Digital image analysis outperforms manual biomarker assessment in breast cancer

Stålhammar, Gustav; Fuentes Martinez, Nelson; Lippert, Michael; Tobin, Nicholas P.; Mølholm, Ida; Kis, Lorand; Rosin, Gustaf; Rantalainen, Mattias; Pedersen, Lars; Bergh, Jonas; Grunkin, Michael; Hartman, Johan

Published in:
Modern Pathology

Link to article, DOI:
10.1038/modpathol.2016.34

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Digital image analysis outperforms manual biomarker assessment in breast cancer

Gustav Stålhammar1,2, Nelson Fuentes Martinez1,3, Michael Lippert4, Nicholas P Tobin5, Ida Mølholm4,6, Lorand Kis7, Gustaf Rosin1, Mattias Rantalainen8, Lars Pedersen4, Jonas Bergh1,5,9, Michael Grunkin4 and Johan Hartman1,5,7

1Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden; 2St Erik Eye Hospital, Stockholm, Sweden; 3Södersjukhuset, Stockholm, Sweden; 4Visiopharm A/S, Hoersholm, Denmark; 5Cancer Center Karolinska, Stockholm, Sweden; 6Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark; 7Department of Clinical Pathology, Karolinska University Hospital, Stockholm, Sweden; 8Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden and 9Department of Oncology, Karolinska University Hospital, Stockholm, Sweden

In the spectrum of breast cancers, categorization according to the four gene expression-based subtypes ‘Luminal A,’ ‘Luminal B,’ ‘HER2-enriched,’ and ‘Basal-like’ is the method of choice for prognostic and predictive value. As gene expression assays are not yet universally available, routine immunohistochemical stains act as surrogate markers for these subtypes. Thus, congruence of surrogate markers and gene expression tests is of utmost importance. In this study, 3 cohorts of primary breast cancer specimens (total \( n = 436 \)) with up to 28 years of survival data were scored for Ki67, ER, PR, and HER2 status manually and by digital image analysis (DIA). The results were then compared for sensitivity and specificity for the Luminal B subtype, concordance to PAM50 assays in subtype classification and prognostic power. The DIA system used was the Visiopharm Integrator System. DIA outperformed manual scoring in terms of sensitivity and specificity for the Luminal B subtype, widely considered the most challenging distinction in surrogate subclassification, and produced slightly better concordance and Cohen’s \( \kappa \) agreement with PAM50 gene expression assays. Manual biomarker scores and DIA essentially matched each other for Cox regression hazard ratios for all-cause mortality. When the Nottingham combined histologic grade (Elston–Ellis) was used as a prognostic surrogate, stronger Spearman’s rank-order correlations were produced by DIA. Prognostic value of Ki67 scores in terms of likelihood ratio \( \chi^2 \) (LR \( \chi^2 \)) was higher for DIA that also added significantly more prognostic information to the manual scores (LR−\( \Delta \chi^2 \)). In conclusion, the system for DIA evaluated here was in most aspects a superior alternative to manual biomarker scoring. It also has the potential to reduce time consumption for pathologists, as many of the steps in the workflow are either automatic or feasible to manage without pathological expertise.

Modern Pathology (2016) 29, 318–329; doi:10.1038/modpathol.2016.34; published online 26 February 2016

Breast cancer is still by far the most common cause of cancer death among women worldwide.1 The World Health Organization suggests a largely morphological classification of this heterogeneous disease,2 whereas categorization according to the four gene expression-based ‘intrinsic’ subtypes ‘Luminal A,’ ‘Luminal B,’ ‘HER2-enriched,’ and ‘Basal-like’ is the method of choice for prognostic and predictive value.3–8 However, gene expression tests are not universally available in clinical practice, as they are still rather expensive and time consuming.9 This has created an opportunity for routine immunohistochemical stains to act as surrogate markers (biomarkers) for the gene expression-based subtypes. As recommended by international expert consensus,3–5 primarily four biomarkers are analyzed during the routine pathological work-up of breast cancer specimens: estrogen receptor-\( \alpha \) (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and the proliferation-associated nuclear protein Ki67. Assessments of these biomarkers are then combined into surrogate subtype classifications, guiding
Table 1 Characteristics of patients and material included in this study, and CONSORT diagram indicating which patients were evaluated for PAM50, clinical, survival, manual, and digital image analysis immunohistochemical data from the breast cancer registry and individual patient journals

<table>
<thead>
<tr>
<th>Cohort no.</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients</td>
<td>195</td>
<td>84</td>
</tr>
<tr>
<td>Mean age at breast cancer diagnosis (years)</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td><strong>Histologic grade</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>31</td>
<td>19</td>
</tr>
<tr>
<td>G2</td>
<td>90</td>
<td>35</td>
</tr>
<tr>
<td>G3</td>
<td>74</td>
<td>30</td>
</tr>
<tr>
<td>Mean tumor diameter (mm)</td>
<td>25</td>
<td>22</td>
</tr>
<tr>
<td>Proportion of patients with axillary lymph node metastasis</td>
<td>29%</td>
<td>38%</td>
</tr>
<tr>
<td>PAM50 Luminal A, n (% of cohort)</td>
<td>123 (63)</td>
<td>24 (29)</td>
</tr>
<tr>
<td>PAM50 Luminal B, n (% of cohort)</td>
<td>50 (26)</td>
<td>17 (20)</td>
</tr>
<tr>
<td>PAM50 HER2-enriched, n (% of cohort)</td>
<td>10 (5)</td>
<td>10 (12)</td>
</tr>
<tr>
<td>PAM50 basal-like, n (% of cohort)</td>
<td>12 (6)</td>
<td>16 (19)</td>
</tr>
<tr>
<td>PAM50 normal-like, n (% of cohort)</td>
<td>0 (0)</td>
<td>17 (20)</td>
</tr>
<tr>
<td>PAM50 subtype not available, n (% of cohort)</td>
<td>27 (12)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>10-Year overall survival</td>
<td>N/A</td>
<td>62%</td>
</tr>
<tr>
<td>Type of material contributed to this study</td>
<td>Full sections (Ki67) and TMA (ER, PR, and HER2)</td>
<td>Full sections (Ki67)</td>
</tr>
</tbody>
</table>

Cohort 1

- Primary tumor paraffin blocks unavailable (n=65)
- Unable to determine PAM50 subtype (n=1)
- Unable to access clinical data and/or manual scores of biomarkers (n=21)
- <100 tumor cells in TMA cores (n=16)

Cohort 2

- Digital imaging failed (n=5)
- Unable to align or analyze image files in DIA software
  - Due to tissue or tissue artifacts (n=2)
  - Due to insufficient amount of tissue for analysis (n=4)
  - Due to missing image identification (n=3)
conclusions about the tumors’ biological characteristics and expected response to therapy.\textsuperscript{3–5,9,10} Congruence of evaluations of these surrogate markers to the gene expression tests are consequently of utmost importance, not least as discrepancies in classification induces dissimilar treatment decisions such as on which patients to give cytotoxic chemotherapy. Unfortunately, assessments of biomarker status struggle with intra- and interobserver variability, as well as discordance with the gene expression tests.\textsuperscript{11,12} This is perhaps especially evident for Ki67\textsuperscript{13–15} as there is no consensus on what tumor region or number of cells to score\textsuperscript{13,16,17} and what cutoff values for the proportion of positive cells (Ki67 index) distinguish highly from lowly proliferative tumors. In fact, even the consensus guidelines that do exist have been considered unreliable outside individual laboratories’ own reference data.\textsuperscript{3,6,17} A threshold proportion of Ki67 positivity within the range of 20 to 29% to distinguish the highly proliferative ‘Luminal B-like’ disease from the lowly proliferative ‘Luminal A-like’ disease is however mentioned,\textsuperscript{5} and at our and several other institutions a cutoff of \( \geq 20\% \) for highly proliferative tumors is commonly used.\textsuperscript{4,18–21} The most recent version of these guidelines mentions that this uncertainty and variability may be reduced by image analysis, but provides no further details on how to apply this to biomarker testing in practice.\textsuperscript{5}

Hence, in this study we aim to contribute with precisely that; we take an equally broad and detailed approach on manual and digital image analysis (DIA) evaluation of biomarkers in invasive breast cancer by comparing a novel system of DIA with the manual immunohistochemical method used in current clinicopathological routine for performance in subclassification and prognostication. Furthermore, we use our three different cohorts to evaluate and suggest methods to improve the concordance to gene expression assays, prognostic power, reproducibility, as well as to reduce time consumption for pathologists.

Materials and methods

Patients and Samples

Two cohorts of primary breast cancer specimens were used for this study, along with a third cohort consisting exclusively of tissue microarrays, as reported in the Supplementary Data (total \( n = 436 \)).

Cohort 1 (\( n = 195 \)) consists of fresh frozen and paraffin-embedded breast cancer tissue from patients who underwent surgery at the Karolinska University Hospital from 1 January 2006 to 31 December 2010. They, along with data on clinically reported manual immunohistochemical and HER2 FISH results, were identified in the population-based Stockholm–Gotland breast cancer registry and individual patient journals after approval from the regional ethical review board. From the paraffin blocks, full sections for DIA Ki67 scoring were prepared as well as a tissue microarray for ER, PR, and HER2 scoring: hematoxylin and eosin-stained slides were used for selection of invasive tumor areas without ductal carcinoma \textit{in situ}, intense inflammation, fibrosis, necrosis, or poor fixation. Then, 4–8 tissue cores (\( \varnothing \ 0.8 \text{ mm} \)) per patient were punched and mounted into a tissue microarray using a semiautomated instrument (Minicore\textsuperscript{3}, Tissue Arrayer, Alphelys, France). After exclusions of patients with incomplete PAM50 gene assay data and/or clinical immunohistochemical data, tissue microarray cores with \( < 100 \) tumor cells,\textsuperscript{22,23} failed digital scanning, and errors in software operation, 195 patients remained for analysis (Table 1).

Cohort 2 (\( n = 84 \)) consists of paraffin-embedded breast cancer tissue from patients who underwent surgery at the Karolinska University Hospital, Stockholm, from 1 January 1994 to 31 December 1996. These were identified in the population-based Stockholm–Gotland breast cancer registry after approval from the regional ethical review board. This cohort has been published previously.\textsuperscript{24,25} The cohort originally included 159 cases of whom 84 had PAM50 data and sufficient paraffin-embedded tumor tissue for glass slide sectioning available, the latter enabling manual scoring of Ki67 by a board-certified pathologist and scanning for DIA (Table 1). A subgroup of 41 tumors classified into Luminal A and Luminal B subtypes was assessed by two additional board-certified pathologists for a brief analysis of interobserver concordance.

A third cohort of 130 consecutive tumor specimens collected at the Department of Pathology, Uppsala University Hospital, Uppsala, Sweden, from 1 January 1987 to 31 December 1989 was also analyzed. Here, ER, PR, HER2, and Ki67 were scored on tissue microarray sections only. Consequently, experimentation with and comparison of different scoring methods was not possible when assessing a heterogeneously distributed biomarker such as Ki67 in this cohort. Full details on the results of manual and DIA scoring, including optimal Ki67 thresholds for the highly vs the lowly proliferational Luminal subtype, congruence to gene expression assays, and overall survival analysis can be found in the Supplementary Data.

Immunohistochemistry

All three cohorts, as well as a separate tissue microarray with 78 tumor cores from 78 random breast cancer tissue specimens that were produced to confirm optimal staining conditions and to allow for calibration of the DIA system, were prepared at the accredited clinical laboratory of the Department of Clinical Pathology, Karolinska University Hospital. The paraffin blocks were cut in 3\( \mu \text{m} \) sections, conditioned in CC1 solution (Ventana Medical Systems, Tucson, AZ, USA) for 36 min (Ki67) to
64 min (PR) and incubated with mouse monoclonal antibodies for CkMNF116 and Ki67 (clone Mib-1) (Dako A/S, Glostrup, Denmark) and rabbit monoclonal primary antibodies (Ventana) for ER (clone SP1), PR (clone 1E2), Ki67 (clone 30-9), and HER2 (clone 4B5) at 35 °C (HER2) or 37 °C (others) for 16 min (Ki67) to 44 min (ER) according to the manufacturer’s instructions, and finally counterstained with hematoxylin (section order in Supplementary Table 6). Cohort 2 was stained with CkMNF116 and Ki67 (clone 30-9) only. Note that ER, PR, and HER2 were stained on tissue microarray slides in all cohorts, as these biomarkers are relative to Ki67 homogenously distributed in breast cancer tissue26–28 and thereby well accepted for analysis in biopsies and tumor cores.29,30

**Gene Expression Assays**

For our first cohort, RNA was extracted from frozen tumor tissue using AllPrep DNA/RNA/Protein mini kit (Qiagen, Hilden, Germany) and assessed to ensure high quality (RIN > 8). Next, 1 μg of RNA was used for rRNA depletion using the Ribo-Zero removal kit (Illumina, San Diego, CA, USA). Stranded RNAseq libraries were then constructed using TruSeq Stranded Total RNA Library Prep Kit (Illumina) at the Science for Life Laboratory (Stockholm, Sweden). Gene-level expression estimates were calculated using HTSeq count version 0.6.1,31 and data were normalized using the TMM method32 in the edgeR package.33 Unaligned RNAseq data from the ‘Cancer Genome Atlas’ breast cancer data set34 were downloaded (n = 1073) and processed through an identical bioinformatics pipeline as the primary data set. A total of 35 observations were excluded as potential outliers based on inspection by PCA. Of the 1038 remaining individuals, 885 had molecular subtype assignments available. Samples classified as ‘Normal-like subtype’ (n = 105) were excluded as the clinical relevance for this subtype has been questioned,35 leaving 780 samples for further analysis. To reduce any potential batch differences between our and the ‘Cancer Genome Atlas’ data sets, the two data sets were preprocessed using the same bioinformatic pipeline and variables were mean centered and scaled to unit variance.

Tumors were then classified according to the PAM50 intrinsic molecular subtype model.7 A nearest shrunken centroid classifier36 was trained on the ‘Cancer Genome Atlas’ data set using the PAM50 gene set.7 Each tumor in our material was then classified into one of the subtypes by application of the nearest shrunken centroid model. Here, it is worth noting that when PAM50 subtyping is applied to a whole tumor, intratumor heterogeneity is not taken into consideration and as such is unlikely to represent each and every subset of clones within the tumor.37–39

**Digital Image Analysis**

After sectioning and staining, all glass slides were digitally scanned at ×20, using a Nano Zoomer 2.0 HT (Hamamatsu Photonics K.K., Hamamatsu, Japan) at the Departments of Clinical Pathology, Danderyd Hospital, Stockholm, and Copenhagen University Hospital, Rigshospitalet, Denmark.

The DIA software used was the Visiopharm integrator system for Windows 7, version 4.6.3.857 (Visiopharm A/S, Hoersholm, Denmark), run on
standard off-the-shelf laptop computers (Apple Cupertino, CA, USA, and Dell Round Rock, TX, USA). The Visiopharm integrator system software utilizes a method for tissue classification based on virtual double staining that automatically distinguishes tumor from stromal tissue. In short, each biomarker slide is aligned with an adjacent 3 μm slide stained with a pancytokeratin marker such as CKMNF 116. This enables exclusion of nonepithelial cells that potentially express the biomarker in question, that is, proliferating Ki67-positive lymphocytes. Thus, only cells that express cytokeratin are eligible for detection of positivity or negativity for the respective biomarker. Individual applications for each biomarker then run the scoring of positive and negative cells itself, with subcellular resolution (Figure 1). Excellent reproducibility with this and similar systems has been shown previously (see specific statistics on reproducibility with the Visiopharm integrator system for each tested biomarker in Supplementary Data).

For Ki67, we evaluated 2 fully automatic and 1 semi-manual DIA methods of scoring. The distinction between fully automatic and semi-manual is that the former needs only the manual actions of importing digitally scanned slide images to the Visiopharm integrator system software, a review of the automatic alignment of biomarker and pancytokeratin slides, and the push of a ‘start’ button, and the latter needs an additional manual definition of a region of interest in which the software runs the analysis. In further detail, the scoring methods tested illustrate three different approaches with regard to what tumor region and number of cells to score:

1. The tumor’s invasive margin (semi-manual),
2. ‘hot spot’ of highest concentration of Ki67-positive tumor cells (fully automatic), and
3. an average Ki67 positivity across the full tumor cross-section (fully automatic). Further description of details in these scoring methods can be found in the Supplementary Data.

**Table 2** Molecular ‘intrinsic’ breast cancer subtypes and surrogate definitions by immunohistochemical profile

<table>
<thead>
<tr>
<th>Intrinsic subtype</th>
<th>Surrogate IHC classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminal A</td>
<td>ER ≥ 1% and/or PR ≥ 20% and HER2 ‘negative’ and Ki67 ‘low’</td>
</tr>
<tr>
<td>Luminal B</td>
<td>1. ER ≥ 1% and/or PR ≥ 20% and HER2 ‘negative’ and Ki67 ‘high’ or</td>
</tr>
<tr>
<td>HER2-enriched</td>
<td>2. ER ≥ 1% and PR &lt; 20% and HER2 ‘negative.’ Any Ki67 or</td>
</tr>
<tr>
<td>Basal-like</td>
<td>3. ER ≥ 1% and/or PR ≥ 1% and HER2 ‘positive.’ Any Ki67</td>
</tr>
<tr>
<td></td>
<td>ER &lt; 1% and PR &lt; 1%, HER2 ‘positive.’ Any Ki67</td>
</tr>
<tr>
<td></td>
<td>ER &lt; 1% and PR &lt; 1%, HER2 ‘negative.’ Any Ki67</td>
</tr>
</tbody>
</table>

% = Proportion of tumor cells stained with the respective biomarker. ‘Positive’ and ‘negative’ = as defined by the American Society of Clinical Oncology and College of American Pathologists recommendations for human epidermal growth factor receptor 2-testing in breast cancer. ‘High’ and ‘low’ = as defined by each laboratory’s own reference data, with threshold generally in the range of 14–29%. 5,19–21

Statistical Methods

In addition to the cutoffs for classification provided by current guidelines, we evaluated cutoffs for Ki67 into ‘high’ and ‘low’ proliferational groups after adjustments by points on receiver operating characteristics curves. For measurement of concordance between manual/DIA surrogate subclassifications to PAM50 gene expression assays, Cohen’s χ statistic was computed. For survival analysis, we used the Kaplan–Meier method, and for hazard of all-cause mortality the Cox regression proportional hazard analysis. Likelihood ratio (LR χ2) and change in LR χ2 (LR Δχ2) were computed for an estimation of the individual scoring methods’ prognostic value and for the relative amount of prognostic information of manual vs DIA Ki67 scores. For cohort 1, which still lacks long-term survival data, Spearman’s rank-order correlations were run to determine the relationship between Ki67 indexes vs Nottingham combined histologic grade (Elston–Ellis), primary tumor diameter, and axillary lymph node status. Differences with a P < 0.05 were considered significant. All P-values were two sided.

The steps in the Visiopharm integrator system workflow requiring manual input were performed by a resident in training (corresponding author). All were blinded to any previous data on biomarker status, clinical and survival parameters, and gene expression assay results.

All statistical analyses were performed using IBM SPSS statistics version 22 (Armonk, NY, USA).

**Results**

**Ideal Tumor Area Fraction**

As described in the Materials and methods, the DIA software evaluated here utilizes a method for automatic exclusion of stroma, lymphocytes, and
other nonepithelial tissue. Thus, the role of the operator is to review the automatic steps of the workflow and to, if desired, manually define regions of interest for the software to process. The operator also has the option to let the software run a fully automatic identification of a tumor’s ‘hot spots’ or a representation of the average biomarker positivity across the full tumor cross-section. The results of the scoring of ER, PR, HER2, and Ki67 are then combined into a surrogate immunohistochemical subclass for the tumor, in the very same way as it is done after manual scoring of the same biomarkers.

To determine the area fraction to score for optimal representation of the average Ki67 score across the full tumor cross-section, a sample fraction study of 20 randomized cases from cohort 1 was conducted. Here, it was determined that scoring 25% of the tumor area was ideal considering variance ($R^2 = 0.991$) and time consumption: scoring 25% took in average $\sim 7$ min per slide on our standard off-the-shelf laptop computers. Scoring smaller areas induced higher variances and scoring larger areas claimed more time: scoring 10% ($R^2 = 0.960$) took $\sim 3$ min, scoring 50% ($R^2 = 0.998$) took $\sim 12$ min, and scoring 100% ($R^2 = 1$) took $\sim 24$ min per slide (see further details of this sample fraction study in Supplementary Data), the latter in stark contrast to the scoring of relatively small invasive margin or ‘hot spot’ tumor areas of $>1000$ cells that took $\sim 1–2$ min each.

**Interobserver Concordance**

For an analysis of interobserver concordance in manual classification of Ki67 ‘high’ vs ‘low,’ a subgroup including PAM50 Luminal A and B tumors from cohort 2 ($n = 41$) was assessed by three independent board-certified pathologists. Applying the $\geq 20\%$ cutoff, interobserver concordance for pathologist 1 and 2 scores of Ki67 clone 30-9 was 80% ($\kappa = 0.57$). Their concordance with pathologist 3 scoring of clone Mib-1 was 66% ($\kappa = 0.10$) and 66% ($\kappa = 0.17$) for pathologist 1 and 2, respectively. Thus, interobserver concordance was moderate when pathologists scored the same Ki67 clone, and very poor when they scored different Ki67 clones (see details in Supplementary Data). This is, as far as comparisons are possible, clearly inferior to the previously published intra- and inter scanner, reagent, and operator reproducibility with the Visiopharm integrator system application for Ki67 scoring ($\kappa = 1.00$, presented in the Supplementary Data).

### Table 3

Sensitivity, specificity, and misclassification percentage for each method of Ki67 scoring using the $\geq 20\%$ cutoff as well as adjusted cutoffs for separation of PAM50 Luminal B from A subtypes after analyses of receiver operating characteristics where maximum sensitivity and specificity were given equal importance (receiver operating characteristics curves and area under the curve in Supplementary Figures 4a and b).

<table>
<thead>
<tr>
<th>Ki67 scoring method</th>
<th>Sensitivity for PAM50 Luminal B vs A</th>
<th>Specificity for PAM50 Luminal B vs A</th>
<th>Proportion misclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DIA invasive margin</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutoff $\geq 20%$</td>
<td>84%</td>
<td>78%</td>
<td>20%</td>
</tr>
<tr>
<td>Cutoff $\geq 20.2%^*$</td>
<td>82%</td>
<td>79%</td>
<td>20%</td>
</tr>
<tr>
<td><strong>DIA hot spot</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutoff $\geq 20%$</td>
<td>90%</td>
<td>65%</td>
<td>24%</td>
</tr>
<tr>
<td>Cutoff $\geq 25.2%^*$</td>
<td>86%</td>
<td>77%</td>
<td>19%</td>
</tr>
<tr>
<td><strong>DIA average</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutoff $\geq 20%$</td>
<td>60%</td>
<td>90%</td>
<td>31%</td>
</tr>
<tr>
<td>Cutoff $\geq 15.5%^*$</td>
<td>80%</td>
<td>83%</td>
<td>19%</td>
</tr>
<tr>
<td><strong>Manual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutoff $\geq 20%$</td>
<td>75%</td>
<td>70%</td>
<td>30%</td>
</tr>
<tr>
<td>Cutoff $\geq 22.5%^*$</td>
<td>74%</td>
<td>75%</td>
<td>29%</td>
</tr>
</tbody>
</table>

* = Adjusted cutoffs.

Manual scores retrieved from patient records.

**Figure 2**

Clustered box plot for Ki67 index (%) by each scoring method in PAM50 Luminal A and B subtypes. Error bars represent 95% confidence interval. Circles represent outliers and asterisks represent extremes. DIA, digital image analysis ($n = 214$).
The difference of Ki67-scores in Luminal A and Luminal B PAM50 subtypes was significant \((P < 0.002)\) by independent-samples Mann–Whitney \(U\)-tests in all evaluated methods, manual and DIA (Figure 2).

In 67 out of the 279 cases in cohort 1 and 2 combined, (24%), the ‘hot spot’ area of highest Ki67 intensity was within 1 mm of the tumor’s invasive margin.

**Subclassification**

To determine which of manual or DIA-generated biomarker scoring outcomes that best corresponded to PAM50 gene expression profiles, manual and DIA scores of Ki67, ER, PR, and HER2 in cohort 1 \((n = 195)\) were combined into surrogate immunohistochemical subtypes according to the specifications in Table 2. For all Ki67 scoring methods, both manual and DIA, we used full-section slides and the cutoffs adjusted to receiver operating characteristics for ‘high’ vs ‘low’ described above. With DIA, we scored ER, PR, and HER2 on tissue microarrays only, whereas the patient records contain data of manual scoring on full sections for all biomarkers.

Still, all tested DIA methods exceeded manual immunohistochemical subtype concordance and Cohen’s \(\kappa\) agreement with PAM50 gene expression assays with 2.2 to 5.5 percentage points (Figure 3).

If Luminal cases were to be grouped together without dichotomization into A and B subtypes, thereby omitting Ki67 as a factor in surrogate immunohistochemical subtype (see details in Supplementary Data), concordance increases further to up to 95.3% \((\kappa = 0.533)\) for DIA and to 87.4% \((\kappa = 0.498)\) for manual scoring. This gain in concordance is however naturally at the expense of the prognostic value of information on ‘high’ vs ‘low’ proliferative activity. It is also points to the fact that accuracy in assessments of ER, PR, and HER2 is generally excellent, with DIA leading to slightly higher concordance to gene expression assays than manual biomarker scoring (details in Supplementary Data).

**Prognostication**

As the first cohort analyzed here \((n = 195)\) still lacks long-term survival data, clinically reported Nottingham combined histologic grade, number of axillary lymph node metastases \((N)\), and largest primary tumor diameters \((Ø)\) were used as prognostic surrogates. Spearman’s rank-order correlation was run to determine the relationship between Nottingham combined histologic grade and Ki67 index measured by DIA of the tumors’ invasive margins, ‘hot spots’ and full tumor cross-section averages, as well as by the manual method used in current clinicopathological routine. This showed a positive and statistically significant correlation for all methods, with the strongest correlation for DIA.
Table 4 Mean overall survival and 95% confidence interval for Ki67 ‘high’ and ‘low’ classified by manual and each digital image analysis method in PAM50 Luminal A and B subtypes only (top), all PAM50 subtypes (middle), as well as for PAM50 Luminal A and B subtypes (bottom, italic).

<table>
<thead>
<tr>
<th>Ki67 scoring method</th>
<th>Population (n)</th>
<th>Mean overall survival Ki67 ‘high’, years (95% CI)</th>
<th>Mean overall survival Ki67 ‘low’, years (95% CI)</th>
<th>Hazard ratio for all-cause mortality Ki67 ‘high’ vs ‘low’ (P)</th>
<th>LR $\chi^2$ (P)</th>
<th>LR $-\Delta\chi^2$ vs manual method (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital image analysis, invasive margin</td>
<td>PAM50 Luminal A and B subtypes only (41)</td>
<td>13.2 (10.6–15.8)</td>
<td>15.1 (13.2–17.1)</td>
<td>2.44 (0.038)</td>
<td>4.488 (0.034)</td>
<td>1.99 (0.158)</td>
</tr>
<tr>
<td>Digital image analysis, hot spot</td>
<td></td>
<td>13.0 (10.3–15.8)</td>
<td>15.2 (13.4–17.1)</td>
<td>3.00 (0.011)</td>
<td>7.215 (0.007)</td>
<td>4.043 (0.044)</td>
</tr>
<tr>
<td>Digital image analysis, average</td>
<td></td>
<td>13.0 (10.3–15.8)</td>
<td>15.3 (13.5–17.1)</td>
<td>2.31 (0.047)</td>
<td>4.129 (0.042)</td>
<td>1.868 (0.172)</td>
</tr>
<tr>
<td>Manual</td>
<td></td>
<td>12.8 (10.3–15.3)</td>
<td>16.2 (14.7–17.7)</td>
<td>3.20 (0.012)</td>
<td>6.35 (0.012)</td>
<td>–</td>
</tr>
<tr>
<td>Digital image analysis, invasive margin</td>
<td>All PAM50 subtypes (84)</td>
<td>11.1 (9.3–12.9)</td>
<td>12.8 (10.9–14.7)</td>
<td>1.56 (0.163)</td>
<td>1.953 (0.162)</td>
<td>2.018 (0.155)</td>
</tr>
<tr>
<td>Digital image analysis, hot spot</td>
<td></td>
<td>10.8 (9.0–12.5)</td>
<td>13.4 (11.7–15.2)</td>
<td>1.79 (0.073)</td>
<td>3.759 (0.053)</td>
<td>0.348 (0.555)</td>
</tr>
<tr>
<td>Digital image analysis, average</td>
<td></td>
<td>10.7 (8.8–12.5)</td>
<td>13.3 (11.6–14.9)</td>
<td>1.56 (0.150)</td>
<td>1.188 (0.276)</td>
<td>0.005 (0.942)</td>
</tr>
<tr>
<td>Manual</td>
<td></td>
<td>10.7 (9.0–12.4)</td>
<td>14.0 (12.3–15.8)</td>
<td>2.12 (0.029)</td>
<td>1.679 (0.195)</td>
<td>–</td>
</tr>
<tr>
<td>Gene expression assay</td>
<td></td>
<td>13.2 (10.9–15.5)</td>
<td>10.6 (8.1–13.1)</td>
<td>2.23 (0.065)</td>
<td>3.57 (0.059)</td>
<td>0.062 (0.803)</td>
</tr>
</tbody>
</table>

LR $\chi^2$ = likelihood ratio $\chi^2$; LR $-\Delta\chi^2$ = likelihood ratio $\chi^2$ change.
Prognostic value as LR $\chi^2$ and relative prognostic value of digital image analysis/PAM50 to manual Ki67 scores as LR $-\Delta\chi^2$. Cutoffs adjusted to receiver operating characteristics.

*Not significant on the 0.05 level.

Discussion

In this study, all tested DIA methods of scoring Ki67 outperformed the manual scores, even our most accurate pathologist’s manual scores, in terms of sensitivity and specificity for the Luminal B subtype. When comparing DIA vs manual scores, a higher correlation of concordance and Cohen’s k agreement were found between DIA and manual scores for the Luminal B subtype, with the highest correlation of concordance and Cohen’s k agreement between Ki67 ‘high’ and ‘low’ groups with all the manual DIA methods. With the strongest correlation of concordance and Cohen’s k agreement between Ki67 ‘high’ and ‘low’ groups, digital image analysis/PAM50 to manual Ki67 scores as LR $-\Delta\chi^2$. Cutoffs adjusted to receiver operating characteristics.

Finally, each DIA method was added separately to the model as a prognostic surrogue, Spearman’s rank-order correlations showed a positive and significant correlation for both manual and DIA methods, with the strongest correlation of concordance and Cohen’s k agreement between Ki67 ‘high’ and ‘low’ groups, digital image analysis/PAM50 to manual Ki67 scores as LR $-\Delta\chi^2$. Cutoffs adjusted to receiver operating characteristics.
Figure 4 Kaplan–Meier curves for overall survival of cases classified into Ki67 'low' (dark) and 'high' (light) with digital image analysis (DIA) and manual methods using cutoffs adjusted to receiver operating characteristics. Left: PAM50 Luminal A and B subtypes only (n=41). Right: All PAM50 subtypes (n=84). 95% Confidence interval in Table 4.
method giving an automatic representation of the average Ki67 positivity across the full tumor cross-section.

When the prognostic value of a Ki67 index determined by each of the manual and DIA scoring methods was tested, all contributed with significant information on overall survival in the PAM50 Luminal A and B subtype tumors, with the highest LR $\chi^2$ for DIA of Ki67 in ‘hot spots’. Furthermore, this method added significantly more prognostic information than the manual scoring method in the same subgroup. This was however not the case when we included all PAM50 subtypes, confirming that the prognostic role for Ki67 is mainly related to the Luminal A and B subtypes.

DIA of Ki67 positivity did yield different scores depending on what tumor area and number of cells was in focus of the analysis. This however did not induce any major differences in performance of subclassification or prognostication, possibly except by DIA of Ki67 in ‘hot spots’ that had a slightly better prognostic value. It should nevertheless be emphasized that in a quarter of the tumors in this study, the ‘hot spot’ was within 1 mm of the tumor’s invasive margin, a fact that should be taken into consideration in the event of future studies of what tumor regions have the highest metastatic potential.

In the tissue microarray cohort (reported in Supplementary Data), DIA matched the pathologist’s manual assessments of all biomarkers for a quite low concordance to gene expression assays and poor sensitivities and specificities for the Luminal B subtype.

As a consequence of these results, we cannot recommend therapeutic decisions or prognostic information based on Ki67 scored on tissue microarrays when full sections are available. In analogy with consensus recommendations, we also found good reasons to support the notion that the distinction of Ki67 into ‘high’ and ‘low’ groups should be done only after cutoffs are adjusted to each laboratory’s own reference data and the scoring method used.

One could argue that DIA is a complicating development in biomarker scoring that may not be sufficiently user friendly for pathologists with many years of experience with manual biomarker scorings. Substantial investments in digital scanning capacity, data storage, software, and training are required at each institution before effective use of the technology can be expected. With a perhaps tempting but excessive automation, DIA could also withdraw direct control over the assessment in terms of what tumor areas and which cells are being scored, potentially leading to dire consequences to patients.

Furthermore, DIA may in itself be a source of variance. Different DIA approaches will inherently classify tumor, nuclei, and membranes differently, and poor performance of the algorithm’s identification of tumor vs nontumor tissue as well as cellular components would be a significant source of error.

To minimize the variance contributed by the DIA software used here, the manufacturer has chosen a single well-tested set of algorithms adherent to Conformité Européenne In Vitro Diagnostics. These have previously been validated on data from multiple sites comprising thousands of tumor samples to ensure that the variance that DIA contributes is kept at a minimum\(^{40}\)(see specific statistics on reproducibility with Visiopharm integrator system for each tested biomarker in Supplementary Data).

When interpreting the results of any method’s concordance to gene expression assays, one should also note that the individual tumor’s PAM50 subtype is based on the average gene expression profile in the very piece of tumor tissue from which RNA was extracted. Thus, presence of substantial intratumor heterogeneity could potentially lead to uncertainty in subtype assignment and consequently affect the immunohistochemical vs PAM50 subtype concordance. In an ongoing study we seek to shed clarity to this subject (unpublished). So far, our preliminary data indicate that intratumor heterogeneity in terms of PAM50 subtype is quite limited and not a common occurrence. Moreover, manual vs DIA immunohistochemical subtype concordance to PAM50 assays would be influenced to an equal degree by the presence of intratumor heterogeneity. We consequently believe that it is not likely to affect the results and conclusions of this study in any significant way.

When summarizing this study, manual assessment of the biomarkers ER, PR, HER2, and Ki67, with an emphasis on the latter, was in most aspects an inferior alternative to DIA. This implicates that with the manual methods of scoring these biomarkers currently used, an avoidable high proportion of patients could receive either potentially harmful treatments such as cytotoxic chemotherapy without benefit or be excluded from the beneficial treatments that a better diagnostic method would indicate.

This is perhaps especially relevant as DIA in many ways is already an accessible, simple option with superior reproducibility. A growing number of ready-to-use systems are offered on the market including the one tested here. Combined with the increasingly efficient and less expensive digital glass slide scanners, digital pathology is set to challenge manual biomarker scoring for the method of choice for the time being until gene expression assays or their equivalent are universally available. In addition to its competitive performance, DIA also provides an opportunity to reduce time consumption for pathologists and allocate precious resources to more qualified tasks. In the fully automatic scoring methods described here, manual input and thereby the sampling bias is reduced to a minimum. An
operator of the Visiopharm integrator system even has the option to define regions of interest on pancytokeratin slides only, thereby avoiding subjective assessments of biomarker positivity in different tumor areas altogether. This implicates that an approach like DIA of the full tumor cross-section average or ‘hot spots’ could allow for biomedical scientists or other laboratory personnel with only a basic understanding of histopathology and immunohistochemistry to manage surrogate immunohistochemical subclassification in breast cancer.

Accordingly, we conclude that DIA is already a viable and competitive, if not superior, alternative for biomarker testing in breast cancer. We strongly encourage further studies to confirm the results found here in larger populations to facilitate implementation and to evaluate the performance of DIA in clinical use. It is with great anticipation that we look forward to the continued technological progress in this matter.

Acknowledgments

We thank the individual members of each of the cancer registries, laboratories, and institutions who have contributed greatly to the manuscript. Specifically, we point out the valuable comments by Dr Göran Elmberger in the planning phase of this project. This work was supported with grants from the Swedish Society of Medicine, the Swedish Society for Medical Research, the Wallström fund, Annéristiftelsen, and Magnus Bergvalls Stiftelse. Johan Hartman was supported by Karolinska Institutets Theme Center in Breast Cancer (BRECT) and Jonas Bergh by the Swedish Cancer Society and the Swedish Research Council. Visiopharm A/S did not offer any financial support or incentives and was not engaged in the scientific content of this paper other than their contribution with technical support and free use of their software.

Author contributions

GS, LP, MG, and JH participated in the study design. GS, NFM, ML, IM, GR, LK, and MR participated in data collection and analysis. GS, NPT, ML, JB, and JH participated in interpretation of the results and writing of the paper. All contributors were involved in approving the final version of the manuscript before submission.

Disclosure/conflict of interest

The authors declare no conflict of interest.

References

Supplementary Information accompanies the paper on Modern Pathology website (http://www.nature.com/modpathol)