OPEN

Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer

David F. Steiner, MD, PhD, Robert MacDonald, PhD,* Yun Liu, PhD,* Peter Truszkowski, MD,* Jason D. Hipp, MD, PhD, FCAP,* Christopher Gammage, MS,* Florence Thng, MS,† Lily Peng, MD, PhD,* and Martin C. Stumpe, PhD**

Abstract: Advances in the quality of whole-slide images have set the stage for the clinical use of digital images in anatomic pathology. Along with advances in computer image analysis, this raises the possibility for computer-assisted diagnostics in pathology to improve histopathologic interpretation and clinical care. To evaluate the potential impact of digital assistance on interpretation of digitized slides, we conducted a multireader multicase study utilizing our deep learning algorithm for the detection of breast cancer metastasis in lymph nodes. Six pathologists reviewed 70 digitized slides from lymph node sections in 2 reader modes, unassisted and assisted, with a washout period between sessions. In the assisted mode, the deep learning algorithm was used to identify and outline regions with high likelihood of containing tumor. Algorithm-assisted pathologists demonstrated higher accuracy than either the algorithm or the pathologist alone. In particular, algorithm assistance significantly increased the sensitivity of detection for micrometastases (91% vs. 83%, $P = 0.02$). In addition, average review time per image was significantly shorter with assistance than without assistance for both micrometastases (61 vs. 116 s, $P = 0.002$) and negative images (111 vs. 137 s, $P = 0.018$). Lastly, pathologists were asked to provide a numeric score regarding the difficulty of each image classification. On the basis of this score, pathologists considered the image review of micrometastases to be significantly easier when interpreted with assistance ($P = 0.0005$). Utilizing a proof of concept assistant tool, this study demonstrates the potential of a deep learning algorithm to improve pathologist accuracy and efficiency in a digital pathology workflow.

Key Words: artificial intelligence, machine learning, digital pathology, breast cancer, computer aided detection

(*Am J Surg Pathol* 2018;00:000–000)

The regulatory approval and gradual implementation of whole-slide scanners has enabled the digitization of glass slides for remote consults and archival purposes.[1] Digitization alone, however, does not necessarily improve the consistency or efficiency of a pathologist's primary workflow. In fact, image review on a digital medium can be slightly slower than on glass, especially for pathologists with limited digital pathology experience.[2] However, digital pathology and image analysis tools have already demonstrated potential benefits, including the potential to reduce inter-reader variability in the evaluation of breast cancer HER2 status.[3,4] Digitization also opens the door for assistive tools based on Artificial Intelligence (AI) to improve efficiency and consistency, decrease fatigue, and increase accuracy.[5]

Among AI technologies, deep learning has demonstrated strong performance in many automated image-recognition applications.[6–8] Recently, several deep learning–based algorithms have been developed for the detection of breast cancer metastases in lymph nodes as well as for other applications in pathology.[9,10] Initial findings suggest that some algorithms can even exceed a pathologist's sensitivity for detecting individual cancer foci in digital images. However, this sensitivity gain comes at the cost of increased false positives, potentially limiting the utility of such algorithms for automated clinical use.[11] In addition, deep learning algorithms are inherently limited to the task for which they have been specifically trained. While we have begun to understand the strengths of these algorithms (such as exhaustive search) and their weaknesses (sensitivity to poor optical focus, tumor mimics; manuscript under review), the potential clinical utility of such algorithms has not been thoroughly examined. While an accurate algorithm alone will not necessarily aid pathologists or improve clinical interpretation, these benefits may be achieved through thoughtful and appropriate integration of algorithm predictions into the clinical workflow.[8]

From the *Google AI Healthcare; and †Verily Life Sciences, Mountain View, CA.

D.F.S., R.M., and Y.L. are co-first authors (equal contribution).

Work done as part of the Google Brain Healthcare Technology Fellowship (D.F.S. and P.T.).

Conflicts of Interest and Source of Funding: D.F.S., R.M., Y.L., P.T., J.D.H., C.G., F.T., L.P., M.C.S. are employees of Alphabet and have Alphabet stock.

Correspondence: David F. Steiner, MD, PhD, Google AI Healthcare, 1600 Amphitheatre Way, Mountain View, CA 94043 (e-mail: davesteiner@google.com).

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.ajsp.com.

In this paper, we present findings from a multireader multicase study of pathologists using a state-of-the-art algorithm,[9] LYmph Node Assistant (LYNA), to review hematoxylin and eosin (H&E)-stained lymph node images for breast cancer metastasis. Although lymph node review is crucial for staging and therapy decisions,[12] these reviews can be both time consuming and mentally fatiguing. This is due to the small size of some metastases as well as morphologic similarities between metastatic cells and elements of normal lymph node histology (such as sinusoidal tissue and histiocytes). Surprisingly, a recent analysis of pathologists performing this task with glass slides under simulated time constraints reported a mean sensitivity of only 38% for detecting micrometastases while using H&E sections alone.[10] Other data also suggest suboptimal sensitivity and inter-reader variability for identifying micrometastases in sentinel lymph node biopsies, with retrospective studies identifying occult metastases in ~10% to 15% of cases and changes in nodal status upon expert review in 24% of cases.[13–15] Although immunohistochemistry (IHC) is often utilized to improve diagnostic sensitivity, such stains are costly and time consuming, and institution-specific practices for the use of IHC vary considerably.

Combining the sensitivity of assistive algorithms with the specificity and expertise of pathologists has the exciting potential to improve the efficiency, accuracy, and consistency of manual reads, particularly in a time-limited clinical setting. The goal of this study was to better understand design elements[16] and potential benefits of a computer assistance tool in pathology. To this end, we conducted a reader study to evaluate the accuracy and speed of pathologists performing assisted reads of lymph nodes for metastatic breast cancer. Our findings demonstrate that use of LYNA to indicate regions of interest during pathologist review can result in significant time savings, particularly for slides containing micrometastases. With appropriate design and implementation, we believe the benefits of an assistive tool such as this may also be applicable to other cancer types (such as colon and lung) for which there is a similar need for accurate and efficient lymph node review.

MATERIALS AND METHODS

Case Enrollment and Validation

Study Images

We obtained archived lymph node tissue blocks from 2 independent clinical partners. All blocks were deidentified and originated from expired clinical archives according to the 10 year CAP requirement.[17] These formalin-fixed paraffin-embedded blocks were cut onto glass slides, and both H&E and IHC (panCK, AE1/AE3) slides were made and digitized (details below). On the basis of review pace and reader fatigue in a pilot study (Supplemental Digital Content 1, http://links.lww.com/PAS/A677), we estimated that 60 to 80 images could be reviewed in a scheduled 3-hour session at a pace comparable to normal clinical review and still allow time for training images, instructions, and breaks as needed. In order to allow statistically adequate evaluation of micrometastasis cases while maintaining a feasible total number of images, the final image set was selected with moderate enrichment for micrometastases. On the basis of deidentified diagnostic pathology reports for the specimens available, we selected 70 images consisting of negative cases, micrometastases, and macrometastases to be further evaluated for reference standard classification. These 70 images were from 67 unique tissue blocks across 50 distinct cases. To verify that the images from the 3 blocks with 2 sections from the same block did not bias our findings, we repeated the analysis after excluding the second image from these blocks to confirm that the findings (such as the results of statistical comparisons) did not change meaningfully.

Reference Standard Classification

To establish a reference standard for images used in this study, digital images of both H&E-stained and IHC-stained slides were reviewed independently by 3 US board certified pathologists having a minimum of 7 years posttraining anatomic pathology experience. Whole-slide level findings were categorized by each reference standard pathologist as negative, micrometastasis (tumor >0.2 mm and ≤ 2 mm), macrometastasis (tumor >2 mm), or isolated tumor cell clusters (ITCs; ≤ 200 scattered tumor cells or tumor cluster ≤ 0.2 mm) based on the largest tumor deposit identified and according to current clinical guidelines.[12,18] All negative cases were also subject to IHC review to confirm the negative classification. The reference standard for discrepant cases was established through additional, adjudicated review and tumor-size measurement using the H&E and IHC images.

Reader Study

Pathologists

A total of 6 US board certified anatomic pathologists from 3 practices participated as readers in this study. These pathologists did not participate in the reference standard classification. Their years of attending pathologist experience ranged from 1 to 15. None had a specialization in breast pathology, and all pathologists self-reported a broad anatomic pathology clinical practice that included review of lymph node specimens from breast cancer cases. None of the pathologists were using digital pathology in routine clinical practice.

Reader Training

To establish familiarity with the digital image viewer and the assistant tool, each reader session began with a review of 5 training images that were not part of the 70 study images. Study administrators also clarified any questions about the functionality of the viewer and the digital assistant tool at this stage. Notably, training images included a mix of examples with both moderate confidence regions highlighted in green and high confidence regions highlighted in cyan (see the "Digital assistant design" section below). An explanation of these 2 categories of regions was provided to pathologists during training image review. In addition, 5 of the 6 pathologists had also previously participated in the prior pilot study that utilized an independent set of images.

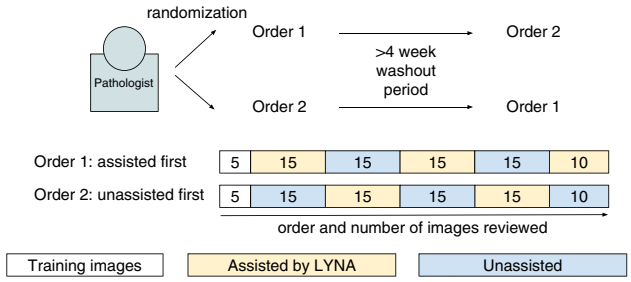

FIGURE 1. Schematic of reader study design. Readers review the same images in the same sequence, but with different modes: algorithm assisted, or unassisted. Readers are randomized to one of the 2 assistance "orders." Each rectangle indicates a set of images; the color of the rectangle indicates the mode (assisted or unassisted), and the number in the rectangle indicates the number of images in that set. Readers reviewed a total of 5 images for familiarization and 70 images for formal review.

Diagnostic Image Classification

Pathologists were instructed to classify each image as negative, ITC, micrometastasis, or macrometastasis based on the largest metastasis identified and without additional information such as deeper levels, IHC, or review of the primary tumor specimen. A digital ruler was available as a feature within the image viewer interface to allow precise measurement of tumor size as needed. Given the limited prognostic significance of isolated tumor cells in sentinel lymph nodes from invasive breast cancer cases[13,14] and because ITCs are excluded from the total positive node count for breast cancer nodal staging,[12] pathologists were not evaluated for detection of tumor cells in ITC cases. For micrometastasis cases, a classification of ITC was considered a false negative. If a pathologist initially reported an equivocal interpretation, they were prompted by the study administrator to select the most likely diagnostic category given the information available.

Study Design

To evaluate performance metrics for both assisted and unassisted reads, the study was designed as a fully crossed, intermodal, multireader study as illustrated in Figure 1. Pathologists interpreted all study images in both modalities (with assistance or without, as illustrated in Fig. 2) in 2 sessions separated by a wash-out period of at least 4 weeks and with ongoing full time clinical practice in the interim. To mitigate bias for possible performance differences at the beginning versus the end of a given session, the complete set of images was divided into blocks of 10 or 15 images, with each block containing a similar distribution of case types. In addition, to reduce possible biases due to seeing an image with assistance in the first session versus the second session, the 6 pathologists were randomized into 2 groups and began the first session either with assistance (mode 1) or without assistance (mode 2).

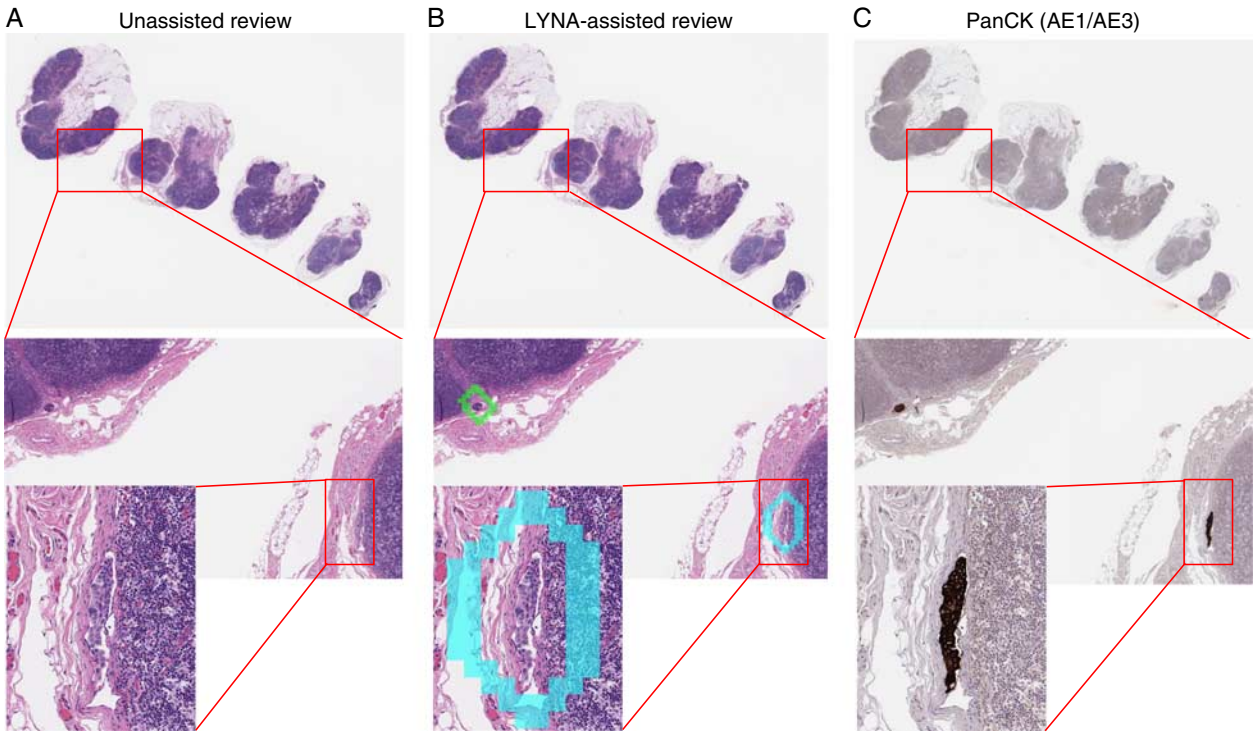

FIGURE 2. Digital image assistance by presenting algorithm predictions as a direct overlay; without assistance (A) versus with assistance (B). Suspicious regions of interest are highlighted in cyan for high confidence and green for moderate confidence (based on the algorithm predictions). In this image, both the high and moderate confidence regions of interests are confirmed as metastatic tumor as indicated by cytokeratin IHC (C). Another image example is shown in Supplemental Figure 2 (Supplemental Digital Content 1, http://links.lww.com/PAS/A677).

In either mode, the order and specific images reviewed were identical; the difference was solely in modality (assisted or not assisted). Ethics review and Institutional Review Board (IRB) exemption was obtained from Quorum Review IRB (Seattle, WA).

Image Review Timing

To simulate clinical review as much as possible despite the constraints of this study, a study administrator instructed pathologists to evaluate all images with an approach and pace as similar to normal clinical workflow as possible. To further encourage review of cases at a realistic pace, pathologists were not told exactly how many cases would be evaluated, but instead were asked to review as many images as possible in the time provided with a goal of at least 60 images. For all cases, the time from opening the image in the viewer to final verbal classification was recorded by the study administrator observing the study.

Obviousness Score

In order to quantify the subjective difficulty of the classification task for each image, pathologists were requested to provide an "obviousness score" from 0 to 100 for each image based on how easily they were able to reach a final diagnosis. To provide additional context for this score, and to encourage a quantitatively useful distribution, readers were instructed to consider this value as a ranking, relative to the last 100 breast cancer lymph node cases reviewed, with 100 representing the easiest case, and 0 representing the most challenging. High scores thus correspond to cases for which the final classification was relatively obvious or easy, and low scores correspond to cases for which the diagnosis was challenging or less obvious. Pathologists were instructed that this score should represent their overall experience with each image, and thus may involve multiple different components including, but not limited to, amount of tissue area to review, image quality, artifact, histomorphologic features of tumor cells, and location of tumor cells.

Slide Preparation and Slide Scanning

Selected tissue blocks were processed by 2 commercial, CAP/CLIA certified histology laboratories. Sections were cut from each block and processed using standard H&E and IHC staining protocols. IHC stains were performed with pan-cytokeratin antibody (AE1/AE3). Before scanning, slides were manually inspected for defects such as air bubbles, and any substandard slides were rejected. Slides that passed the quality check were scanned with a Leica AT2 system at a resolution of 0.25 μm/pixel. Resulting whole-slide images (WSI's) were manually reviewed for overall image quality. During this quality control process, images containing defects such as missing tissue or large out of focus regions were rescanned.

Equipment Used for Viewing

Images were presented on an 27″ Healthcare LED monitor (HP HC270) through an internet browser based image viewer running on a desktop computer. Navigation across the WSI was exclusively translation and zoom, controlled using a standard computer mouse.

Digital Assistant Development and Deep Learning Algorithm

We developed the deep learning algorithm, LYNA, as described by Liu and colleagues using the Camelyon16 challenge data set.[9,10] Briefly, we trained a small version of the Inception V3 deep learning architecture on WSI's of digitized pathology slides, partitioned into 32×32 μm tiles, and their corresponding labels indicating whether that tissue area contained tumor (0 or 1). By adjusting the weights of the deep learning algorithm to reduce the difference between its predicted output and the label, the algorithm gradually learned to distinguish normal from tumor image tiles. We then used the algorithm to predict for each tissue area, the likelihood of it containing tumor. This generated a 2D "heatmap" for each slide indicating the likely areas of tumor. We then created a list of predicted tumor regions using this heatmap (more details provided in the Supplemental Digital Content 1, http://links.lww.com/PAS/A677, "Methods" section).

Digital Assistant Design

We evaluated several approaches to display algorithm predictions and found that an outline of predicted tumor regions presented as a direct overlay on the H&E image provided a more intuitive interface than presenting a likelihood heatmap for the full image in an adjacent window (Supplemental Fig. 1, Supplemental Digital Content 1, http://links.lww.com/PAS/A677). For the overlay, we expanded the outline by 32 μm beyond the actual boundary predicted by LYNA to avoid obscuring the tumor-benign boundary in the underlying image. To help users prioritize regions for review, we categorized the regions into high confidence (high specificity) and moderate confidence (high sensitivity) regions based on the prediction output of LYNA. These 2 prediction categories were outlined in cyan and green, respectively (Fig. 2 and Supplemental Fig. 2, Supplemental Digital Content 1, http://links.lww.com/PAS/A677). These colors were chosen to stand out from the background pink and purple of H&E-stained tissue.

In our work, we observed that if a slide contained multiple false-positive regions, these regions were often of the same type of morphologic or histologic feature, such as out of focus macrophages, histiocytes, giant cells or the dark zone of germinal centers.[9] True positive metastases in these slides, if present, typically had a predicted output higher than the vast majority of false positives in the same slides. To reduce reader distractions and alert-fatigue caused by potential false positives we limited the number of moderate confidence regions outlined to those with the highest algorithm scores. If fewer than 3 high confidence regions were present, the total number of high and moderate confidence regions combined was limited to 3. If there were 3 or more high confidence regions, all of those regions were displayed and no moderate confidence regions were displayed (Supplemental Table 1, Supplementary Digital Content 1, http://links.lww.com/PAS/A677).

TABLE 1. Case Composition

Category	No. Images (n [%])
Negative	24 (34)
Isolated tumor cells	8 (11)
Micrometastasis	19 (27)
Macrometastasis	19 (27)

Statistical Analysis

Measurements of time, accuracy, and numeric score between assisted and unassisted modes were analyzed using mixed-effects models. Models were generated with pathologists and images treated as random effects and the assistance modality and session (mode 1 or mode 2) treated as fixed effects. For statistical significance evaluation, *P*-values were obtained using the Likelihood Ratio Test with the anova function in R, comparing the full model to a null model with the fixed effect of interest (eg, assistance mode) removed. For statistical significance of accuracy, binomial mixed-effect models across all observations were generated using the glmer function. All models were generated using the lme4 package in R and each category (eg, negative or micrometastases) was modeled separately.

RESULTS

Reference Standard and Study Images

All study images were reviewed by 3 reference standard pathologists as described in the "Materials and methods" section above. For all slides, digitized images of both H&E-satined and IHC-stained sections were reviewed to establish the reference standard classification. The final category classification of images used in this study is summarized in Table 1. Three cases had initially discrepant classifications among reference standard pathologists, all of which involved the identification of tumor foci near the border of 2 size categories. These discrepancies were resolved easily with adjudicated measurement or tumor cell counts (additional details in Supplemental Table 2, Supplementary Digital Content 1, http://links.lww.com/PAS/A677).

Classification Accuracy

The overall mean sensitivity for detection of metastases by the pathologists in this study was 94.6% (95% confidence interval [CI], 93.4%-95.8%). To evaluate the impact of the assisted read on accuracy, we analyzed performance by case category and assistance modality (Fig. 3A). For micrometastases, sensitivity was significantly higher with

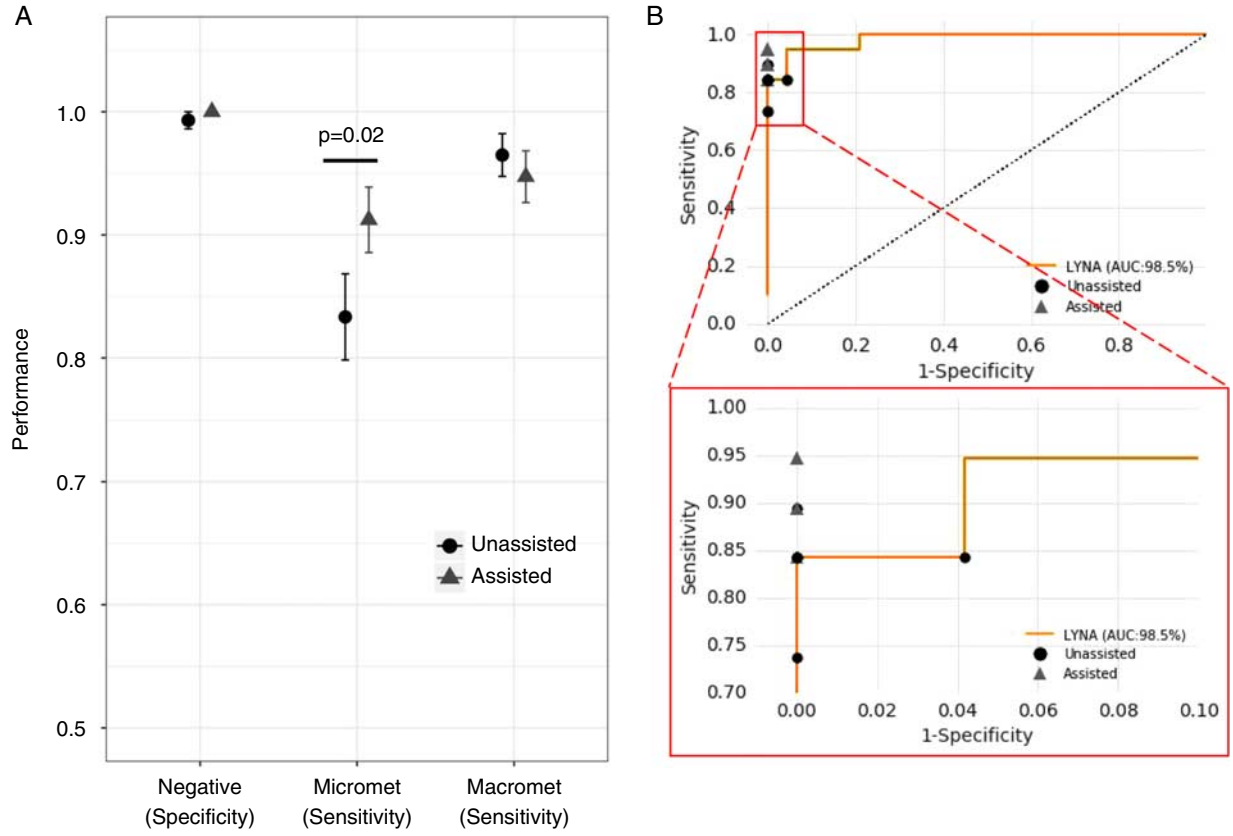

FIGURE 3. Improved metastasis detection with algorithm assistance. A, Data represents performance across all images by image category and assistance modality. Error bars indicate SE. The performance metric corresponds to corresponds to specificity for negative cases and sensitivity for micrometastases (micromet) and macrometastases (macromet). B, Operating point of individual pathologists with and without assistance for micrometastases and negative cases, overlayed on the receiver operating characteristic curve of the algorithm. AUC indicates area under the curve.

computer assistance than without, with a mean sensitivity across images of 91.2 (95% CI, 86.0%-96.5%) for assisted reads and 83.3 (95% CI, 76.4%-90.2%) for unassisted reads ($P = 0.023$). In addition, assistance resulted in increased sensitivity for the detection of micrometastases for 5 of 6 pathologists in this study (Supplemental Fig. 3, Supplemental Digital Content 1, http://links.lww.com/PAS/A677). Removing the single micrometastasis image for which another section from the same block was presented earlier in the sequence of study images did not meaningfully change this result ($P = 0.035$ for increased sensitivity of micrometastases with assistance vs. without assistance). No significant difference in accuracy between assistance modes was observed for negative images or macrometastases (Fig. 3A).

Sensitivity and specificity tradeoffs are often evaluated by generating receiver operating characteristic curves for readers. However, the specificity of assisted reads in this study was 100%; the only false-positive interpretation corresponded to an unassisted image review. In this unassisted false positive, the pathologist self-reported low confidence for their interpretation and verbally indicated that they would order IHC, but ultimately, given the requirement in this study to give an interpretation using H&E alone, the pathologist favored involvement by metastatic disease. With nearly perfect specificity and thus no performance tradeoff to evaluate with a receiver operating characteristic, we focused our analysis on sensitivity.

The overall performance of the LYNA algorithm has been described previously.[9] On this set of previously unevaluated images, the overall area under the curve was 99.0%, and when including only micrometastases and negative images, the area under the curve was similar at 98.5%. Although some pathologists in the unassisted mode were less sensitive than LYNA, all pathologists performed better than the algorithm alone in regards to both sensitivity and specificity when reviewing images with assistance (Fig. 3B).

One micrometastasis image was interpreted as negative by all 6 pathologists, both with and without assistance. This image was one of the 3 images that required reference standard adjudication and was ultimately labeled as a micrometastasis due to the presence of >200 individual tumor cells, albeit without a contiguous tumor deposit (Fig. 4A). The algorithm did highlight tumor cells in this image and many pathologists commented that they observed ITCs or that they would request IHC, but none reported a final interpretation of micrometastasis. In addition, all of the false negatives in the macrometastasis category corresponded to a single image and this image was misclassified as negative by all 6 pathologists in the assisted mode and by 4 pathologists in the unassisted mode. Interestingly, the algorithm did highlight tumor on this image with the moderate confidence outline, but was still interpreted as negative by all 6 pathologists in this study with assistance (Fig. 4B). As ITCs are excluded from the total positive node count for nodal stage categorization and their clinical significance is less well established, the small number of ITC cases were not included in the accuracy analysis.

Although there were no false positive final interpretations with assistance, 7 of the 24 negative images did have moderate confidence prediction outlines and one negative image had a high confidence outline. The non-metastatic histologic features most commonly flagged by the algorithm involved histiocytes, giant cells or the dark zone of germinal centers and the high confidence prediction outlined fat necrosis (Fig. 4C), all of which were correctly identified as benign by the pathologists in this study.

Image Review Efficiency

Average time of review per image was significantly shorter with assistance than without assistance for both micrometastases ($P = 0.002$) and negative images ($P = 0.018$); (Table 2 and Fig. 5A). This time benefit was especially pronounced for micrometastases with nearly every image reviewed faster on average with assistance than without (Fig. 5B). For micrometastases, the mean time of review per image with assistance was 61 versus 117 seconds without assistance. The mean time of review per image for negative images with assistance was 111 versus 136 seconds without assistance. Similar to classification accuracy, the removal of 3 images from blocks with 2 sections did not meaningfully change these results (<2 s change in the average review times; negative $P = 0.021$, micrometastasis $P = 0.002$ for time per image difference between assisted and unassisted reviews). For ITCs and macrometastases, review times per image were slightly shorter on average with assistance, but these differences were not statistically significant. CIs and a summary of these results are reported in Table 2 and visualized in Supplemental Figure 4 (Supplemental Digital Content 1, http://links.lww.com/PAS/A677).

We further compared the change in review time between assistance modalities for each individual image across pathologists. Two individual images with macrometastases showed large time differences between assistance modes: one image was reviewed faster with assistance and one image was slower with assistance. The single macrometastasis that was reviewed noticeably faster with assistance exhibited tumor foci with bland morphology and is the same image with false-negative interpretations discussed in the Classification accuracy section (Fig. 4B). The single macrometastasis that consistently took longer on average to review with assistance was notable for involving tumor with significant fibrotic change, interpreted by several pathologists as possible treatment effect. In addition, the algorithm did not completely outline the contiguous tumor region for this metastasis, and some regions containing tumor cells were classified as moderate confidence by the algorithm (Supplemental Fig. 5, Supplemental Digital Content 1, http://links.lww.com/PAS/A677).

Time differences between the 2 sessions of the crossover study design were also analyzed using mixed-effects models. In addition to the significant effects of assistance for negative cases and micrometastases as described above, the second reader session was independently associated with a shorter average time per image than the first session for micrometastases

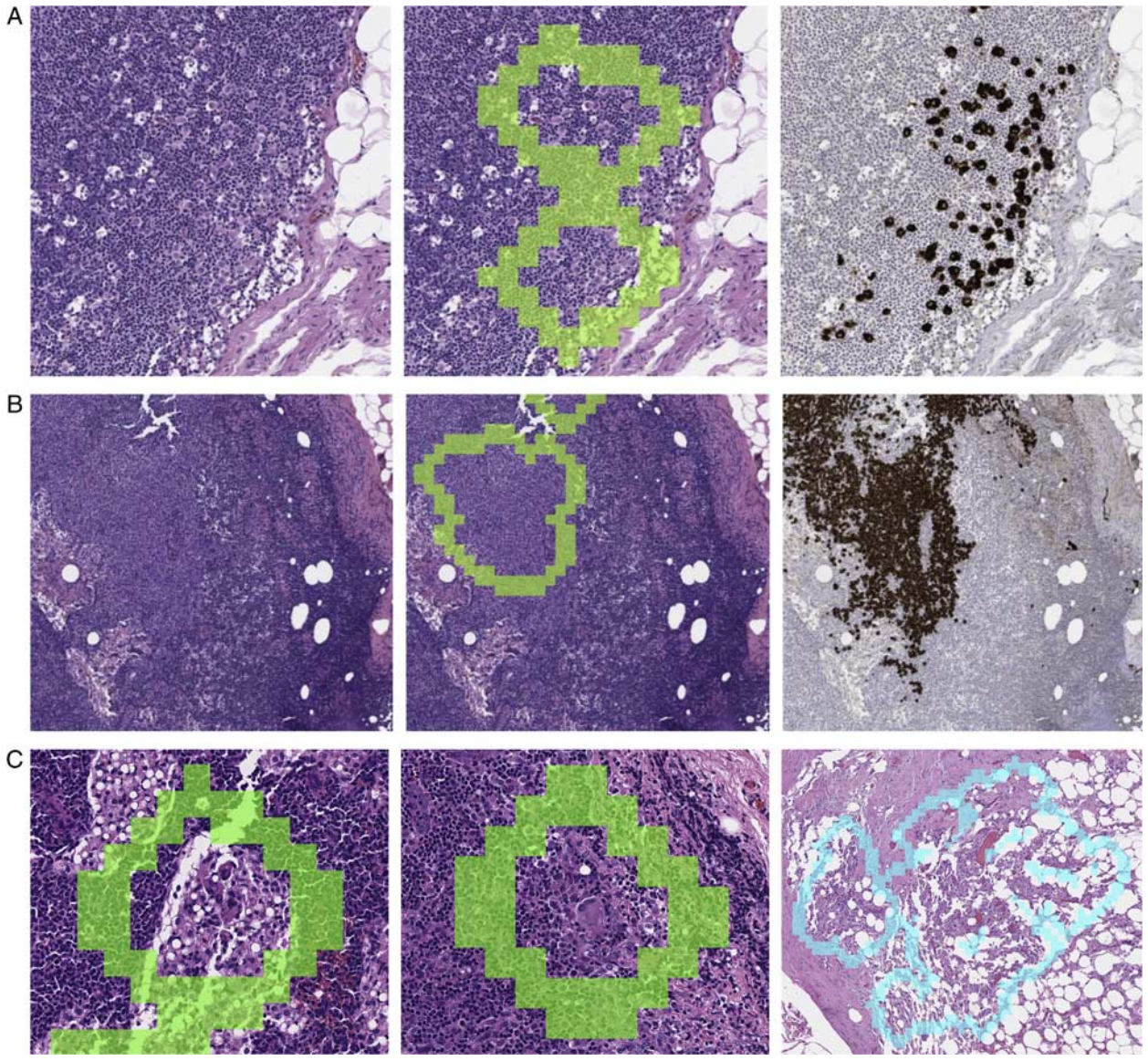

FIGURE 4. False-negative interpretations and false-positive algorithm predictions. A, Lymph node with > 200 dispersed tumor cells but classified as negative by all readers with or without assistance; selected field of view shows the most concentrated focus of tumor cells. Left: without algorithm overlay, middle: with algorithm overlay; LYNA algorithm highlights small areas within this region with moderate confidence. Right: cytokeratin IHC as reference for presence of tumor cells; the final reference standard classification as micrometastasis for this image was reached based on counting > 200 tumor cells on IHC. B, Left: stain quality and "bland" morphology led to poor visual contrast between a small macrometastasis and benign lymphoid tissue. Middle: tumor focus was outlined with moderate confidence, albeit incompletely circumscribed. Right: cytokeratin IHC for reference. Notably, despite the region highlighted by the algorithm, the tumor in this section was missed by pathologists both with and without assistance. C, Representative examples from independent cases in which LYNA falsely highlights histiocytes in the sinus (left), a giant cell (middle), and fat necrosis (right).

(12 s, $P = 0.04$), negative cases (23 s, $P < 0.01$), and ITCs (22 s, $P = 0.03$).

Evaluating Subjective Difficulty of Image Review

To test the hypothesis that computer assistance can affect the subjective difficulty of image review and to gain additional insight into the potential impact of assisted reads, pathologists were instructed to report an "obviousness score" (Materials and methods) following their interpretation for each image. This approach, based in part on a scoring system described by Gallas et al,[19] was aimed at addressing how easily pathologists were able to reach their final image classification.

The average obviousness score for macrometastases (91.6) was the highest of the categories, and the average obviousness score for ITC cases (52.9) was the lowest of the

TABLE 2. Average Review Times by Image Category and Assistance Modality

	Average Review Time (95% CI) (s)		
Category (n images)	**Unassisted**	**Assisted**	***P***
Negative (24)	**137 (126-148)**	**111 (101-121)**	**0.018**
Isolated tumor cells (8)	145 (123-166)	124 (104-145)	0.21
Micrometastasis (19)	**117 (102-133)**	**61 (54-69)**	**0.002**
Macrometastasis (19)	39 (25-55)	34 (21-47)	0.46

Bold values indicates statistically significant differences.

TABLE 3. Average Obviousness Scores to Assess the Difficulty of Each Case by Image Category and Assistance Modality

	Average Obviousness Score (95% CI)		
Category (n images)	**Unassisted**	**Assisted**	***P***
Negative (24)	67.5 (63.6-71.3)	72.0 (68.7-75.3)	0.29
Isolated tumor cells (8)	55.6 (47.7-63.5)	50.4 (42.2-58.6)	0.47
Micrometastasis (19)	**63.1 (58.3-67.9)**	**83.6 (80.3-86.9)**	**0.0005**
Macrometastasis (19)	90.1 (86.4-93.7)	93.1 (90.0-96.1)	0.16

Bold values indicates statistically significant differences.

individual categories. Regarding assisted versus unassisted reads, digital assistance was associated with an increase in the obviousness score for micrometastases without any significant differences for the other categories (Table 3).

DISCUSSION

Recent studies have described the ability of deep learning algorithms to perform on par with expert pathologists for isolated diagnostic tasks.[10] Underlying these exciting advances, however, is the important notion that these algorithms do not replace the breadth and contextual knowledge of human pathologists and that even the best algorithms would need to integrate into existing clinical workflows in order to improve patient care. In this proof-of-concept study, we investigated the impact of a computer assistance tool for the interpretation of digitized H&E slides, and show that a digital tool developed to assist with the identification of lymph node metastases can indeed augment the efficiency and accuracy of pathologists.

In regards to accuracy, algorithm assistance improved the sensitivity of detection of micrometastases from 83% to 91% and resulted in higher overall diagnostic accuracy than that of either unassisted pathologist interpretation or the computer algorithm alone. Although deep learning algorithms have been credited with comparable

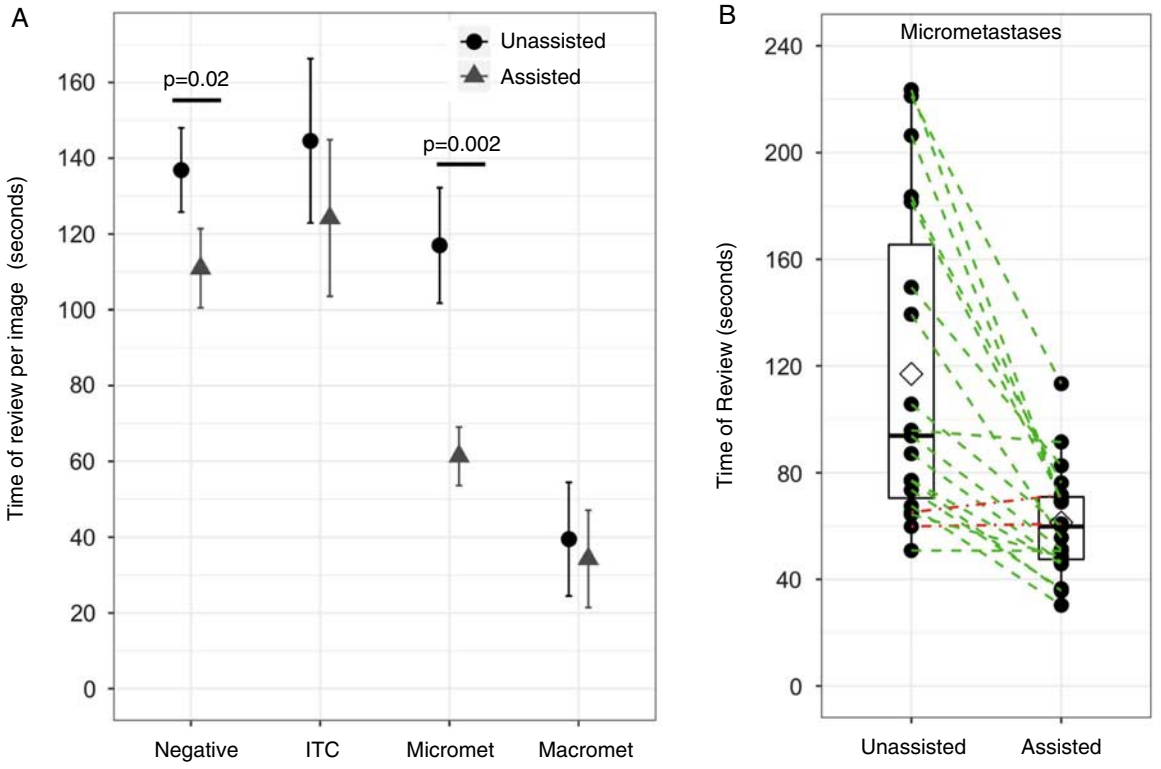

FIGURE 5. Average review time per image decreases with assistance. A, Average review time per image across all pathologists analyzed by category. Black circles are average times with assistance, gray triangles represent average times without assistance. Error bars indicate 95% confidence interval. B, Micrometastasis time of review decreases for nearly all images with assistance. Circles represent average review time for each individual micrometastasis image, averaged across the 6 pathologists by assistance modality. The dashed lines connect the points corresponding to the same image with and without assistance. The 2 images that were not reviewed faster on average with assistance are represented with red dot-dash lines. Vertical lines of the box represent quartiles, and the diamond indicates the average of review time for micrometastases in that modality. Micromet indicates micrometastasis; macromet, macrometastasis.

or superior diagnostic performance to pathologists, our results suggest that combining the unique strengths of computers and human experts may provide an even more promising opportunity. Pathologists understand the clinical setting and diagnostic workflow, allowing them to contextualize the therapeutic implications of false positives and false negatives in order to optimize their diagnostic operating point, sometimes even on a case by case basis. In contrast, algorithms that exceed the sensitivity of pathologists often do so at the cost of increased false positives. By using an algorithm to surface the most pertinent clinical information and allowing pathologists to review the findings, thoughtfully designed assistant tools have the potential to maximize both sensitivity and specificity while also allowing for the identification of any findings that are not interpretable by the algorithm alone.

The baseline sensitivity (without assistance) for micrometastases in our study (83%) was considerably higher than that reported recently for a similar task (38% on average, range: 15% to 63%).[10] Although pathologists in our study were instructed to review images at a pace similar to that of their routine clinical review, one possible factor contributing to the sensitivity discrepancy is the time spent per image. Pathologists in the prior study reviewed images with "flexible" time constraints that averaged to ~1 minute per slide, compared with 1.5 minutes in our study. Because the prior study did not report the time spent by pathologists on each category of images, we were unable to directly compare the time spent on cases with small or no tumor foci. As supporting evidence that time of review can affect sensitivity, the prior study also reported a sensitivity of 89% for review of micrometastases without time constraints. Another possibility is that the utilization of IHC in different practices or in different geographic regions results in a different diagnostic threshold for calling micrometastases using H&E alone. The pathologists in our study self-reported IHC utilization for roughly 10% of clinical cases, which may represent a different baseline experience than the pathologists in the prior study. Different levels of experience with digital pathology may also contribute to sensitivity differences, although there were no clear differences in this regard between the two studies. The sensitivity in actual clinical practice for micrometastases on H&E alone may be hard to evaluate in the era of IHC use, but original studies suggest that 12% to 29% of cases called negative using single-level H&E were in fact positive on review of IHC.[20] This suggests a substantial false-negative rate for the relatively infrequent micrometastases using H&E alone, perhaps even higher than the rate observed in our study. Although the 6 pathologists in our study represent a range of experience and clinical practice (none were breast specialists), they might represent above-average performance. If a reader study was performed in a particularly time-limited clinical setting with a broad range of pathologists, the accuracy benefits of an assistive tool may be more pronounced.

The most significant benefits of algorithm assistance observed in this study were for efficiency, with a time savings of 19% for negative cases and 52% for micrometastases. These observations are perhaps unsurprising for micrometastases, where we expect an accurate algorithm to help pathologists locate small tumor foci. However, the increased efficiency for negative cases is particularly notable given the majority of cases in clinical practice are negative. Extrapolating our data to clinical cases consisting of roughly 75% negative cases and 5% to 10% micrometastases,[21] these results would suggest a potential overall time savings of ~20%. We hypothesize that this time benefit for negative cases requires both an accurate algorithm and the establishment of trust in the algorithm's performance. This trust was likely developed through the use of the algorithm assistance during the study sessions. Taken together with our pilot study observations that a side-by-side heatmap display (Supplementary Digital Content 1, http://links.lww.com/PAS/A677) did not provide a similar time benefit, we reason that thoughtful user interface design is critical in order to avoid distracting users with extraneous information. Still, larger studies will be important to further validate the observed impact of digital assistance on efficiency and accuracy, especially for negative cases and in actual clinical workflows.

In addition to the time benefits associated with the assistance tool, we also observed independent, statistically significant time differences between the first and second sessions of this cross over study. One likely possibility is that the study participants became more familiar with all aspects of the viewer interface and the specific task such that their review was shorter. This is also consistent with the efficiency gains reported with increased digital pathology experience.[2] The faster review in the second session was not associated with any significant differences in accuracy between sessions, either overall or as a function of assistance modality.

The "obviousness score" described in this study was intended to provide insights into the subjective perception of the task with and without assistance and to build further on the objective measurements of accuracy and efficiency. Despite considerable inter-reader variability, these scores suggest that algorithm assistance increased the perceived ease of image review, specifically for micrometastases. While inter-reader variability in the use of quantitative scores is a known challenge in multiple-reader multiple-case studies,[22] future studies could provide more extensive training to improve inter-reader consistency for this type of score. Subjective metrics such as this, although often challenging to calibrate and incorporate, may be an important aspect of evaluating impact and value as digital tools in pathology and medicine continue to be developed and implemented.

The observed time savings for negative cases raises the important consideration of over-reliance on algorithm assistance, with the theoretical possibility of less thorough review for some cases.[23,24] In our study, pathologists were informed of the differences between the high confidence and moderate confidence regions, and to expect the possibility of false positives and false negatives in the assistance. Because the algorithm identified at least one moderate confidence region in every metastasis-positive image in this data set, we could not formally evaluate for

false negative pathologist interpretations due to over-reliance. Although it may be possible to tune the information presented by an algorithm to approach 100% sensitivity, the possibility of over-reliance on negative predictions warrants further investigation across larger data sets for which histologic diversity may result in unexpected algorithm false negatives. In regards to pathologist over-reliance on positive predictions, we did observe one image where pathologists ignored moderate-confidence regions that were indeed tumor (verified by IHC) and called the case negative. In this case, pathologists reviewing the image without assistance also failed to detect the metastasis, and thus it was an instance where the assistance tool "outperformed" pathologists, but the assistance had neither a clear positive nor negative effect. The only algorithmic false positive "high confidence" outline corresponded to a region of fat necrosis (Fig. 4C). This region was universally identified as such by the pathologists and ignored, and no false positive final diagnoses occurred among the 420 images reviewed in assisted mode. Seven additional tumor-negative images contained moderate confidence outlines but were also correctly recognized as nontumor by the readers, highlighting the important role of trained pathologists in this assisted read scenario. These findings suggest that despite limited experience with the assistant tool, pathologists were able to calibrate their reliance on the tool in order to use it effectively, and without evidence of over-reliance in this set of images.

In addition to an assisted primary review as evaluated in this study, digital assistance could also be integrated into clinical workflows in other ways: as a "screening" tool to triage definitively negative and/or positive cases, or as a "second read" for difficult cases following primary pathologist review. The screening approach is similar to the FDA-approved use of computer assistance for cervical cytology specimens,[25,26] while the second read approach could mitigate over-reliance or assistive bias during primary review. For lymph node staging, a tool that reduces the review burden for negative cases or triages challenging cases for IHC before initial pathologist review may reduce reporting delays and provide cost savings through efficiency and accuracy gains. Although the potential benefits are exciting, thoughtful consideration of the limitations along with clear instructions and definitions for intended use will be critical to the success and safety of assistive tools in pathology.

Several of the notable limitations to this study stem from the artificial constraints introduced by having pathologists review images in isolation, without the context of the actual clinical workflow. For example, pathologists would typically have access to the other slides comprising the case (including the primary tumor), additional levels, or IHC. In addition, unexpected findings such as lymphomas, infections, or other lymph node pathologies may be present in clinical cases, but were not evaluated in this study and represent a key limitation to stand-alone computer algorithms in diagnostic pathology. Future work can more thoroughly explore algorithms to identify other lymph node pathologies as well as nodal metastases for other cancer types. The algorithm used in this study detects metastatic tumor in lymph nodes, without the ability to interpret the positioning of these foci relative to the nearest lymph node. Correspondingly, extensions of this algorithm could also indicate specific diagnostic features of these detected tumor foci, such as extracapsular extension or lymphovascular invasion by identifying the location of each focus relative to the associated lymph node. Assistive tools for lymph node review may also facilitate easier measurement of tumor deposits, allowing more efficient and accurate reporting for prognosis and clinical management, and potentially even contribute to refined guidelines for nodal staging across different cancer types.

Ultimately, studies to demonstrate the clinical utility and value for assistive tools in digital pathology will require prospective clinical evaluation and thorough evaluation of the clinical task in question. For nodal staging, inclusion of the complete set of lymph node slides for a given case may be a useful next step,[27] potentially followed by the evaluation of the complete case. Given the findings reported here, understanding the actual clinical significance of increased lymph node micrometastasis detection is also an important consideration. While several studies have demonstrated prognostic significance for lymph node micrometastases relative to node-negative disease,[13,28,29] the reported outcomes differences can be small, institutional management practices may vary, and the ultimate impact on clinical decision making depends on the final staging and other patient-specific variables. Studies to evaluate the impact of assisted reads on the use of IHC, refined categories for prognosis, or presorting of cases based on algorithm predictions may demonstrate still additional value for "intelligent" tools in digital pathology.

In summary, this study directly demonstrates some of the potential benefits of assisted reads in pathology, including specific gains in efficiency and accuracy. As such, this study is a useful first step in understanding how assistive tools in pathology can be best designed and utilized, both to improve clinical care and to allow more time and mental energy for tasks that require invaluable human experience and expertise.

ACKNOWLEDGMENTS

The authors thank Greg Corrado and Philip Nelson for their advice and guidance in enabling this work, Craig Mermel for helpful comments on the manuscript, David Miller and Sunny Jansen for their input in study design and data analysis, and James Wren for logistical support. The authors thank members of the Google AI Pathology team for software infrastructure support, logistical support, and slide digitization services. Gratitude also goes to Sara Gabriele, T Saensuksopa, and Melissa Strader for insights into the user interface design. Deepest appreciation goes to pathologists Kathy Brady, Chris Kim, and 7 other pathologists for determining the reference standard for the images or reviewing images as a reader in the pilot or formal study.

REFERENCES

1. 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.
2. Mills AM, Gradecki SE, Horton BJ, et al. Diagnostic efficiency in digital pathology: a comparison of optical versus digital assessment in 510 surgical pathology cases. *Am J Surg Pathol*. 2018;42:53–59.
3. Gavrielides MA, Gallas BD, Lenz P, et al. Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. *Arch Pathol Lab Med*. 2011;135:233–242.
4. Wolff AC, Hammond MEH, Hicks DG, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J Clin Oncol*. 2013;31:3997–4013.
5. Acs B, Rimm DL. Not just digital pathology, intelligent digital pathology. *JAMA Oncol*. 2018;4:403–404.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
7. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115:211–252.
8. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 25*. San Francisco, CA: Curran Associates Inc; 2012:1097–1105.
9. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. 2017. Available at: http://arxiv.org/abs/1703.02442. Accessed March 9, 2017.
10. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318:2199–2210.
11. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep*. 2016;6:26286.
12. Badve SS, Beitsch PD, Bose S, et al. Breast cancer staging aystem: AJCC Cancer Staging Manual, 8th ed. 2017 Available at: https://cancerstaging.org/references-tools/deskreferences/Documents/AJCC%20Breast%20Cancer%20Staging%20System.pdf. Accessed May 1, 2018.
13. Weaver DL, Ashikaga T, Krag DN, et al. Effect of occult metastases on survival in node-negative breast cancer. *N Engl J Med*. 2011;364:412–421.
14. Giuliano AE, Hunt KK, Ballman KV, et al. Axillary dissection vs no axillary dissection in women with invasive breast cancer and sentinel node metastasis: a randomized clinical trial. *JAMA*. 2011;305:569–575.
15. Vestjens JH, de Boer M, van Diest PJ, et al. Prognostic impact of isolated tumor cells in breast cancer axillary nodes: single tumor cell(s) versus tumor cell cluster(s) and microanatomic location. *Breast Cancer Res Treat*. 2012;131:645–651.
16. Fine JL. 21(st) century workflow: a proposal. *J Pathol Inform*. 2014;5:44.
17. Rabinovitch A. The College of American Pathologists laboratory accreditation program. *Accredit Qual Assur*. 2002;7:473–476.
18. Lester SC, Bose S, Chen Y-Y, et al. Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Arch Pathol Lab Me*. 2009;133:1515–1538.
19. Gallas BD, Chan H-P, D'Orsi CJ, et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad Radiol*. 2012;19:463–477.
20. Apple SK. Sentinel lymph node in breast cancer: review article from a pathologist's point of view. *J Pathol Transl Med*. 2016;50:83–95.
21. Rutledge H, Davis J, Chiu R, et al. Sentinel node micrometastasis in breast carcinoma may not be an indication for complete axillary dissection. *Mod Pathol*. 2005;18:762–768.
22. Dendumrongsup T, Plumb AA, Halligan S, et al. Multi-reader multi-case studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy: systematic review with a focus on quality of data reporting. *PLoS One*. 2014;9:e116018.
23. Taylor P, Potts HW. Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. *Eur J Cancer*. 2008;44:798–807.
24. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318:517–518.
25. Wilbur DC, Black-Schaffer WS, Luff RD, et al. The Becton Dickinson FocalPoint GS Imaging System: clinical trials demonstrate significantly improved sensitivity for the detection of important cervical lesions. *Am J Clin Pathol*. 2009;132:767–775.
26. Biscotti CV, Dawson AE, Dziura B, et al. Assisted primary screening using the automated ThinPrep Imaging System. *Am J Clin Pathol*. 2005;123:281–287.
27. Litjens G, Bandi P, Bejnordi BE, et al. Cancer metastases in lymph nodes challenge 2017 (CAMELYON17). 2017. Available at: https://camelyon17.grand-challenge.org/. Accessed May 1, 2018.
28. de Boer M, van Deurzen CHM, van Dijck JAAM, et al. Micrometastases or isolated tumor cells and the outcome of breast cancer. *N Engl J Med*. 2009;361:653–663.
29. de Boer M, van Dijck JAAM, Bult P, et al. Breast cancer prognosis and occult lymph node metastases, isolated tumor cells, and micrometastases. *J Natl Cancer Inst*. 2010;102:410–425.