Standardized Ki67 diagnostics using automated scoring - clinical validation in the GeparTrio breast cancer study

Frederick Klauschen1†, Stephan Wienert12, Wolfgang D. Schmitt1, Sibylle Loibl3, Bernd Gerber4, Jens Blohmer5, Jens Huober6, Thomas Rüdiger7, Erhard Erbstößer8, Keyur Mehta3, Bianca Lederer3, Manfred Dietel1, Carsten Denkert1* and Gunter von Minckwitz3,9*

1 Institute of Pathology, Charité University Hospital Berlin, Charitéplatz 1, 10117 Berlin, Germany
2 VMscope GmbH, Charitéplatz 1, 10117 Berlin, Germany
3 German Breast Group, c/o GBG Forschungs GmbH, Martin-Behaim-Str. 12, 63263 Neu-Isenburg
4 University Women’s Hospital Rostock, Suedring 81, 18059 Rostock, Germany
5 St. Gertrauden Hospital, Paretzer Str. 12, 10713 Berlin, Germany
6 University Women’s Hospital Ulm, Prittwitzstr. 43, 89075 Ulm, Germany
7 Klinikum Karlsruhe, Moltkestr. 90, 76133 Karlsruhe, Germany
8 AMEOS Klinikum Halberstadt, Gleimstr. 5, 38820 Halberstadt, Germany
9 University Women’s Hospital Frankfurt, Theodor-Stern Kai 7, 60590 Frankfurt, Germany
* These authors contributed equally

correspondence to:
Frederick Klauschen, MD PhD
Institut f. Pathologie
Charité Universitätsmedizin Berlin
Charitéplatz 1
10117 Berlin
Germany
Phone: +49-30-450536-053
Fax: +49-30-450536-900
frederick.klauschen@charite.de

Conflict of interest statement
The authors declare no conflict of interest. S.W.’s Ph.D. thesis was partly funded by VMscope GmbH, the company that is distributing a licensed version of the Ki67 Quantifier described here.

Translational Relevance
Clinical decision-making in breast cancer increasingly depends on the quantitative evaluation of molecular and immunohistochemical markers, such as Ki67 for tumor proliferation assessment. However, recent studies have shown a limited reliability of the established manual Ki67 scoring in breast cancer diagnostics in a multicenter setting. As a solution we present a novel automated Ki67 scoring system, which we validate using data from the prospective neoadjuvant GeparTrio trial. Our approach is readily deployable in routine diagnostics and may thus help solve the recently reported reliability issues in Ki67 scoring.
Abstract

Purpose
Scoring proliferation through Ki67-immunohistochemistry is an important component in predicting therapy response to chemotherapy in breast cancer patients. However, recent studies have cast doubt on the reliability of “visual” Ki67 scoring in the multicenter setting, particularly in the lower – yet clinically important - proliferation range. Therefore, an accurate and standardized Ki67-scoring is pivotal both in routine diagnostics and larger multi-center studies.

Experimental Design
We validated a novel fully-automated Ki67 scoring approach that relies on only minimal a priori knowledge on cell properties and requires no training data for calibration. We applied our approach to 1082 breast cancer samples from the neoadjuvant GeparTrio trial and compared the performance of automated and manual Ki67 scoring.

Results
The three groups of autoKi67 as defined by low (≤15%), medium (15.1-35%) and high (>35%) automated scores showed pCR rates of 5.8%, 16.9%, and 29.5%, respectively. AutoKi67 was significantly linked to prognosis with overall and progression free survival p-values p_{OS}<0.0001 and p_{PFS}<0.0002, compared to p_{OS}<0.0005 and p_{PFS}<0.0001 for manual Ki67 scoring. Moreover, automated Ki67 scoring was an independent prognosticator in the multivariate analysis with p_{OS}=0.002, p_{PFS}=0.009 (autoKi67) vs. p_{OS}=0.007, p_{PFS}=0.004 (manual Ki67).

Conclusions
The computer-assisted Ki67 scoring approach presented here offers standardized means of tumor cell proliferation assessment in breast cancer that correlated with clinical endpoints and is deployable in routine diagnostics. It may thus help to solve recently reported reliability concerns in Ki67 diagnostics.
**Introduction**

The prognostic and predictive utility of the proliferation marker Ki67(1) has been studied extensively in breast cancer(2-8). Unlike other immunohistochemical markers, such as ER and PR, Ki67 requires a more precise quantification to stratify patients into different risk groups, which is usually done manually in routine diagnostics. However, Ki67 scoring by visual inspection has been shown to suffer from a considerable inter- and intra-observer variability(9, 10), which is likely partly due to the fact that only a limited number of cells are evaluated in routine diagnostics. In this context, automated image analysis may help to streamline and standardize Ki67 scoring, for which several approaches have been proposed in the recent years by academic groups(11-13) and companies. However, the previous approaches have not been validated using a prospective-retrospective trial design considered necessary to reach a sufficient level of evidence to determine clinical utility when validation studies rely on archived tissue samples(14). Here, we present a novel robust and easy to use Ki67 scoring approach that follows the International Ki67 in Breast Cancer Working group guidelines(9) and is based on a cell detection method and image analysis framework we recently developed(15, 16). We validate our approach by applying it to 1082 breast cancer cases from the prospective neoadjuvant GeparTrio trial(17, 18).

**Methods**

**Patient characteristics and treatment in the GeparTrio breast cancer cohort**

Overall, 2090 breast cancer patients had been recruited in the GeparTrio breast cancer cohort who had received no prior treatment. 2072 patients started chemotherapy with TAC (docetaxel-doxorubicin-cyclophosphamide), out of which 1166 tissue specimens were available here. For detailed patient characteristics see table S1.

**Specimen characteristics**

Core needle biopsies of breast tumors from patients acquired in a routine diagnostic setting that were formalin-fixed and paraffin-embedded were used in this study. Tissue specimens were fixated at room temperature in 10% neutral-buffered formalin immediately after biopsies were acquired (max. time to fixation 3min). Time of fixation was 16-72h.
Subsequently, specimens were paraffin-embedded and further processed on the same or the following day.

**Assay methods**

**Immunohistochemical staining and image acquisition**

1166 digital images were taken from Ki67 stained whole core biopsies using a conventional light microscope (Leica DMRB and Olympus BX51) and microscope camera (both JVC KY-F75U). Ki67 immunohistochemistry was performed using clone MIB-1, 1:50, DAKO, Hamburg, Germany on an autostainer, Ventana, Tucson, USA, using the standard protocol “xt ultraview dab v3”. After visual inspection by an experienced pathologist 84 images were considered not suitable for automated image analysis because of poor specimen or image quality.

**Automated Image Analysis**

The automated image analysis approach for computer-assisted Ki67 scoring we present here is based on a cell segmentation method developed in (16), of which an implementation is available in the open source image analysis framework CognitionMaster(15). To extend the approach to Ki67 immunohistochemistry assessment the following modifications were necessary: While the original cell detection imposed very loose restrictions on cell nucleus sizes we improved breast cancer cell detection by estimating the range of nucleus sizes in a pre-segmentation (0.95 percentile of the area of convex objects is used as target size). Moreover, we replaced Otsu’s method (19) previously used to determine the threshold to distinguish cell nuclei from image background, because it is based on the concept that an image is essentially separable into two intensity classes. However, typical histological images are characterized by the presence of more than two intensity classes, such as no tissue, tissue but no cell nuclei, Ki67 negative cell nuclei and Ki67 positive cell nuclei each showing a certain haematoxylin intensity (Figure 5 and supplemental Fig. 1). In the modified cell nucleus detection method presented here, initially, all objects with a minimum haematoxylin intensity and positive intensity gradient (objects with higher haematoxylin intensity compared to the adjacent objects) are selected. Consecutively, the histograms for the (temporary) image background and foreground are computed (Figure 5 and supplemental Fig. 1C). Then, objects are iteratively removed from the foreground that fit more into the image background. Finally, pixels that were not detected by the original
method, but can be associated with a cell nucleus via the background and foreground histogram are set to “foreground” and are then grouped via connected component analysis. To classify cell nuclei as Ki67 negative or positive, first, the IHC- and counter-stain are separated through color deconvolution(20). Consecutively, the optimal threshold is computed by testing all possible thresholds (1..255). The F₁ score (the harmonic mean of precision and recall) and a plausibility score are used to evaluate each threshold and finally the threshold with the best combination of F₁ score and plausibility is used. For a more detailed description of the threshold finding method please see supplemental methods online. Finally, a cell nucleus is classified as Ki67 positive if at least five percent of the nucleus area are above the threshold. Another important aspect of Ki67 scoring is the classification of cell nuclei in tumor and non-tumor. Here, we use the MaximumDistanceToBorder feature available in CognitionMaster framework to distinguish between tumor and stromal cells. To this end, the 0.95 percentile of all cell nuclei is used to define the target value typical of tumor cells and objects below 33 percent of the target value are marked as non-tumor. Furthermore, objects that have a value smaller than 50 percent of the target value and an aspect ratio lower than 0.5 are marked as non-tumor, too. Please note that this routine is sometimes does not classify all cell nuclei into tumor and non-tumor correctly. However, first, regions can be interactively refined by the pathologist to exclude problematic region, if necessary, but more importantly, note that our automated Ki67 scoring validation was successful without any such manual intervention.

**Manual and Automated Ki67 scoring**

The manual Ki67 scoring was performed by one out of three pathologists (central counting) for each tumor who examined the complete slide of the core biopsy and who selected one area for exact counting. In cases with a heterogenous distribution of Ki67, an area with the highest expression (the so-called hot spot) was selected. The selected area was documented in a microphotograph, and this microphotograph was used for both exact manual counting of 200 adjacent invasive tumor cells (compare (3)) as well as for automated image analysis. Already a weak nuclear staining was regarded sufficient for Ki67 positivity. For the automated scoring a software tool based on the image analysis method described in the previous section was implemented using the CognitionMaster platform that may be used in single- or multi-case mode. Multi-case batch processing was performed for 1082 out 1166
Ki67-stained immuno-histological images (84 images were excluded due to poor image quality). These images were identical to those that we used for manual scoring acquired at 200x and a resolution of 1240x800 pixels by pathologists using a standard microscope – microscope camera set-up. The software automatically sequentially loads the image files and saves visual and numeric result representations. Processing time was 1502 minutes for all 1082 images, with an average processing time of 1 minute 23 seconds per image. All processing shown here was performed on a standard PC running Microsoft Windows 7.

Study design

GeparTrio breast cancer study cohort

The GeparTrio trial (NCT00544765) is a prospective randomized phase III study comparing neoadjuvant vinorelbine-capecitabine vs. docetaxel-doxorubicin-cyclophosphamide chemotherapy in early nonresponsive breast cancer. Overall, 2090 patients were centrally confirmed, 2072 patients started chemotherapy with TAC. For overall 1166 patients Ki67 immunohistochemistry was assessed in core biopsies. Clinical endpoints were pathologic complete remission (pCR, defined as no residual histological evidence of tumor after chemotherapy at the time of surgery), overall survival (OS, defined as survival time after diagnosis) and disease-free survival (DFS, defined as time interval from therapy to disease recurrence). The multivariate analysis included variables age, stage, nodal status, histology, grade, hormone receptor and Her2 status. Sample size was defined by the original GeparTrio trial; here, all patients were included for whom tissue samples were available. Clinical data on pathological complete response (pCR) was available for 1166 patients. Clinical data on disease-free and overall survival was available for 1101 patients. Patients had given prior consent for tissue collection and biomarker analysis. Clinical follow-up data has been updated since the previous report on the GeparTrio cohort in (3). For the Consort statement see Fig. 1, further details see supplement.

Statistical analysis methods and Ki67 cutoff selection

A statistical analysis plan was prepared prior to the start of this study and the pathologists performing the scoring analysis were blinded for the clinical data. Clinical data were available only for the collaborators at the German Breast Group to whom the scoring results were sent for further analysis: Time-to-event outcome parameters were estimated using Kaplan-Meier product-limit method and treatment groups were compared by log-rank test.
Cox proportional hazards models were used to calculate hazard ratios (HR) including 95% confidence intervals (CI). All statistical tests were two-sided by default. $\chi^2$ test was used for trends for pCR rates. Multivariate analysis adjusted for age ($\leq$40 yrs), cT1-3 vs. cT4, cN0 vs. CN1, lobular histology vs. others, G1,2 vs. G3, HER2 and HR status. For consistency purposes and to ensure comparability we chose the same cut-off points that were previously used for the manual analysis(3), where they had been determined in a systematic way using the Cutoff Finder algorithm(21) based on the results for pCR, DFS and OS as well as results from previous studies including the St. Gallen recommendations (4, 22, 23): Cases were grouped as $\leq$15% Ki67-positive tumor cells (low), 15.01%-35% (intermediate) and $>$35% (high). Statistical, and in particular survival analysis was performed using SPSS Statistics software package (Version 21, IBM). See supplement for the detailed statistical analysis plan (SAP).

Results

Baseline clinical data

The clinicopathological characteristics of the study cohort are shown in supplementary Table S1 (see also (17, 18)). The median follow-up was 82.4 months for DFS and 91.0 months for OS. For the flow of patients through the study see Consort statement in Fig. 1.

An exemplary visualization of the automated Ki67 scoring results is shown in Figure 2.

Correlation between manual and automated Ki67 scores and intra-class correlation

We performed computer-assisted automated Ki67 scoring based on digitized images, for which manual Ki67 scores were determined previously. The results show a correlation between manual and automated Ki67 scores of $r_{\text{Pearson}}=0.89$ (Figure 3). The intraclass correlation between automated and manual scores was $r_{\text{ICC}}=0.93$ ($p<0.001$).
Comparison of tumors with low, intermediate and high Ki67 expression for pCR, DFS and OS by univariate and multivariate analyses

The pCR rates of the groups with low, intermediate and high Ki67 were 5.9%, 15.6% and 31.6% (manual: 3.3%, 13.0% and 29.6%) ($P < 0.001$, $\chi^2$ test for trends), the odds ratios of the three groups were 1.00, 2.95 and 7.36 (manual: 1.0, 4.4 and 12.4) ($P < 0.001$, logistic regression). These findings remain statistically significant after adjusting for clinicopathological parameters ($P < 0.001$, logistic regression).

For autoKi67, disease-free survival (DFS) and overall survival (OS) were also significant on univariate (DFS: $p = 0.00011$, OS: $p = 0.00005$, Figure 4) and multivariate analysis (DFS: $p<0.01$, OS: $p = 0.002$) for the three groups. Corresponding significance levels for manual Ki67 scores in univariate analysis were $p<0.0001$ (DFS) and $p<0.0005$ (OS) and for multivariate testing $p<0.01$ (DFS) and $p<0.01$ (OS). Hazard ratios for autoKi67 low, intermediate and high were HR(DFS)=1, 1.6, 1.8 and HR(OS)=1, 1.9, 2.2. Corresponding hazard ratios for manual scoring for DFS were HR(DFS)=1, 1.7, 1.9 and for OS HR(OS)=1, 1.5 and 2.2, respectively (see table 1 for confidence intervals).

Discussion

In this study, we present a novel computer-assisted Ki67 scoring method for breast cancer. We validate the approach in a large cohort of breast carcinomas from a prospective neoadjuvant clinical trial and compare its performance to manual scoring results presented earlier(3). While several automated Ki67 evaluation approaches have been proposed (e. g. (11-13)), our approach is the first whose performance is compared to thorough manual scoring and evaluated using tissue samples from a large prospective clinical trial. Because of the size of the trial of over 1000 patients for whose tumors manual and automated Ki67 scoring were performed, a broad variety of breast cancer samples is covered demonstrating the robust performance of our approach across morphologies. Unlike existing methods, we designed our automated Ki67 scoring in a way that no training, calibration or user interaction is needed. The approach by Mohammed et al. (13), for instance, requires observer-specified (fixed) intensity thresholds to classify nuclear Ki-67
staining, while Konsti et al. (12) use a similar method that requires the user to define intensity thresholds also to classify hematoxylin into cell nuclei and background. However, defining parameters prior to analysis is very susceptible towards morphological variability typical of many tumors or technical variation and therefore requires constant adjustment. Other approaches for automated immunohistology evaluation such as, for instance, described or applied in (11, 24, 25) rely on training of certain cell and tissue features, a procedure that is usually time consuming and often requires re-training of the method prior to analyzing new images. Our approach is capable of providing reliable results without any training, calibration or interaction performed by the user, because we combine very generic assumptions on cell features with automated adaptive threshold search to detect Ki67-positive and negative cells. This minimizes detection biases due to variation of morphological features or staining variation between cases. Selecting or excluding certain tissue regions or cells is optional, but was not used in this study.

Despite the described technical advantages over previous approaches it is neither the intention nor within the scope of this study to show that our approach performs better than other automated scoring systems. The aim of this study was to present and validate an easy-to-use automated Ki67 scoring method in a prospective-retrospective clinical study approach using samples from a large prospective clinical trial. This validation, that has to our knowledge not been performed with any other automated Ki67 analysis system, is in our opinion essential to show the clinical utility and to promote reliable quantitative histological diagnostics.

Interestingly, the overall results are relatively similar between manual and automated approaches, which might appear surprising given the high variability of manual Ki67 scoring reported previously (10). However, this is likely due to the fact that the manual scoring results were obtained by using a standardized stringent monocentric evaluation with exact counting of cells rather than semiquantitative estimations. Therefore, automated Ki67 scoring can be expected to perform more robustly than manual approaches in routine diagnostics where a considerable intra- and interobserver variability has been observed for manual scoring (9). But even if manual scoring performed similarly also in routine settings, would our automated approach have several advantages over manual scoring: Whereas exact manual counting of about 500 cells takes between about 6 and 10 min, our approach requires about 30 to 40 sec to score Ki67 in a typical routine diagnostic field of view. Similary, the automated system is capable of evaluating the complete cohort of over 1,000 cases within a day, which is not feasible manually.

As a caveat, despite the fact that the international Ki67 consortium has shown that interobserver variation is one main reason for the limited reproducebility of Ki67 scoring (9) it is certainly not the only cause and automated methods will not be the ultimate solution. Selection of
representative tumor regions, for instance, particularly in tumors showing a heterogeneous Ki67 expression with “hot spots” or strong variability in staining protocols will need to be addressed separately. However, the reduction of observer-related variation is an important step and automated methods are also necessary to study questions like tumor heterogeneity, because large amounts of image data need to be processed in a standardized manner to address such questions.

Regarding the broader use of our approach in routine diagnostics, recent implementation also at other sites appears to be confirming the robustness of our method and another systematic study addressing automated Ki67 scoring performance under routine diagnostic conditions including the analysis of samples processed in different laboratories is currently under way.

To summarize, our approach allows for a robust and standardized fully automated Ki67 scoring that may therefore contribute to a solution to the widely critized variations observed in current manual evaluation that have fundamentally put the utility of Ki67 as a prognostic and predictive biomarker in breast cancer diagnostics into question.

Author contributions

FK, CD and GM directed the study. GM, SL, BG, JB, JH, TR and EE provided the clinical data and BL and KM performed parts of the survival analysis. FK and SW designed the algorithm analyzed the results. SW implemented the algorithm. FK, CD, WS, MD and SW tested and evaluated the software performance and wrote the paper.

Figure Legends

Figure 1: Consort statement indicating the flow of samples through the study.

Figure 2: Exemplary visualization of Ki67 scoring result. A, C: original images showing Ki67-DAB stain and Hematoxylin background stain; B, D: corresponding analysis results (green:
tumor cells classified as Ki67 negative; red: Ki67 positive tumor cells; black: stromal cells automatically detected and excluded from analysis).

Figure 3: Scatterplot between automated versus manual scoring. The automated scoring showed good correlation to the manual counting results (Pearson correlation coefficient=0.89, Fig. 2).

Figure 4: Survival analysis results for automated and manual scoring for the three Ki67 subgroups <15% (blue), 15.01 to 35% (green) and >35% (yellow). A, B: Disease-free survival (DFS). C, D: Overall survival (OS).

Figure 5: Image analysis flow chart: Hematoxylin-DAB breast cancer Ki67 immunohistology (A). Contours of the initial segmentation of all available contours (B). HDAB intense and convex objects (C) are used to estimate the target size of cell nuclei to select optimal contours from the initial segmentation (D). Finally, objects are classified into non-tumor (black), Ki67 negative tumor cells (green) and Ki67 positive tumor cells (red) and used for Ki67 scoring (E).

References


Figure 1: Consort statement indicating the flow of samples through the study.
Figure 4

(A) Survival Functions

Cum Survival vs DFS_period

AUTO

(B) Survival Functions

Cum Survival vs DFS_period

MANUAL

(C) Survival Functions

Cum Survival vs OS_period

AUTO

(D) Survival Functions

Cum Survival vs OS_period

MANUAL
Figure 5

- A: Original Image
- B: Full Contour Search
- C: Target Size Estimation
- D: Optimized Cell Nuclei Segmentation
- E: Classification & Counting
Table 1: Univariate and multivariate of automated and manual Ki67 for overall and disease free survival with hazard ratios.

<table>
<thead>
<tr>
<th>All cases (N=1082)</th>
<th>auto</th>
<th>manual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DFS [95% CI]</td>
<td>HR(DFS)</td>
</tr>
<tr>
<td>Ki67 &lt;15%</td>
<td>7.2 [6.9 7.4]</td>
<td>1</td>
</tr>
<tr>
<td>p-value (univar.)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>p-value (multivar.)</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
Standardized Ki67 diagnostics using automated scoring - clinical validation in the GeparTrio breast cancer study

Frederick Klauschen, Stephan Wienert, Wolfgang Schmitt, et al.

Clin Cancer Res Published OnlineFirst December 11, 2014.

Updated version  Access the most recent version of this article at: doi:10.1158/1078-0432.CCR-14-1283

Supplementary Material  Access the most recent supplemental material at: http://clincancerres.aacrjournals.org/content/suppl/2014/12/12/1078-0432.CCR-14-1283.DC1

Author Manuscript  Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.