Regular Article

Academic Pathology: Volume 6
DOI: 10.1177/2374289519859841
journals.sagepub.com/home/apc

SAGE

Agreement in Histological Assessment of Mitotic Activity Between Microscopy and Digital Whole Slide Images Informs Conversion for Clinical Diagnosis

Bih-Rong Wei, PhD[1,2], Charles H. Halsey, DVM, PhD[1], Shelley B. Hoover, BS[1], Munish Puri, PhD[1], Howard H. Yang, PhD[1], Brandon D. Gallas, PhD[3], Maxwell P. Lee, PhD[1], Weijie Chen, PhD[3], Amy C. Durham, MS, VMD[4], Jennifer E. Dwyer, MS[1], Melissa D. Sánchez, VMD, PhD[4], Ryan P. Traslavina, DVM[5], Chad Frank, MS, DVM[6], Charles Bradley, VMD[4], Lawrence D. McGill, DVM, PhD[7], D. Glen Esplin, DVM, PhD[7], Paula A. Schaffer, DVM, MS[6], Sarah D. Cramer, DVM, PhD[8], L. Tiffany Lyle, DVM, PhD[9], Jessica Beck, DVM[10], Elizabeth Buza, DVM[4], Qi Gong, MS[3], Stephen M. Hewitt, MD, PhD[11], and R. Mark Simpson, DVM, PhD[1]

Abstract

Validating digital pathology as substitute for conventional microscopy in diagnosis remains a priority to assure effectiveness. Intermodality concordance studies typically focus on achieving the same diagnosis by digital display of whole slide images and conventional microscopy. Assessment of discrete histological features in whole slide images, such as mitotic figures, has not been thoroughly evaluated in diagnostic practice. To further gauge the interchangeability of conventional microscopy with digital display for primary diagnosis, 12 pathologists examined 113 canine naturally occurring mucosal melanomas exhibiting a wide range of mitotic activity. Design reflected diverse diagnostic settings and investigated independent location, interpretation, and enumeration of mitotic figures. Intermodality agreement was assessed employing conventional microscopy (CM40×), and whole slide image specimens scanned at 20× (WSI20×) and at 40× (WSI40×) objective magnifications. An aggregate 1647 mitotic figure

[1] Laboratory of Cancer Biology and Genetics, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[2] Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD, USA
[3] Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD, USA
[4] Department of Pathobiology, University of Pennsylvania, Philadelphia, PA, USA
[5] Section of Infections of the Nervous System, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA
[6] Department of Microbiology, Immunology, and Pathology, Colorado State University, Fort Collins, CO, USA
[7] Animal Reference Pathology, Salt Lake City, UT, USA
[8] Cancer and Inflammation Program, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[9] Women's Malignancies Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[10] Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[11] Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

Corresponding Author:
R. Mark Simpson, Laboratory of Cancer Biology and Genetics, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA.
Email: ms43b@nih.gov

count observations were available from conventional microscopy and whole slide images for comparison. The intraobserver concordance rate of paired observations was 0.785 to 0.801; interobserver rate was 0.784 to 0.794. Correlation coefficients between the 2 digital modes, and as compared to conventional microscopy, were similar and suggest noninferiority among modalities, including whole slide image acquired at lower 20× resolution. As mitotic figure counts serve for prognostic grading of several tumor types, including melanoma, 6 of 8 pathologists retrospectively predicted survival prognosis using whole slide images, compared to 9 of 10 by conventional microscopy, a first evaluation of whole slide image for mitotic figure prognostic grading. This study demonstrated agreement of replicate reads obtained across conventional microscopy and whole slide images. Hence, quantifying mitotic figures served as surrogate histological feature with which to further credential the interchangeability of whole slide images for primary diagnosis.

Keywords
cancer grading, informatics, prognosis, reproducibility study, validation, digital pathology, technology adoption, training

Received March 08, 2019. Received revised May 15, 2019. Accepted for publication May 19, 2019.

Introduction

Over the last decade, diagnostic applications in digital pathology have become increasingly possible. The technology represents a major innovation with potential to significantly impact a large swath of the health-care enterprise; however, adoption of digital pathology has been less rapid than anticipated. The protracted momentum appears due in part to cost and the lack of use cases that demonstrate pathologists can accomplish tasks with whole slide images (WSIs) at a level equal to, or better than, the conventional optical microscope (noninferiority). Digital pathology practice requires WSI scanning instruments, computers, high-quality display monitors, and server solutions for data storage and computational image processing. Reliable diagnostic implementation of digital pathology must ensure that the entire tissue section has been scanned and digitized appropriately and that the WSI created permits uncompromised specimen review and interpretation.[1] There is concern that the workflow required for digital pathology can lead to increased turnaround time, burdening case management compared to conventional microscopy (CM).[1] Furthermore, digital microscopy may constrain specimen visualization. For example, unlike a CM, a single plane WSI does not permit the specimen to be focused in the z-axis plane, possibly impacting accuracy in primary diagnosis.[2,3] Therefore, validating the conversion to digital microscopy from CM in clinical diagnosis remains a priority to assure effectiveness, and documenting reproducibility is important for such goals.

Evaluating the substitution of WSI for CM has largely focused on whether the same, or similar, diagnosis can be reached using each of the viewing modalities (intraobserver and/or interobserver concordance between modalities).[4,5] Histopathological diagnoses using WSI have been considered noninferior to CM.[5-7] Additionally, agreement between CM and WSI has been demonstrated when scoring/interpreting immunohistochemistry (IHC), such as anti-Ki-67 and anti-HER2, when the targets are readily visible.[8,9] By contrast, evaluation designed to validate examination of discrete histological feature details, such as mitotic figures (MFs) and microorganisms, for primary diagnosis by digital display, is not well established.[10] The number of MF (mitotic counts) in tissue sections relates to tumor proliferative activity and can provide a clinically relevant, prognostically useful tumor grading biomarker in surgical pathology. Mitotic activity evaluation generally requires pathologists to determine the tissue area with most numerous MF (mitotic hot spot), followed by detailed scrutiny of nuclear morphology to count MF, in order to determine a mitotic activity index.[3,11-15] Assessing tumor mitotic activity provides a challenging, yet quantifiable, histological feature identification task with which to further evaluate pathologist performance between CM and digital display of WSI. It is plausible that inferiority of one or the other modality would be more likely revealed in an intermodality comparison that requires the location, interpretation, and enumeration of MF.

In order to more comprehensively substantiate the interchangeability of digital and CM in diagnosis, agreement in assessing MF histological feature detail was examined in this multi-institutional study employing a clinical practice paradigm using both CM and WSI. In this study, 12 observers, including academic and clinical laboratory-based pathologists, evaluated a series of canine spontaneous oral mucosal melanoma biopsies, exhibiting a wide range of mitotic activity. Observer performance was appraised for correlation between CM and digital WSI, the latter on specimens scanned at 20× as well as 40× magnification. In addition, the utility of CM and digital modes was assessed for prognostic grading performance. Detection and enumeration of MF were correlated among modalities, and with patient survival, developing a foundation for the interchangeability of MF evaluation by digital display with CM.

Materials and Methods

Study Pathologists

The multi-institutional study comparing CM of glass slide specimens with digital display of WSIs was performed at 4

institutions. Two academic departments, a national reference laboratory, and National Institutes of Health (NIH)-based pathologists were represented. Participating pathologists from these institutions included 1 postresidency pathology fellow and 11 specialty certified pathologists with 3 to 33 years of clinical diagnostic experience (8 pathologists, 1-10 years; 2 pathologists, 11-20 years; 2 pathologists >20 years). Pathologists' self-report of their experience with digital pathology indicated they generally lacked substantial or, in several cases, any experience with clinical diagnosis on computer display, with one exception; 1 pathologist used the digital platform routinely in diagnostic practice.

Specimens

The tumor specimens used in this study were spontaneous, naturally occurring oral mucosal melanomas obtained from dogs in the course of clinical veterinary patient care. This malignancy represented a high-fidelity human cancer model and replicates the histopathology of human cutaneous melanoma.[16-18] Study specimens included formalin-fixed, paraffin-embedded tissue blocks of surgically excised spontaneous canine melanomas, sourced from participating institution archives at the Colorado State University and University of Pennsylvania. Patient survival data for a subset of cases were provided from these institutions, and in collaboration with Dr Michael Goldschmidt, University of Pennsylvania, and Dr EJ Ehrhart, Colorado State University. As a consequence of veterinary patient care management of client-owned pet dogs, this retrospective use of archived diagnostic specimens is not subject to prospective research animal use approvals. Tumors had been resected with intent to cure; patients received no treatment prior to surgery. Specimens were anonymized for this study. Single paraffin-embedded specimens from each case were processed in the same laboratory. For each case, one 5 μm-thick section was mounted on a glass slide, rehydrated, bleached to remove melanin pigment, and subsequently stained with hematoxylin and eosin (H&E). Melanin quenching involved immersing deparaffinized, rehydrated slides in 0.25% aqueous potassium permanganate solution for 1 hour, washing in running tap water, followed by decolorizing for 5 minutes in 5% aqueous oxalic acid, washing, and final rinsing in deionized water (Histoserv, Inc, Gaithersburg, Maryland).

Original histopathological diagnoses from the submitting institutions were reverified (by C.H.H.), and all cases included following this review had characteristic melanoma features.[18] Immunohistochemistry for melanoma differentiation antigens Melan A, PNL2, and Trp-2 was performed on serial sections by the National Cancer Institute (NCI) following methods described previously,[18] in order to further establish the melanocytic origin of all study specimens (data not shown). Each case was assigned a randomly generated 4-digit identification number to replace any institutional identifying information. One hundred thirteen H&E-stained glass slides were subsequently optically scanned as WSI in batch scan mode at both 20× (0.5 μm per pixel) and 40× (0.25 μm per pixel) using an AT2 digital slide scanner (Leica Biosystems, Vista, California). Scanned image files were reviewed to ensure proper quality for examination, including appropriate focus and tissue inclusion. This preliminary evaluation of image scans led to rescanning 3 slides at both 20× and 40× (2.65% rescan rate).

Specimen Assessment Protocol

Pathologists at each study location reviewed the same slides and image files for assigned specimens. Specimen were randomized into 3 case groups (n = 37 or 38 per group), and assignments for evaluation using 3 microscopy modalities were made according to a split plot study design.[19] Each pathologist reviewed all patient specimens: 2/3 of the total cases by CM, 2/3 by WSI scanned at 20× (WSI20×), and 1/3 by WSI scanned at 40× (WSI40×; Figure 1 and Supplemental Table 1). Each pathologist group evaluated 1 case group with all 3 modalities. For example, each observer in observer group 1 evaluated case groups 1 and 2 with CM40×, case groups 2 and 3 using WSI20×, and case group 2 using WSI40× (Supplemental Table 1). In this example, group 1 observers evaluated case group 2 with all 3 modalities (Figure 1). This approach promotes efficient use of cases, each observer's time, and the total number of observations from a study.[20]

For each slide (glass or WSI), pathologists examined specimens and identified areas of most numerous MF (mitotic hot spots) as they would in diagnostic practice according to experience and preference, consistent with the standard of care.[21] Evaluation and subsequent enumeration of MF were made to include 10 consecutive but nonoverlapping fields at maximum resolution, defined uniquely for the 3 modalities as using the 40× high-power objective lens by CM for glass slides (CM40×), by computer display of WSI at the 20× setting for 20× scanned images and at the 40× setting for WSI scanned at 40×. Data collection included a balanced mix of the order of modalities. The case order was randomized for each pathologist, for each modality. A washout period of at least 1 week was structured between examinations of patient groups and modality uses, to minimize potential case recall bias. Pathologist participation anonymity regarding study outcomes was maintained.

Pathologists were provided with all study materials. The single set of glass slides used for the study (no recuts or duplicates) was shipped on a rotation to the various institutions. Pathologists within an institution completed their assigned cases and subsequently shipped the glass slides to the next center. Image files, image viewing software (Aperio ImageScope v12.0.1.5027, Leica Biosystems, Vista, California), cell counter software application (see Record of Mitotic Figure Counts, Collation, and Quality Review), assigned cases, randomized read sequence, and study protocol were provided on external hard drives for each pathologist individually; these materials could all be loaded onto pathologists' personal computers. Standard desktop display monitor resolution varied somewhat among pathologists (median 92.195, range 86.273-102.460 pixels per inch; Supplemental Table 2). Prior to initiation of the study, pathologists attended a training webinar,

Figure 1. Split plot study design for comparing mitotic activity fine histological feature assessment between CM and digital WSI, depicted graphically according to assigned case groups and observer groups. One hundred thirteen melanoma biopsy cases were divided into 3 groups (n = 37 or 38 per case group), and each pathologist examined all 113 patients, by observer group, represented accordingly as assigned cases (gray boxes) and no assigned cases (white boxes). Three viewing modalities: pathologist's conventional microscope (CM) 40× objective lens, whole slide image (WSI) scanned at 20× (WSI20×), and WSI scanned at 40× (WSI40×) were employed. Assessment of digital image files was conducted on the pathologist's personal display monitor. The design provided for each case to be read by at least 8 pathologists in CM40× and WSI20×, and 4 pathologists in WSI40×.

including interactive discussion, during which the study protocol was reviewed, and examples of MF intended for inclusion were illustrated. Instructions on recording MF counts using a cell counter were discussed (see "Methods" section). A study director confirmed that pathologists possessed operational files and had standardized the color calibration of their digital display monitors (Spyder4PRO, Datacolor, Lawrenceville, New Jersey).

Table 1. Intraobserver and Interobserver, Intermodality Concordance Analyses.

Modality Comparison	Intraobserver	Interobserver
CM40×/WSI20×	0.785	0.784
CM40×/WSI40×	0.801	0.786
WSI20×/WSI40×	0.798	0.794

Abbreviations: CM, conventional microscopy; WSI, whole slide image.

Record of Mitotic Figure Counts, Collation, and Quality Review

When reading WSI, pathologists annotated their individual images using the annotation function in ImageScope to record the regions of interest (ROIs) corresponding to each of 10 individual high-power fields of view (FOV) where MF counting was conducted. The size of a circular ROI was equivalent to area of 1 high-power FOV corresponding to the pathologist's personal microscope used for CM40× (Supplemental Figures S1 and S2). A custom graphical user interface (GUI) counting application was created using MATLAB Coder (MathWorks, 2012, Natick, Massachusetts) to standardize MF count recording for both CM and digital WSI (Supplemental Figure S3). The GUI program and instructions for downloading and navigating the application were included in the external hard drive for each pathologist. Graphical user interface operation was in 3 functional steps: (1) registration entry of observer identification, case identification number, and modality; (2) incremental tally of MF as they were observed within each of 10 FOV in the MF cell counting application by computer mouse click on the "count" key; this was accompanied by an audible sound for each count registration made; and (3) exporting data after counts have been registered for 10 FOV by clicking the "Export" key; total MF count summation and export of the case data to each pathologist's spreadsheet was executed through the GUI.

The data set for analysis, including 2260 mitotic activity count entries from 20,600 FOV, was transposed as a relational database in R programming (R Foundation for Statistical Computing, Vienna, Austria). The accuracy of the patient-identification codes metadata, the appropriate performance for each case by each participant, and observer compliance with the specimen evaluation protocol in each case were inspected. Issues, including presence of duplicate entries and errors in nomenclature, were identified and corrected appropriately. Review of pathologist-annotated ROI on WSI revealed deviation from the read protocol in 4 instances. After excluding these data (613 MF count values [27% of the total]), there were 1647 total aggregate mitotic activity count entries.

Data Analysis

Analyses were carried out through multiple means. Processing in R (R Foundation for Statistical Computing) utilized the NIH High Performance Computing Biowulf cluster and helix systems (https://hpc.nih.gov/, accessed April 24, 2019). Mitotic figure count relationships between observers using the same modality were established using Spearman correlation (interobserver/intramodality). Mitotic figure counts derived from different modalities (intraobserver/intermodality) were compared using Spearman rank correlations and linear regression analyses. The linear association was evaluated by the slope β, P value, and R^2 of each linear model. The 95% prediction bands for future observations about the regression lines were calculated and plotted. We also calculated the intra- and interobserver rank-based concordance rates for paired observations.[22] As with many measurements, especially counts, we found that the variance of the counts grew with the average. Therefore, MF values were transformed by the function $\log_{10}$ (MF counts +1).

Two-factor analysis of variance (ANOVA) was used to analyze the (fixed) effects of observers and modalities on MF counts. When the transformed input data $\log_{10}$ (MF counts +1) were used in the 2-factor analysis, a qq-plot analysis of the residuals provided appropriate support for an assumption of normality in the ANOVA (data not shown). Clinical utility regarding the relationship of MF counts made by CM and WSI20× to melanoma patient survival was assessed using Kaplan-Meier survival analysis and the log-rank test.

Results

Study Performance

Pathologist performance at identifying and counting MF, from H&E-stained specimens of canine spontaneous mucosal melanomas, was evaluated using 3 microscopy modalities (CM40×, WSI20×, and WSI40×) based on the clinical paradigm of counting 10 contiguous high-power FOV for each case. Pathologists were divided into 3 groups (primarily based upon their 4 institutions, providing efficient study of the same slides for all reviews). To maintain reasonable workloads while conserving statistically performing replicate reads, the study employed a split-plot design with the total 113 cases divided randomly into 3 assigned case groups (Figure 1). The design provided the necessary overlap for replicate reads of each patient in each modality. As a result, each case was read by at least 8 pathologists for CM40× and WSI20×, or 4 pathologists in the case of WSI40× (Figure 1 and Supplemental Table 1).

Pathologists used their routine clinical microscopes and computer display monitors to examine cases in their respective diagnostic sign-out environments. Case groups of melanoma tissue slides were shipped serially to the various pathologist groups for evaluation by CM. Pathologists were instructed to evaluate specimens by locating the tumor mitotic hot spot and then begin recording MF numbers in each of 10 consecutive, adjacent, nonoverlapping FOV at maximum resolution. The CM40× modality, employing a 40× objective lens for identifying and quantifying MF, was considered the routine standard of care. Although distinct, the method to assess MF as well as the total area examined has similarities to the American Joint Committee on Cancer guidelines for cutaneous melanomas.[23] The canine melanomas evaluated were of oral mucosa origin and previous efforts have rigorously demonstrated the histomorphologic similarities of these tumors to those in humans arising on squamous epithelial surfaces.[16-18] Although lacking an ultraviolet injury signature at the molecular level, canine melanoma biology retains similarities to human melanomas.

A separate examination was conducted for these same case/slides through the visualization of the digital WSI files on computer display. Each pathologist had their own copy of WSIs that had been created by scanning the single glass slide per case. The slides were scanned at both 20× (0.5 μm/pixel) and 40× (0.25 μm/pixel) resolution. Similar to CM, pathologists were directed to locate and then annotate the mitotic hot spot, as well as the subsequent 9 FOV in succession when reading WSI. Using a circle annotation tool, an ROI equivalent to the area of 1 high-power FOV of their personal CM was drawn first at the identified hot spot and subsequently in the neighboring 9 contiguous areas (Supplemental Figures S1 and S2). Hence, for each pathologist, the tumor area assessed on each patient was constant across all modalities and depended upon each pathologists' personal microscope (for most pathologists, total 2.37 mm^2 for 10 FOVs; Supplemental Table 2). In many cases, mitotic hot spot ROI selection varied among observers for a given case (Supplemental Figure S2; unpublished data), a factor that appeared capable of influencing MF count values recorded. Impact of different FOV choices on observed variances among pathologists is a subject of ongoing study.

All pathologists recorded counts of MF in the identical manner for both CM and WSIs (Supplemental Figure S3). Twelve pathologists initiated and 10 completed enumeration of tumor MF (data were excluded from observers D and I; and see Supplemental Table 1). Primary causes for censoring some data included protocol deviations, that is, not completing all assigned reads, or failure to place annotated ROI in WSI files according to study protocol (Supplemental Table 1, and Materials and Methods). Six pathologists indicated qualitative opinions regarding WSI examination user ergonomics. One considered WSI and CM modalities to be similar ergonomically, while 2 considered WSI to add to observer workflow and 3 preferred WSI due to perceived visual enhancement and reduced operator strain.

Mitotic Figures Feature Agreement

Agreement across conventional and digital microscopy modalities was evaluated using various analyses in an attempt to determine if one or more modalities might be clearly superior for the task of identifying and quantifying MF. The design also allowed examination of individual pathologist performance across the 3 modalities. Overall, comparisons included (1) between-observer, within-modality, (2) between-observer, between-modality, and (3) within-observer, between-modality (Figure 1 and Supplemental Table 1).

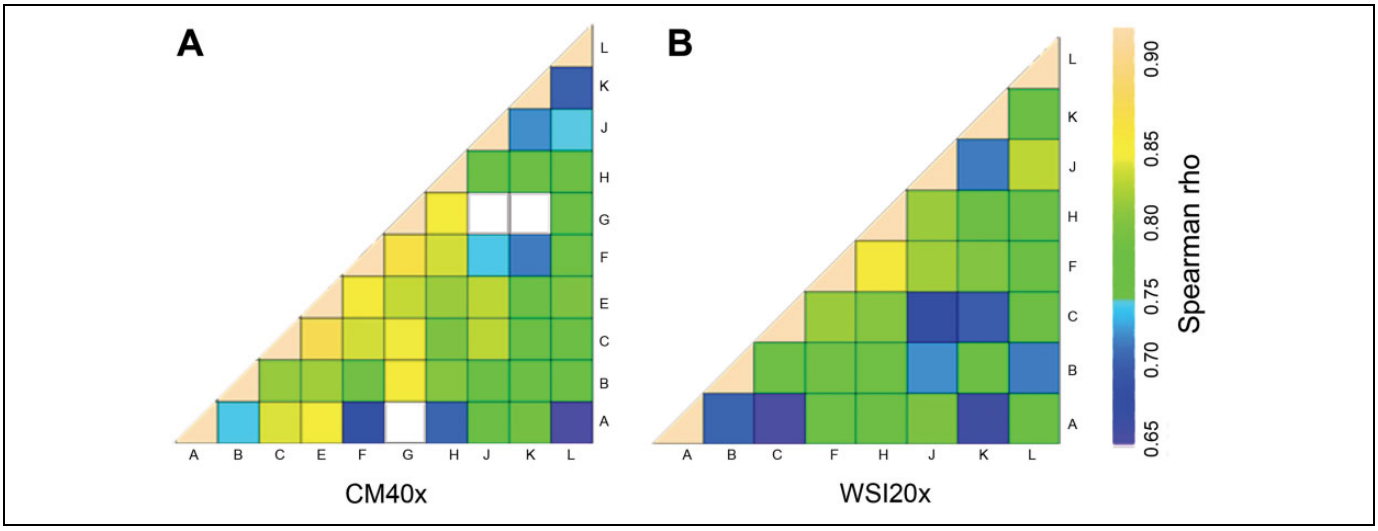

Figure 2. Pair-wise interobserver agreement of mitotic figure assessment for each pathologist (designated A-L) within modalities (A) CM40× and (B) WSI20×. Each cell in the heatmaps is the Spearman rank correlation coefficient analyses (ρ values) between each pair of pathologists. White squares indicate a read protocol fault, which were excluded. n = 37 to 75 for each comparison. CM indicates conventional microscopy; WSI, whole slide image.

We first ascertained the agreement between observers within the same modality, based upon MF counts. Using Spearman correlation analyses, MF identification and quantitation by CM40× or WSI demonstrated substantial interobserver agreement among pathologists. Pair-wise comparisons of pathologists' reads on same cases revealed interobserver correlation coefficients >0.65 in CM40× (Figure 2A), WSI20× (Figure 2B), and WSI40× (data not shown). Furthermore, 14 (33%)/42 interobserver correlations exhibited greatest agreement (ρ > 0.85) in CM40×, compared to approximately 5 (18%)/28 having comparable agreement for WSI20×. This difference may be due to pathologists' greater familiarity with CM and less experience in navigating WSI.

Agreement between CM and WSI modalities was assessed in 3 comparisons: CM40× versus WSI20×, CM40× versus WSI40×, and WSI20× versus WSI40×. Individually, all pathologists (A-L) achieved significant rank-order correlations for MF quantification in intermodality comparisons ($P < .001$; Figure 3), although intermodality correlations varied among pathologists. This indication of interchangeability between modalities for individual pathologists was also evident on scatter plots of all cases and all observers (Figure 4). Each point in these plots corresponds to one observer evaluating one case in both modalities; each case has multiple entries from different observers. Intermodality comparisons assessed using log-transformed data in all 3 combinations were characterized by similar regression slopes (0.75-0.78) and R^2 values (range, 0.63-0.66; Figure 4). Corroborating the Spearman correlation analyses, intermodality concordance coefficients spanned narrow ranges (0.785-0.801, intraobserver; and 0.784-0.794, interobserver; Table 1). Collectively, these analyses indicated results achieved among the 3 modalities for the assessment of MF were similarly concordant.

Further analysis into the relative agreement among individual pathologists was examined using 2-way ANOVAs. For each pathologist, a fitted value (mean of the observer's MF counts on all cases read) was plotted against the residuals (differences between each case MF count and the fitted value) by modality (Supplemental Figure S4). The distribution of residuals is generally uniform among the observers both within and across all 3 modalities, without obvious outliers. For each pathologist, the fitted values across 3 modalities were similar. Thus, observer performance characteristics among modalities were comparable by this analysis as well.

Utility for Clinical Prognosis

We next investigated whether the prognostic utility of identifying and enumerating MF for tumor grading would be sustained when pathologists transferred from CM to digital WSI microscopy. A clinically predictive MF cut point count was determined from CM data of all study pathologists to define significant differences in survival. Patient survival follow-up was available for a subset of 66 dogs through the contributing institutions (disease-specific survival up to 1 year following diagnosis). The majority of total mitotic counts from all cases were ≤20 in 10 CM FOV. Therefore, to accomplish mitotic count cut point determination, continuous MF count values (X = 1-20) from CM counts were serially applied as putative cut points to divide all cases into short- and longer-term survivor groups. For each cut point in turn (X), Kaplan-Meier survival curves of cases with MF ≥ X versus cases with MF < X were plotted and *P* values were computed; this was performed for each pathologist (Supplemental Figure S5). At a cut point of ≥10 MF, survival curves generated from 9 (90%) of 10 pathologists appropriately divided the cases into a high- or low-survival prognosis ($P < .05$; Figure 5A). When applying the same cut point of ≥10 developed by CM40× to mitotic counts obtained using WSI20×, 6 (75%) of 8 pathologists successfully predicted prognostic outcome (Figure 5A). Example Kaplan-

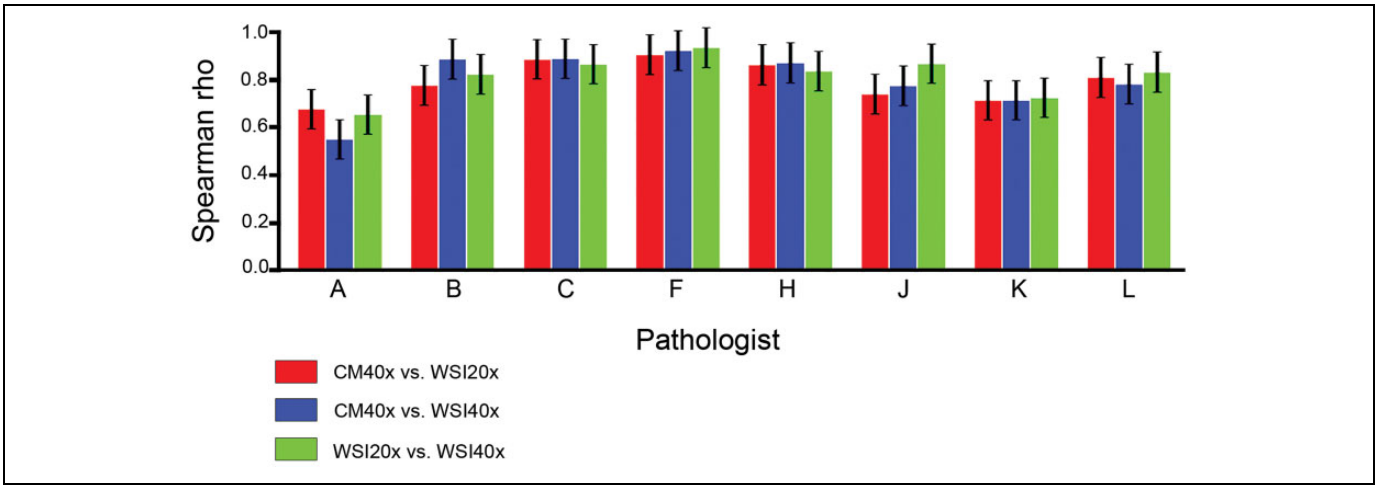

Figure 3. Intermodality agreement of mitotic activity assessment for each of 8 pathologists. Each bar indicated the Spearman correlation between each pair of modalities. Variable, but significant correlation ($P < 0.001$) was achieved under all microscopy modality comparisons by each pathologist, when assessing the same patients. The range of the intermodality correlations is (0.54, 0.93) and the mean correlation is 0.8, n = 37 to 38 for each comparison. Error bars represent standard errors of means.

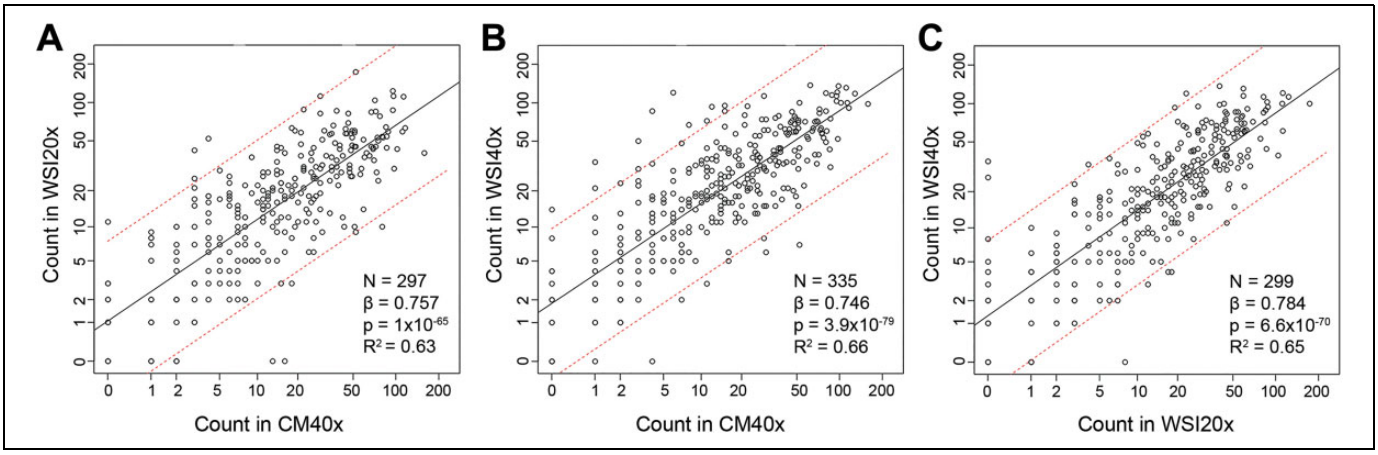

Figure 4. Intraobserver tumor mitotic figure counts, in paired comparisons of each of 3 modalities, are shown as scatter plot displays of regression analyses of the total MF counts (n +1) (log_{10} transformation). Each data point is an individual pathologist's total MF value from 10 FOV of the same patient using the 2 modalities. X and Y axes labeled for n, instead of n+1. For each comparison, (A) CM40× versus WSI20×, n = 271; (B) CM40× versus WSI40×, n = 312; and (C) WSI20× versus WSI40×, n = 278, there is a significant correlation between the modalities indicated (solid line of best fit, $P < 0.0001$). Dashed lines represent the 95% prediction band. Indicated slope (β) and R^2 values are similar for A, B, and C, and >0.63, indicating the strong linear relationship, supporting the interchangeability of modalities. CM indicates conventional microscopy; FOV, fields of view; MF, mitotic figure; WSI, whole slide image.

Meier survival analyses for a representative pathologist is shown (Figure 5B and C). Among the cases read by each observer with WSI40×, only up to 23 of these dogs had survival data available. This small data set was not sufficient to fully establish significant survival differences in WSI40× mode (data not shown). The findings supported evidence of clinically acceptable agreement in the prognostic utility of MF count values from WSI.

Discussion

Studies of diagnostic accuracy using WSI typically assess the ability of pathologists to agree on tissue diagnosis or IHC expression scoring.[24-29] The assessment of fine histological features, such as eukaryotic nuclear structures, cytoplasmic organelles, or microorganisms in tissue sections using WSI, has not been sufficiently authenticated in diagnostic practice. Although assessment of discrete histological features is inherent in the intermodality diagnostic concordance achieved in a number of studies by others,[24-29] enumerating MF provided an objective metric with which to judge intermodality performance agreement more precisely. In this manner, analyzing pathologist performance assessing mitotic activity in digital mode is critical to further establish the utility of digital pathology for primary diagnosis. To address this, we developed a

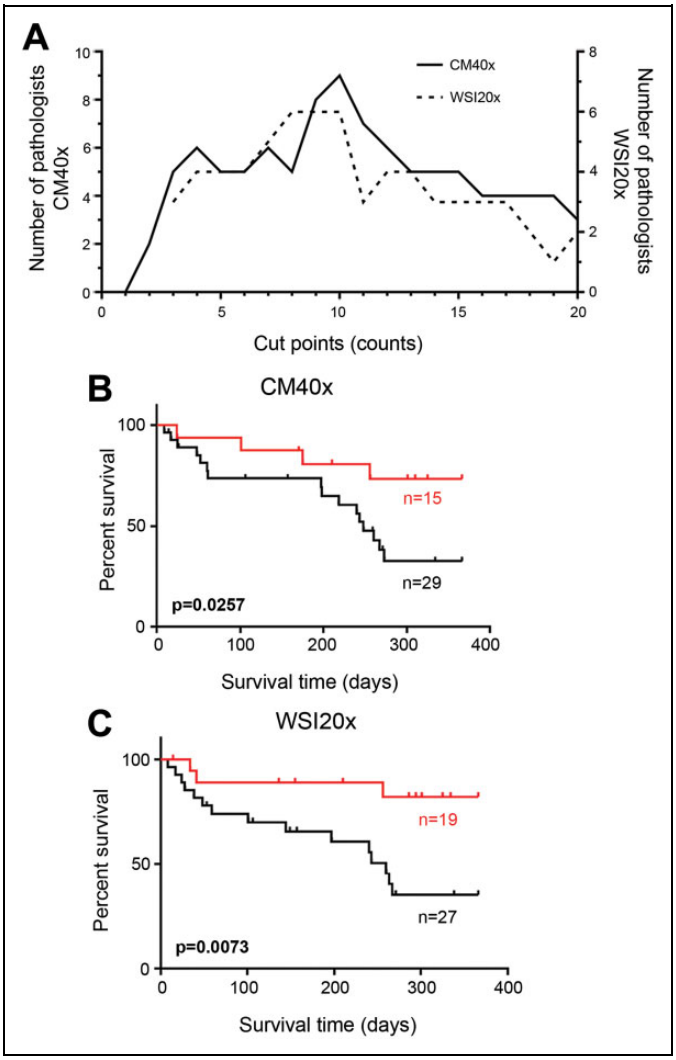

Figure 5. Utility of mitotic activity counting for estimating patient survival prognosis, established using CM, is transferable to WSI. A, Number of pathologists whose MF counts lead to an optimal prognostic prediction at each assumed cut point (x-axis) (CM40×, solid line; WSI20× dashed line). The cut point 10 MF, in 10 FOV established by CM (see also Supplemental Figure S5), was subsequently applied and validated for predicting survival differences using digital mode, for the same and different patients ($P < 0.05$). B, C Example Kaplan-Meier survival analyses for a representative pathologist displayed for reads obtained by conventional (B, CM40×) and digital microscopy (C, WSI20×) using the cut point count of 10 mitotic figures. MF ≥10 black line; MF <10 red line, n = number of patients. CM indicates conventional microscopy; FOV, fields of view; MF, mitotic figures; WSI, whole slide image.

clinical research framework including study design/execution, data collection protocols (hardware and software), and data analysis methods/software. Within this framework, we compared the proficiency of assessing and enumerating MF using 3 modalities to assess the interchangeability of digital WSI and CM in the routine diagnostic setting. The limited range of correlation coefficients amid mitotic activity assessments achieved was indicative of noninferiority among CM and WSI modalities.

These results were further augmented by finding 6 of 8 pathologists successfully predicted significant differences in patient survival prognosis using digital mode (WSI20×). Such performance in digital mode, grounded upon an MF count threshold developed with this patient population, was supportive of task accuracy and represents a first such comparison of clinical prognostic utility between CM and WSI. The extent of intermodality agreement achieved among pathologists evaluating mitotic activity by CM and WSI contributes to understanding the diagnostic interchangeability of the modalities for such a task. The multiobserver, multicase design facilitates extrapolation of these findings to other pathologists and other cases,[19] although it would be useful to confirm prognostic fitness with validation studies in additional tumor types and for other discrete histological features. The findings will be useful for future studies to continue validating the use of WSI to examine discrete histological feature details, such as MF and microorganisms, which are important components of primary diagnoses and patient management.

Detailed assessment of discrete histological features using WSI has been previously examined in related studies that suggested digital display may not be inferior to CM.[3,10] However, the approach was more controlled. In particular, MF counts in a previous breast cancer study were obtained by 3 pathologists who shared the same microscope for traditional MF counting on glass slide specimens.[3] Each observer used the same digital display monitor that was focused to a uniform WSI FOV for MF enumeration, using high-resolution WSI (WSI40×). Intraclass correlation coefficients were considered almost perfect by the authors (0.879 and 0.924, for microscopy and WSI, respectively.[3] Weighted κ coefficients ranged from 0.83 to 0.95 in another study evaluating intermodality detection of MF from melanomas and gastric biopsies.[10] Similarly, MF counting in preselected ROI on uniformly shared display monitors in a separate study also constituted the approach for melanoma mitotic activity enumeration.[30] The directed focus for pathologists in these studies likely contributed to the reproducibility achieved across conventional and digital modalities. Highly controlled methodologies in which the same FOV by microscope or digital display is employed are decidedly informative; however, these approaches are manifestly less reflective of the variety of routine diagnostic evaluations. During the current study by contrast, pathologists were required to independently localize mitotic hot spots from the entirety of the series of glass slide specimens and WSI, counting MF, and recording data successfully using microscopes and digital displays they manipulate daily. This approach was reflective of diverse clinical practice environments and provided a realistic setting to test the translatability of CM to diagnosis by digital display. In this regard, previous findings are insufficient alone to judge the interchangeability for primary diagnosis.[3,10,30] Consequently, the present study design, focused on interpretive locating, identifying, and quantifying MF as it is reduced to actual practice, is a particular strength.

Assessment of both MF and *Helicobacter species* microorganisms was the subject of another related study.[10] These

discrete microscopic features were detected in WSI using either a consumer grade or medical application display monitor. This finding has parallels in the present study, in which a variety of consumer display monitor characteristics were represented. While not formally controlled for, stipulation of minimum monitor size and use of standard color calibration for displays was considered to reduce some variability. Color calibration of monitors has been demonstrated to benefit pathologist performance, although the precise influence of digital display characteristics and color calibration on the quality of WSI review remains to be determined.[10,30] Effectively a multitude of factors, including display monitor characteristics, pathologist ability or experience, and resolution of specimen scans, can all impact the ability to make assessments of WSI.[1,2,31] Common WSI viewing software, standardized display color calibration, and uniform means of obtaining and electronically recording MF counts with the software application developed for this study (the latter used for both CM and digital WSI reads) controlled for some variables across settings. These features may have practical benefit employed in clinical practice. Furthermore, uniform viewing software, color calibration, and counting application were considered to contribute to the tallied MF counts recorded from the >20,000 FOV, as well as the reproducibility, in the present study.

Of particular note are the comparisons made with different WSI resolution, that is, optical scans of identical specimens obtained at both WSI20× and WSI40×. Digital pathology concordance studies have been conducted using a variety of resolutions corresponding to either 20× or 40× scans, and in some cases scan resolution is not clearly indicated.[27,28,32,33] Contention that WSI20× scans introduce problems with MF visualization[28] was not objectively or subjectively confirmed in the present study. By contrast, this study showed that MF count values from WSI40× were correlated with both CM40× and WSI20×. In addition, inter- and intraobserver rank-based concordance[22] across the 2 digital modes (WSI20× vs WSI40×) were found to be similar (0.794 and 0.798, respectively), findings that function to support the ability to interpret most MF at either WSI scan resolution in practice, and despite an inability to make focus adjustments of the WSI in z-axis direction. These observations indicate that certain aspects of tissue specimen examination conducted at lesser image projection/resolution on display monitors may not be adversely impacted from limiting scan resolution at image acquisition.

The interpretive nature of pathologist performance is inherently variable. The range of intramodality ρ correlation values in the visual assessment of mitotic activity documented among the different observers reinforced previously recognized degrees of uncertainty in counting MF microscopically.[34,35] Sources of MF count variability, both for individual pathologists and among pathologists, are several,[34,35] and do not appear to be entirely circumvented using WSI. In fact, variability among pathologists working in digital mode in the present study was evident to a greater degree than in CM. For example, greater correlation coefficients generally were demonstrated for the intramodality, interobserver comparisons by CM, as compared to WSI20× (Figure 2). These analyses were interpreted to indicate pathologists were somewhat more adept in CM. The finding was not surprising, as pathologists self-reported having limited digital pathology experience. The absence of z-axis focus was not considered disadvantageous in evaluating WSI by 4 of 6 responding pathologists, while the remaining 2 observers felt depth of focus might play a role in limiting MF identification, in some instances.

Regardless, the practice of enumerating MF in H&E-stained tissue sections by CM generally lacks consistency and can be conditional upon training and experience. For example, in a recent study, only 21 of 92 MF were unanimously identified as MF by all 5 participating pathologists examining the same 40 high-power microscopic FOV in bleached, H&E-stained melanomas.[36] In the current study, the same process for bleaching melanomas was not considered to produce an adverse impact on tissue review. Irrespective of such tissue treatment or not, it is recognized that false positives and negatives can result from a failure to distinguish MF from pyknotic nuclei, apoptotic bodies, or other distortions of chromatin pattern.[11] Furthermore, other limitations, such as cellular level variations in color, intensity, and morphological shape/size, can all contribute to counting variability. These collective circumstances result in an error-prone, tedious, and time-consuming task. The exercise can be poorly reproducible with interobserver variability resulting in discordant inter- and intraobserver mitotic count values.[34,35,37]

In addition to substantial lack of consensus on what constitutes an individual MF, the degree of agreement on the specific area in the tissue chosen by observers for MF mitotic hot spot location, and its potential impact on examining prognostic mitotic count thresholds, is not well understood. Thus, most perceptible sources of discordance appear to more likely be a greater function of interpretive disparities, rather than due to intermodality properties; or at the least, such disparities may be responsible for confounding cross-technology performance issues in comparison studies. Academic training programs can view these as ongoing challenges to study further and address in preparation for routine adoption of examination using digital display. Example approaches for continued investigation of visual evaluations in digital mode can be partly exemplified in the mitotic hot spot ROI image annotation overlays (unpublished data), and the ongoing Evaluation Environment for Digital and Analog Pathology project organized with several collaborators (https://nciphub.org/groups/eedapstudies, accessed April 24, 2019). Proficiencies learned in examining WSI microscopically by digital display will permit more seamless task integration with future computational pathology and informatics tools, along with clinical data streams from other sources, in computer-assisted diagnostics.

How validation assessments of discrete histological features in digital WSI are conducted can be important to academic programs in evaluating the best means to train pathologists in digital pathology for primary diagnosis. The extent of replicate observations obtained from diverse diagnostic settings can be instrumental in performance assessments. These findings,

derived in the context of a wide dynamic range of MF count values in mucosal melanoma, contributed to further confidence in the agreement estimates in this study.[38] Notwithstanding this level of agreement, it is noteworthy that evidence, such as variability in making ROI annotations by some pathologists in digital mode, supports a previously published assertion that specific training in digital pathology would stand to enhance performance.[32] Further investigation is also necessary to help determine a most appropriate reading environment (eg, display resolution, size, and calibration) in an effort to further achieve improved interobserver concordances. In addition, technology appears to have an expanded role to play in improving accuracy for such tasks. In the current study, pathologists welcomed the use of the on-screen counting application for recording counts as a means to improve consistency in recording and transcribing. Furthermore, technology, such as automated computer-assisted mitotic hot spot mapping decision support, is emerging to more efficiently address the impediments or challenges in achieving mitotic hot spot topographic location consensus among pathologists.[39]

Authors' Note

This work utilized the computational resources of the NIH High Performance Computing Biowulf cluster and the NCI Laboratory of Cancer Biology and Genetics High-Dimension Data Analysis Group. Specimens and patient survival data were provided as a consequence of collaborations with Drs Michael Goldschmidt, University of Pennsylvania, and EJ Ehrhart, Colorado State University, in addition to those from coauthors. The research represents a work of the US Government; however, statements herein are the judgments and opinions of the authors and do not necessarily represent official US Government policies. The mention of commercial entities, or commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such entities or products by the US Department of Health and Human Services or its constituent agencies.

Halsey is now with Pfizer Drug Safety Research and Development, Groton, CT, USA. Sánchez is now with Antech Diagnostics, Framingham, MA, USA. Traslavina is now with Antech Diagnostics, Lacey, WA, and the Comparative Pathology Branch, US Army Medical Research Institute of Chemical Defense, Aberdeen Proving Ground, MD, USA. Cramer is now with Tox Path Specialists, LLC., subsidiary of StageBio, Frederick, MD, USA. Lyle is now with Department of Comparative Pathobiology, Purdue University, West Lafayette, IN, USA. Buza is now with Gene Therapy Program, Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

Acknowledgments

The authors thank Dr David W. Gardiner, Animal Reference Pathology, Inc, for advice on study performance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Intramural Research Program, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland. Support was also provided through the NIH Comparative Biomedical Scientist Training Program. Qi Gong was supported by fellowship administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and the US Food and Drug Administration.

ORCID iD

Brandon D. Gallas https://orcid.org/0000-0001-7332-1620
R. Mark Simpson https://orcid.org/0000-0001-8852-1297

Supplemental Material

Supplemental material for this article is available online.

References

1. Pantanowitz L, Sinard JH, Henricks WH, et al. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med*. 2013;137:1710-1722.
2. Jansen I, Lucas M, Savci-Heijink CD, et al. Histopathology: ditch the slides, because digital and 3D are on show. *World J Urol*. 2018;36:549-555.
3. Al-Janabi S, van Slooten HJ, Visser M, van der Ploeg T, van Diest PJ, Jiwa M. Evaluation of mitotic activity index in breast cancer using whole slide digital images. *PLoS One*. 2013;8:e82576.
4. Williams BJ, Hanby A, Millican-Slater R, Nijhawan A, Verghese E, Treanor D. Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study on digital pathology validation and training. *Histopathology*. 2018;72:662-671.
5. Mukhopadhyay S, Feldman MD, Abels E, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (Pivotal Study). *Am J Surg Pathol*. 2018;42:39-52.
6. Goacher E, Randell R, Williams B, Treanor D. The diagnostic concordance of whole slide imaging and light microscopy: a systematic review. *Arch Pathol Lab Med*. 2017;141:151-161.
7. Krishnamurthy S, Mathews K, McClure S, et al. Multi-institutional comparison of whole slide digital imaging and optical microscopy for interpretation of hematoxylin-eosin-stained breast tissue sections. *Arch Pathol Lab Med*. 2013;137:1733-1739.
8. Gavrielides MA, Conway C, O'Flaherty N, Gallas BD, Hewitt SM. Observer performance in the use of digital and optical microscopy for the interpretation of tissue-based biomarkers. *Anal Cell Pathol (Amst)*. 2014;2014:157308.
9. Nunes C, Rocha R, Buzelin M, et al. High agreement between whole slide imaging and optical microscopy for assessment of HER2 expression in breast cancer: whole slide imaging for the assessment of HER2 expression. *Pathol Res Pract*. 2014;210:713-718.

10. Norgan AP, Suman VJ, Brown CL, Flotte TJ, Mounajjed T. Comparison of a medical-grade monitor vs commercial off-the-shelf display for mitotic figure enumeration and small object (*Helicobacter pylori*) detection. *Am J Clin Pathol*. 2018;149:181-185.
11. Bonert M, Tate AJ. Mitotic counts in breast cancer should be standardized with a uniform sample area. *Biomed Eng Online*. 2017;16:28.
12. Olar A, Wani KM, Sulman EP, et al. Mitotic index is an independent predictor of recurrence-free survival in meningioma. *Brain Pathol*. 2015;25:266-275.
13. Donizy P, Kaczorowski M, Leskiewicz M, et al. Mitotic rate is a more reliable unfavorable prognosticator than ulceration for early cutaneous melanoma: a 5-year survival analysis. *Oncol Rep*. 2014;32:2735-2743.
14. Coindre JM. Grading of soft tissue sarcomas: review and update. *Arch Pathol Lab Med*. 2006;130:1448-1453.
15. Azzola MF, Shaw HM, Thompson JF, et al. Tumor mitotic rate is a more powerful prognostic indicator than ulceration in patients with primary cutaneous melanoma: an analysis of 3661 patients from a single center. *Cancer*. 2003;97:1488-1498.
16. Gillard M, Cadieu E, De Brito C, et al. Naturally occurring melanomas in dogs as models for non-UV pathways of human melanomas. *Pigment Cell Melanoma Res*. 2014;27:90-102.
17. Hernandez B, Adissu HA, Wei BR, Michael HT, Merlino G, Simpson RM. Naturally occurring canine melanoma as a predictive comparative oncology model for human mucosal and other triple wild-type melanomas. *Int J Mol Sci*. 2018;19:E394.
18. Simpson RM, Bastian BC, Michael HT, et al. Sporadic naturally occurring melanoma in dogs as a preclinical model for human melanoma. *Pigment Cell Melanoma Res*. 2014;27:37-47.
19. Chen W, Gong Q, Gallas BD. Paired split-plot designs of multireader multicase studies. *J Med Imaging (Bellingham)*. 2018;5:031410.
20. Obuchowski NA, Gallas BD, Hillis SL. Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. *Acad Radiol*. 2012;19:1508-1517.
21. Smedley RC, Spangler WL, Esplin DG, et al. Prognostic markers for canine melanocytic neoplasms: a comparative review of the literature and goals for future investigation. *Vet Pathol*. 2011;48:54-72.
22. Kim J-O. Predictive measures of ordinal association. *Am J Sociol*. 1971;76:891-907.
23. Gershenwald JE, Scolyer RA. Melanoma staging: American Joint Committee on Cancer (AJCC) 8th edition and beyond. *Ann Surg Oncol*. 2018;25:2105-2110.
24. Wang H, Sima CS, Beasley MB, et al. Classification of thymic epithelial neoplasms is still a challenge to thoracic pathologists: a reproducibility study using digital microscopy. *Arch Pathol Lab Med*. 2014;138:658-663.
25. Rodriguez-Urrego PA, Cronin AM, Al-Ahmadie HA, et al. Interobserver and intraobserver reproducibility in digital and routine microscopic assessment of prostate needle biopsies. *Hum Pathol*. 2011;42:68-74.
26. Jen KY, Olson JL, Brodsky S, Zhou XJ, Nadasdy T, Laszik ZG. Reliability of whole slide images as a diagnostic modality for renal allograft biopsies. *Hum Pathol*. 2013;44:888-894.
27. Al-Janabi S, Huisman A, Vink A, et al. Whole slide images for primary diagnostics of gastrointestinal tract pathology: a feasibility study. *Hum Pathol*. 2012;43:702-707.
28. Campbell WS, Lele SM, West WW, Lazenby AJ, Smith LM, Hinrichs SH. Concordance between whole-slide imaging and light microscopy for routine surgical pathology. *Hum Pathol*. 2012;43:1739-1744.
29. Rizzardi AE, Johnson AT, Vogel RI, et al. Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. *Diagn Pathol*. 2012;7:42.
30. Krupinski EA, Silverstein LD, Hashmi SF, Graham AR, Weinstein RS, Roehrig H. Observer performance using virtual pathology slides: impact of LCD color reproduction accuracy. *J Digit Imaging*. 2012;25:738-743.
31. Malarkey DE, Willson GA, Willson CJ, et al. Utilizing whole slide images for pathology peer review and working groups. *Toxicol Pathol*. 2015;43:1149-1157.
32. Nielsen PS, Lindebjerg J, Rasmussen J, Starklint H, Waldstrom M, Nielsen B. Virtual microscopy: an evaluation of its validity and diagnostic performance in routine histologic diagnosis of skin tumors. *Hum Pathol*. 2010;41:1770-1776.
33. Bertram CA, Gurtner C, Dettwiler M, et al. Validation of digital microscopy compared with light microscopy for the diagnosis of canine cutaneous tumors. *Vet Pathol*. 2018;55:490-500.
34. Ikenberg K, Pfaltz M, Rakozy C, Kempf W. Immunohistochemical dual staining as an adjunct in assessment of mitotic activity in melanoma. *J Cutan Pathol*. 2012;39:324-330.
35. Voss SM, Riley MP, Lokhandwala PM, Wang M, Yang Z. Mitotic count by phosphohistone H3 immunohistochemical staining predicts survival and improves interobserver reproducibility in well-differentiated neuroendocrine tumors of the pancreas. *Am J Surg Pathol*. 2015;39:13-24.
36. Gallas BD, Jamal B, Qi G, et al. *A Reader Study on a 14-Head Microscope, Pathology Informatics Summit*. Pittsburgh, PA. 2018. https://nciphub.org/groups/eedapstudies/wiki/Presentation:AReaderStudyona14headMicroscope. Accessed December 10, 2018.
37. Rezanko T, Akkalp AK, Tunakan M, Sari AA. MIB-1 counting methods in meningiomas and agreement among pathologists. *Anal Quant Cytol Histol*. 2008;30:47-52.
38. Preiss D, Fisher J. A measure of confidence in Bland-Altman analysis for the interchangeability of two methods of measurement. *J Clin Monit Comput*. 2008;22:257-259.
39. Puri M, Hoover SB, Hewitt SM, et al. Automated computational detection, quantitation, and mapping of mitosis in whole-slide images for clinically actionable surgical pathology decision support. *J Pathol Inform*. 2019;10:4.