Mod Pathol 36 (2023) 100033

MODERN PATHOLOGY

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Journal homepage: https://modernpathology.org/

Research Article

Noninferiority of Artificial Intelligence–Assisted Analysis of Ki-67 and Estrogen/Progesterone Receptor in Breast Cancer Routine Diagnostics

Niklas Abele[a,*,‡], Katharina Tiemann[b], Till Krech[c,k], Axel Wellmann[d], Christian Schaaf[e], Florian Länger[f], Anja Peters[g], Andreas Donner[h], Felix Keil[i], Khalid Daifalla[j], Marina Mackens[b], Andreas Mamilos[i], Evgeny Minin[k], Michel Krümmelbein[b], Linda Krause[l], Maria Stark[l], Antonia Zapf[l], Marc Päpper[j], Arndt Hartmann[a], Tobias Lang[j]

[a] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Institut für Pathologie, Erlangen, Germany;* [b] *Institute of Hematopathology Hamburg, Hamburg, Germany;* [c] *Institute of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany;* [d] *Institute of Pathology Celle, Celle, Germany;* [e] *Department of Internal Medicine II, Klinikum rechts der Isar of the TU Munich, Munich, Germany;* [f] *Institut für Pathologie, Medizinische Hochschule Hannover, Hannover, Germany;* [g] *Institut für Pathologie, Städtisches Klinikum Lüneburg gGmbH, Lüneburg, Germany;* [h] *Zentrum für Pathologie, Zytologie und Molekularpathologie Neuss, Neuss, Germany;* [i] *Institute of Pathology, University of Regensburg, Regensburg, Germany;* [j] *Mindpeak, Hamburg, Germany;* [k] *Institute of Pathology, Clinical Center Osnabrueck, Osnabrueck, Germany;* [l] *Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*

ARTICLE INFO

Article history:
Received 12 January 2022
Revised 19 September 2022
Accepted 22 September 2022

Keywords:
digital pathology
mammary carcinoma
surgical pathology

ABSTRACT

Image analysis assistance with artificial intelligence (AI) has become one of the great promises over recent years in pathology, with many scientific studies being published each year. Nonetheless, and perhaps surprisingly, only few image AI systems are already in routine clinical use. A major reason for this is the missing validation of the robustness of many AI systems: beyond a narrow context, the large variability in digital images due to differences in preanalytical laboratory procedures, staining procedures, and scanners can be challenging for the subsequent image analysis. Resulting faulty AI analysis may bias the pathologist and contribute to incorrect diagnoses and, therefore, may lead to inappropriate therapy or prognosis. In this study, a pretrained AI assistance tool for the quantification of Ki-67, estrogen receptor (ER), and progesterone receptor (PR) in breast cancer was evaluated within a realistic study set representative of clinical routine on a total of 204 slides (72 Ki-67, 66 ER, and 66 PR slides). This represents the cohort with the largest image variance for AI tool evaluation to date, including 3 staining systems, 5 whole-slide scanners, and 1 microscope camera. These routine cases were collected without manual preselection and analyzed by 10 participant pathologists from 8 sites. Agreement rates for individual pathologists were found to be 87.6% for Ki-67 and 89.4% for ER/PR, respectively, between scoring with and without the assistance of the AI tool regarding clinical categories. Individual AI analysis results were confirmed by the majority of pathologists in 95.8% of Ki-67 cases and 93.2% of ER/PR cases. The statistical analysis provides evidence for high interobserver variance between pathologists (Krippendorff's α, 0.69) in conventional immunohistochemical quantification. Pathologist agreement increased slightly when using AI support (Krippendorff α, 0.72). Agreement rates of pathologist scores with and without AI assistance provide evidence for the reliability of immunohistochemical scoring with the support of the investigated AI tool under a large number of environmental variables that influence the quality of the diagnosed tissue images.

Introduction

Advances in breast cancer treatment and survival are in part due to more accurate assessment of tumor physiology including

‡ The present work was performed in (partial) fulfillment of the requirements for obtaining the degree "Dr. med."

* Corresponding author.
E-mail address: niklas.abele@uk-erlangen.de (N. Abele).

https://doi.org/10.1016/j.modpat.2022.100033

tumor cell proliferation and hormone receptor status (estrogen receptor [ER] and progesterone receptor [PR]) and the resulting patient stratification for targeted and individually tailored therapy,[1] next to progress in molecular pathology, early detection, and other areas. The manual analysis of immunohistochemical (IHC) slides, however, is time-consuming and prone to inter- and intraobserver variability.[2-5] This might result in different and suboptimal treatment decisions. By contrast, valid analyses with low variability have the potential to improve patient outcomes.[6,7]

Image analysis (IA) assistance with artificial intelligence (AI) methods has been one of the great promises to support pathologists in diagnosis over recent years.[8-10] IA tools process histopathological images and detect and classify target structures such as tumor cells in an automated way. When used as an assistance tool, IA results are presented to pathologists for further evaluation and the final diagnosis, in contrast to fully automated analysis without human intervention. Traditional IA methods are based on conventional computer vision techniques with manually predefined analysis rules. By contrast, AI-based methods use machine learning techniques such as deep learning to learn analysis rules autonomously from image examples. AI-based methods have the potential to achieve IA accuracy levels that are hard to obtain with conventional methods and allow for wide-spread clinical routine usage.[11]

Recent studies on using IA tools for IHC assessment based on traditional computer vision techniques have shown promising results to improve reproducibility across institutions.[6,7,12] Nonetheless, adoption in routine diagnostics has been slow, and IHC assessment is almost always performed manually in a pathology practice. A major reason for this is the instability of many IA systems in a "real-life" setting with a high number of environmental variables. These variables introduce considerable variability into tissue images and include preanalytical differences in tissue fixation, section thickness and preparation and analytical differences in staining procedures, whole-slide image (WSI) scanners and antibody clones resulting in varying staining intensities, staining colors, and other parameters. Although slides with subpar quality might still be considered usable for analysis with conventional methods, they regularly cause difficulties for automated solutions. Missing IA robustness in these circumstances can be hazardous to patient safety as failing software may bias pathologists to incorrect conclusions. To compensate for missing robustness, many IA tools, including both conventional and AI-based systems, require extensive adaptation on representative data when used in a new laboratory context, necessitating additional data collection, preparatory analyses, and site-specific validation studies.

In this study, reliability of using an AI-based IA assistance tool (Mindpeak Breast Ki-67 RoI and ER/PR RoI) in the analysis of Ki-67 proliferation rate and ER/PR was investigated. The goal of this study was to validate this reliability in a setting representative for routine practice, including significant variability in tissue images due to using multiple staining procedures and WSI scanners. The AI tool was not adapted to individual experimental conditions by a human operator. To the authors' knowledge, no study with similar coverage of variability or quantity has been published to date. Agreement rates for scoring with and without AI assistance in clinical categories were investigated to assess the reliability of using this AI assistance tool in a routine diagnostic setting in a scenario without context-specific AI adaptation.

Table 1
Overview of the sample distribution according to original diagnostic score groups as available for the study

Staining	Score, %	Samples per group
Ki-67	0-5	18
	6-10	18
	11-25	18
	26-100	18
PR	<1	8
	1-9	29
	10-100	29
ER	<1	8
	1-9	29
	10-100	29

Significant fractions of cases are close to the clinical cutoffs used for the statistical analysis.
ER, estrogen receptor; PR, progesterone receptor.

Materials and Methods

Samples and Patients

The study design involved cropped sections of scanned core biopsy specimens of 204 female patients with invasive breast cancer. This follows the recommendation of the American Society of Clinical Oncology/College of American Pathologists Guidelines to prefer core biopsies of tumors for testing if they are representative of the tumor (grade and type) at resection.[13] The samples comprised routine diagnostic cases from the archive of the Institute of Hematopathology Hamburg (HPH), Hamburg, Germany, diagnosed between 2016 and 2020, together with broad binning intervals of quantitative scores to allow for case stratification. Neither patient data nor survival data were available for this study. Thus, the initially reported quantitative scores for Ki-67 and ER/PR were not included in the study design. Ethical approval of the study was granted by the Medical Association Hamburg, Germany.

Sample selection was stratified to reflect average frequencies of carcinoma types and a representative distribution of proliferation levels (Table 1), including a significant fraction of cases close to clinical cutoffs, following recommendations for Ki-67 and ER/PR evaluation.[14] Within stratification buckets, samples were chosen at random without additional post hoc manual quality filtering, thus including samples with artifacts and subpar quality. This study focused on the 2 most common carcinoma types in invasive breast cancer (other types were excluded): 168 (82%) samples of invasive ductal carcinoma (no special type) and 36 (18%) of invasive lobular carcinoma were included. Staining intensities for ER/PR cases were not investigated in this study and ignored in case selection.

Tissue Preparation, Immunohistochemistry, and Slide Scanning

Tissue preparation and IHC staining were performed at the Institute of Hematopathology Hamburg, Hamburg, Germany, and Vivantes Klinikum Neukölln, Berlin, Germany, according to the consensus criteria established by the International Ki-67 Working Group.[14]

Three different automatic sample preparation stainers for IHC staining were used: Roche Ventana Benchmark, Leica Bond III (both at the Institute of Hematopathology), and Dako Omnis (at Vivantes). The following antibodies were used for staining: Ki-67,

Table 2
Experimental variables

Variable	Count	Comment
Cases	204	Balanced for clinical scores
Slide preparation sites	2	Institute of Hematopathology Hamburg, Vivantes Klinikum Neukölln
Pathologist sites	8	
Pathologists	10	
Scanners/microscope cameras	6	3DHistech P1000, Roche DP200, Leica Aperio GT 450, Hamamatsu S360, Philips IntelliSite UFS, and Basler camera on Nikon microscope
Staining machines	3	Roche Ventana Benchmark, Leica Bond III, Dako Omnis
Regions of interest per case	20	10 pathologists × 2 rounds
Total no. of observations	4080	204 cases × 10 pathologists × 2 rounds

clone MIB1 (Dako); ER, clone SP1 (Ventana) and EP1 (Dako); and PR, clones 16 (Leica), 1E2 (Ventana), and 1294 (Dako). The effect of different antibodies for the same IHC staining was beyond the scope of this study. In total, 72 samples were immunohistochemically stained for Ki-67, 66 for ER, and 66 for PR.

Five different WSI scanners and one microscope camera were used to generate digital images of tissue slides for this study: 3DHistech P1000, Roche Ventana DP200, Leica Aperio GT 450, Hamamatsu S360, Philips IntelliSite Ultra-Fast Scanner, and microscope camera Basler 1920-40uc on a Nikon Eclipse Ni microscope. Table 2 provides an overview of the experimental variables. Figure 1 demonstrates resulting image variability in representative areas of interest; 34 slides were scanned per scanner/microscope camera with a resolution of approximately 0.25 μm per pixel (precise numbers vary slightly per scanner). Samples were stratified across stainers and scanners to have a representative coverage across combinations (Table 3).

Digital Image Analysis

An AI assistance tool for IHC quantification based on deep learning with convolutional neural networks[11] was used in this study (Mindpeak Breast Ki-67 RoI and Mindpeak ER/PR RoI) (Fig. 2). This AI tool detects individual cells in immunohistochemically stained tissue and distinguishes tumor cells from nontumor cells. It uses a neural network architecture consisting of 21 convolutional layer blocks with a rectified linear unit activation function and batch normalization and separate branches for cell detection and classification.[15] Neural network weights in both branches were learned simultaneously using gradient descent with the AdamW optimizer[16] in an AI learning phase before this study. For a given tissue image, the neural network identifies cells by predicting for every pixel whether it belongs to a cell, using the cell detection branch. The cell classification branch is used to predict for every pixel a cell class (tumor, nontumor, and background) and to assign detected cells after cell detection to classes. In addition, the AI tool distinguishes positive from negative tumor cells based on the respective intensity of the used 3,3'-diaminobenzidine staining. The AI tool proposes a global brown (3,3'-diaminobenzidine staining) intensity threshold per case based on a two-component Gaussian mixture model.[17] Users can modify this threshold to fulfill their interpretation criteria. The AI tool was not adapted to experimental conditions. It was used in the same configuration across experimental conditions, clones, and participants.

Study Design

AI-assisted IHC-based quantification of Ki-67–, ER-, and PR-stained breast cancer core biopsies from clinical routine within a realistic setting of scoring cases under varying preanalytical

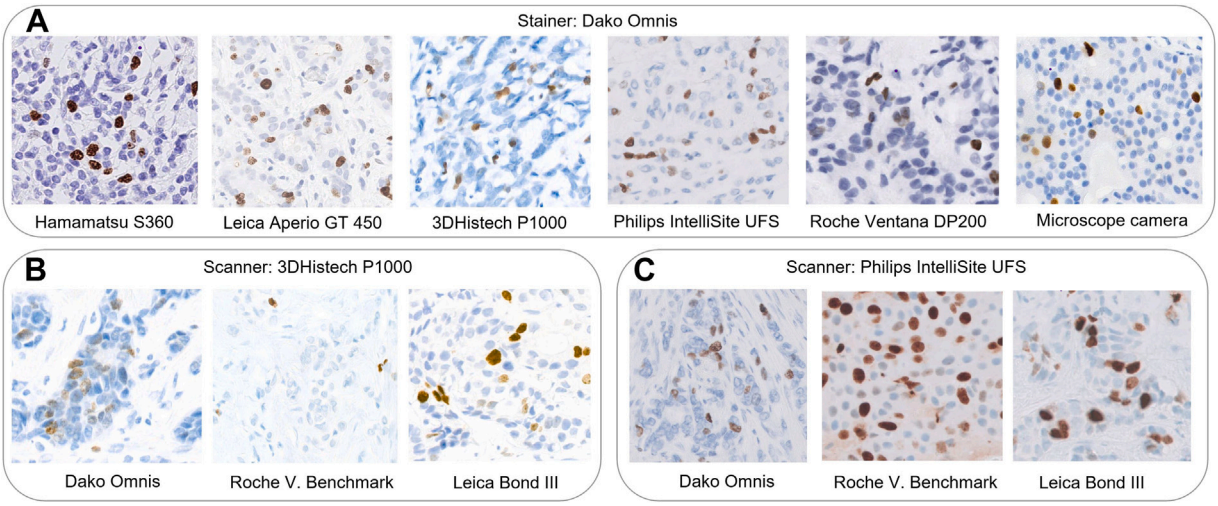

Figure 1.
Cutouts from a selection of study tissue images shown at 10× magnification (1.0 μm per pixel). (A) Tissue images stained with Dako Omnis and scanned with various scanners. Tissue images scanned with (B) P1000 and (C) Philips and stained with various stainers.

Table 3
Stratification of samples across stainers, scanners, and staining

Staining	Stainer/scanner	Philips	P1000	Ventana DP200	Leica 450	Hamamatsu S360	Microscope
Ki-67	Dako Omnis	4	4	5	3	3	5
	Roche Ventana Benchmark	4	4	3	5	4	4
	Leica Bond III	4	4	4	4	5	3
PR	Dako Omnis	2	4	4	3	4	5
	Roche Ventana Benchmark	4	2	5	4	4	3
	Leica Bond III	5	5	2	4	3	3
ER	Dako Omnis	5	4	3	4	2	4
	Roche Ventana Benchmark	3	3	4	3	6	3
	Leica Bond III	3	4	4	4	3	4

ER, estrogen receptor; PR, progesterone receptor.

conditions (stainers, scanners, multisite, and antibodies) was investigated in this study. The study goal was to assess the intraobserver agreement between scoring a case with and without the AI assistance tool.

To ensure focus on relevant tissue for cases where multiple cores and on-slide control tissue were present on WSIs, an area of 9 mm^2 was cropped beforehand. A size of 9 mm^2 was deemed sufficiently large enough to reflect the real-world diagnostic setting where pathologists have to choose the most relevant region of interest (ROI) in tissue slides by themselves. All 9-mm^2 areas were cropped from WSIs of tissue slides by consensus between 3 expert board-certified pathologists (each with >15 years of experience on breast cancer diagnosis) (Fig. 3). These 3 experts were not part of the 10 participating study pathologists. One of the experts selected an initial area of 9 mm^2. Then, the other 2 experts were asked if they agree. For all cases without an initial agreement, consensus on a (sometimes modified) suitable area was achieved in personal discussion between the 3 experts.

These areas contained more than 15,000 cells on average. Subsequently, 10 study participants were asked to provide individual quantitative scores for all cases using these digital images of 9 mm^2 area. Quantitative scores are percentages of positive tumor cells in relation to all tumor cells (proliferation score in Ki-67) in individually defined ROIs suitable for scoring. Participants defined ROIs containing about 300 to 500 malignant invasive cells.[14] If there was not a sufficient number of suitable cells, fewer cells were taken into account. Participants had to choose ROIs individually within cases for scoring, potentially resulting in varying scores per case.[18] Scores were provided in percentages with up to 1 decimal place (eg, 17.3%). Participants were asked to follow the standard operation procedure in their clinical routine for scoring based on national guidelines by the German Cancer Society (Deutsche Krebsgesellschaft) and recommendations of the Working Group Gynecologic Oncology (Arbeitsgemeinschaft Gynäkologische Onkologie), either by counting individual cells or semiquantitative methods. Participants were asked to use the same amount of time as they would use in routine. No other time restrictions were given. The relative freedom in individual scoring methods was given to participants to ensure an accurate representation of actual clinical routine and diminish response bias.

In this study, 10 investigator pathologists of varying degrees of experience (experience, 11.95 ± 10.9 years; range, 1.5-30 years), coming from 8 different institution sites (4 university clinics/public hospitals and 4 private institutes), participated. The study consisted of 2 scoring rounds: (1) without AI assistance; (2) with AI assistance, blinded to the respective result of the previous round (Fig. 3). In each round, pathologists had to choose ROIs to perform the scoring. Between the rounds, there was a washout period of 1 month where participants did not see any study cases. In both rounds, participants saw cases in different random sequential orders to ensure blinding. The same participant could choose different ROIs in rounds 1 and 2. In round 1, ROIs were not outlined explicitly because only human scoring was involved. In round 2, participants had to outline ROIs explicitly to perform the AI analysis. Pathologists were asked to view the results of the AI analysis and decide on a final score, which could differ from the provided result of the AI analysis. The AI tool was not adapted to experimental conditions: there was no human operator who preconfigured the AI tool to the data at hand to achieve acceptable IA results. Instead, the investigated AI tool was used in the same configuration across experimental conditions and participants.

All pathologists were blinded to information about particular antibodies, stainers, and scanners used in the study and about

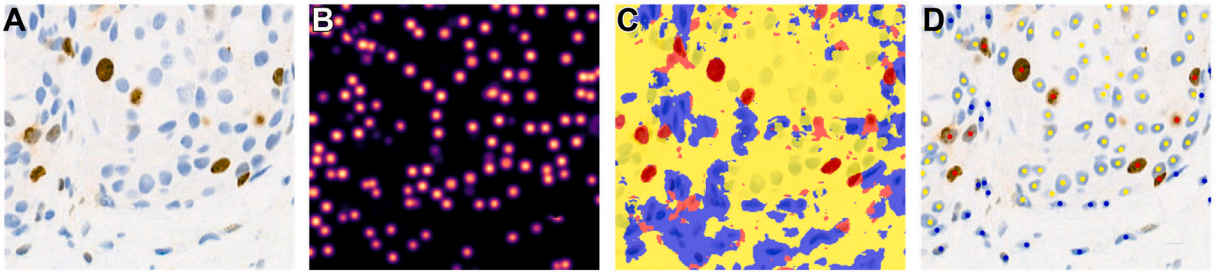

Figure 2.
The artificial intelligence assistance tool is based on a convolutional neural network that combines cell detection and tissue classification in a single quantification result: (A) input image; (B) results of cell detection branch (yellow, red, and black indicate high, medium, and low cell probability, respectively); (C) results of tissue classification branch (red, positive tumor; yellow, negative tumor; blue, nontumor); (D) results of detected cells (red and yellow indicate positive and negative tumor cells, respectively; blue indicates nontumor cells).

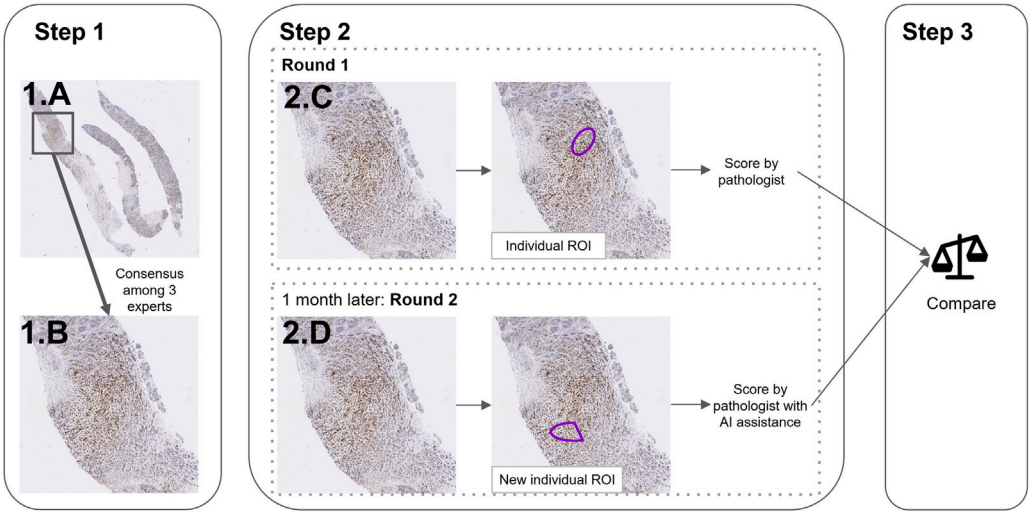

Figure 3.
A realistic scoring setting was achieved in the study by the following 3 steps: Step 1: (1.A) tissue areas of 9 mm^2 for all 204 cases of mammary carcinoma were determined by a consensus of 3 expert board-certified pathologists. (1.B) The resulting 9-mm^2 areas (typically approximately 15,000 cells) were used for subsequent scoring. Step 2: all 10 participants scored all images by selecting regions of interest (ROIs) (purple) within 9-mm^2 tissue areas twice with a washout time of 1 month. (2.C) In round 1, pathologists scored individually selected ROIs manually. (2.D) In round 2, pathologists scored new individually selected ROIs with artificial intelligence (AI) assistance. Step 3: The scores of both rounds were compared.

individual cases to avoid bias. Only the staining (Ki-67, ER, or PR) was revealed per case. All investigators scored all samples.

A specifically designed study viewer tailored to the individual study steps was used for this study (Fig. 4A). This study viewer is an online tool that can be accessed on a website using an internet browser. Each study investigator had their separate account to the viewer and worked from an individual working place including variable computer screens. Participants were able to provide comments per case in a separate text box.

Statistical Analysis

The goal of the statistical analysis was to investigate whether the investigators' assessment is not inferior when using AI assistance. For each investigator, there are 2 sets of scores: with and without the support of the AI tool, resulting in 2040 paired assessments (10 investigators $\times$ 204 cases). The 2 coprimary hypotheses were that investigators' assessment is leading to sufficient agreement between with and without usage of AI assistance, separately for Ki-67 and ER/PR. The coprimary hypotheses were investigated using confusion matrices comparing categorized results of scoring with using the AI tool to those of manual scoring without the AI tool.

For calculating confusion matrices, measurements were dichotomized using the following cutoff values:

- ER/PR: <1% (<1% is negative class, ≥1% positive class), as agreed by the American society of clinical oncology[13];
- Ki-67: ≤20% (≤20% is negative class, >20% is positive class), as agreed by the Saint Gallen Conference.[19]

For every case, it was measured for each pathologist whether their categorized scores in round 1 (pathologist) and round 2 (pathologist + AI assistance) agree. Overall agreement across cases and pathologists was calculated, and 95% logit CIs were computed. The agreement rate was calculated separately for Ki-67 and ER/PR. Sufficient agreement, that is, noninferiority, was shown if the lower bound of the CI was above the noninferiority margin of 75%. This margin was decided on following a common convention in the regulatory literature on AI assistance for the licensing of similar products.[20] The overall study hypothesis is shown only if both agreements for Ki-67 and ER/PR are noninferior. In a secondary analysis, the agreement rate was analyzed in a stratified view by scanner, stainer, and antibody. Krippendorff's α[21] was used to measure interobserver reliabilities between investigators within the same image, ranging from 0 to 1 (0, perfect disagreement; 1, perfect agreement). For Krippendorff's α, 95% CIs were calculated based on bootstrapping. Furthermore, the confirmation of the initial result of the AI assistance tool by the human pathologist in round 2 was analyzed with 95% CIs based on bootstrapping. This is an indicator of agreement of the pathologist with the AI tool. For the statistical analysis, software packages R (version 3.5.3), Python 3.8.12, and SciPy 1.7.1 were used.

Accuracy of Artificial Intelligence Without Human Intervention

Two additional experiments were performed to assess the accuracy of the AI tool without human intervention to bring the results of the main study into context. AI accuracy without human intervention could not be assessed directly from data of the main study: The individual ROIs from both rounds were chosen on the spot and without further external review by each pathologist; a comparison of them would thus be biased to the human interpretation of the right ROI to select. The same AI tool as in the main study was used, again without any calibration to experimental conditions.

In the first additional experiment, the scoring results (percentages of positive tumor cells, as described earlier) of 2 pathologists and the AI tool without human intervention were compared in a fixed set of ROIs. Regions of 200 to 300 tumor cells suitable for IHC quantification, potentially also including nontumorous tissue areas, were defined by an experienced cancer expert in a subset of the study samples (44 Ki-67 and 54 ER/PR) and used as ROIs. Cases were selected to represent the spectrum of

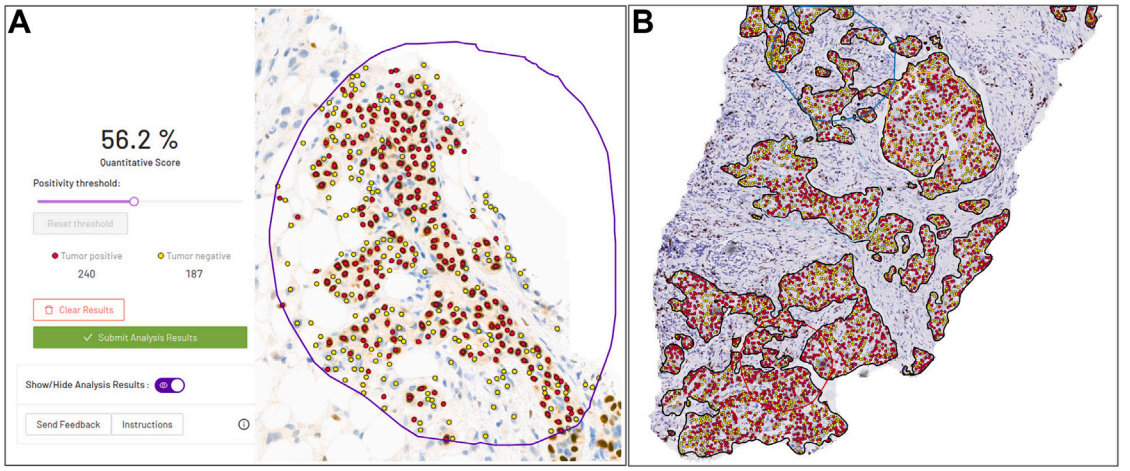

Figure 4.
The study viewer allowed participants to view, navigate, and zoom in and out of the 9-mm^2 tissue slide for each case. (A) A region of interest was individually selected for scoring (purple circle). In round 2, the artificial intelligence result for the chosen region of interest was presented to participants and visualized with red/yellow dots for positive/negative tumor cells, respectively. Pathologists had to enter their—potentially different —final score in a separate text field (denoted by "Quantitative score"). (B) In the additional experiment without human intervention, the artificial intelligence identified all areas of tumor (outlined in black) and detected and classified the tumor cells. For the hotspot and the International Ki-67 in Breast Cancer Working Group method, 4 regions with 400 cells each were further selected fully automatically (red, highest proliferation; orange and dark and light blue, the other groups in descending order).

proliferation regimes, including a significant number of cases close to cutoffs used to categorize the scores, with cases per proliferation regime chosen at random. The ROIs in these cases naturally differed from the ROIs individually chosen per pathologist in the main study. Two pathologists with multiyear experience in routine breast cancer diagnostics were asked to follow standard operating procedures and count individual cells. The washout period between the main study and this additional experiment was 3 months. For calculating agreement measurements, the above-described dichotomizing cutoffs (Ki-67, 20%; ER/PR, 1%) were used.

In the second additional experiment, the AI analysis was run on the complete 9 mm^2 Ki-67 tissue images of the main study without human intervention. Even though the investigated AI tool

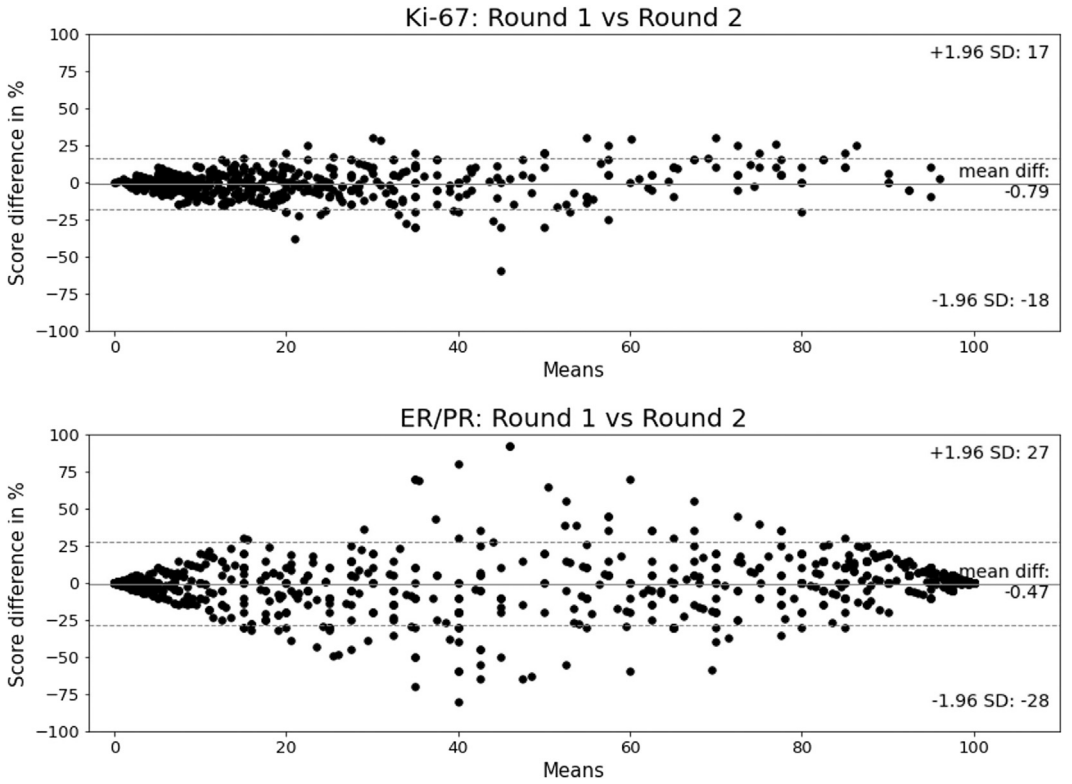

Figure 5.
Bland-Altman plots showing the absolute differences of values (in percentage) between the readings in rounds 1 and 2 against the respective mean values, derived from 718 pairs for Ki-67 and 1315 pairs for estrogen receptor (ER)/progesterone receptor (PR).

Table 4
Overall results

Staining	Total no.	No. agreed	Agreement rate, %	95% CI
Ki-67	718	629	87.6	85.0-89.8
ER/PR	1315	1176	89.4	87.6-91.0

ER, estrogen receptor; PR, progesterone receptor.

is intended as an assistance tool for manually prespecified ROIs, this allows for the detection of positive and negative tumor cells across the whole tissue slide and is uninfluenced by a human user (Fig. 4B). Three automatic scoring methods based on detection results without any human intervention were investigated: (i) the percentage of positive tumor cells across all detected tumor cells; (ii) the score in the most proliferating hotspot; (iii) the (unweighted) global score according to the recently proposed whole section protocol by the International Ki-67 in Breast Cancer Working Group (IKWG 4R).[4] The latter is a more formalized method of scoring field selection and considers regions for scoring that are high, medium, low, or negative in relation to the overall percentage positivity. For both the hotspot and the IKWG 4R scoring method, an automated technique for scoring field selection based on proliferation scores was implemented. This method groups all detected tumor cells into potential scoring regions of a fixed cell number, calculates a positivity score for each region, and selects final regions according to the scoring method: the most proliferating hotspot in the hotspot method or up to 4 scoring regions in the IKWG 4R method.

Results

Overall Robustness

Overall 2033 paired assessments were entered into the analysis. Three investigators could not provide a score for 7 cases in round 1 without AI assistance because of an error in the study software (2 for investigator 2, 3 for investigator 4, and 2 for investigator 9). Differences between the 2 assessments (with and without AI assistance) are visualized in detail for all pairs in Figure 5 and Supplementary Figure S1. The average intraobserver agreement between rounds 1 and 2 for Ki-67 was 87.6% (95% CI 85.0-89.8) and for ER/PR it was 89.4% (95% CI, 87.6-91.0) (Table 4). For both CIs, the lower bound of the CI is above the aforementioned noninferiority margin of 75%. Therefore, the 2 coprimary hypotheses lead to a significant result.

During the study, in 14 cases (7%), at least 1 participant pathologist raised a concern regarding scoring difficulties. Such concerns were due to staining problems (9 cases), artifacts in tissue preparation such as squeezed tissue or air bubbles under the cover slip (4), or unfamiliar image appearance for a participant (1 case). This rate is due to the level of variance in experimental variables that this study was intended to cover. Agreement rates between initial AI results and pathologists slightly increased when excluding these problematic cases, as detailed further.

Table 5
Results per staining and stainer

Staining	Stainer	Total no.	No. agreed	Agreement rate, %	95% CI
ER/PR	Dako	439	389	88.6	85.3-91.3
ER/PR	Leica	439	394	89.7	86.5-92.3
ER/PR	Roche	437	393	89.9	86.7-92.4
Ki-67	Dako	240	221	92.1	87.9-94.9
Ki-67	Leica	240	212	88.3	83.6-91.8
Ki-67	Roche	238	196	82.4	77.0-86.7

ER, estrogen receptor; PR, progesterone receptor.

Robustness Across Stainers

For Ki-67 and ER/PR, lower bounds of the CIs of the agreements comparing categorized results of scoring with and without AI assistance stratified by stainer were all larger than the noninferiority margin of 75% (Table 5 and Fig. 6).

Robustness Across Scanners

For Ki-67 and ER/PR, lower bounds of the CIs of the agreements comparing categorized results of scoring with and without AI assistance stratified by scanner were all larger than the noninferiority margin of 75% (Table 6 and Fig. 6).

Interobserver Reliability

The effect of AI assistance on interobserver (interrater) reliability was investigated (Supplementary Fig. S1). To reemphasize, the agreement of scores was calculated based on the case level (not on the level of ROIs). For each case, pathologists selected individual ROIs for scoring, with potentially varying proliferation levels. Without AI assistance, Krippendorff's α was 0.69 (95% CI, 0.65-0.73). With AI assistance, Krippendorff's α slightly increased to 0.72 (95% CI, 0.68-0.76). CIs overlap.

Human Confirmation of the Artificial Intelligence Result

It was investigated how often in round 2 (with AI assistance) participants confirmed the proposed AI assistance result regarding the clinical category. Across all 2033 readings, participants kept the initial AI result in 92.3% (95% CI, 90.5-93.8) of scorings for Ki-67 and 89.0% (95% CI, 86.9-90.7) of scorings for ER/PR. Because participants chose individual ROIs for AI assistance in each case, the 2033 readings refer to 2033 different ROIs. Once cases where participants raised concerns about image quality (1 Ki-67, 13 ER/PR cases) were removed from the analysis, confirmation rates on the level of individual readings increased from 92.3% to 93.8% (95% CI, 91.7-95.2) for Ki-67 and from 89% to 91.5% (95% CI, 89.7-92.9) for ER/PR across the resulting individual 1893 readings.

This analysis was broken down to case level to relate this to interobserver variance. Most of the participants (at least 6 of 10) agreed with the proposed AI result in 95.8% (95% CI, 94.4-96.9) of Ki-67 and 93.2% (95% CI, 91.5-94.6) of ER/PR cases (Tables 7 and 8). All 10 participants agreed with proposed AI results in about two-thirds of cases (Ki-67, 68.0% [95% CI, 65.0-70.8]; ER/PR, 63.6% [95% CI, 60.6-66.5%]).

Cases with low agreement rates (5 participants or less agreed with the proposed AI result) were investigated in more detail. For Ki-67, the case with the lowest agreement had only 3 of the 10 pathologists agree with the AI result. Some participants reported image quality problems in this case because of air under the cover slip, impairing AI analysis accuracy. For ER/PR, there was 1 case where no participant agreed with the AI result. In this case, 8 participants reported slide image problems (such as very bad

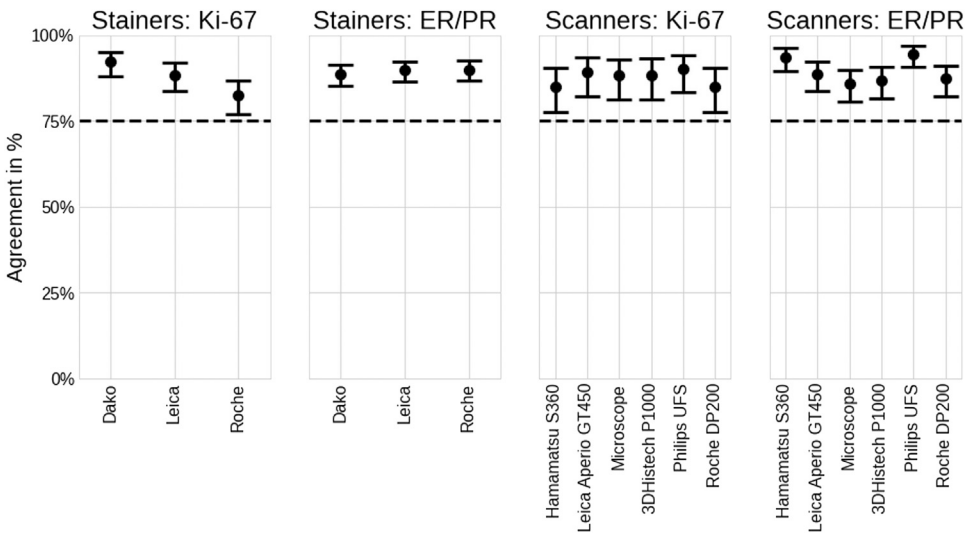

Figure 6.
Percentage of agreements of a pathologist's categorized scores in round 1 (pathologist) and round 2 (pathologist + artificial intelligence assistance), stratified by stainers and scanners. ER, estrogen receptor; PR, progesterone receptor.

staining and air bubble artifacts). Similarly, in other cases with low AI–human agreement, participants reported image problems and artifacts (such as no clear nuclear staining and weak staining). Such challenging cases had deliberately not been excluded from the study beforehand to ensure a realistic study design.

Artificial Intelligence Accuracy Without Human Intervention

For the first additional experiment without human intervention based on predefined ROIs, for Ki-67, each pathologist exhibited a higher agreement with the results of the AI without human intervention than the agreement among the 2 pathologists (Table 9, Fig. 7, and Supplementary Fig. S2). For ER/PR, the highest agreement was between AI without human intervention and pathologist 2. All agreements were above 75%. The distribution of cases included a relevant fraction of challenging cases close to thresholds (Table 10).

For the experiment based on analyzing complete Ki-67 tissue images without any human intervention (ie, beyond predefined ROIs as normally required by the AI tool), 1 case had to be removed because it was not suitable for full-slide processing because of strong clipping and background artifacts. In 2 cases, the AI failed to automatically find scoring regions (hotspots) with a sufficient number of tumor cells because of an extremely low tumor cell count, thus producing results only for the all-tumor-cell method but not for the IWKG 4R and the hotspot methods. AI results without human intervention were within the range of the conventional scores by pathologists using the hotspot method in 85.5% of cases. Furthermore, 92.8% of the AI results for the IWKG 4R method fell in the range of the conventional scores and 87.1% of the AI results when all tumor cells were considered (Fig. 8).

Discussion

IA with AI has been one of the great promises over recent years in pathology, but only few image AI systems have entered clinical routine already. A major reason is the missing robustness of many AI systems,[22] which can be hazardous for patient safety because failing AI analysis may bias pathologists to incorrect diagnoses. Only a few studies have addressed this problem so far.[23]

Other studies investigated IHC scoring with AI tools in contexts with less variance. Bankhead et al[7] studied fully automated IHC

Table 6
Results per staining and scanner

Staining	Scanner	Total no.	No. agreed	Agreement rate, %	95% CI
ER/PR	Hamamatsu S360	218	204	93.6	89.5-96.2
ER/PR	Leica Aperio GT450	220	195	88.6	83.7-92.2
ER/PR	Microscope	220	189	85.9	80.7-89.9
ER/PR	P1000	219	190	86.8	81.6-90.6
ER/PR	Philips	220	208	94.5	90.6-96.9
ER/PR	Roche DP200	218	190	87.2	82.0-91.0
Ki-67	Hamamatsu S360	120	102	85.0	77.4-90.3
Ki-67	Leica Aperio GT450	119	106	89.1	82.1-93.6
Ki-67	Microscope	119	105	88.2	81.1-92.9
Ki-67	P1000	120	106	88.3	81.3-93.0
Ki-67	Philips	120	108	90.0	83.2-94.2
Ki-67	Roche DP200	120	102	85.0	77.4-90.3

ER, estrogen receptor; PR, progesterone receptor.

Table 7
Agreement rates of pathologists with the results of the artificial intelligence tool in round 2 with 95% CIs across readings (N = 2033)

Staining	Agreement rates across all readings, % (95% CI)
Ki-67	92.3 (90.5-93.8)
ER/PR	89.0 (86.9-90.7)

Note that per case pathologists scored individually chosen ROIs, resulting in 10 different ROIs per case.
ER, estrogen receptor; PR, progesterone receptor; ROIs, regions of interest.

scoring with AI in a single-stainer, single-scanner setting with predefined ROIs on tissue microarrays. Acs et al[6] investigated 3 AI systems for Ki-67 scoring support that require initial AI training steps by the operator in a single-stainer, single-scanner setting on tissue microarrays. Rimm et al[12] did a follow-up study involving 14 laboratories and 10 different AI systems (9 of them requiring an initial AI training step by the operator), 7 scanners, and 1 stainer. All studies state the need for further clinical validation and formal comparisons across stainers, laboratories, and/or scanners. This study intended to contribute to fulfilling this need.

Thus, the primary goal of this study was to investigate the reliability of using an AI system as a diagnostic decision support tool in a setting that is close to the reality of routine clinical pathology. The results of the statistical analysis confirmed the reliability when scoring with the AI assistance tool under investigation. To the authors' knowledge, this is the largest study to date with regard to a number of environmental conditions, including 10 participant pathologists from 8 sites, 6 WSI scanners/microscopes, 3 staining machines, routine tissue slides, and participant-specific selection of scoring ROIs in tissue images. In contrast to other studies, the AI assistance tool used in this study does not require manual fine-tuning to the provided image data or an initial training phase by the pathologist.[6,12] It was not adapted to individual experimental conditions during this study but used in the same configuration across all tissue images. The AI tool distinguishes between tumor and nontumor cells, which facilitates its usage in the tumor microenvironment. For this reason, a precise preselection of tumor areas, as in other studies,[6,7,12] was not necessary in the main study. This capability was particularly relevant in an additional set of experiments that investigated the AI without human intervention: the AI was run on complete tissue slide images and in predefined ROIs, including in both cases large areas of nontumorous tissue.

The statistical analysis showed that pathologists reached a sufficient agreement (defined as at least 75%) when scoring by themselves and when using AI assistance. It can be concluded that it is safe to use this AI tool for diagnostic assistance. However, in a few (9) cases with intraobserver disagreement between rounds 1 and 2, scores changed by more than 50% (Supplementary Fig. S3). On reviewing these cases, this was mostly due to significant differences in what was preselected manually by pathologists as the region of invasive tumor and at what intensity of staining a cell was considered positive by them.

Table 8
Agreement rates of pathologists with the results of the artificial intelligence tool in round 2 with 95% CIs across cases (number of agreeing pathologists per case)

No. of pathologists agreeing with AI	Staining	Cases, % (95% CI)
Majority (at least 6 of 10)	Ki-67	95.8 (94.4-96.9)
	ER/PR	93.2 (91.5-94.6)
All (10 of 10)	Ki-67	68.0 (65.0-70.8)
	ER/PR	63.6 (60.6-66.5)

Note that per case pathologists scored individually chosen ROIs, resulting in 10 different ROIs per case.
AI, artificial intelligence; ER, estrogen receptor; PR, progesterone receptor; ROIs, regions of interest.

Table 9
First additional experiment of artificial intelligence accuracy without human intervention: agreement rates

	Ki-67, % (95% CI)	ER/PR, % (95% CI)
P1/AI	81.8 (79.3-84.0)	81.5 (79.0-83.8)
P2/AI	88.6 (86.5-90.4)	90.7 (88.7-92.4)
P1/P2	75.0 (72.2-77.6)	87.0 (84.8-88.9)

AI, artificial intelligence; ER, estrogen receptor; P1, pathologist 1; P2, pathologist 2; PR, progesterone receptor.

Further analysis revealed that pathologists typically agree with the proposed AI result and confirm it in their final diagnosis. Most pathologists agreed with the proposed AI assistance results in 95.8% of Ki-67 and 93.2% of ER/PR cases. This indicates the potential of relying on automated cell counting with AI assistance after manual definition of ROIs. For some cases, pathologists reported problems with the corresponding tissue image due to preprocessing or scanning. When excluding such cases in the statistical analysis, agreement rates increased.

Ideally, AI assistance biases the pathologist in a positive way: an increase in interobserver reliability was found, although the improvement was not statically significant. The realistic study design based on the assessment of cases (instead of a limitation of scoring in small predefined ROIs) with individual choice of ROIs and including tissue images with suboptimal image quality due to scanning, staining, or preanalytics resulted in considerable interobserver variances, complicating the detection of interobserver effects. The interobserver reliability when using AI assistance (0.72; 95% CI, 0.68-0.76) is similar to the interobserver reliability in the conventional analysis observed in the literature.[6]

A recent large-scale nationwide 5-year study in Sweden on variability in breast cancer biomarker assessment found "unacceptable interlaboratory and intralaboratory variability" in Ki-67 assessment, stating the need for the adoption of new technologies in practice.[5] Moreover, the study by IKWG 4R[24] on assessment of Ki-67 without AI assistance showed that a higher level of interobserver concordance can be achieved with more formalized methods of scoring field selection with median scoring times of 6 to 9 minutes. AI assistance tools similar to the one investigated in this study have the potential to substantially reduce such scoring times and further increase the interobserver concordance.

To put the results of the main study into context, additional experiments on the scoring accuracy of the AI tool without human

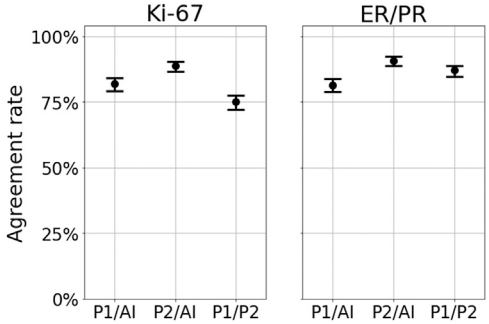

Figure 7.
AI accuracy without human intervention for predefined regions of interest: agreement rates in categorized scores with 95% CIs. AI, artificial intelligence; ER, estrogen receptor; P1, pathologist 1; P2, pathologist 2; PR, progesterone receptor.

Table 10
First additional experiment of artificial intelligence accuracy without human intervention: distribution of cases (based on scores of pathologist 2).

Staining	Score, %	Cases
Ki-67	<10	12
	10-40	21
	>40	11
ER/PR	<1	14
	1-9	10
	≥10	30

ER, estrogen receptor; PR, progesterone receptor.

intervention were performed. First, in a separate set of predefined ROIs the AI results were compared with the results of 2 human pathologists. The highest agreement rates for both Ki-67 and ER/PR were between the AI and one of the pathologists. AI–human agreement was on the same level as human–human agreement, indicating that the AI tool without human intervention achieves the same level of accuracy as pathologists (Supplementary Fig. S2).

Second, the AI was run on complete tissue images for Ki-67 quantification without human intervention, using 3 different scoring methods (hotspot, IKWG 4R, and all tumor cells). The results provided insights into differences between methods and potential challenges in scoring with conventional methods: When the selection of the hotspot was performed by the AI alone, the single connected region of 400 tumor cells with the highest proliferation was selected by the AI. In cases where tumor detection by the AI was flawless, this resulted in scores higher than the average scores acquired by the conventional method. This might be due to humans having difficulties in finding hotspot areas with the highest proliferation. In cases where proliferation was heterogeneous, the IKWG method applied by the AI resulted, as expected, in slightly lower scores than those using the hotspot method. Of interest, for most cases, the analysis of all tumor cells on the scan resulted in even lower scores than the IKWG method produced. This raises the question whether the IKWG method can always reflect the relative proliferation across a whole slide. However, in contrast to the results discussed in this study, the intended use of the investigated AI tool does require human supervision for quality control of the analyzed image because AI results on complete images can fail due to subpar image quality: for 2 cases, no results were produced for the hotspot and the IKWG methods (Fig. 8); for 8 cases, results were outside the range of human-only results from round 1 of the main study. On reviewing these cases, the following reasons for AI failure could be identified: folded tissue areas with staining artifacts were sometimes chosen as hotspot areas by the AI, resulting in erroneously high proliferation results; dirt and debris were detected as positive staining; and faintly stained/low-contrast negative tumor cells could be missed. Although the main study with human supervision was run on the same samples, these issues had less impact and were largely avoided in the main study because of the selection of relevant areas without image artifacts by a human pathologist before the analysis.

Thus, both additional experiments on using the AI without human intervention showed high accuracy of the AI but supported the initial premise of using this kind of AI-based tool as assistance tools with human observation and intervention.

Using AI as an assistance tool may have several benefits: assuring accuracy, reproducibility, increasing intraobserver and interobserver reliability, and improvements in turnaround times. Experimentally assessing turnaround time is nontrivial and will require participants to have significant experience in using the AI tool under investigation. In this study, participants were given instructions for using the AI tool, but most of them did not have experience in using the tool beyond these instructions. Furthermore, participants were allowed to use semiquantitative methods for manual assessment, which may be less accurate but hard to accelerate by means of AI assistance. Experimentally assessing turnaround time in a controlled way is left for future work.

The focus of the study design to reflect clinical routine and incorporate a wide array of preanalytical variables introduced necessarily some limitations on the results. The selection of different ROIs by each participant in each round of the main study might have caused some of the more extreme discrepancies between rounds 1 and 2. Because the validation of noninferiority regarding clinical outcomes is tied to the relevant cut offs, study results were dichotomized for the statistical analysis of the main study, although additional continuous results are provided (Fig. 5). Comparing results between pathologists on a per-case basis but

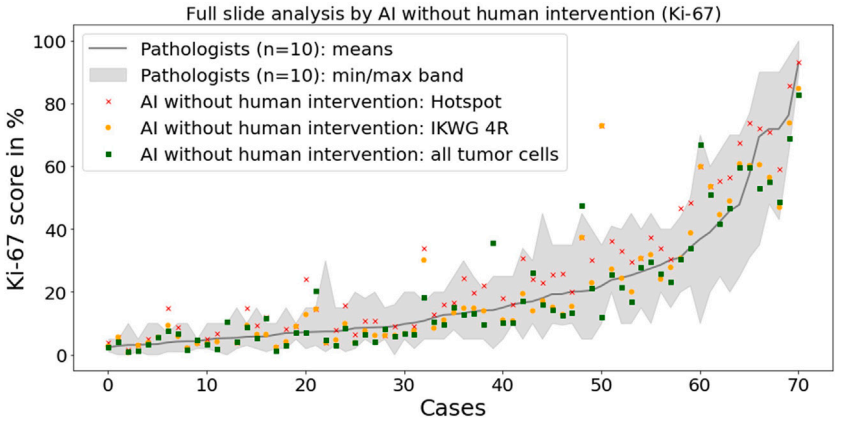

Figure 8.
Full-slide analysis by the artificial intelligence (AI) tool without any human intervention. Three automated scoring methods based on the tumor cells detected by the AI are compared with pathologist scores. The gray band shows the range of pathologist scores (min/max scores) without AI assistance. Cases are shown sorted by mean pathologist scores for better visualization (actual case order within the study was different, with case orders varying across participants). The analysis of all tumor cells resulted in higher scores than the hotspot method in a few cases where the AI detected isolated small highly proliferative groups of tumor cells that were too dispersed to be included in a single hotspot. IKWG, International Ki-67 in Breast Cancer Working Group.

not on a per-ROI level made it more difficult to assess the benefits of AI assistance regarding interobserver variability. A more restrictive, though less representative image selection excluding images with subpar quality due to staining artifacts, air bubbles under the cover slip, and unfamiliar stainings caused by the variety of preanalytical variables would have reduced scoring difficulties in some cases for both pathologists and the AI tool. This would have led to even higher accuracy levels as sometimes observed in restrictive, less clinically representative study settings.

In future work, accuracy improvements when using AI assistance will be studied in more detail. This will require a different and more controlled study design at the price of a less realistic scenario where ROIs are predefined for assessment since ROI selection is a major concern in standardizing Ki-67 evaluation.[25] While in the main part of this study there was no absolute ground truth to compare AI results with due to individually chosen ROIs, further experiments with controlled ROIs will serve to investigate whether acceptance or rejection of AI results will correlate with closeness to gold standard results. Another future avenue of research is to investigate whether the use of AI diagnostic tools helps to increase diagnostic speed and reduce interobserver variance in clinical routine. Because the study results indicate reliable AI assistance results within ROIs, a promising next step is to explore AI assistance for finding ROIs. Ethical issues for using AI in pathology will need to be considered for successful practical adaptation of AI tools.[26] Ultimately, this will increase overall accuracy and allow the selection of the most suitable individualized treatment for the patient.

In conclusion, this study investigated the application of an AI-based assistance tool for IA in IHC quantification in a realistic study set representative of clinical routine, including a large-scale number of environmental conditions such as 10 participant pathologists from 8 sites, 6 WSI scanners/microscopes, and 3 staining machines. The potential and accuracy and the limitations of AI tools with and without human intervention were investigated. The results showed the safety of using the assistance tool with a statistical significance.

Author Contributions

N.A., T.L., K.D., M.P., and A.W. conceptualized and designed the study. K.T., T.K., A.W, F.L., A.P., A.D., F.K., M.M., A.M., E.M., and M.K. acquired and annotated the data. L.K., M.S., and A.Z. performed the statistical analysis. N.A., T.L., C.S., and A.H. wrote, reviewed, and revised the paper. All authors read and approved the final paper.

Data Availability

The datasets used and/or analyzed during this study are mostly available from the corresponding author upon reasonable request. Restrictions apply for the availability of the remaining data that were used under license for this study and for which data protection restrictions apply. However, these data are available from the corresponding author upon reasonable request with permission of the involved institutions.

Funding

No external funding was received for this study.

Declaration of Competing Interest

N. Abele and C. Schaaf are advisors to Mindpeak but were not compensated by Mindpeak for participation in this study. There is no other potential conflict of interest to disclose.

Ethics Approval and Consent to Participate

Ethical approval for this study was granted by the Medical Association Hamburg, Germany.

Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.modpat.2022.100033.

References

1. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406:747–752.
2. Varga Z, Diebold J, Dommann-Scherrer C, et al. How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast- and Gynecopathologists. *PLoS One*. 2012;7:e37379.
3. Polley MY, Leung SC, McShane LM, et al. An international Ki67 reproducibility study. *J Natl Cancer Inst*. 2013;105:1897–1906.
4. Nielsen TO, Leung SCY, Rimm DL, et al. Assessment of Ki67 in breast cancer: updated recommendations from the International Ki67 in Breast Cancer Working Group. *J Natl Cancer Inst*. 2021;113:808–819.
5. Acs B, Fredriksson I, Rönnlund C, et al. Variability in breast cancer biomarker assessment and the effect on oncological treatment decisions: a nationwide 5-year population-based study. *Cancers (Basel)*. 2021;13:1166.
6. Acs B, Pelekanou V, Bai Y, et al. Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. *Lab Invest*. 2019;99:107–117.
7. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep*. 2017;7:16878.
8. Serag A, Ion-Margineanu A, Qureshi H, et al. Translational AI and deep learning in diagnostic pathology. *Front Med (Lausanne)*. 2019;6:185.
9. Homeyer A, Lotz J, Schwen LO, et al. Artificial intelligence in pathology: from prototype to product. *J Pathol Inform*. 2021;12:13.
10. 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.
11. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
12. Rimm DL, Leung SCY, McShane LM, et al. An international multicenter study to evaluate reproducibility of automated scoring for assessment of Ki67 in breast cancer. *Mod Pathol*. 2019;32:59–69.
13. Allison KH, Hammond MEH, Dowsett M, et al. Estrogen and progesterone receptor testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists guideline update. *Arch Pathol Lab Med*. 2020;144:545–563.
14. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst*. 2011;103:1656–1664.
15. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. *Med Image Comput Comput Assist Interv*. 2013;16:411–418.
16. Loshchilov I, Hutter F. Decoupled weight decay regularization. Paper presnted at International Conference on Learning Representations; May 6-9, 2019; New Orleans, Louisiana.
17. McLachlan GJ, Peel D. *Finite Mixture Models*. Wiley; 2000:419.
18. Robertson S, Acs B, Lippert M, Hartman J. Prognostic potential of automated Ki67 evaluation in breast cancer: different hot spot definitions versus true global score. *Breast Cancer Res Treat*. 2020;183:161–175.
19. Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Ann Oncol*. 2013;24:2206–2223.
20. Roche Ventana Medical Systems. *I*. k121033. Federal Drug Administration; 2012.
21. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. Sage; 2018.

22. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021;27:775–784.
23. Bejnordi BE, Veta M, van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318:2199–2210.
24. Leung SCY, Nielsen TO, Zabaglo LA, et al. Analytical validation of a standardised scoring protocol for Ki67 immunohistochemistry on breast cancer excision whole sections: an international multicentre collaboration. *Histopathology*. 2019;75:225–235.
25. Christgen M, von Ahsen S, Christgen H, Länger F, Kreipe H. The region-of-interest size impacts on Ki67 quantification by computer-assisted image analysis in breast cancer. *Hum Pathol*. 2015;46:1341–1349.
26. Chauhan C, Gullapalli RR. Ethics of AI in pathology: current paradigms and emerging issues. *Am J Pathol*. 2021;191:1673–1683.