J Pathol Inform

Editor-in-Chief:
Anil V. Parwani , Columbus, OH, USA
Liron Pantanowitz, Pittsburgh, PA, USA

OPEN ACCESS
HTML format

For entire Editorial Board visit : www.jpathinformatics.org/editorialboard.asp

Original Article

A multisite validation of whole slide imaging for primary diagnosis using standardized data collection and analysis

Katy Wack[1,2], Laura Drogowski[2], Murray Treloar[3], Andrew Evans[4], Jonhan Ho[5], Anil Parwani[6], Michael C. Montalto[2,7]

[1]Western Oncolytics, LLC, Pittsburgh, PA 15238, [2]Work peformed while at Omnyx, LLC. Pittsburgh, PA 15222, USA, [3]Dynacare, Bowmanville, Ontario L1C 3K5, [4]University Health Network, Toronto General Hospital, Toronto, Ontario M5G 2C4, Canada, [5]Department of Dermatology, University of Pittsburgh, Pittsburgh, PA 15213, [6]The Ohio State University Wexner Medical Center, Columbus, OH 43210, [7]Department of Translational Medicine, Bristol-Myers Squibb, etc. Princeton, NJ 08543, USA

E-mail: *Dr. Katy Wack - katy.wack@westernoncolytics.com
*Corresponding author

Received: 06 September 2016 Accepted: 28 October 2016 Published: 29 November 2016

Abstract

Context: Text-based reporting and manual arbitration for whole slide imaging (WSI) validation studies are labor intensive and do not allow for consistent, scalable, and repeatable data collection or analysis. **Objective:** The objective of this study was to establish a method of data capture and analysis using standardized codified checklists and predetermined synoptic discordance tables and to use these methods in a pilot multisite validation study. **Methods and Study Design:** Fifteen case report form checklists were generated from the College of American Pathology cancer protocols. Prior to data collection, all hypothetical pairwise comparisons were generated, and a level of harm was determined for each possible discordance. Four sites with four pathologists each generated 264 independent reads of 33 cases. Preestablished discordance tables were applied to determine site by site and pooled accuracy, intrareader/intramodality, and interreader intramodality error rates. **Results:** Over 10,000 hypothetical pairwise comparisons were evaluated and assigned harm in discordance tables. The average difference in error rates between WSI and glass, as compared to ground truth, was 0.75% with a lower bound of 3.23% (95% confidence interval). Major discordances occurred on challenging cases, regardless of modality. The average inter-reader agreement across sites for glass was 76.5% (weighted kappa of 0.68) and for digital it was 79.1% (weighted kappa of 0.72). **Conclusion:** These results demonstrate the feasibility and utility of employing standardized synoptic checklists and predetermined discordance tables to gather consistent, comprehensive diagnostic data for WSI validation studies. This method of data capture and analysis can be applied in large-scale multisite WSI validations.

Key words: Digital pathology, discordance, validation, whole slide image

Access this article online

Website: www.jpathinformatics.org

DOI: 10.4103/2153-3539.194841

Quick Response Code:

INTRODUCTION

Large-scale, multisite, pivotal clinical trials will be required by the Food and Drug Administration (FDA) to test the clinical performance of whole slide imaging (WSI) devices for primary diagnostic use in the United States.[1] There are many aspects of validation designs that must be considered for a robust study, including tissue type,

This article may be cited as:
Wack K, Drogowski L, Treloar M, Evans A, Ho J, Parwani A, *et al.* A multisite validation of whole slide imaging for primary diagnosis using standardized data collection and analysis. J Pathol Inform 2016;7:49.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2016/7/1/49/194841

number of sites, number of readers, experience of readers, number of cases, complexity of cases, procedure type, period of washout, arbitration, and analysis methods. General guidance documents on study designs have been published, however the details in such documents are not standardized and all recommendations may not apply to large-scale studies such as those required for regulatory registrations.[2-4]

In a typical WSI validation study design, two or more diagnoses are provided per case by a single or multiple readers. For each case, one diagnosis is made on the reference method (i.e., light microscope) and the other on the test method (WSI). Since the main end point is the concordance between these two reads, it is critical to determine whether each paired diagnosis is the same or different. The most favored approach to determine differences is a manual arbitration method, in which a single or panel of pathologists/clinicians review each paired diagnosis for each case and determine whether the diagnoses are the same or different.[2,5-14] If differences are observed, the arbitrator or panel further determines whether these are considered major or minor errors. It is generally accepted that major errors are those that have a significant impact on patient management and minor errors do not.

While manual arbitrations in WSI validation studies are generally accepted, this method of data analysis is subjective and prone to variability. The main source of variability stems from definitions of "major error" or "significant change in patient management", which are subject to interpretation. Additional challenges with respect to the manual interpretation of discordance and its clinical significance can arise if study participants are able to use descriptive/nonstandardized diagnostic terminology. This would be particularly true in atypical/borderline cases if a different terminology was used between WSI and glass reads and the arbitrator was not clear on the meaning of a given descriptive diagnosis. Some studies have made strong attempts at a detailed definition of error, which is aimed at reducing variability in arbitration for general quality studies.[15,16] However, these methods are not often cited as a standard method in WSI validation studies. Further, interpretation of error definitions, regardless of how clear, can be still subjective and thus not uniformly applied throughout the entire study or across different studies. We and others at the University of Pittsburgh Medical Center have informally examined the process of arbitration and found that multiple arbitrators can have different opinions of what is considered a significant impact on patient management even with clear definitions (name removed for blinded review purposes, personal communication date April 30, 2016). Further, even if clear definitions are employed, it is possible that the definitions are applied differently on similar discrepancies over time, merely due to intra-arbitrator error. The limitations of subjective manual arbitration make it difficult to establish repeatable data from validation studies. Most importantly, this method of arbitration is difficult to scale for large multicenter studies, such as those mandated by the FDA for primary diagnosis. Such studies could have well over 100,000 pairwise comparisons when all primary and secondary analyses are considered.

To overcome the limitations of manual arbitration, we have developed a novel method for data capture and evaluation of pairwise comparisons which lends itself to standardization and scalability across the entire study population. This method can provide repeatable data collection and analysis regardless of the size or complexity of the study and could be used to directly compare results across sites and studies. To our knowledge, this is the first study to use standardized, codified College of American Pathologists (CAP) checklists and predetermined discordance tables to consistently capture and analyze diagnostic concordance data. We used this method in a multisite noninferiority pilot study to generate preliminary data of accuracy, precision, and reproducibility of WSI-based primary diagnosis.

METHODS AND STUDY DESIGN

Data Capture and Case Report form Checklists

Diagnostic reporting options included benign, atypical, and malignant categories, each with subcategorical options [Figure 1]. A fourth "no diagnosis" category was included with deferral categories. These included deferrals to subspecialist, to additional stains, if the tissue was nondiagnostic, if the histology was insufficient for diagnosis, and finally deferral to glass if the pathologist felt that the whole slide image was insufficient for diagnosis. Options for deferral to glass, subspecialist, and or additional stains were also included under each categorical diagnosis to allow the pathologist to categorize as far as they judged possible. A separate lymph node and margin evaluation reporting form was also codified to capture specific diagnostic classifications of case parts consisting of lymph nodes or margins only. For lymph node evaluations, categories include benign, reactive, lymphoproliferative, and metastatic cancer, each with subcategorical evaluations. For case parts that consist of margins, cancer and dysplasia are assessed and included measurement evaluations. In addition, a case report form was included for the scoring of special and immunohistochemistry stains and the request of rescans (for higher magnification or for image quality).

The validation study spanned seven organ groups, represented by 13 case report form checklists created from 9 cancer protocols [Table 1]. In addition, there

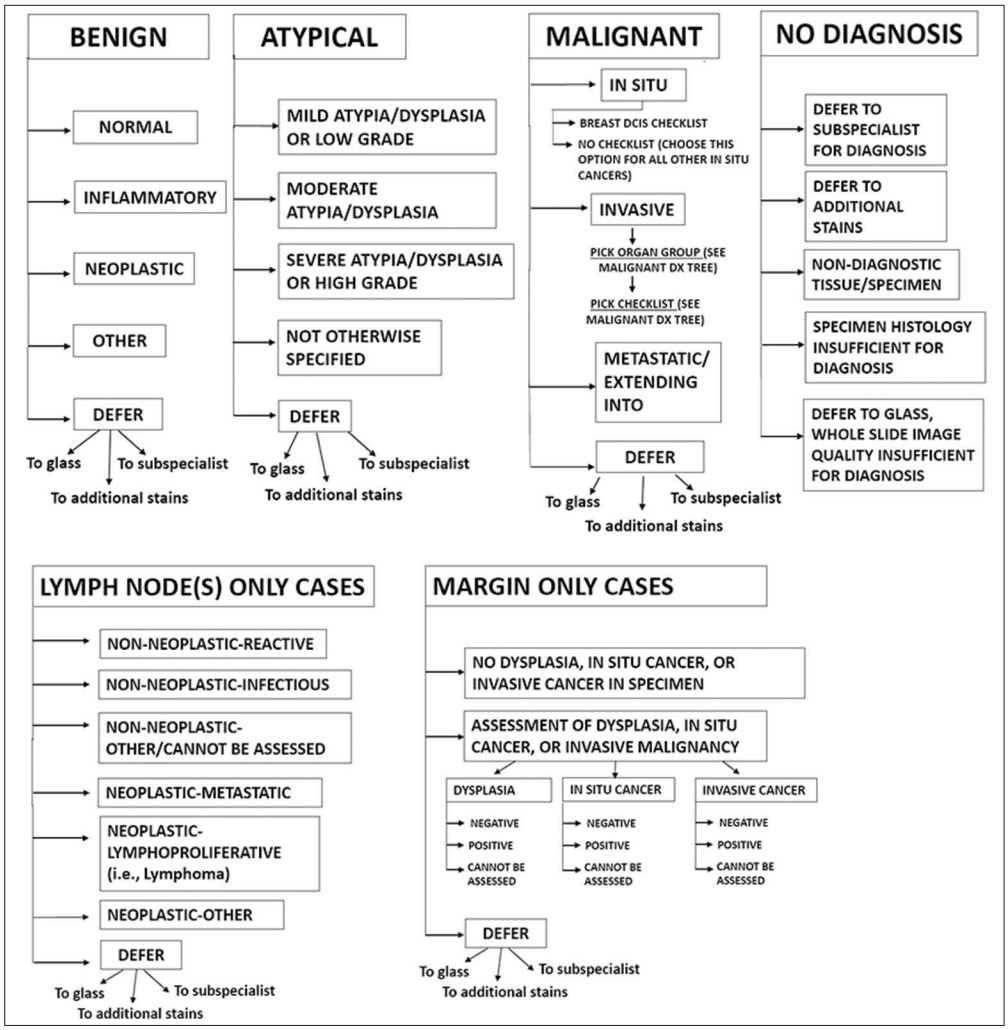

Figure 1: Flow diagrams of categorical diagnostic case report form. Readers were allowed to choose from one of the six high-level categorical diagnostic options. Each case report form included deferrals. A selection of "*in situ*" or "malignant" led the reader to additional detailed cancer case report form checklists based on the College of American Pathologists protocols

was one categorical case report form checklist used for all organs (i.e., benign, atypical, and malignant) and one lymph node checklist for a total of 15 case report form checklists. CAP protocols were modified to remove gross evaluation and codified to capture specific diagnostic information. Cancer protocols that have questions for both resection and biopsy sample types were separated into multiple forms such that sample types could be analyzed separately [Table 1]. The resulting case report form checklists were coded into a custom electronic data capture (EDC) system software for reporting. The study participants were first presented with the categorical and subcategorical forms [Figure 2a] and when selected, the appropriate CAP checklist was provided [Figure 2b]. Complete checklists can be found in the supplementary files. Specific edit checks were included in the reporting software to ensure all necessary questions to be answered before allowing the submission of a diagnosis. The EDC underwent extensive verification and validation prior to use.

Generation of Discordance Tables

Predetermined discordance tables were created for each case report form checklist. Discordance tables were generated by making pairwise comparisons between all possible discrete answers within and across on all diagnostic case report forms at each level of granularity. An expert general pathologist with 35 years of experience (including acting as Peer Assessor for Quality Assurance, Chief of Staff, and Director of Laboratory and Genetic Services for a large hospital system, and awarded several Distinguished Pathology Honors) reviewed each pairwise comparison and determined a level of harm based on predetermined definitions which included potential changes in patient management [Table 2].[15,16] Levels of harm were first assigned for each procedure type and diagnosis on the categorical/subcategorical level [Figure 3a] and then on the specific cancer checklist level [Figure 3b]. Complete discordance tables can be found in the supplementary files. The original sign-out diagnosis was designated as ground truth (GT) and each

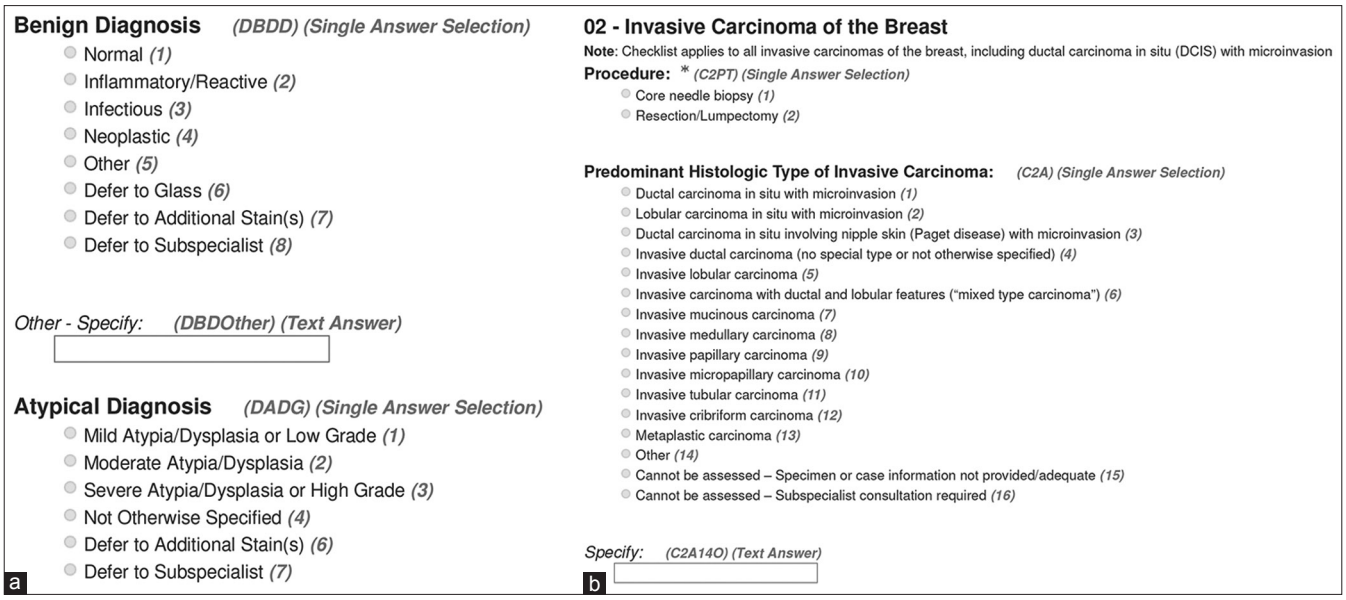

Figure 2: Standardized case report forms. (a) An example of a categorical case report form used for "benign" or "atypical" diagnoses. (b) An example of a modified College of American Pathologists cancer protocol case report form for invasive breast. Each field was given a unique variable identifier which is indicated in brackets

Table 1: The College of American Pathologists cancer protocols and resulting case report form checklists

Organ	CAP cancer protocol*	Codified study checklist**
Breast	DCIS-breast	Breast DCIS-biopsy
	Invasive breast	Invasive breast carcinoma-biopsy
Lung	Lung	Lung-biopsy
		Lung-resection
Colon	Colon and rectum	Colon-biopsy
		Colon-resection
	Colon net	Colon-neuroendocrine tumors (net)-biopsy
		Colon-net-resection
Skin	Squamous cell carcinoma	Squamous cell carcinoma-biopsy
	Melanoma	Melanoma-biopsy
Endometrium	Endometrium	Endometrium-resection
Uterine cervix	Uterine cervix	Uterine cervix-biopsy
		Uterine cervix-resection

*7th ed..ition; 2013, **Each CAP protocol was modified to separate biopsy and resections if appropriate. There was an additional 1 categorical checklist for all organ types and 1 for lymph node. Total of 15 case report form checklists. CAP: College of American Pathologists, DCIS: Ductal carcinoma *in situ*

possible hypothetical error was systematically entered for the test method.

Potential harm for discordances in biopsy cases was assessed separately from harm in resection cases as treatment plans differ. The pathologist who assigned harm consulted with clinicians or other pathologists for complex cases as needed, and assumptions and justifications were recorded to maximize consistency. There were 15 discordance tables created from each case report form checklist. A total of approximately 13,000 side-by-side comparisons were evaluated. Every discordance table underwent a review for logic and consistency in assignments and justifications by an experienced general pathologist, and subspecialists were consulted where appropriate. The most likely clinical outcome was always chosen for determining harm.

Site Selection and Training

Four independent clinical trial sites were chosen with four reading pathologists at each site for a total of 16 readers. The four sites comprised two reference laboratories (MIRACA Life Sciences, Dallas, TX; Tricore Life Sciences, Albuquerque, NM) and two academic hospital laboratories (SUNY, Syracuse, NY, and UCLA). Almost 50% of the reading pathologists had over 5 years of general pathology experience and 50% under 5 years of experience (with a range of 1 years to 27 years' experience), including a wide range of fellowship areas [Table 3]. A best attempt was given to have equal amounts of average experience as well as fellowship training. It was logistically not possible to have perfect equity in types of fellowship training. All pathologists underwent WSI software training (Omnyx DPS 1.3.0.84) and compared eight cases side by side using glass and WSIs. Training was designed for users who were not experienced in the field of digital pathology to "self-calibrate" between WSI and microscope reading. All readers were also trained on the protocol and the method of data collection using the coded case report form checklists in EDC. For training, archived samples of

Table 2: Definitions of harm for discordances

Severity	Level of harm	Definition
Minor	A	No harm
		Will not result in harm
		No change in prognosis or a change in prognosis that is unlikely to result in a change in treatment according to standards of care
	B	Minimal harm (Grade 1 [15])
		Further unnecessary noninvasive diagnostic test (s) performed (e.g., blood test or noninvasive radiologic examination)
		Delay in diagnosis or therapy of <6 months
		Minor morbidity due to (otherwise) unnecessary further diagnostic effort (s) or therapy predicated on the presence of (unjustified) diagnosis
Major	C	Moderate harm (Grade 2 [15])
		Further unnecessary invasive diagnostic test (s) (e.g., tissue biopsy, re-excision, angiogram, radionuclide study, or colonoscopy)
		Delay in diagnosis or therapy of >6 months
		Major morbidity lasting <6 months due to (otherwise) unnecessary further diagnostic efforts or therapy predicted on the presence of (unjustified) diagnosis
	D	Severe harm (Grade 3 [15])
		Loss of life, limb, or other body parts, or long-lasting morbidity (lasting >6 months)

Table 3: Pathologists' experience

Site*	Reader	Experience (years)	Average experience (SD), years	Fellowship training
1	1	1	13 (10.3)	GI
	2	20		Cytopathology
	3	26		None
	4	5		GI
2	1	27	14.4 (9.6)	General surgical
	2	2.5		Neurology
	3	20		Cytopathology
	4	8		Surgical pathology, cytopathology
3	1	23	15 (9.1)	Surgical pathology, cytopathology
	2	4		Cytopathology
	3	25		General surgical
	4	8		Gynecological
4	1	8	7.75 (7.5)	GI
	2	20		none
	3	2		Cytopathology
	4	1		GI, liver

*Sites 2, 4: Academic medical centers, Sites 1, 3: Reference laboratories. SD: Standard deviation, GI: Gastrointestinal

de-identified, signed-out cases (comprising 1 part each) were procured from SUNY-Syracuse University Hospital (Syracuse, NY) and used to train at all sites.

Sample Enrollment

A single site procured 36 cases, and 9 cases were distributed to each site [Table 4]. Power calculations and sample size were not predetermined because the intent of this study was to establish the methods for a large-scale multisite study, which will have preestablished sample size calculations. Single case parts from paraffin-embedded, formalin-fixed tissue samples from solid organs only were included in the study. Exclusion criteria included case parts from frozen preparation, immunofluorescence samples, case parts with more than one primary tumor, fluid-based specimens, and any case with slides that were damaged, missing, or failed a quality control inspection. The case distribution across diagnostic categories was 11% benign, 22% atypical, and 67% malignant. The case set was enriched for challenging and malignant cases [Table 4]. Challenging cases include small foci of micrometastases, borderline malignant melanoma, and small foci of atypical ductal hyperplasia (ADH) in breast cases. The cases covered the following organs: skin (melanoma), lung, colon, breast, endometrium, uterine cervix, lymph nodes, and thyroid. Malignant cases using codified CAP checklists and discordance tables analysis covered the following categories: melanoma, lung, endometrium, breast ductal carcinoma *in situ* (DCIS), invasive breast carcinoma, and uterine cervix carcinoma. Procedure types included biopsies, excisions, and resections. Cases of the same organ, procedure, and diagnosis category were procured for every site, with the exception of one site reading a rare lymphoma case in the lung instead of in the Hodgkin's lymphoma case of the cervical lymph node. Institutional Review Board approval was given for this study without the need for informed consent, and all applicable harmonized good clinical practice guidelines were followed.

Reading

Following training, each pathologist read nine cases in random order (glass or digital first), using both light microscopy and WSI, with a 2-week washout period between modalities. Discordances were assessed using the predetermined tables as described above. The original sign-out diagnosis for each case was used as GT and was coded using the case report form checklists in the

CATEGORICAL DIAGNOSIS	GT DESCRIPTION	TEST DESCRIPTION	CALCULATED DISCORDANT STATEMEN	DISCORDANCE	JUSTIFICATIONS/ASSUMPTIONS
BENIGN	NORMAL	MILD/LOW	MINOR	A	
BENIGN	INFLAMMATORY/ REACTIVE	MILD/LOW	MINOR	A	
BENIGN	INFECTIOUS	MILD/LOW	MINOR	A	
BENIGN	NEOPLASTIC	MILD/LOW	MINOR	A	
BENIGN	OTHER	MILD/LOW	MINOR	A	
ATYPICAL	MILD/ LOW	MILD/LOW	CONCORDANT	CONCORDANT	
ATYPICAL	MODERATE	MILD/LOW	MAJOR	C	Delayed diagnosis and treatment for moderate or severe atypia (equivalent to ADH)
ATYPICAL	SEVERE	MILD/LOW	MAJOR	C	Delayed diagnosis and treatment for moderate or severe atypia (equivalent to ADH)
ATYPICAL	NOS	MILD/LOW	MINOR	A	
MALIGNANT	IN SITU: NO MATRIX	MILD/LOW	MAJOR	C	Delay in diagnosis greater than 6 months and additional unnecessary invasive diagnostic test
MALIGNANT	IN SITU: MATRIX	MILD/LOW	MAJOR	C	Delay in diagnosis or treatment greater than 6 months
MALIGNANT	INVASIVE: NO MATRIX	MILD/LOW	MAJOR	C	Delay in diagnosis greater than 6 months
MALIGNANT	INVASIVE: MATRIX	MILD/LOW	MAJOR	C	Delay in diagnosis greater than 6 months
MALIGNANT	METASTATIC/ EXTENDING INTO	MILD/LOW	MAJOR	C	Delay in diagnosis or treatment greater than 6 months
BENIGN	NORMAL	MODERATE	MAJOR	C	Unnecessary excision
BENIGN	INFLAMMATORY/ REACTIVE	MODERATE	MAJOR	C	Unnecessary excision
BENIGN	INFECTIOUS	MODERATE	MAJOR	C	Unnecessary excision
BENIGN	NEOPLASTIC	MODERATE	MINOR	A	
BENIGN	OTHER	MODERATE	MINOR	B	Delay in diagnosis less than 6 months
ATYPICAL	MILD/ LOW	MODERATE	MAJOR	C	Unnecessary excision
ATYPICAL	MODERATE	MODERATE	CONCORDANT	CONCORDANT	
ATYPICAL	SEVERE	MODERATE	MINOR	A	
ATYPICAL	NOS	MODERATE	MINOR	B	Potential unnecessary treatment for prevention of invasive cancer (tamoxifen etc.)
MALIGNANT	IN SITU: NO MATRIX	MODERATE	MINOR	B	Excise both
MALIGNANT	IN SITU: MATRIX	MODERATE	MINOR	B	Excise both
MALIGNANT	INVASIVE: NO MATRIX	MODERATE	MAJOR	C	Unnecessary invasive 2nd procedure (lymph node biopsy) when the ground truth is known
MALIGNANT	INVASIVE: MATRIX	MODERATE	MAJOR	C	Unnecessary invasive 2nd procedure (lymph node biopsy) when the ground truth is known
MALIGNANT	METASTATIC/ EXTENDING INTO	MODERATE	MAJOR	C	Delay in diagnosis or treatment greater than 6 months
BENIGN	NORMAL	SEVERE	MAJOR	C	Unnecessary excision

a

02-INVASIVE BREAST-A1-PREDOMINANT HISTO TYPE

GT DESCRIPTION	TEST DESCRIPTION	CALCULATED DISCORDANCE STATEMENT	DISCORDANCE	JUSTIFICATIONS/ASSUMPTIONS
DUCTAL CARCINOMA IN SITU WITH MICROINVASION	INVASIVE LOBULAR CARCINOMA	MINOR	B	For DCIS or LCIS, axillary node is not sampled. However, here assume "Microinvasion" leads to node sampling, which is the same management as for Invasive carcinoma.
LOCULAR CARCINOMA IN SITU WITH MICROINVASION	INVASIVE LOBULAR CARCINOMA	MINOR	B	For DCIS or LCIS, axillary node is not sampled. However, here assume "Microinvasion" leads to node sampling, which is the same management as for Invasive carcinoma.
DUCTAL CARCINOMA IN SITU INVOLVING NIPPLE SKIN (PAGET DISEASE) WITH MICROINVASION	INVASIVE LOBULAR CARCINOMA	MAJOR	C	For DCIS or LCIS, axillary node is not sampled. However, here assume "Microinvasion" leads to node sampling, which is the same management as for Invasive carcinoma.
INVASIVE DUCTAL CARCINOMA (NO SPECIAL TYPE OR NOS)	INVASIVE LOBULAR CARCINOMA	MINOR	A	

b

Figure 3: Discordance table examples. (a) An example of a discordance table for "categorical" case report form checklist for breast. There were over 300 comparisons for this table. (b) An example of predominant histologic type comparisons in a discordance table for invasive breast cancer. There were over 700 comparisons for this table

same manner as the reviewers' diagnoses in the EDC. At each site, the four pathologists read the same nine cases (a maximum of 36 paired reads per site).

Analysis

Of 36 cases enrolled, three cases were removed for missing data due to software error, resulting in 33 total cases used for error analysis. If one reviewer had missing data, the entire case was removed from the analysis. The average difference in agreement between glass and digital was calculated by subtracting the average agreement (concordance or minor discordance, across four readers) for glass reads of a case compared to GT from the average agreement for digital reads of a case compared to GT. This difference was then averaged across all cases and 95% confidence intervals (CIs) were estimated (average difference ± 1.96^{x} standard error on the mean). The average agreement was calculated as the number of concordant plus minor discordant reads

Table 4: Cases enrolled and analyzed

Organ	Procedure	Diagnosis category	Original sign-out diagnosis	Site*
Lymph node (breast, sentinel)	Excision	Malignant	Small foci of metastasis	1, 2, 3, 4
Breast	Biopsy	Atypical	Small focus of ADH	1, 2, 3, 4
Breast	Biopsy	Malignant	Invasive ductal carcinoma	1, 2, 3, 4
Uterine cervix	Biopsy	Atypical	Cervical intraepithelial neoplasia 2, 3	1, 2, 3, 4
Endometrium	Biopsy	Malignant	Adenocarcinoma	1, 2, 3, 4
Colon/rectum	Biopsy	Benign	Inflammatory bowel disease	1, 2, 3, 4
Skin	Biopsy/excision	Malignant	Melanoma	1, 2, 3, 4
Lymph node (neck)	Excision	Malignant	Hodgkin's lymphoma	2, 3, 4
Lung	Resection	Malignant	Lymphoma	1
Lung	Resection	Malignant	Poorly differentiated carcinoma	1, 2, 3, 4

*Each site received the same type of cases with the exception of site 1 which used lung resection for a lymphoma. *n*=36 cases total.
ADH: Atypical ductal hyperplasia

out of the total for each site and across all sites. For each agreement, Pearson's correlation coefficient was calculated to assess correlation between modalities per case, and *P* value was generated from the correlation coefficient, where $P < 0.05$ was considered to be statistically significant. The number and percentage of completely concordant, minor discordant, and major discordant cases were calculated for each site and across sites for both glass and digital reads.

The categorical intrareader/intermodality agreements were calculated and results were pooled. Cases for which the reader chose to defer to additional stains or subspecialist consultation (for both modalities) were not included in this categorical analysis. The number and category of deferrals were later calculated for both WSI and glass. The average intrareader/intermodality agreement was calculated, and the exact Clopper–Pearson 95% CIs were determined.

Finally, the average categorical interreader/intramodality agreements (across the six inter-reader pairs at each site) were calculated for both glass and WSI and the results were then pooled across sites. Again, cases that were designated as deferrals were not included in this categorical analysis. Exact Clopper–Pearson 95% CIs were determined, and for the pooled data, a weighted kappa with 95% CIs was calculated.

RESULTS

Thirty-three cases (four readers per case, resulting in 132 independent reads for each modality) were fully analyzed using preestablished discordance tables where discordances were classified as major or minor. The average major discordance rate for glass across sites was 12.1% (16 major errors out of 132 total reads) and 11.4% (15 major errors out of 132 total reads) for WSI [Table 5]. Major discordances [Table 6] occurred on challenging cases, regardless of modality, including a micrometastasis in a lymph node, a small focus of ADH in a breast biopsy, and a difficult melanoma case sent for expert consult. In 12 cases with major errors, the error was in a single modality in 5 cases: Three on glass and two on digital [Table 6]. For all the other cases with major errors, the same reason for error was indicated on both modalities. The average difference in error rate between WSI and glass was 0.75% with an upper bound of 3.23% (95% CI). There were also similar numbers of total minor errors and completely concordant cases between the two modalities, within each and across all sites [Table 7].

Discordance tables were used to calculate intrareader/intermodality categorical diagnostic agreement for each reader, the average for each site, and across sites. The average percentage agreement for each site was as follows: 89.3% for site 1, 90.6% for site 2, 96.9% for site 3, and 87.5% for site 4. The average percentage agreement across sites between glass and WSI was 91.6% (0.851, 0.959; 95% CI [exact Clopper–Pearson intervals]) when each diagnostic category was pooled [Table 8].

To compare the relative reproducibility of each modality, interpathologist/intramodality percentage agreement and weighted kappa statistic were calculated for each site [Table 9] and across sites [Table 10]. The average inter-reader agreement across sites for glass was 0.753 (0.683, 0.810; 95% CI [Clopper–Pearson]), with an average weighted kappa of 0.675 (0.587, 0.763; 95% CI). The average inter-reader agreement across sites for WSI was 0.792 (0.725, 0.849; 95% CI [Clopper–Pearson]), with an average weighted kappa of 0.718 (0.633, 0.803; 95% CI). These results indicate no significant difference between glass and WSI for inter-reader agreement across diagnostic categories.

To be consistent with real-life conditions and to determine whether readers could understand when a deferral was appropriate for either modality, the case report forms had options for deferrals in every category

of diagnosis. For glass, there were a total of 20 deferrals. Ten deferrals were for additional stains and 10 were deferrals to a subspecialist. For WSI reads, there were a total of 16 deferrals. There were six deferrals for additional stains and ten deferrals to a subspecialist. There were no deferrals for image quality indicated.

Table 5: Accuracy of whole slide image versus glass per site and pooled

	Average percentage agreement with GT	Correlation coefficient, *P*
Site 1		
Glass	0.893	0.3523, 0.066
WSI	0.929	
Site 2		
Glass	0.938	0.8028, <0.001
WSI	0.906	
Site 3		
Glass	0.833	0.9103, <0.001
WSI	0.806	
Site 4		
Glass	0.861	0.4601, 0.005
WSI	0.917	
Across sites		
Glass	0.879	0.6715, <0.001
WSI	0.886	

WSI: Whole slide image, GT: Ground truth

DISCUSSION

We describe a standardized approach for WSI validation which aims to reduce the variability of determining error between two reads and to streamline the laborious process of traditional manual arbitrations. Our method employs codified case report form checklists for both categorical and malignant diagnoses based on the CAP cancer protocols. We tested the feasibility of this approach in a multisite validation study which examines accuracy and intra- and inter-reader reproducibility. For this pilot study, we created 15 synoptic checklists, systematically identified all hypothetical pairwise discordances, and assigned the levels of harm based on published definitions. Using these "discordance tables," we were able to analyze 264 paired reads for error, 132 paired reads for intrareader/intermodality, and 768 paired reads for inter-reader/intramodality (384 glass and 384 WSI). We applied identical error criteria for all comparisons across all sites, readers, and cases. This method is a valid approach for studies, in which there are many different analysis methods applied and where scalability across many sites and readers is required.

A recent study conducted by Snead *et al.* examined over 3000 WSI cases for noninferiority to glass reads.[14] This is one of the largest studies to date and required one pairwise comparison per case requiring approximately

Table 6: Major discordance numbers and details for each site and modality

		Number and classification of major discordances per case			
Site	**Case**	**WSI**	**Glass**	**GT (original sign-out diagnosis)**	**Major discordance details**
1	1	1	1	Endometrium biopsy, well-differentiated adenocarcinoma	All (both WSI and glass) called it severe atypia
	2	1	1	Breast biopsy, small focus of ADH	All (both WSI and glass) called it benign inflammatory/fibrocystic change
	3	0	1	Skin excision, invasive melanoma. Breslow thickness of 0.5 mm	Glass called it melanoma *in situ*
2	4	2	2	Skin excision, invasive melanoma (per consult), Breslow thickness of 0.22	All (both WSI and glass) called it melanoma *in situ*
	5	1	0	Breast biopsy, small foci of ADH	WSI called it benign inflammatory
3	6	2	2	Endometrium biopsy, small foci of adenocarcinoma, Grade 1	All (both WSI and glass) called it moderate-severe dysplasia
	7	2	2	Breast sentinel lymph node, micrometastasis	All (both WSI and glass) called it nonneoplastic reactive
	8	3	2	Breast biopsy, small foci of ADH	All (both WSI and glass) called it benign neoplastic
4	9	1	0	Endometrium biopsy, adenocarcinoma, Grade 2	WSI called it severe dysplasia
	10	2	3	Cervix biopsy, cervical intraepithelial neoplasia 2	All (WSI and glass) called it benign inflammatory
	11	0	1	Breast biopsy, small foci of ADH	Glass called it benign inflammatory
	12	0	1	Breast sentinel lymph node, tiny foci of micrometastasis	Glass called it benign reactive
Total		15	16		

ADH: Atypical ductal hyperplasia, WSI: Whole slide image, GT: Ground truth

Table 7: Major, minor, and concordance for each site and pooled

	Major (%)	Minor (%)	Concordant (%)
Site 1			
Glass-GT	3 (11)	13 (46)	12 (43)
WSI-GT	2 (7)	14 (50)	12 (43)
Site 2			
Glass-GT	2 (6)	13 (41)	17 (53)
WSI-GT	3 (9)	11 (34)	18 (56)
Site 3			
Glass-GT	6 (17)	15 (42)	15 (42)
WSI-GT	7 (19)	14 (39)	15 (42)
Site 4			
Glass-GT	5 (14)	20 (56)	11 (31)
WSI-GT	3 (8)	19 (53)	14 (39)
Across sites			
Glass-GT	16 (12)	61 (46)	55 (42)
WSI-GT	15 (11)	58 (44)	59 (45)

WSI: Whole slide image, GT: Ground truth

Table 8: Pooled average intra-reader glass to whole slide image comparisons

All readers/all sites (glass-digital)	Intrareader/intermodality agreement*		
	Benign	Atypical	Malignant
Benign	25	3	0
Atypical	2	18	2
Malignant	0	3	67
Agreement (95% CI)	0.917 (0.852, 0.959)		

*3 cases (across 4 sites, 12 total comparisons) are not included, as readers chose to defer without categorizing (to additional stains or subspecialist consultation), on both modalities. CI: Confidence interval

Table 9: Average categorical inter-reader agreements by site

Site	Average categorical agreement for glass (95% CI)	Average categorical agreement for digital (95% CI)
1	0.762 (0.606, 0.879)	0.810 (0.659, 0.915)
2	0.750 (0.604, 0.864)	0.896 (0.748, 0.953)
3	0.750 (0.600, 0.860)	0.750 (0.600, 0.860)
4	0.799 (0.618, 0.901)	0.709 (0.540, 0.834)

CI: Confidence interval

3000 arbitrations. Even for such a large study, these sample sizes are manageable to arbitrate. However, for complex multisite, multireader studies, such as those required by the FDA, there would be more pairwise comparisons with one to two orders of magnitude. For example, in a study where the case size is 2000 split evenly among four sites with four readers at each site, there will be 16 readers reading 500 cases each. In a noninferiority design, each modality is compared to a separate GT. Thus, there would be approximately 16,000 pairwise ($16 \times 500 \times 2$) comparisons for the primary end point. In addition, precision and reproducibility (e.g., inter- and intra-reader) will also be required per modality. Such a study could easily generate over 100,000 pairwise comparisons, which is logistically challenging and time consuming and further increases the likelihood of intra-arbitrator variability.

There are relatively few studies that employ a noninferiority design for validation of WSI.[5,14] Bauer *et al*. were the first to do so and reported a difference in error rate of 0.66% between WSI and glass.[5] In the study by Snead *et al*., a difference of 0.1% between WSI and glass was reported.[14] Under our study conditions, we showed the difference in error rates between glass and WSI to be 0.75% with an upper bound of 3.23% (95% confidence). Thus, even given our low sample size, our data are consistent with previous studies of similar design. However, we acknowledge a key limitation of this study that is the low sample size itself, which greatly underpowers the study. Nonetheless, the methods developed in this report will serve as a foundation to expand into a much larger study, and it is important to validate the design prior to commencing such a large-scale study.

In addition to calculating the overall agreement of glass and WSI with GT, we performed a correlation analysis comparing glass and WSI diagnosis on a case-by-case basis for each site and overall. This agreement included all fields in the checklists, including the CAP cancer checklists, where applicable. The use of detailed, standardized checklists allows for this very thorough side-by-side comparison, which can highlight any difference between a test and reference diagnosis and help determine how likely the diagnoses, using two different modalities, will be the same. This is an important analysis that can capture site-by-site modality differences beyond comparing the overall concordance or agreement. For example, site 3 had an overall lower average agreement with GT for both glass and digital than site 4 but they had a higher correlation coefficient (0.9103 compared to 0.4601), indicating that glass and digital diagnoses of site 3 were more likely to be in agreement than for site 4 (although site 4 still has a significant positive correlation between modalities).

Unlike the two previous noninferiority studies which show an inherent glass-to-GT error of <2%, our data show an inherent glass-to-GT error of 12.4%. There are several possible reasons for this discrepancy. Both the former studies used cases that were previously reported by the study participants (either all or a large fraction of cases) as GT which would naturally reduce the error between the test methods and GT. Further, in our study, all cases were selected from a single institution and it is possible that the inter-institution variability could have contributed to an increased error. Finally, our samples

Table 10: Average categorical agreements pooled

	Average inter-reader-glass (across sites)			Average inter-reader-digital (across sites)		
	Benign	Atypical	Malignant	Benign	Atypical	Malignant
Pooled pairs						
Benign	30	8	4	30	8	4
Atypical	9	14	12	7	18	13
Malignant	1	8	89	1	4	93
Agreement (95% CI)	0.753 (0.683, 0.810)			0.792 (0.725, 0.849)		
Weighted kappa (95% CI)	0.675 (0.587, 0.763)			0.718 (0.633, 0.803)		

For each site, there were six inter-reader pairs. Pairs that included a "no diagnosis" defer to subspecialist or additional stains were not included in the categorical agreement analysis. CI: Confidence interval

were purposely enriched for complex cases and most readers did not have the required specialty training needed for high accuracy. Others have shown that inter-reader error is substantially greater when cases are enriched as opposed to random.[17] For example, a report by Elmore *et al*. demonstrated only a 75% consensus among a panel of experts when reviewing cases enriched for DCIS and ADH of the breast.[18] Thus, our inherent error is not outside the norm given the study conditions.

A unique aspect of our study was the option in the case report forms to defer a diagnosis for various reasons, including image quality. An important safety consideration of WSI is that a pathologist can always use a microscope if he/she is unsure of a WSI diagnosis. We intentionally wanted to test whether a reader would recognize the need to defer WSI-based diagnoses in the same manner as on glass-based reads. Our data show that deferrals were similar for both WSI (16) and glass (20), and there were no deferrals due to image quality, despite the enrichment of the study set with particularly challenging cases. Thus, it is clear that a reader can determine whether deferrals are needed when reading cases by WSI and they do not require deferrals at a greater rate on WSI.

An important limitation in this study is the fact that discordance tables were created by one person. It is likely that there is variability in the interpretation of assigned error depending on who creates the tables. However, the main advantage of this approach is that the tables serve as a framework. Such a framework could be used by a working group or panel for a given study or from an accredited organization such as the CAP or American Society of Clinical Oncology to establish more finalized standard assignments of harm for any given discordance. Further, the framework itself can be used as a mechanism to examine the sensitivity of results. For example, an investigator may decide to analyze the data using two different assigned harm levels for a single discordance and then re-run the analysis to see whether such changes impact the outcome. The change would be clearly understood by the study participants and reviewers such that robust interpretations of the data could be established.

Another limitation of our study is that our case report forms did not have high levels of granularity for the categorical diagnoses of benign and atypical. This is because it would be logistically difficult to create such checklists and further, if it were possible, they would be unique to this study and not an accepted standard. However, to ensure that a large portion of cases were tested to a high level of granularity, we used CAP cancer protocols for all malignant cases. This level of granularity allows for a deep understanding of the reasons for error. For example, two diagnoses being compared may both indicate an invasive malignancy, but assessment of grade or pathological stage could differ resulting in an unnecessary treatment or missed diagnosis that could cause harm to a patient. Finally, there are some cases that do not "fit" intuitively into one distinct diagnostic category (i.e., atypical small acini that are suspicious for but not diagnostic of malignancy or a biologically benign brain tumor that is malignant due to location). This occasionally may introduce variation in reporting but can be minimized by efficient training to introduce "rules of categorization" and proficiency testing before a validation study begins. Another possibility is to require the use of standardized checklists but include a free text field to add additional information that can be filled out in addition to the categorization, which may be used in arbitration of questionable discrepancies.

CONCLUSION

To our knowledge, this is the first multisite study to intentionally involve pathologists with a broad range of professional experience (1–28 years) and expertise with respect to the presence or absence of fellowship training across a number of different subspecialty areas. In addition, none of these pathologists had prior training in the use of WSI for diagnostic purposes. Cases were purposely chosen to be challenging with a breadth of benign, atypical, and malignant as well as biopsies, resections, and excisions. This is also the first study to use standard synoptic reporting to eliminate variability in data collection and analysis and which measured accuracy

as well as reproducibility. Our methods are feasible and easily adapted to large-scale studies such as those required for medical device registrations.

Financial Support and Sponsorship
This work was supported by Omnyx LLC and was presented in the form of a poster at the 2016 USCAP Meeting in Seattle, WA.

Conflicts of Interest
There are no conflicts of interest.

REFERENCES

1. Parwani AV, Hassell L, Glassy E, Pantanowitz L. Regulatory barriers surrounding the use of whole slide imaging in the United States of America. J Pathol Inform 2014;5:38.
2. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, *et al.* Validating whole slide imaging for diagnostic purposes in pathology: Guideline from the College of American Pathologists Pathology and Laboratory Quality Center. Arch Pathol Lab Med 2013;137:1710-22.
3. Evans AJ, Krupinski EA, Weinstein RS, Pantanowitz L. 2014 American Telemedicine Association clinical guidelines for telepathology: Another important step in support of increased adoption of telepathology for patient care. J Pathol Inform 2015;6:13.
4. Canadian Association of Pathologists Telepathology Guidelines Committee, Bernard C, Chandrakanth SA, Cornell IS, Dalton J, Evans A, *et al.* Guidelines from the Canadian Association of Pathologists for establishing a telepathology service for anatomic pathology using whole-slide imaging. J Pathol Inform 2014;5:15.
5. Bauer TW, Schoenfield L, Slaw RJ, Yerian L, Sun Z, Henricks WH. Validation of whole slide imaging for primary diagnosis in surgical pathology. Arch Pathol Lab Med 2013;137:518-24.
6. Bauer TW, Slaw RJ. Validating whole-slide imaging for consultation diagnoses in surgical pathology. Arch Pathol Lab Med 2014;138:1459-65.
7. Bauer TW, Slaw RJ, McKenney JK, Patil DT. Validation of whole slide imaging for frozen section diagnosis in surgical pathology. J Pathol Inform 2015;6:49.
8. Buck TP, Dilorio R, Havrilla L, O'Neill DG. Validation of a whole slide imaging system for primary diagnosis in surgical pathology: A community hospital experience. J Pathol Inform 2014;5:43.
9. Campbell WS, Hinrichs SH, Lele SM, Baker JJ, Lazenby AJ, Talmon GA, *et al.* Whole slide imaging diagnostic concordance with light microscopy for breast needle biopsies. Hum Pathol 2014;45:1713-21.
10. Gilbertson JR, Ho J, Anthony L, Jukic DM, Yagi Y, Parwani AV. Primary histologic diagnosis using automated whole slide imaging: A validation study. BMC Clin Pathol 2006;6:4.
11. Houghton JP, Ervine AJ, Kenny SL, Kelly PJ, Napier SS, McCluggage WG, *et al.* Concordance between digital pathology and light microscopy in general surgical pathology: A pilot study of 100 cases. J Clin Pathol 2014;67:1052-5.
12. Ordi J, Castillo P, Saco A, Del Pino M, Ordi O, Rodríguez-Carunchio L, *et al.* Validation of whole slide imaging in the primary diagnosis of gynaecological pathology in a University Hospital. J Clin Pathol 2015;68:33-9.
13. Reyes C, Ikpatt OF, Nadji M, Cote RJ. Intra-observer reproducibility of whole slide imaging for the primary diagnosis of breast needle biopsies. J Pathol Inform 2014;5:5.
14. Snead DR, Tsang YW, Meskiri A, Kimani PK, Crossman R, Rajpoot NM, *et al.* Validation of digital pathology imaging for primary histopathological diagnosis. Histopathology 2016;68:1063-72.
15. Raab SS, Grzybicki DM, Janosky JE, Zarbo RJ, Meier FA, Jensen C, *et al.* Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. Cancer 2005;104:2205-13.
16. Raab SS, Nakhleh RE, Ruby SG. Patient safety in anatomic pathology: Measuring discrepancy frequencies and causes. Arch Pathol Lab Med 2005;129:459-66.
17. Raab SS, Grzybicki DM, Mahood LK, Parwani AV, Kuan SF, Rao UN. Effectiveness of random and focused review in detecting surgical pathology error. Am J Clin Pathol 2008;130:905-12.
18. Elmore JG, Pepe MS, Weaver DL. Discordant interpretations of breast biopsy specimens by pathologists – Reply. JAMA 2015;314:83-4.