Digital Whole Slide Imaging Compared With Light Microscopy for Primary Diagnosis in Surgical Pathology

A Multicenter, Double-Blinded, Randomized Study of 2045 Cases

Alexander D. Borowsky, MD; Eric F. Glassy, MD; William Dean Wallace, MD; Nathash S. Kallichanda, MD; Cynthia A. Behling, MD; Dylan V. Miller, MD; Hemlata N. Oswal, MD; Richard M. Feddersen, MD; Omid R. Bakhtar, MD; Arturo E. Mendoza, MD; Daniel P. Molden, MD; Helene L. Saffer, MD; Christopher R. Wixom, MD; James E. Albro, MD; Melissa H. Cessna, MD; Brian J. Hall, MD; Isaac E. Lloyd, MD; John W. Bishop, MD; Morgan A. Darrow, MD; Dorina Gui, MD, PhD; Kuang-Yu Jen, MD, PhD; Julie Ann S. Walby, MD; Stephen M. Bauer, MD; Daniel A. Cortez, MD; Pranav Gandhi, MD; Melissa M. Rodgers, MD; Rafael A. Rodriguez, MD; David R. Martin, MD; Thomas G. McConnell, MD; Samuel J. Reynolds, MD; James H. Spigel, MD; Shelly A. Stepenaskie, MD; Elena Viktorova, PhD; Robert Magari, PhD; Keith A. Wharton Jr, MD, PhD; Jinsong Qiu, PhD; Thomas W. Bauer, MD

• ***Context.***—**The adoption of digital capture of pathology slides as whole slide images (WSI) for educational and research applications has proven utility.**

Objective.—**To compare pathologists' primary diagnoses derived from WSI versus the standard microscope. Because WSIs differ in format and method of observation compared with the current standard glass slide microscopy, this study is critical to potential clinical adoption of digital pathology.**

Design.—**The study enrolled a total of 2045 cases enriched for more difficult diagnostic categories and represented as 5849 slides were curated and provided for diagnosis by a team of 19 reading pathologists separately as WSI or as glass slides viewed by light microscope. Cases were reviewed by each pathologist in both modalities in randomized order with a minimum 31-day washout between modality reads for each case. Each diagnosis was compared with the original clinical reference diagnosis by an independent central adjudication review.**

Results.—**The overall major discrepancy rates were 3.64% for WSI review and 3.20% for manual slide review diagnosis methods, a difference of 0.44% (95% CI, −0.15 to 1.03). The time to review a case averaged 5.20 minutes for WSI and 4.95 minutes for glass slides. There was no specific subset of diagnostic category that showed higher rates of modality-specific discrepancy, though some categories showed greater discrepancy than others in both modalities.**

Conclusions.—**WSIs are noninferior to traditional glass slides for primary diagnosis in anatomic pathology.**

(*Arch Pathol Lab Med.* 2020;144:1245–1253; doi: 10.5858/arpa.2019-0569-OA)

Accepted for publication January 3, 2020.

Published online February 14, 2020.

Supplemental digital content is available for this article at www.archivesofpathology.org in the October 2020 table of contents.

From the Department of Pathology and Laboratory Medicine, University of California, Davis, Sacramento (Borowsky, Bishop, Darrow, Gui, Jen, Walby); the Affiliated Pathologists Medical Group, Rancho Dominguez, California (Glassy, Kallichanda); UCLA Pathology & Lab Medicine, University of California, Los Angeles (Wallace); the Pacific Rim Pathology Lab and Sharp Healthcare, San Diego, California (Behling, Mendoza, Molden, Saffer, Wixom); Intermountain Central Laboratory, Salt Lake City, Utah (Miller, Albro, Cessna, Hall, Lloyd); the Pathology Department, Lucent Pathology Partners Mercy San Juan Hospital, Carmichael, California (Oswal, SM Bauer, Cortez, Rodgers, Rodriguez); the Histology Lab, TriCore Reference Laboratories, Albuquerque, New Mexico (Feddersen, Martin, McConnell, Reynolds, Spigel, Stepenaskie); Scripps Clinic Torrey Pines, La Jolla, California (Bakhtar, Ghandi); Beckman Coulter, Inc., Miami, Florida (Viktorova, Magari); Leica Biosystems Imaging, Inc., Danvers, Massachusetts (Wharton); Leica Biosystems, Vista, California (Qiu); and the Department of Pathology and Laboratory Medicine, Hospital for Special Surgery, Weill Cornell Medical College, New York, New York (TW Bauer). Wallace is now with the Department of Pathology, at Keck School of Medicine of University of Southern California, Los Angeles. Glassy, Wallace, Kallichanda, Behling, Miller, Oswal, Feddersen, Bakhtar, Mendoza, Molden, Saffer, Wixom, Albro, Cessna, Hall, Lloyd, Bishop, Darrow, Gui, Jen, Walby, SM Bauer, Cortez, Gandhi, Rodgers, Rodriguez, Martin, McConnell, Reynolds, Spigel, and Stepenaskie contributed equally to the study. Borowsky and TW Bauer contributed equally in directing and reporting the study.

This work was sponsored by Leica Biosystems Imaging, Inc. and supported by the Center for Genomic Pathology, Davis, California (a 501(c)(3) organization affiliated with UC Davis).

Glassy received consulting fees/honoraria for other Leica projects. Wallace was a full-time employee of the University of California when he performed this work; he received an honorarium for his participation. Viktorova and Magari are full-time employees of Beckman Coulter, Inc. Wharton is a stock holder and employee at Leica Biosystems and holds stock with Novartis. Qiu is a full-time employee of Leica Biosystems. TW Bauer is a consultant to Leica Biosystems and has received payments related to meeting with United States Food and Drug Administration and digital pathology validation. The other authors have no relevant financial interest in the products or companies described in this article.

Corresponding author: Alexander D. Borowsky, MD, Department of Pathology and Laboratory Medicine, University of California, Davis, 4400 V St, Sacramento, CA 95817 (email: adborowsky@ucdavis.edu).

Histopathology interpretation remains an essential, foundational method for disease diagnosis in medical practice.[1] Standard histopathology assessment is initiated through interpretation of hematoxylin and eosin–stained glass microscope slides prepared from formalin fixed tissue and evaluated by trained anatomic pathologists using a transmission light microscope, typically outfitted with turret-mounted objectives that magnify the tissue approximately ×20 to ×400 (×2 to ×40 objective lens plus ×10 to ×12 eyepiece magnification). Special histochemical stains and immunohistochemical (IHC) stains augment interpretation of the hematoxylin and eosin–stained slides to provide a "primary diagnosis" that becomes part of the medical record to serve as a basis of the patient's therapy.

Methods and technologies to convert glass slides into digital whole slide images (WSI) that can be viewed on digital monitors over a similar range of magnifications have been tested for utility in a variety of applications.[2–5] Previous studies have compared pathologists' use of WSI with microscopy for primary diagnosis in surgical pathology.[6–10] Some studies tested a relatively small number of cases, others have had relatively short "washout" times between interpretations, some have been limited to a specific tissue type/organ, and only a few have compared diagnoses of each modality with a reference diagnosis. Guidelines for validating WSI for diagnosis have been proposed.[11,12] Mukhopadhyay et al[13] reported the results of a randomized study of 1992 cases with 16 reading pathologists, in which use of one specific whole slide imaging and display system was found to be noninferior to use of a microscope. A study of 900 cases with 9 pathologists used and compared multiple platforms, including manufacturer's platforms used in both the Mukhopadhyay study as well as this study, and did not find platform dependent differences in diagnostic discrepancy rates.[14] Additional validation studies on multiple platforms will contribute to the confidence in the generalizability of WSI for clinical use.[15]

The purpose of this study was to determine if primary diagnoses rendered by pathologists viewing WSIs are noninferior to diagnoses rendered by viewing glass slides with light microscopy. The Aperio AT2 DX system (Leica Biosystems, Inc., Vista, California) evaluated in this study included a scanner with a 0.75 numerical aperture ×20 objective in a line scan format with multipoint autofocusing. Pathologists access the WSI database to retrieve case image sets that are viewed on workstations with Dell (Round Rock, Texas) medical grade monitors. The ImageScope browser software (Leica Biosystems, Inc.) includes tools for zoom, rotation, and navigation, including a navigation history pane that indicates unexamined areas of each WSI. We hypothesized that the Aperio AT2 DX system's WSI used for primary diagnosis is noninferior to traditional microscopy. To test this hypothesis, we compared concordances of diagnoses from over 2000 cases interpreted by pathologists with WSI and light microscopy with the case's original diagnosis, referred to in this study as the "reference" diagnosis. We show, using predetermined endpoints derived from consensus recommendations, primary diagnoses from practicing pathologists using digital WSI are noninferior to diagnoses generated using standard glass slide microscopy.

METHODS

Design

This multicenter, double-blinded, randomized study was conducted at 5 sites (1 academic hospital and 4 community and reference practice groups in California, Utah, and New Mexico). The study involved the retrospective evaluation of archival slides and recut sections of tissue blocks previously used for patient care; no human subjects were prospectively enrolled. The protocol and its amendments were approved by institutional review boards at each site. This noninferiority study (study design shown in Figure 1) started in July 2016 with the initiation of the laboratory information system search for the first case. Data analysis was completed in December 2018.

Pathologists in the study were US board-certified in anatomic pathology and licensed practitioners in their state and included 27 active surgical pathology practitioners. Each pathologist served in only one of the following roles: curating pathologists assembled and curated study sets (2 pathologists per site); reading pathologists provided study diagnoses for both glass slide and WSI modalities (3 or 4 per site); and adjudication pathologists compared reading pathologist diagnosis with reference diagnoses with adjudicate discrepancies (3 per study, for central review).

Case Screening and Admission Into the Study

Case Curation.—Guidelines for the types of cases and organ systems included in the study were based on discussions with the US Food and Drug Administration, and were intended to be representative of cases commonly encountered in routine practice, with an over-representation of cases from more difficult diagnostic categories (Table 1). Each study site performed one or more searches of the respective laboratory information system to identify cases for inclusion. Briefly, all cases were obtained consecutively by date of original collection, which was at least 1 year prior to the initiation of the study. Curating pathologists reviewed the archival slides, including relevant special stains and IHC studies, to determine adequacy for the reference diagnosis. If routine hematoxylin and eosin slides were not available, recut sections were obtained. If special stains or IHC studies were not available but needed for the diagnosis, they were not repeated and the case was excluded. One of 2 curating pathologists at each site confirmed adequate clinical data were provided, and selected slides required for the diagnosis, staging, and pertinent negative findings for each case. In some cases, all of the slides for a case were included, but a representative subset of slides considered sufficient for determination of the reference diagnosis could be selected.

The selection of slides from each clinical case followed predefined rules. For example, a multiple part case, such as simultaneously obtained prostate biopsies from different areas of the prostate, might have been obtained in a search for the category of "prostate core biopsy cancer." In such a case, the first sequentially obtained specimen meeting the diagnostic category (eg, prostate adenocarcinoma) was selected as follows: if the first and second core sites were originally diagnosed as benign, they were excluded; if the third core site was diagnosed as cancer, it was chosen for study inclusion even if subsequent biopsy sites contained higher grade or a greater volume of cancer, to avoid biasing the case cohort with the largest or most prominent example of a lesion or diagnosis. In this example, the result of this rule ensured that prostate cores with varying degrees of abnormality and hence challenge to diagnose were randomly entered into the study set. In some such multipart cases, additional site slides could be included if they were useful in interpretation of that first sequential site.

Case Curation Verification.—Once enough cases for each of the predetermined, US Food and Drug Administration–recommended organ site categories were obtained at each site, the curated cases were de-identified. De-identified data from the diagnostic report entered into the study database included the patient's age and sex, race (if provided), organ site, procedure, gross description, block designation, clinical context metadata, and

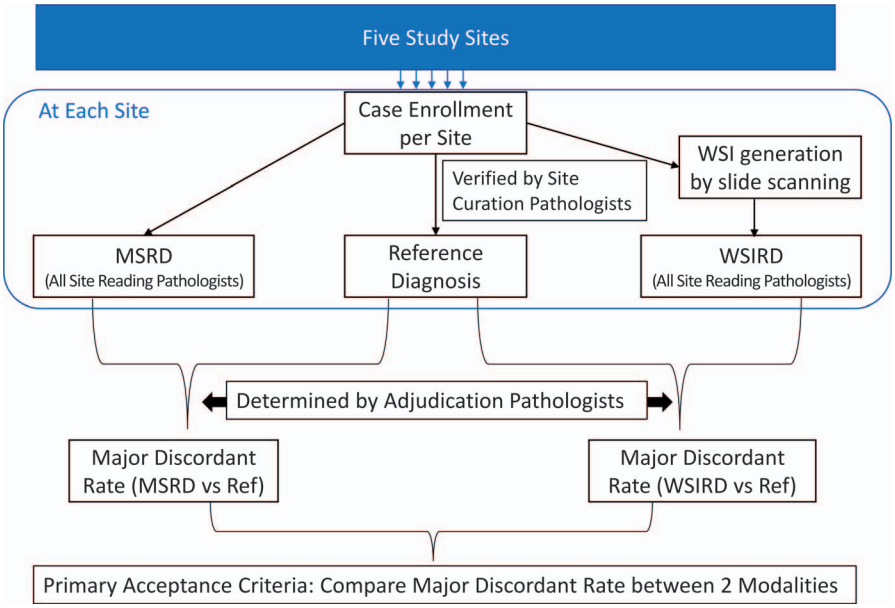

Figure 1. *Overall study design. Each of 5 sites enrolled cases, curated, scanned, and provided 3 or 4 reading pathologists' diagnoses for each case with a minimum 31-day washout between modalities. All sites' data were reviewed centrally by the adjudication pathologists. Abbreviations: MSRD, light microscopy slide review diagnosis; Ref, reference; WSI, whole slide image; WSIRD, whole slide image review diagnosis.*

reference diagnosis. The de-identified cases and report data were reviewed by the second curation pathologist at each site to confirm diagnostic adequacy of the slides and data selected, and to confirm curated cases met the study selection criteria. In total, 2045 cases represented by 5849 slides, including 4331 hematoxylin and eosin slides and 1518 special stained slides (and 5849 corresponding WSIs) were verified and entered into this study (Table 1). The average number of slides per case was 2.86 with a distribution of 1115 cases with 1 slide (median = 1) and 930 with 2 or more, including 107 cases having 10 or more slides and the largest having 51 slides.

Case Slide Scanning

Slides were relabeled with randomized identification barcodes and scanned at the same site using the site's Aperio AT2 DX system at ×20 magnification. A trained laboratory technician(s) ensured that slides were free of cracks, bubbles, and artifacts, and were clean and devoid of hand annotations ("Sharpie" dots, circles, or writing) using ethanol wipes and tissue paper as needed. The technician verified, and quality checked scanned images, and rescanned slides if the images did not pass the quality check or if the reading pathologist requested a rescan. Scanned slides were organized into the appropriate case files using slide management software.

Case Reading

Each case was evaluated twice (once by WSI and once by light microscopy) by all site reading pathologists (3–4 reading pathologists per site), with a minimum required washout period between modality reads of 31 days (actual average of 58 days). After deferrals, 15 031 diagnoses were entered, corresponding to a median 7 reads per case (median 4 WSI reads and 4 glass microscope slide reads per case). Reading pathologists were asked to provide a complete diagnosis from the available slides per case, and for cancer resection specimens, they were also asked to complete a synoptic report (College of American Pathologists synoptic reporting templates from the year/date of the original reference diagnosis were provided). For each reading pathologist, the case order and assignment of either glass slide or WSI for first review of each case was randomized. Reading pathologist diagnoses were time and date stamped, permitting an estimated time per case analysis, and ensuring that the second modality read was performed after the minimum 31-day washout. De-identified report data except for the reference diagnosis were available to each reading pathologist during case review. Reading pathologists were masked to the diagnoses made by others and to their own first modality diagnosis. Reading pathologists could request slide rescanning and could request a ×40 objective scan on any WSI. Following each sites' standard of practice, reading pathologists could defer cases if a diagnosis could not be made based on available information. Reasons for case deferral were recorded and analyzed.

Diagnosis Adjudication

Because primary diagnoses are qualitative in nature, to compare WSI and light microscopy for diagnostic accuracy, the study team, including the study director and the curating pathologists and excluding the reading and adjudicating pathologists, created a "lookup" table of synonymous terms and predetermined major and minor discordances (see supplemental digital content at www.archivesofpathology.org in the October 2020 table of contents). In this way, the determination of concordance by the adjudicating pathologists could be assessed as objectively as possible. Following reads, a separate panel of expert case adjudicators reviewed each reading pathologist's diagnosis from WSI or microscopy and compared it with that case's reference diagnosis, as described below. The adjudicators were able to refer to the lookup table as an aid to determine concordance.

Two of 3 adjudicating pathologists, always from different institutions or sites than the reading pathologists, were randomly assigned to compare each reading pathologist's diagnosis with the reference diagnosis in accordance with predefined rules. Each adjudicator was masked to other adjudications, as well as study modality of each diagnosis. Adjudicators scored each reading pathologist's diagnosis compared with the reference diagnosis with one of the following four possible adjudicator scores: concordant, minor discrepancy (defined as different diagnoses but with no impact on patient management) and major discrepancy (defined as different diagnoses associated with different patient management), or if the adjudicator could not reach a determination, the adjudicator could defer the case.

If the 2 initial adjudicators' scores for a reading pathologist's diagnosis agreed, with determinations of either "no major discrepancy" (concordant and/or minor discrepancy) or major discrepancy, a consensus determination was reached (Figure 2). If the first 2 adjudicator's scores disagreed, the reading pathologist's diagnosis was reviewed by a third adjudicator who was masked to the first 2 scores and to whether their determination was an initial review or a "tie-breaker" review. If consensus was still not reached due to at least one "defer" determination, or if the consensus was for deferral, the 3 adjudicators met as a panel, blinded to all previous adjudicator's scores, including their own, to reach a

Table 1. Distribution of Cases by Organ Type

Organ	Description	Site Enrolled					
		IMH	PRP	UCD	Dignity	TriCore	Total
Breast	Benign/atypical CNB	10	14	19	9	5	**57**
	Benign/atypical lumpectomy	10	12	4	9	9	**44**
	In situ carcinoma CNB	10	10	15	7	9	**51**
	In situ carcinoma lumpectomy	10	16	10	14	3	**53**
	Invasive carcinoma CNB	10	14	15	11	3	**53**
	Invasive carcinoma lumpectomy	10	13	10	7	6	**46**
Prostate	Benign core Bx	24	33	37	5	27	**126**
	Benign resection	6	8	5	12	4	**35**
	Adenocarcinoma Bx	24	25	24	11	32	**116**
	Adenocarcinoma resection	6	11	0	6	0	**23**
Lung/bronchus/larynx/ oral cavity/nasopharynx	Benign/Inflammatory Bx only	5	6	3	7	0	**21**
	Dysplasia Bx only	5	5	1	3	6	**20**
	Carcinoma Bx	6	7	6	7	1	**27**
	Carcinoma resection	6	2	2	6	17	**33**
Colorectal	Benign/inflammatory Bx	10	14	12	4	14	**54**
	Adenomas including severe dysplasia Bx	10	13	13	9	6	**51**
	Adenocarcinoma endoscopic Bx	8	16	9	9	0	**42**
	Adenocarcinoma resection	2	5	3	3	0	**13**
GE junction	R/O Barrett/dysplasia Bx	10	13	15	17	1	**56**
	Nonneoplastic/inflammatory Bx	10	9	9	7	9	**44**
Stomach	Inflammatory including R/O *H. pylori* Bx	10	13	17	7	3	**50**
	Polyps/neoplastic Bx	7	13	7	11	0	**38**
	Polyps/neoplastic resection	3	5	3	2	0	**13**
Skin	Nonneoplastic/inflammatory Bx	9	16	11	10	6	**52**
	Squamous/basal cell neoplasms Bx	10	11	14	13	17	**65**
	Melanocytic lesions Bx	16	17	14	17	0	**64**
Lymph node	For presence/absence of metastasis	15	14	14	9	23	**75**
	Nonneoplastic	5	6	4	3	8	**26**
Bladder	Benign/inflammatory/Nonneoplastic Bx	5	7	6	3	1	**22**
	Dysplasia Bx	5	4	4	0	5	**18**
	Noninvasive carcinoma (TUR or Bx)	5	7	8	1	0	**21**
	Carcinoma TUR/Bx	3	7	4	11	0	**25**
	Carcinoma resection	2	3	7	2	0	**14**
Gynecological	Endometrial Bx/curetting	8	14	16	9	0	**47**
	Hysterectomy for endometrial or cervical cancer	2	4	10	2	0	**18**
	Cervix Bx/curetting (Bx, ECC)	5	7	10	4	0	**26**
	Cervix Bx/curetting (cone/LEEP)	5	8	5	11	0	**29**
	Ovary benign/nonneoplastic	4	7	7	3	0	**21**
	Ovary neoplastic	6	6	1	8	5	**26**
Liver/BD, neoplasm	Core Bx	8	16	9	8	0	**41**
	Wedge Bx or resection	2	2	3	1	2	**10**
Endocrine	Pancreas	10	12	10	11	7	**50**
	Thyroid	6	5	9	6	4	**30**
	Parathyroid	2	5	3	3	0	**13**
	Adrenal	2	2	2	2	1	**9**
Brain/neuro	Nonneoplastic	2	3	0	2	0	**7**
	Neoplastic Bx	5	9	2	5	2	**23**
	Neoplastic resection	5	7	16	3	0	**31**
Kidney, neoplastic	All comers (consecutive cases)	10	5	13	12	7	**47**
Salivary gland	All comers (consecutive cases)	10	14	12	11	2	**49**
Hernial/peritoneal	All comers (consecutive cases)	2	1	3	2	0	**8**
Gallbladder	All comers (consecutive cases)	2	4	3	4	0	**13**
Appendix	All comers (consecutive cases)	2	5	4	2	0	**13**
Soft tissue tumors	All comers (consecutive cases)	4	5	24	4	1	**38**
Anus/perianal	Bx	10	4	13	7	14	**48**
Total		**399**	**514**	**500**	**372**	**260**	**2045**

Abbreviations: BD, bile duct; Bx, biopsy; CNB, core needle biopsy; Dignity, Dignity Health, Carmichael, California; ECC, endocervical curettage; GE, gastroesophageal; IMH, InterMountain Health, Salt Lake City, Utah; LEEP, loop electrosurgical excision procedure; PRP, Pacific Rim Pathology, San Diego, California; R/O, rule out; TriCore, TriCore Reference Lab, Albuquerque, New Mexico; TUR, transurethral resection; UCD, University of California, Davis Health, Sacramento, California.

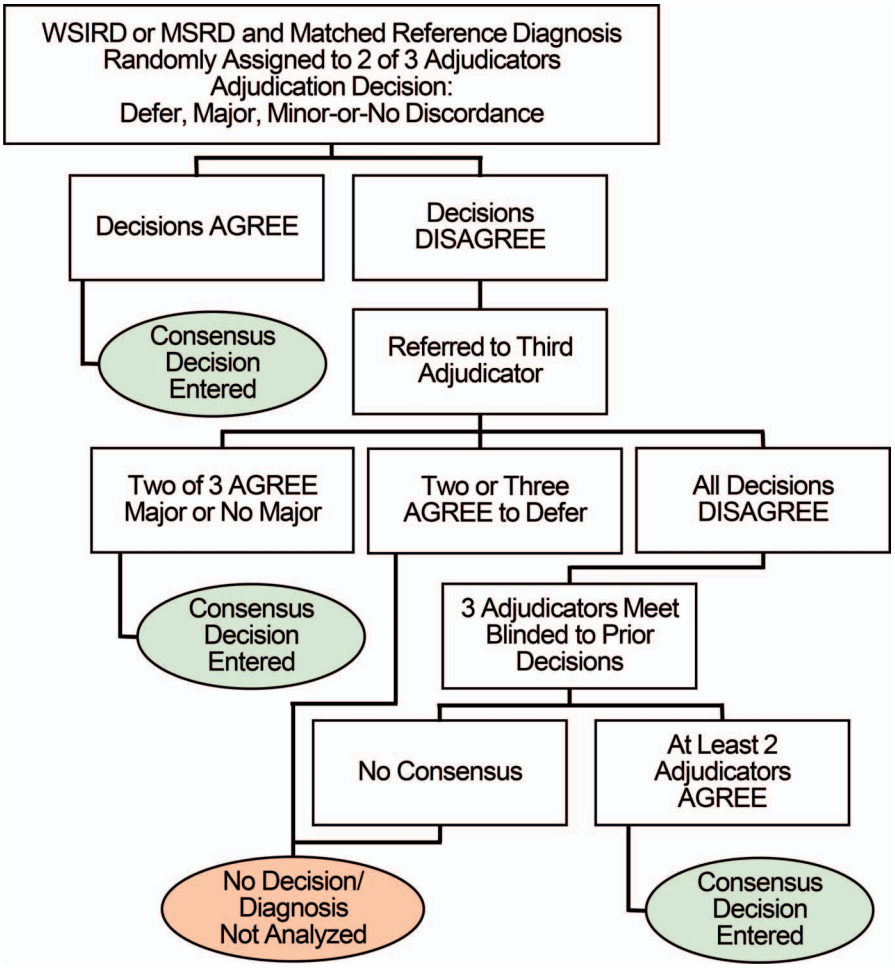

Figure 2. *Adjudication procedure. Reading pathologist's diagnosis and matched reference diagnosis were randomly assigned to 2 of the 3 adjudication pathologists for comparison. Logic for additional review is shown. Note: Only 17 of 15 507 diagnoses were referred to a consensus meeting and only 2 were excluded. Abbreviations: MSRD, light microscopy slide review diagnosis; WSIRD, whole slide image review diagnosis.*

consensus on the case. If they could not, no determination was entered, and the diagnosis was not used in the analysis.

Statistical Methods

A generalized linear mixed-effect model was built to account for the fixed effects presented in the study, such as modality and organs, and random effects, such as pathologists and sites. A modeling-based approach considered the appropriate variance–covariance matrix enabling the proper estimation of the major discrepancy rates for each modality, the difference between the major discrepancy rates between modalities, and their standard errors.

To appropriately estimate modality differences, the significance of the interaction effect between modality and organ type was first tested using Model 1 as follows: $\text{logit}(Y) = \mu + \beta y + \delta + \beta\delta + \alpha + \gamma + \rho + \beta\rho + \varepsilon$. The observed binary variable, Y, had a binomial distribution. $Y = 0$ when WSI review (WSIR) diagnosis or light microscopy slide review (MSR) diagnosis had no major discrepancy with the reference diagnosis, and $Y = 1$ when they had a major discrepancy. β represented the marginal modality effect. There were 2 modalities, WSIR diagnosis and MSR diagnosis. Modality was a fixed effect. δ was the organ fixed effect. $\beta\delta$ was the modality-by-organ interaction, a fixed effect as well. α was the effect of cases within each organ. Cases within organ effects were considered to be random and normally distributed as $\alpha \sim N(0, \sigma_\alpha^2)$. γ was a site effect, a random effect and normally distributed as $N(0, \sigma_\gamma^2)$. ρ were pathologists within each site, a random effect and normally distributed as $N(0, \sigma_\rho^2)$. $\beta\rho$ were pathologist within site by modality interaction, a random effect. ε was the experimental error.

Because interaction between modality and organ was not significant, Model 2 was used for final estimations of the modality effects and their standard errors as $\text{logit}(Y) = \mu + \beta y + \delta + \alpha + \gamma + \rho + \beta\rho + \varepsilon$. Parameter definitions and related assumptions were the same as in Model 1.

The Generalized Linear Mixed Model (GLIMMIX) package from SAS 9.4 was used to fit the model. Estimates of two-sided 95% CI of each modality overall major discrepancy rate as well as the difference in overall major discrepancy rates (WSIR diagnosis major discrepancy rate − MSR diagnosis major discrepancy rates) were derived from this analysis. Observed estimates of the major discrepancy rates of each modality, calculated as a proportion of the major discrepancy outcomes to the total number of evaluated slides, were also calculated and reported.

The following acceptance criteria were predefined:

1. The performance of WSIR would be considered noninferior to MSR if the upper bound of the two-sided 95% CI of the difference between the overall major discrepancy rates of WSIR diagnoses and MSR diagnosis was 4% or less.
2. The upper bound of the two-sided 95% CI of the overall major discrepancy rate of the WSIR diagnoses (relative to the reference diagnosis) was 7% or less.

RESULTS

Overall Major Discrepancy Rates

The overall major discrepancy rates estimated by the logistic regression model were 3.64% and 3.20% for WSIR and MSR diagnosis methods, respectively, a difference of 0.44% (95% CI, –0.15 to 1.03%; Table 2). WSIR is noninferior to MSR because this upper limit of the 95% CI

Table 2. Overall Major Discrepancy Rates for the Two Modalities and the Difference Between the Overall Major Discrepancy Rates

	WSIRD				MSRD				Difference in Major Discrep Rates (WSIRD – MSRD)	
	Major Discrep	Total	Major Discrep Rate, %	Model, 95% CI	Major Discrep	Total	Major Discrep Rate, %	Model, 95% CI	%	Model, 95% CI
Observed	280	7509	3.73	-	247	7522	3.28	-	0.45	-
Model			3.64	(3.21, 4.12)			3.20	(2.80, 3.65)	0.44	(−0.15, 1.03)

Abbreviations: Discrep, discrepancy; MSRD, manual slide review diagnosis; WSIRD, whole slide image review diagnosis.

at 1.03% was less than the predetermined noninferiority threshold of 4%.

Major Discrepancy Rates by Organ Site

The highest major discrepancy rates were from urinary bladder biopsies (9.47% by MSR and 10.4% by WSIR; Table 3), followed by lung, endocrine, soft tissue, and breast, with rates ranging from 3.5% to 5.25%. Overall, differences in major discrepancy rates between modalities by organ site were small, ranging from −1.14% (salivary gland) to +2% (skin). Review of the modality-dependent major discrepancy diagnoses and rates did not identify an organ site or biopsy type for which WSIR presented diagnostic challenges not present in MSR.

Unbalanced Diagnoses

"Unbalanced" case diagnoses in which one modality diagnosis was a major discrepancy and the other modality diagnosis was concordant with the reference diagnosis, following diagnosis adjudications, were evaluated. The rate of WSIR discordance with MSR concordance across tissue/organ type ranged from 0.0% (0 unbalanced diagnoses of 306 cases) for kidney to 4.9% (15 of 306) for breast. Similarly, the rate of MSR discordance with WSIR concordance ranged from 0.0% (0 of 302) for skin to 4.3% (13 of 302) for breast. Thus, unbalanced diagnoses represented a small percentage of the total number of cases, and they occurred with similar frequency for both modalities. When the specific diagnoses were examined by tissue/organ type, there were no apparent trends in any tissue/organ type that suggested WSIR was more prone to major discrepancies compared with MSR. Consistent with these observations, the modality-by-organ interaction was not statistically significant (see Table 3).

Table 3. Major Discrepancy Rates by Organ

Organ Type	Major Discrepancy Rate, %		Difference in Major Discrepancy Rates (WSIRD – MSRD), %
	MSRD	WSIRD	
Anus/perianal	2.79	3.95	1.16
Appendix	0.00	0.00	0.00
Bladder	9.47	10.40	0.93
Brain/neuro	2.54	3.09	0.55
Breast	3.53	4.29	0.76
Colorectal	2.46	2.46	0.00
Endocrine	4.57	4.04	−0.53
Gastroesophageal junction	3.16	3.69	0.54
Gallbladder	0.00	0.00	0.00
Gynecological	3.18	4.28	1.10
Hernia/peritoneal	0.00	0.00	0.00
Kidney	1.69	1.14	−0.56
Liver/bile duct	0.53	1.59	1.06
Lung	3.68	5.24	1.55
Lymph node	1.87	1.09	−0.78
Prostate	3.44	3.00	−0.44
Salivary gland	1.69	0.55	−1.14
Skin	2.72	4.74	2.03
Soft tissue	4.83	4.23	−0.60
Stomach	2.09	3.15	1.06

Abbreviations: MSRD, manual slide review diagnosis; WSIRD, whole slide image review diagnosis.

When we examined unbalanced diagnoses by reading pathologist, none of our 19 reading pathologists showed an increasing tendency for major discrepancies by WSIR as compared with MSR. Among the 2045 cases evaluated, there were no cases in which three or more reading pathologists had a diagnosis with a major discrepancy by one modality but had a concordant diagnosis by the other modality. There were 4 cases (4 of 2045, 0.2%; 1 case each of liver, gastroesophageal junction, colorectum, and bladder) in which 2 of 4 reading pathologists at the same site had a WSIR diagnosis with a major discrepancy but a concordant diagnosis by MSR. Similarly, there were 3 cases (3 of 2045, 0.15%; 1 case each of prostate, breast, and bladder) in which 2 of 4 reading pathologists at the same site had an MSR diagnosis with a major discrepancy but a concordant diagnosis by WSIR. Thus, no tissue/organ types were disproportionately prone to discordance by WSIR as compared with MSR, and among our cohort of reading pathologists, major discrepancies were no more likely to occur with WSIR as compared with MSR.

Instrument Performance

Of 5849 slides processed, 99.2% (5801 of 5849) were successfully scanned with the Aperio AT2 DX System on the first attempt; system-detected errors were generated for 48 slides that were not initially scanned. Errors occurred with similar frequency across all sites; 47 errors were due to prescanning operation issues, such as slides not properly prepared, and 1 system error (mechanical motion control error due to slide not loaded in proper position). After the issues were corrected, all 48 slides (100%) were successfully rescanned.

Among the 5849 slides initially scanned, 39 slides required rescanning due to issues identified during the WSI quality check for tissue cropping, focus, image segment defects, and stripes/stitching artifacts; the WSIs from these 39 slides were not provided to the reading pathologists. Of these, 20 were attributed to slide quality, 16 to digital image quality, and 3 to human error (eg, manual over cropping). These 39 slides were rescanned once, and all images from the rescan were

deemed acceptable by the scanning technician, and then provided to the reading pathologists. In addition, 3 cases (1 slide per case) were rescanned due to reading pathologists' request to scan at a higher magnification (×40) and were provided to the requesting reading pathologist only. One of 3 cases was a brain biopsy. Two of 3 were gastric biopsies, and reasons for ×40 scan were related to evaluation of *Helicobacter pylori*. All (100%; 3 of 3) rescans were deemed acceptable by the reading pathologists.

Read Times

Case reading time was defined as the time a pathologist completed a case diagnosis to the completion of the next case. Case reading times included WSIR or MSR, reading of all available clinical information, completion of cancer protocols, diagnosing the case, and recording the diagnosis in the data capture system.

Reading times longer than 30 minutes were considered to not reflect the actual reading time because such instances generally occurred due to external factors distracting a reading pathologist from the task and were thus excluded from the reading time calculations. Overall, 91.8% (7118 of 7756) of MSR reads and 92.3% (7156 of 7751) of WSIR reads were completed in less than 30 minutes.

Case reading times for WSIR and MSR diagnoses were similar. The mean reading time for MSR diagnosis was 4.95 minutes and the mean reading time for WSIR diagnosis was 5.20 minutes. The difference between the mean reading times for WSIR diagnosis and MSR diagnosis was 0.25 minutes (15 seconds).

Case Deferrals

There were 357 cases deferred by reading pathologists among the 15 562 total diagnoses; 271 of these were among the 7781 WSI case reads and 258 were among the 7781 MS, with 172 cases deferred using both modalities. The latter number included 106 skin biopsies (106 of 172; 62%) with a deferral reason indicating that these would normally be read by a dermatopathologist in that practice setting. Similarly, 27 brain cases were deferred due to consultation by neuropathologists in that site's standard practice. Additional reasons for a reading pathologist to defer in both modalities included need for consultation (30 cases) and need for additional IHC (9 cases). Of the remaining 99 cases deferred by WSIR only and 86 cases deferred by MSR only, reasons for deferral were similar, with the majority (75.6% [75 of 99] for WSIR and 73.7% [73 of 99] for MSR) related to request for a specialty consultation. Additional single-modality deferral reasons included additional IHC needed (14 WSI and 13 MSR), and additional clinical data or the results of a prior biopsy needed (10 WSI and 9 MSR). In one case, a slide with a special stain (Congo-red), although negative for congophilic material, was felt by some reading pathologists to require a polarized light examination (unavailable in the study) to be fully evaluated and was therefore excluded for all reading pathologists. Overall, no systematic differences were apparent for deferrals made by WSIR versus MSR.

Within-Observer Agreement Between the WSIR and MSR Diagnoses

Comparing consensus scores from WSIR with MSR, 96.1% (7137 of 7423 diagnoses) agreed between the 2 modalities (Table 4). Although the primary predetermined endpoint included comparison only to the reference (original pathology report) diagnosis, diagnoses adjudicated as discordant compared with the reference diagnosis by one or both modalities for an individual reading pathologist were further compared to determine individual reading pathologist concordance. As a result, 401 of 2045 pairs of WSIR/MSR matched cases had at least one reading pathologist's diagnosis adjudicated as discordant. The majority of these (254 of 401), were read the same way by an individual reading pathologist (within-observer concordance with interobserver discordance) but 147 of 401 showed individual reading pathologist differences. These were statistically equal by modality with 83 of 7781 (1%) by WSIR and 64 of 7781 (0.9%) by MSR. While there was some variation between individual reading pathologists' discordance rates, no individual reading pathologist had a preponderance of discordances in one modality over the other (WSIR versus MSR).

Table 4. Comparison of Adjudication Outcomes From Both Modalities[a]

	MSRD		
	Major Discrepancy	No Major Discrepancy	Total
WSI			
Major discrepancy	**109**	161	270
No major discrepancy	125	**7028**	7153
Total	234	7189	**7423**

Abbreviations: WSI, whole slide images; MSRD, manual slide review diagnosis.

[a] Bolded values represent the primary agreements/comparators.

DISCUSSION

This study demonstrated that clinical diagnoses made by pathologists via WSIR using the Leica Biosystems Aperio AT2 DX system are not inferior to the traditional MSR method for a large collection of pathology cases with diverse tissues/organs and sample types. This study was designed based on parameters established by previous research,[6,13] College of American Pathologists recommendations for a study comparing WSIR with MSR,[11] and US Food and Drug Administration recommendations.[12] Critical design parameters included assessment of more than 2000 cases; admission of consecutive clinical cases with limited exclusion criteria to minimize case selection bias; overrepresentation of a predefined set of tissue site and diagnostic categories along a morphologic spectrum (eg, dysplasia, atypia, and carcinoma in situ); multiple study sites, including an academic hospital and community and reference practices in 3 different states; multiple reading pathologists per site, with all cases read in both modalities by site reading pathologists; and at least a 31-day washout minimum between modality reads for each reading pathologist. In addition, health authority and professional association input helped establish the study's predetermined endpoints.[11,12] As the primary recommended endpoint, the upper bound of two-sided 95% CI of the difference between overall major discrepancy rates of WSIR diagnoses compared with MSR diagnoses should be 4% or less; and a secondary endpoint, the upper bound of the two-sided 95% CI of the overall major discrepancy rate for WSIR diagnosis (compared with reference diagnosis) should be 7% or less. The current study met both endpoints, with a difference between overall major discrepancy rates of 0.44% (upper bound 95% CI = 1.03);

and the overall major discrepancy rate for WSIR diagnosis being 3.73% (upper bound 95% CI = 4.12). Overall, our data are similar to a study conducted on the Phillips IntelliSite Pathology Solution platform, which showed an identical upper bound difference of 1% with a slightly higher discrepancy rate of 4.9%.[13] One possible reason for the higher WSIR diagnosis discrepancy in that study might have been the mandatory requirement for the participating pathologists to provide a diagnosis on every case without the option to defer a case that would normally be sent for additional consultation, though 11 reads in that study were given "no diagnosis." Our study allowed pathologists to simulate their practice environments by deferring cases on which they would normally consult a colleague, although the fraction of cases deferred was less than 3.5% for both modalities.

It is useful to consider differences between glass slides viewed with a microscope and WSIs viewed on flat digital monitors. First, all pathologists working today trained and have practiced using light microscopy; this is the format they are most familiar with and likely the most comfortable. For this reason alone, use of WSIs might initially be associated with lower confidence in diagnosis. We anticipate that proper training and sufficient practice using WSI as it is widely adopted should mitigate this concern in the same way that smartphone-based maps have largely replaced paper maps. Second, despite tissue sections mounted on slides being very thin (approximately 4–5 μm), appreciation of some morphologic features important for diagnosis, such as chromatin pattern, require focusing the microscope though the entire thickness of the tissue section. However, WSI typically—and in the system tested here—captures a single focal plane. The AT2 DX system used does have a "Z-stack" feature, though it was not used by the study pathologists. Finally, it is possible that optical artifacts or digital sampling errors might result in image degradation and compromised interpretation. Note, however, that reading pathologists in this study could defer any case, glass slide or WSI, they determined to be inadequate for diagnosis. In addition to concordance of each modality's diagnoses, we also showed there was a very little difference in case deferral rates for each modality.

The terminology used by pathologists to make primary diagnoses are both categoric and descriptive/qualitative in nature. Thus, experts are required to determine whether pairs of diagnoses are concordant or discordant, and the consequences of any discordance on clinical management. The subjectivity of this process was reduced in our study through the creation and use of a consensus "lookup" table of synonymous terms and predetermined major and minor discordances, and by adjudication design using at least 2 independent determinations in agreement. While the table was originally intended to serve as a basis of automated (computational) concordance determination, we found that the number and diversity of terms reading pathologists used in their diagnostic practice rendered the lookup table impractical as a stand-alone method of concordance determination. Nevertheless, construction of the lookup table was useful because the adjudicators referred to it when judging diagnostic concordance. Clearly, adjudication decisions for some diagnoses are difficult, especially in known "gray areas" where diagnostic criteria are controversial and known to be a source of high-interobserver variability. For example, diagnosis of breast atypia, low- versus high-grade urothelial lesions, and different Gleason grade patterns of prostate cancer have relatively high rates of interobserver variability.[16–22] This observation was also revealed in a review of reading pathologist discordances by organ or biopsy type, where such "gray area" diagnoses were sometimes adjudicated as concordant and sometimes as discordant. Although a previous WSI validation study reported that adjudication by a subject expert determined some of the WSI interpretations were actually "better" than the original diagnosis,[6] such a comparison to a "gold standard" diagnosis by a subspecialty trained expert was not performed in the present study. In other words, our study did not attempt to determine the most "correct" diagnosis of each case, as no re-review of cases by adjudicators or consensus panels was performed.

In this study, all cases were selected based upon the original reference diagnosis; thus, all cases were originally considered adequate for diagnosis by light microscopy. To ensure appropriate cases were entered into the study, candidate cases were selected and reviewed by the primary curation pathologist at each site, then checked for adequacy by a secondary curation pathologist. In contrast to routine practice, the study design did not allow the reading pathologist to obtain additional slides (re-cuts or special stains), nor did it permit consultation with colleagues or experts. Instead, reading pathologists were instructed to reject or "defer" any case that was not adequate for diagnosis and to defer any case for which they would normally obtain a specialist's opinion. An exception was made for reading pathologists who would normally obtain a review/concurrence from a colleague as a part of practice protocol. For example, many practices employ a review/concur policy for all initial cancer diagnoses. For our study, such policies were waived such that reading pathologists were asked to provide their diagnosis unless they needed a second opinion to do so. Another exception was created for WSIs that for whatever reason, in the reading pathologists' judgement, did not display properly on the viewing monitor. Here, a request for rescan or higher resolution (×40) rescan was allowed. Importantly, only three cases (3 slides) were requested to be rescanned ×40 for higher resolution. Thus, the Aperio system using a 0.75 numeric aperture ×20 objective lens was shown here to be noninferior to a standard light microscope for primary diagnosis in surgical pathology, even though almost all such microscopes in the study were outfitted with a turret mounted ×40 objective. An exception may be the situation of *H. pylori* evaluation when performed on hematoxylin and eosin only (without special stains). No ×20 rescans were requested by the reading pathologists for image quality, confirming the overall high quality of the WSI case sets.

Some diagnoses with major discrepancies are well known to be intrinsically challenging, with known high-interobserver variability, such as breast atypia compared with nonatypical hyperplasia and/or low-grade ductal carcinoma in situ or urinary bladder biopsy for noninvasive lesions.[16–18] Such cases were overrepresented in our case sets compared with frequencies in routine practice. Regardless of diagnostic modality, the highest major discrepancy rates in our study were from tissue sites known to have relatively high-interobserver variability, including urinary bladder biopsies (9.47% by MSR and 10.4% by WSIR), followed by lung, endocrine, soft tissue, and breast, with rates ranging from 3.5% to 5.25%. The highest discrepancy rates in the Mukhopadhyay et al study[13] were similarly from organ systems known to have relatively high-interobserver vari-

ability including prostate (11.3% by microscopy and 12% by WSI), brain, gynecologic tract, liver/bile ducts, and urinary bladder (>5%).

Overall, our data show that pathologists who routinely perform MSR can make primary histologic diagnoses from WSIR with a level of accuracy that is noninferior to the current standard of practice, viewing glass slides by light microscopy. The time to review each case was similar between the two modalities, and the reading pathologists found the WSI modality to be comfortable and completely adequate, including review of special histochemical and IHC stains, with only rare exceptions (eg, need for polarized light examination). This study, along with one previous similar[13] and several smaller studies,[2,6,7,10,14] shows that interpretation of pathology images as single plane WSI provides diagnostic quality equal to the centuries-old method of viewing glass slide mounted tissue sections through a light microscope.

The authors thank Kim Oleson, Daniel Schrock, Kevin Liang, Mark Cohn, James Wells, Craig Fenstermaker, Lothar Wieczorek, Yang Bai, MSc, Sandeep Gill, Kelly Paine, Wilda McDonough, Michael Lung, Thom Jessen, Heather Sandoval, Tiffany Shaw, Neil E. Hubbard, PhD, Kevin Kneeland, Jill Funk, Ryan Chiang, Jennifer Welch-Doiron, Ahmed Reza Yousefi, Christine Kishi, Shilpa Saklani, and Estella Kanevsky, MS, for support during this study.

References

1. Cardiff RD, Gregg JP, Miller JW, Axelrod DE, Borowsky AD. Histopathology as a predictive biomarker: strengths and limitations. *J Nutr*. 2006;136(10):2673S–2675S.

2. Bauer TW, Slaw RJ. Validating whole-slide imaging for consultation diagnoses in surgical pathology. *Arch Pathol Lab Med*. 2014;138(11):1459–1465.

3. Jones NC, Nazarian RM, Duncan LM, et al. Interinstitutional whole slide imaging teleconsultation service development: assessment using internal training and clinical consultation cases. *Arch Pathol Lab Med*. 2015;139(5):627–635.

4. Malarkey DE, Willson GA, Willson CJ, et al. Utilizing whole slide images for pathology peer review and working groups. *Toxicol Pathol*. 2015;43(8):1149–1157.

5. Pantanowitz L, Szymas J, Yagi Y, Wilbur D. Whole slide imaging for educational purposes. *J Pathol Inform*. 2012;3:46.

6. Bauer TW, Schoenfield L, Slaw RJ, Yerian L, Sun Z, Henricks WH. Validation of whole slide imaging for primary diagnosis in surgical pathology. *Arch Pathol Lab Med*. 2013;137(4):518–524.

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

8. Evans AJ, Chetty R, Clarke BA, et al. Primary frozen section diagnosis by robotic microscopy and virtual slide telepathology: the University Health Network experience. *Hum Pathol*. 2009;40(8):1070–1081.

9. Molnar B, Berczi L, Diczhazy C, et al. Digital slide and virtual microscopy based routine and telepathology evaluation of routine gastrointestinal biopsy specimens. *J Clin Pathol*. 2003;56(6):433–438.

10. Snead DR, Tsang YW, Meskiri A, et al. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology*. 2016;68(7):1063–1072.

11. Pantanowitz L, Sinard JH, Henricks WH, et al. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med*. 2013;137(12):1710–1722.

12. Technical performance assessment of digital pathology whole slide imaging devices; guidance for industry and food and drug administration staff. FDA Web site. https://www.fda.gov/media/90791/download. Accessed January 24, 2020.

13. 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(1):39–52.

14. Tabata K, Mori I, Sasaki T, et al. Whole-slide imaging at primary pathological diagnosis: validation of whole-slide imaging-based primary pathological diagnosis at twelve Japanese academic institutes. *Pathol Int*. 2017;67(11):547–554.

15. Saco A, Ramirez J, Rakislova N, Mira A, Ordi J. Validation of whole-slide imaging for histolopathogical diagnosis: current state. *Pathobiology*. 2016;83(2-3):89–98.

16. Elmore JG, Nelson HD, Pepe MS, et al. Variability in pathologists' interpretations of individual breast biopsy slides: a population perspective. *Ann Intern Med*. 2016;164(10):649–655.

17. Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*. 2015;313(11):1122–1132.

18. Lee EW, Deng FM, Melamed J, et al. Grading variability of urothelial carcinoma: experience from a single academic medical center. *Can J Urol*. 2014;21(4):7374–7478.

19. Kweldam CF, van Leenders GJ, van der Kwast T. Grading of prostate cancer: a work in progress. *Histopathology*. 2019;74(1):146–160.

20. Oyama T, Allsbrook WC Jr, Kurokawa K, et al. A comparison of interobserver reproducibility of Gleason grading of prostatic carcinoma in Japan and the United States. *Arch Pathol Lab Med*. 2005;129(8):1004–1010.

21. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol*. 2001;32(1):81–88.

22. Allsbrook WC Jr, Mangold KA, Johnson MH, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. *Hum Pathol*. 2001;32(1):74–80.