that "nonrandom review"—that is, systematized viewing of one modality first, followed by the other—showed no differences with randomly ordered modalities on intraobserver agreement.[15] This conclusion greatly facilitates rapid validation. A retrospective cohort of previously diagnosed cases with glass slides can be put together instantaneously and digitally scanned, with case review started in a relatively short period of time. Importantly, this allows for full assessment of intraobserver agreement with the requisite washout period. This phase took 5 days to complete in our project.

The use of a retrospective case cohort raises concerns about the potential for selection bias. We utilized a cohort larger than the CAP-recommended 60 cases and performed subsampling to analyze subsets that were in the CAP-recommended range of 60 to 90 cases.

DOI: 10.1093/ajcp/aqaa280

Table 3
Major and Minor Discrepancies by Subspecialty, Adjudicated in Glass vs Digital Diagnosis

Disagreement Category	Site and Specimen Type	Original Diagnosis (Glass)	Digital Diagnosis	Count, n
Breast				
Major	Biopsy	Atypical ductal hyperplasia	Usual ductal hyperplasia	1
Minor	Biopsy	Invasive ductal carcinoma ESBR grade 1	Invasive ductal carcinoma ESBR grade 2	5
Dermatologic				
Minor	Penile skin	Fungal elements mentioned	Presence of fungal elements not mentioned	1
	Not specified	Mild atypia	Moderate atypia	1
Gastrointestinal				
Major	Gastric biopsy	Necroinflammatory debris	Not identified	1
	Colonic biopsy	Traditional serrated adenoma	Not identified	1
Minor	Esophageal biopsy	Epithelial atypia, indefinite for dysplasia, favor low-grade dysplasia	Epithelial atypia, indefinite for dysplasia, favor reactive	1
	Gastric biopsy	Fundic gland polyp	Polypoid colonic mucosa	1
	Colonic biopsy	Hyperplastic polyp	Tubular adenoma	1
	Rectal biopsy	Leiomyoma of muscularis propria	Not identified	1
	Colonic biopsy	Lymphocytic colitis identified	Not identified in same part; identified in other biopsies	1
	Appendectomy	Mild acute appendicitis	No appendicitis	1
	Multiple part	Tubular adenoma	Tubular adenoma not identified; different part in case shows adenocarcinoma	1
Genitourinary				
Major	Bladder biopsy	"Benign urothelium with focal urethritis"	"Suspicious for low-grade papillary urothelial carcinoma"	1
	Prostate biopsy	Atypical small acinar proliferation	Prostatic adenocarcinoma Gleason grade 3 + 4 = 7	1
	Prostate biopsy	Benign	Suspicious for carcinoma	1
Minor	Prostate biopsy	Prostatic adenocarcinoma Gleason Grade 3 + 3 = Score 6 (GG 1)	Prostatic adenocarcinoma Gleason grade 3 + 4 = 7 (GG 2)	1
Gynecologic				
Major	Cervix biopsy	Low-grade squamous intraepithelial lesion	Reactive change	1
	Oophorectomy	Mucinous cystadenoma	Borderline mucinous tumor	1
Head and Neck				
Minor	Oral biopsy	Invasive squamous cell carcinoma	"Atypical cells, cannot exclude squamous cell carcinoma	1
	Oral biopsy	Necroinflammatory debris	Not identified	1
	Oral biopsy	"Residual invasive squamous cell carcinoma"	"Atypical cells in necrotic tissue, cannot exclude squamous cell carcinoma"	1
	Oral resection	Perineural invasion identified	Perineural invasion not identified	1
Thoracic				
Major	Lung resection	Micropapillary predominant lung adenocarcinoma	Acinar predominant lung adenocarcinoma	1

ESBR, Elston-Ellis modification of Scarff-Bloom-Richardson; GG, Grade Group.

Table 4
Intraobserver Agreement (Concordance) Among 5 Study Pathologists[a]

	Pathologist 1	Pathologist 2	Pathologist 3	Pathologist 4	Pathologist 5
Base concordance	29/35 (82.8)	32/33 (93.9)	23/32 (71.8)	32/34 (94.1)	28/37 (75.7)
Base disagreement rate	6/35 (11.7)	2/33 (6.06)	9/32 (28.1)	2/34 (5.8)	9/37 (24.3)
Concordance counting major disagreements	33/35 (94.2)	32/33 (96.9)	30/32 (93.7)	32/34 (94.1)	35/37 (94.6)
Major disagreement rate	2/35 (5.8)	1/33 (3.03)	2/32 (6.2)	2/34 (5.8)	2/37 (5.4)

[a]Data reflect all cases examined by each pathologist (range, 32-37). Concordant cases are listed over total cases, and the percentage agreement is in parentheses.

A larger cohort is gathered more easily and obviates the need for micromanagement in assembling sample sets to represent cases encountered in routine practice. By selecting cases for all study participants within the same 5-month period (October 2019 to February 2020), we ensured that the cohort would include a representative

DOI: 10.1093/ajcp/aqaa280

Table 5

Mean Intraobserver Agreement (Concordance) and 95% CIs in 1000 Samples (n = 18)[a]

	Pathologist 1	Pathologist 2	Pathologist 3	Pathologist 4	Pathologist 5
Mean concordance, %	82.96	93.61	72.20	94.18	75.72
95% CI	82.56-83.36	93.38-93.84	71.77-72.62	93.95-94.42	75.26-76.17
Mean concordance counting major discrepancies, %	94.28	97	90.49	94.18	94.59
95% CI	94.03-94.53	96.82-97.17	90.20-90.78	93.95-94.42	94.35-94.83

CI, confidence interval.

[a]Samples were drawn from the total cases (range, 32-37) examined by each study pathologist. Minor discrepancies were adjudicated and resolved in discussion with a second expert, as needed.

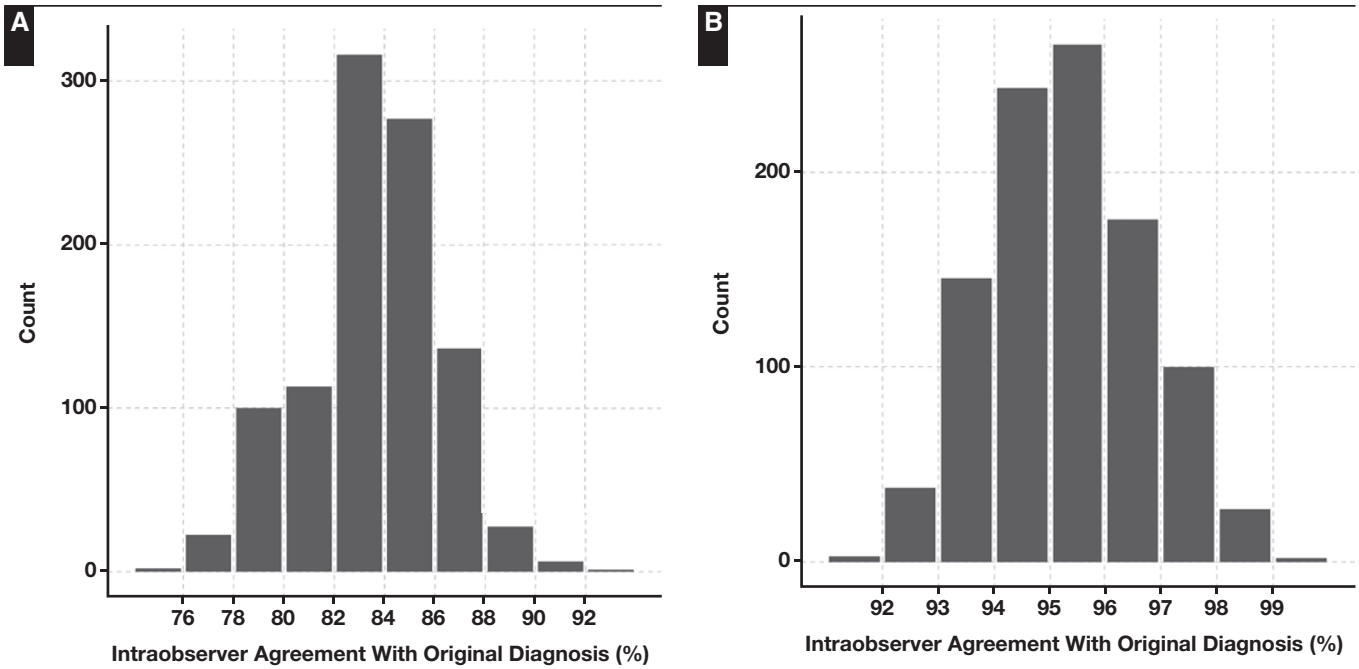

Figure 3 Distribution of percentage of agreement in 90 cases (1,000 samples) drawn from total (n = 171). **A**, Concordance rate counting minor and major diagnostic discrepancies, with a mean of 83.66 (95% confidence interval [CI], 83.49-83.83; range, 75.56-92.22). **B**, Distribution of concordance counting only major discrepancies, with a mean of 94.58 (95% CI, 94.48-94.68; range, 90-100).

cross-section of diagnoses. In this paradigm, both the reviewing pathologist and the investigator are unaware of the cases that count toward final assessment, accomplishing double-blinding. Double-blinding of a similar fashion can be implemented in forward-looking (prospective) validation studies.

In this study, mean intraobserver agreement is tightly constrained with close 95% CIs. For the sample set (n = 90), the mean concordance counting major disagreements was 94.7%, with a 95% CI of 94.6% and 94.8% (Figure 3B). This establishes the robustness of the level of agreement between glass and digital methods and is indistinguishable from the CAP-recommended threshold of 95%. However, the range of possible intraobserver agreement counting all discrepancies was wider: 75.5% to 92.2% (Figure 3A). This result implies that if were intraobserver agreement to be assessed at any fixed interval in real life, depending on the number of cases assessed and level of stringency of assessment, there would be runs in which the level of concordance were lower than that observed in the validation exercise. In other words, the system may appear to perform "worse" than expected. However, this should not be construed as evidence for suboptimal reproducibility or diagnostic performance on digital platforms. Investigators must not expect static concordance of the level observed in the validation study process. Moreover, this result implies that validation studies are better conducted with a prespecified range of concordance in mind rather than a single fixed target figure.

Several factors affect and govern intraobserver agreement in histopathology, and these have been discussed extensively in the digital pathology validation literature.[24-32]

DOI: 10.1093/ajcp/aqaa280

Several factors were pertinent to the design and implementation adopted in the present study.

First, a significant proportion of the base disagreement observed in the study occurred in forms of semiquantitative assessment of morphologic features (ie, dysplasia grading). Discrepancies in dysplasia assessment and grading can have 2 broad contributory factors. Davidson et al[33] observed a 27% intraobserver disagreement using glass slides in assigning a Nottingham grade to cases of invasive breast carcinoma. Similar figures have long been obtained in studies examining intraobserver agreement in grading of cervical intraepithelial neoplasia[34] and Gleason grading of prostatic acinar adenocarcinoma.[35] In fact, in the case of cervical biopsy dysplasia grading, the cause for the higher-than-expected discrepant grading and reduced reproducibility was pinned on the classification system, and a simpler 2-tier system was globally adopted as a result.[36,37] These results are now understood to be domains in diagnostic pathology that inherently exhibit a degree of intra- and interobserver disagreement regardless of the diagnostic modality (ie, even when using glass slides alone).[38,39]

Nevertheless, evidence suggests that WSI examination may pose challenges with interpretation in dysplasia grading and identification of small objects in tissue. This has been noted by others as well; Bauer and Slaw reported improved neutrophil detection in GI biopsies[27] when scanned at 400×. Appreciation of chromatin details is another important area in which digital pathology performs differently compared with conventional light microscopy. We found that assessment of nuclear chromatin likely influenced the interpretation of at least 2 GI pathology cases, leading to minor discrepancies. In both instances, relative hyperchromasia was not properly gauged on WSI slides, leading to underdiagnosis of tubular adenomas on GI luminal biopsies. Similar findings have been reported in the literature.[40] A systematic analysis of digital vs glass discrepancies reported in the literature (39 studies) found that differences in dysplasia diagnosis were the most frequently encountered discrepancy.[41]

Second, we sought to include cases from a relatively long period going back in time (weeks to months). This approach provided for representation of routine practice and obviated the problem of pathologist memory of their cases, which can be surprisingly long in selected instances,[27] particularly with unique or rare diagnoses. However, the longer the duration between the initial diagnosis (glass or diagnosis) and validation, the greater the likelihood that the participant's diagnostic thresholds and techniques have subtly (or dramatically) shifted. At these scales, concordance (intraobserver agreement) behaves more like interobserver agreement, which displays greater variability in reproducibility studies.

Third, there is no reason to suspect that pathologists are not subject to responder or recall bias in retrospective studies. Recall bias is a form of systematic error that is classically described as occurring in epidemiologic research owing to study participants' greater recollection and thoroughness of past events compared with control participants.[42] In this context, pathologists are analogous to study participants, and it would be impossible for them not to allow awareness of participating in a study to affect their interpretation of digital slides, which they may examine in greater detail or spend more time on (thereby reducing equivalence between the glass and digital slides interpretative process).

A notable observation that has bearing on scanning quality assurance is the high rate of skipped areas in tissue scanning that we encountered in breast specimens (36/88 rescanned slides). This likely occurs owing to the difficulty in identifying the tissue plane during image capture in fat-rich tissue, which can be devoid of visual detail that aids in autofocusing. Stemming from experience in research scanning, we previously incorporated into routine practice the quality assurance step of quick verification of the presence of all tissue fragments on scanned whole slides in our laboratory. We found that the "Show Scan area outline" and "Slide preview image" functions in 3DHistech Caseviewer (Figure 1, supplemental data) were highly effective in performing a quick screen for missing tissue fragments or areas. Based on our validation experience, we required that slides that failed the quality check because of skipped areas be rescanned by selecting a scanning profile that increased the number of focus points. The significance of independent whole-slide thumbnail images or views in WSI diagnosis emerges recurrently in several studies,[41,43] mirroring our experience in the present study. The functionalities of manually reviewing the whole-slide thumbnails and adjusting focus points to compensate for skipped or missed areas are of critical importance in validation and primary digital diagnosis independent of the digital platform used. An open-source quality control tool (HistoQC) that automates the process of scanning WSI for blurred areas and artifacts has been described recently,[44] although it is unclear if it can be used to identify missing areas.

None of the participants in the study were trained or had significant prior experience in the use of digital pathology for diagnosis, although they used WSI for research and clinical case conferences on an intermittent basis. Prior training with WSI and program interfaces could improve diagnostic performance. One study, for instance, found improvement in concordance over time among pathologists interpreting uterine cervical biopsies.[45]

DOI: 10.1093/ajcp/aqaa280

Given each of the factors listed, the exact proportion of the total disagreement between glass and digital modalities attributable to digital slides is unclear. Nevertheless, it is highly likely that the diagnostic performance of digital modalities is underestimated rather than overestimated. As presently carried out, factors inherent in studies of measurement of intra- and interobserver agreement—and retrospective study designs in general—are misattributed to and adversely affect measured performance of digital slides.

Conclusions

We described a method for rapid validation of digital pathology for primary digital diagnosis using minimum resources that fully complies with CAP recommendations. In a broader sense, there continues to be a need to evolve better and standardized methods for anatomic pathology validation and measurement of diagnostic performance of digital WSI.

Corresponding author: Anand Rajan KD, MBBS; anand-rajand@uiowa.edu.

Acknowledgments: The validation study was carried out as part of a quality improvement project. The University of Iowa institutional review board determined the validation to be exempt from a full review by the board.

References

1. Evans AJ, Bauer TW, Bui MM, et al. US Food and Drug Administration approval of whole slide imaging for primary diagnosis: a key milestone is reached and new questions are raised. *Arch Pathol Lab Med.* 2018;142:1383-1387.
2. Mukhopadhyay S, Feldman MD, Abels E, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (Pivotal Study). *Am J Surg Pathol.* 2018;42:39-52.
3. College of American Pathologists. Remote sign-out of cases with digital pathology FAQs. 2020. https://www.cap.org/covid-19/remote-sign-out-faqs. Accessed November 2020.
4. Isaacs M, Lennerz JK, Yates S, et al. Implementation of whole slide imaging in surgical pathology: a value added approach. *J Pathol Inform.* 2011;2:39.
5. Abels E, Pantanowitz L. Current state of the regulatory trajectory for whole slide imaging devices in the USA. *J Pathol Inform.* 2017;8:23.
6. Parwani AV, Hassell L, Glassy E, et al. Regulatory barriers surrounding the use of whole slide imaging in the United States of America. *J Pathol Inform.* 2014;5:38.
7. Asa SL, Evans A. Issues to consider when implementing digital pathology for primary diagnosis. *Arch Pathol Lab Med.* 2020;144:1297.
8. Guo YR, Cao QD, Hong ZS, et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak—an update on the status. *Mil Med Res.* 2020;7(11):1-10.
9. Rubin O. "It will bankrupt my department": how private labs are paying the price for testing. *ABC News* 2020. https://abcnews.go.com/Health/bankrupt-department-private-labs-paying-price-testing/story?id = 69707226. Accessed May 7, 2020.
10. McKee GA. US hospitals function like businesses. That's why they are struggling amid COVID pandemic. *USA TODAY* 2020. https://www.usatoday.com/story/opinion/2020/07/05/coronavirus-hospitals-businesses-public-option-health-care-column/3266503001. Accessed July 26, 2020.
11. Centers for Medicare and Medicaid Services. Clinical Laboratory Improvement Amendments (CLIA) laboratory guidance during COVID-19 public health emergency 2020. QSO-20-21-CLIA. https://www.cms.gov/medicareprovider-enrollment-and-certificationsurveycertificationgeninfopolicy-and-memos-states-and/clinical-laboratory-improvement-amendments-clia-laboratory-guidance-during-covid-19-public-health. Accessed May 17, 2020.
12. US Food and Drug Administration. Center for Devices and Radiological Health. Office of Product Evaluation and Quality. Enforcement policy for remote digital pathology devices during the coronavirus disease 2019 (COVID-19) public health emergency. 2020. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enforcement-policy-remote-digital-pathology-devices-during-coronavirus-disease-2019-covid-19-public. Accessed May 7, 2020.
13. Borowsky AD, Glassy EF, Wallace WD, et al. Digital whole slide imaging compared with light microscopy for primary diagnosis in surgical pathology. *Arch Pathol Lab Med.* 2020;144:1245-1253. doi:10.5858/arpa.2019-0569-OA.
14. Allen TC. Comes digital pathology. *Arch Pathol Lab Med.* 2015;139:972.
15. Pantanowitz L, Sinard JH, Henricks WH, et al; College of American Pathologists Pathology and Laboratory Quality Center. 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.
16. Cheng CL, Tan PH. Digital pathology in the diagnostic setting: beyond technology into best practice and service management. *J Clin Pathol.* 2017;70:454-457.
17. Everett GD, McLeod R Jr. *Software Testing: Testing Across the Entire Software Development Life Cycle.* Hoboken, NJ: Wiley; 2007.
18. Wack K, Drogowski L, Treloar M, et al. A multisite validation of whole slide imaging for primary diagnosis using standardized data collection and analysis. *J Pathol Inform.* 2016;7:49. doi:10.4103/2153-3539.194841.
19. Raab SS, Nakhleh RE, Ruby SG. Patient safety in anatomic pathology: measuring discrepancy frequencies and causes. *Arch Pathol Lab Med.* 2005;129:459-466.
20. Roy JE, Hunt JL. Detection and classification of diagnostic discrepancies (errors) in surgical pathology. *Adv Anat Pathol.* 2010;17:359-365.
21. Nielsen PS, Lindebjerg J, Rasmussen J, et al. Virtual microscopy: an evaluation of its validity and diagnostic performance in routine histologic diagnosis of skin tumors. *Hum Pathol.* 2010;41:1770-1776.
22. Lowe A, Chlipala E, Elin J, et al. *Validation of Digital Pathology in a Healthcare Environment.* San Diego, CA: Digital Pathology Association; 2011.

DOI: 10.1093/ajcp/aqaa280

23. Henriksen J, Kolognizak T, Houghton T, et al. Rapid validation of telepathology by an academic neuropathology practice during the COVID-19 pandemic. *Arch Pathol Lab Med.* 2020;144:1311-1320.
24. Al-Janabi S, Huisman A, Vink A, et al. Whole slide images for primary diagnostics of gastrointestinal tract pathology: a feasibility study. *Hum Pathol.* 2012;43:702-707.
25. Al-Janabi S, Huisman A, Willems SM, et al. Digital slide images for primary diagnostics in breast pathology: a feasibility study. *Hum Pathol.* 2012;43:2318-2325.
26. Bauer TW, Schoenfield L, Slaw RJ, et al. Validation of whole slide imaging for primary diagnosis in surgical pathology. *Arch Pathol Lab Med.* 2013;137:518-524.
27. Bauer TW, Slaw RJ. Validating whole-slide imaging for consultation diagnoses in surgical pathology. *Arch Pathol Lab Med.* 2014;138:1459-1465.
28. Bauer TW, Slaw RJ, McKenney JK, et al. Validation of whole slide imaging for frozen section diagnosis in surgical pathology. *J Pathol Inform.* 2015;6:49. doi:10.4103/2153-3539.163988.
29. Buck TP, Dilorio R, Havrilla L, et al. Validation of a whole slide imaging system for primary diagnosis in surgical pathology: a community hospital experience. *J Pathol Inform.* 2014;5:43. doi:10.4103/2153-3539.145731.
30. Campbell WS, Lele SM, West WW, et al. Concordance between whole-slide imaging and light microscopy for routine surgical pathology. *Hum Pathol.* 2012;43:1739-1744.
31. Hanna MG, Reuter VE, Hameed MR, et al. Whole slide imaging equivalency and efficiency study: experience at a large academic center. *Mod Pathol.* 2019;32:916-928.
32. Williams BJ, Hanby A, Millican-Slater R, et al. Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study on digital pathology validation and training. *Histopathology.* 2018;72:662-671.
33. Davidson TM, Rendi MH, Frederick PD, et al. Breast cancer prognostic factors in the digital era: comparison of nottingham grade using whole slide images and glass slides. *J Pathol Inform.* 2019;10:11. doi:10.4103/jpi.jpi_29_18.
34. McCluggage WG, Walsh MY, Thornton CM, et al. Inter- and intra-observer variation in the histopathological reporting of cervical squamous intraepithelial lesions using a modified Bethesda grading system. *Br J Obstet Gynaecol.* 1998;105:206-210.
35. Melia J, Moseley R, Ball RY, et al. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology.* 2006;48:644-654.
36. Genest DR, Stein L, Cibas E, et al. A binary (Bethesda) system for classifying cervical cancer precursors: criteria, reproducibility, and viral correlates. *Hum Pathol.* 1993;24:730-736.
37. Kurman RJ, Malkasian GD Jr, Sedlis A, et al. From papanicolaou to Bethesda: the rationale for a new cervical cytologic classification. *Obstet Gynecol.* 1991;77:779-782.
38. Krishnan L, Karpagaselvi K, Kumarswamy J, et al. Inter- and intra-observer variability in three grading systems for oral epithelial dysplasia. *J Oral Maxillofac Pathol.* 2016;20:261-268.
39. Kujan O, Khattab A, Oliver RJ, et al. Why oral histopathology suffers inter-observer variability on grading oral epithelial dysplasia: an attempt to understand the sources of variation. *Oral Oncol.* 2007;43:224-231.
40. Damaceno Araujo AL, Aristizabal Arboleda LP, Palmier NR, et al. The performance of digital microscopy for primary diagnosis in human pathology: a systematic review. *Virchows Archiv.* 2019;474:269-287.
41. Williams BJ, DaCosta P, Goacher E, et al. A systematic analysis of discordant diagnoses in digital pathology compared with light microscopy. *Arch Pathol Lab Med.* 2017;141:1712-1718.
42. Coughlin SS. Recall bias in epidemiologic studies. *J Clin Epidemiol.* 1990;43:87-91.
43. Fraggetta F, Yagi Y, Garcia-Rojo M, et al. The importance of eslide macro images for primary diagnosis with whole slide imaging. *J Pathol Inform.* 2018;9:46.
44. Janowczyk A, Zuo R, Gilmore H, et al. HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clin Cancer Inform.* 2019;3:1-7.
45. Ordi J, Castillo P, Saco A, et al. Validation of whole slide imaging in the primary diagnosis of gynaecological pathology in a university hospital. *J Clin Pathol.* 2015;68:33-39.

DOI: 10.1093/ajcp/aqaa280