DTU Library

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

[Link back to DTU Orbit](#)

318

Digital image analysis outperforms manual biomarker assessment in breast cancer

Gustav Stålhammar[1,2], Nelson Fuentes Martinez[1,3], Michael Lippert[4], Nicholas P Tobin[5], Ida Mølholm[4,6], Lorand Kis[7], Gustaf Rosin[1], Mattias Rantalainen[8], Lars Pedersen[4], Jonas Bergh[1,5,9], Michael Grunkin[4] and Johan Hartman[1,5,7]

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

Cohort no.	1	2
No. of patients	195	84
Breast cancer diagnosis, period	2006–2010	1994–1996
Mean age at breast cancer diagnosis (years)	58	58
Histologic grade		
G1	31	19
G2	90	35
G3	74	30
Mean tumor diameter (mm)	25	22
Proportion of patients with axillary lymph node metastasis	29%	38%
PAM50 Luminal A, *n* (% of cohort)	123 (63)	24 (29)
PAM50 Luminal B, *n* (% of cohort)	50 (26)	17 (20)
PAM50 HER2-enriched, *n* (% of cohort)	10 (5)	10 (12)
PAM50 basal-like, *n* (% of cohort)	12 (6)	16 (19)
PAM50 normal-like, *n* (% of cohort)	0 (0)	17 (20)
PAM50 subtype not available, *n* (% of cohort)	27 (12)	0 (0)
10-Year overall survival	N/A	62%
Type of material contributed to this study	Full sections (Ki67) and TMA (ER, PR, and HER2)	Full sections (Ki67)
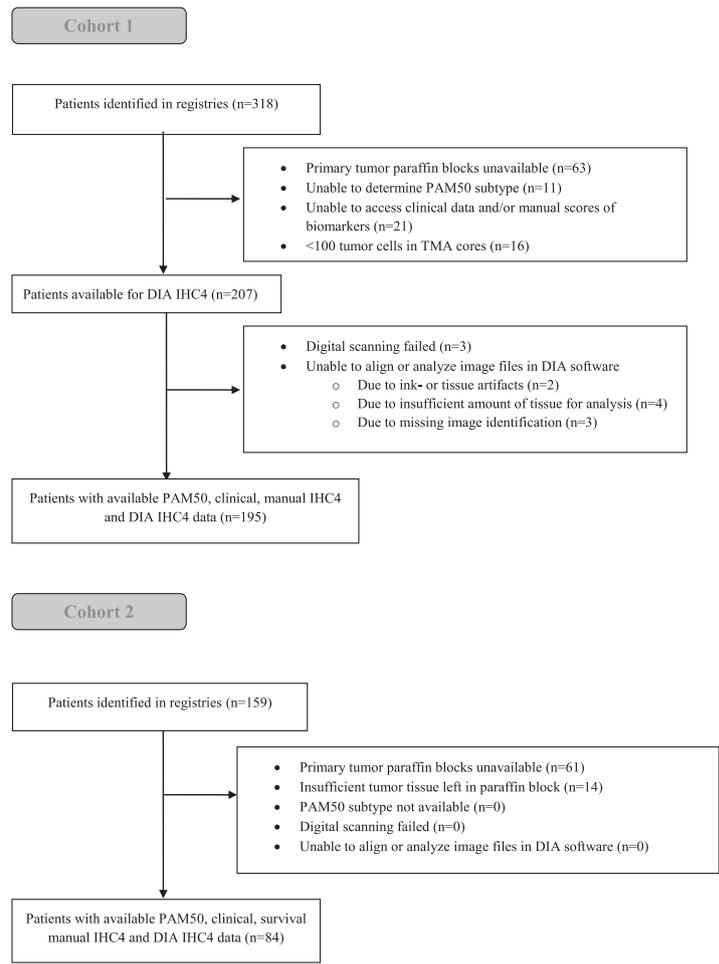
conclusions about the tumors' biological characteristics and expected response to therapy.[3–5,9,10] Congruence of evaluations of these surrogate markers to the gene expression tests are consequently of utmost importance, not least as discrepancies in classification induces dissimilar treatment decisions 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.[11,12] This is perhaps especially evident for Ki67[13–15] as there is no consensus on what tumor region or number of cells to score[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.[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,[5] and at our and several other institutions a cutoff of $\geq 20\%$ for highly proliferative tumors is commonly used.[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.[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 clinico-pathological 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 *in situ*, intense inflammation, fibrosis, necrosis, or poor fixation. Then, 4–8 tissue cores (Ø 0.8 mm) per patient were punched and mounted into a tissue microarray using a semiautomated instrument (Minicore 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,[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.[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 µm sections, conditioned in CC1 solution (Ventana Medical Systems, Tucson, AZ, USA) for 36 min (Ki67) to
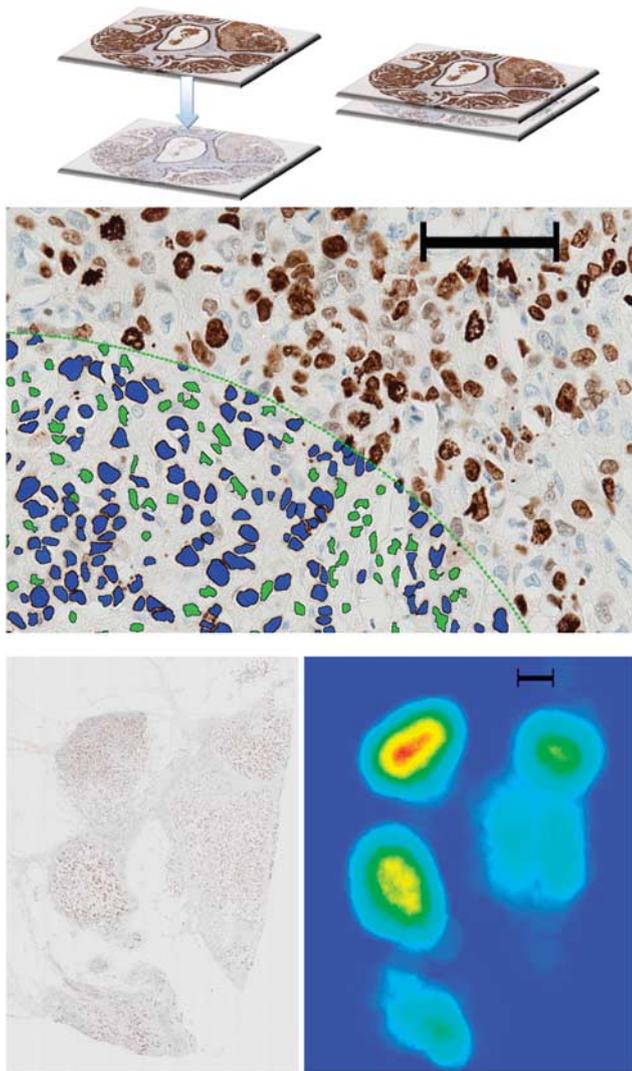
Figure 1 Top: Illustration of the alignment of two adjacent slides stained with a pancytokeratin marker such as CkMNF116 and a biomarker (ER, PR, or Ki67), respectively. Middle: Green dotted line marks part of a region of interest, scored for Ki67 index. Blue polygons mark nuclei positive for both Ki67 and CkMNF116. Green polygons mark nuclei positive for CkMNF116 but negative for Ki67. The proportion of blue polygons to the sum of blue and green polygons constitutes the Ki67 index. Bottom: Illustration of heat map function where the Visiopharm integrator system software has analyzed the digitally scanned glass slide (left) for tumor area with highest concentration of cells stained by both the pancytokeratin marker and Ki67, marked in red (right). Scale bar, middle = 50 μm. Scale bar, lower = 500 μm.

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 tissue[26–28] 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 method[32] in the edgeR package.[33] Unaligned RNAseq data from the 'Cancer Genome Atlas' breast cancer data set[34] 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 classifier[36] 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

Table 2 Molecular 'intrinsic' breast cancer subtypes and surrogate definitions by immunohistochemical profile

Intrinsic subtype	Surrogate IHC classification
Luminal A	ER ≥ 1% and/or PR ≥ 20% and HER2 'negative' and Ki67 'low'
Luminal B	1. ER ≥ 1% and/or PR ≥ 20% and HER2 'negative' and Ki67 'high' or 2. ER ≥ 1% and PR < 20% and HER2 'negative.' Any Ki67 or 3. ER ≥ 1% and/or PR ≥ 1% and HER2 'positive.' Any Ki67
HER2-enriched	ER < 1% and PR < 1%. HER2 'positive.' Any Ki67
Basal-like	ER < 1% and PR < 1%. HER2 'negative.' Any Ki67

% = 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.[30] 'High' and 'low' = as defined by each laboratory's own reference data,[3,6,17] with threshold generally in the range of 14–29%.[4,5,19–21]

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[40] (Figure 1). Excellent reproducibility with this and similar systems has been shown previously[41–45] (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.

Surrogate Subclassification

The assessments of ER, PR, HER2, and Ki67 by both manual and DIA methods were combined and compared for classification into surrogate immunohistochemical subtypes for each tumor using definitions recommended by international international expert consensus[3–5,6,9,22,30,46] (Table 2).

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 κ statistics were 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 χ^2 (LR χ^2) and change in LR χ^2 (LR– $\Delta\chi^2$) 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[47]), 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

Table 3 Sensitivity, specificity, and misclassification percentage for each method of Ki67 scoring using the ≥ 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)

Ki67 scoring method	Sensitivity for PAM50 Luminal B vs A	Specificity for PAM50 Luminal B vs A	Proportion misclassified
DIA invasive margin			
Cutoff ≥ 20%	84%	78%	20%
Cutoff ≥ 20.2%*	82%	79%	20%
DIA hot spot			
Cutoff ≥ 20%	90%	65%	24%
Cutoff ≥ 25.2%*	86%	77%	19%
DIA average			
Cutoff ≥ 20%	60%	90%	31%
Cutoff ≥ 15.5%*	80%	83%	19%
Manual			
Cutoff ≥ 20%	75%	70%	30%
Cutoff ≥ 22.5%*	74%	75%	29%

Manual scores retrieved from patient records.
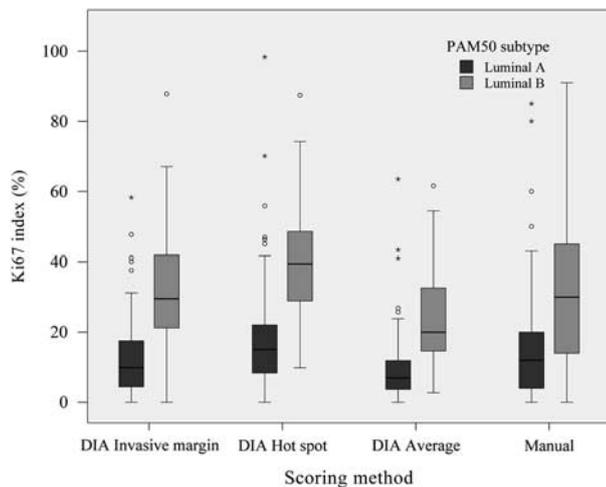* = Adjusted cutoffs.

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 ($n = 214$).

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 ~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 ~3 min, scoring 50% ($R^2 = 0.998$) took ~12 min, and scoring 100% ($R^2 = 1$) took ~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 ~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 ≥ 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).
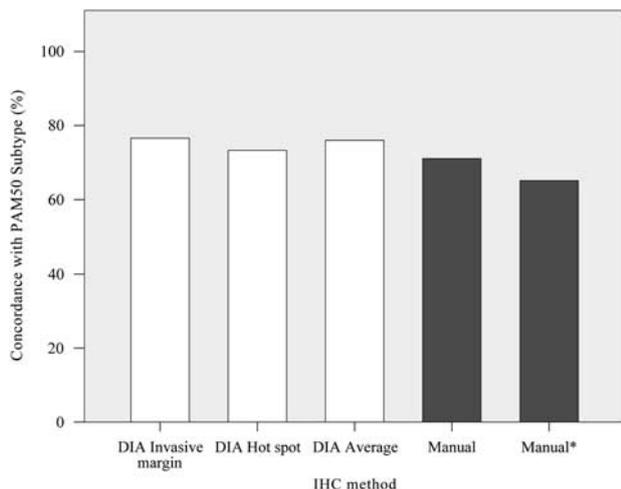
Figure 3 Comparison of manual *vs* digital image analysis (DIA) surrogate immunohistochemical subtype concordance to PAM50 gene expression assays. Concordance specified as proportion (%) of cases classified into identical subtypes (Luminal A, Luminal B, HER2, or Basal) with manual or DIA immunohistochemical methods and PAM50. Data on ER, PR, HER2, and Ki67 scores from patient records were combined for manual immunohistochemical subtype (according to Table 2). DIA Ki67 scores on full sections were combined with DIA ER, PR, and HER2 scores of the same tumors on tissue microarray. Cutoffs for Ki67 'high' after analysis of receiver operating characteristics. Gray bars indicate results of manual scores, and white bars indicate DIA methods ($n = 279$). DIA ER, PR, and HER2 on tissue microarrays+DIA Ki67 *invasive margin* immunohistochemical subtype *vs* PAM50 subtype concordance: 76.6%. Cohen's κ: 0.510. DIA ER, PR, and HER2 on tissue microarrays+DIA Ki67 *hot spot* immunohisto-chemical subtype *vs* PAM50 subtype concordance: 73.3%. Cohen's κ: 0.469. DIA ER, PR, and HER2 on tissue microarrays+DIA Ki67 *Average* immunohistochemical subtype *vs* PAM50 subtype concordance: 76.0%. Cohen's κ: 0.502. *Manual* immuno-histochemical subtype *vs* PAM50 subtype concordance: 71.1%. Cohen's κ: 0.453. *Manual** immunohistochemical subtype with a classical cutoff of $\geq 20\%$ for Ki67 'high' *vs* PAM50 subtype concordance: 65.1%. Cohen's κ: 0.392.

Thresholds for Ki67 'High' *vs* 'Low'

With a $\geq 20\%$ cutoff for Ki67 'high' in all PAM50 Luminal A and B tumors from both cohorts ($n = 214$), DIA produced distinctions that matched or were more accurate than the manual method, depending on what tumor region was scored. Individual analyses of receiver operating characteristics for each scoring method where maximum sensitivity and specificity for the PAM50 Luminal B subtype were given equal importance (see Supplementary Figures 4a and b) yielded cutoffs ranging from ≥ 15.5 to 25.2% in this subset. When applying these adjusted cutoffs, all DIA methods outperformed manual scores in terms of sensitivity and specificity for the Luminal B subtype. It is noteworthy that the method aiming for a representation of the average Ki67 score across the full tumor section had the lowest cutoff adjusted to receiver operating characteristics of $\geq 15.5\%$, reflecting a sampling without focus on highly proliferative areas (Table 3).

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 κ 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' proliferational 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)

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

LR χ^2 = likelihood ratio χ^2; LR $-\Delta\chi^2$ = likelihood ratio χ^2 change.
Prognostic value as LR χ^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.
[a]Not significant on the 0.05 level.

of full tumor cross-section averages ($r_s = 0.575$, $P < 0.001$) and the weakest for the manual scores ($r_s = 0.459$, $P < 0.001$). Ki67 index was however not significantly correlated to neither N nor Ø for any method, manual or DIA (see details in Supplementary Tables 9 and 10).

For the second cohort, we compared the differences in mean overall survival and Cox regression hazard ratios for all-cause mortality for patients with tumors classified into Ki67 'high' and 'low' with each Ki67 scoring method. Mean survival years were significantly higher and hazard ratios significantly lower for patients classified into the Ki67 'low' vs 'high' groups by all scoring methods in the subgroup with PAM50 Luminal A and B tumors. When including all the cohort patients regardless of PAM50 subtype, differences in mean survival between Ki67 'low' vs 'high' was generally lower for all scoring methods and hazard ratios generally not significant (Table 4 and Figure 4).

When each of the Ki67 scoring methods were tested for its individual prognostic value by Cox regression LR χ^2 in the subgroup with PAM50 Luminal A and B tumors only ($n = 214$), all DIA methods as well as the manual method contributed with significant information on overall survival with the highest LR χ^2 for DIA of Ki67 in 'hot spots.' However, when this analysis was repeated for all the patients in the cohort regardless of PAM50 subtype, none contributed with significant information on overall survival.

Finally, each DIA method was added separately to manual Ki67 scoring to determine whether they added any prognostic value. LR– $\Delta\chi^2$ was used to measure and compare the relative amount of information. Here, DIA of Ki67 in 'hot spots' added significantly more prognostic information in the subgroup with PAM50 Luminal A and B tumors only (LR $-\Delta\chi^2$ 4.043, $P = 0.044$), whereas LR $-\Delta\chi^2$ for the other DIA methods were not significantly better (Table 4).

Discussion

In this study, all tested DIA methods of scoring Ki67 outperformed even our most accurate pathologist's manual scores in terms of sensitivity and specificity for the Luminal B subtype. When comparing DIA vs manual immunohistochemical surrogate concordance and Cohen's κ agreement with PAM50 gene expression assays, all tested DIA methods were superior to the manual method.

Furthermore, the manual and DIA methods essentially matched each other for prognostication of hazard ratio for all-cause mortality in tumors with a 'high' vs 'low' Ki67 index. When histological grade was used as a prognostic surrogate, Spearman's rank-order correlations showed a positive and significant correlation for both manual and DIA methods, with the strongest correlation for the DIA
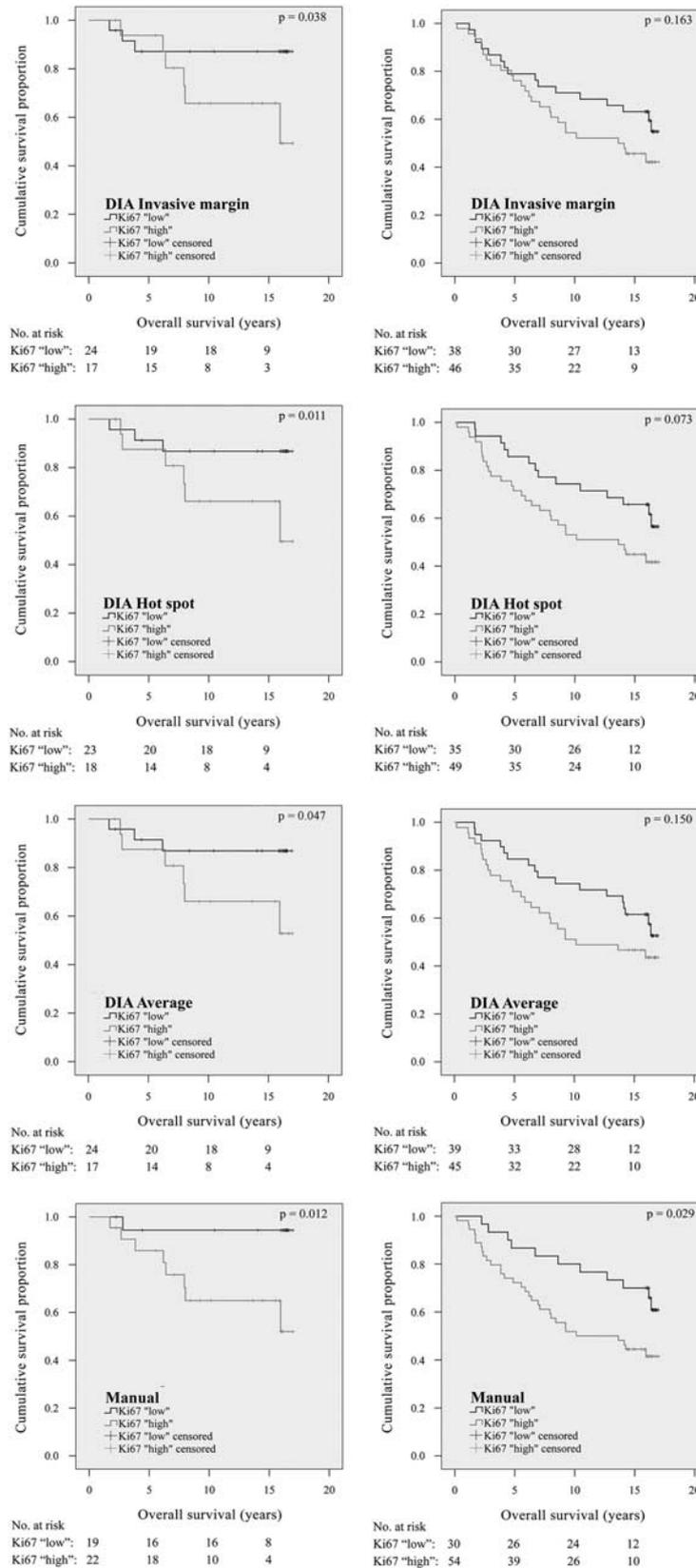
Figure 4 Kaplan–Meier curves for overall survival of cases classified into Ki67 'low' (dark) and 'high' (light) with digital image analysis (DIA) and manual methods using cutoffs adjusted to receiver operating characteristics. Left: PAM50 Luminal A and B subtypes only (*n* = 41). Right: All PAM50 subtypes (*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 χ^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 consequentially 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

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

1 Schnitt SJ, Lakhani SR, Anderson BO *et al.* Human cancers by organ site, breast cancer. In: Stewart BW, Wild C (eds). Chapter 5.2. World Cancer Report 2014, pp 508–526.

2 Ellis IO, Schnitt SJ, Sastre-Garau X *et al.* Tumours of the breast. In: Kurman RJ, Carcangiu ML, Herrington CS *et al.* (eds). Chapter 1. WHO Classification of Tumours of Female Reproductive Organs. International Agency for Research on Cancer. IARC press: Lyon, 2014, pp 9–110.

3 Goldhirsch A, Wood WC, Coates AS *et al.* Strategies for subtypes-dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol 2015;22:1736–1747.

4 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 2015;24:2206–2223.

5 Coates AS, Winer EP, Goldhirsch A *et al.* Tailoring therapies - improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. Ann Oncol 2015;26:1533–1546.

6 Guiu S, Michiels S, Andre F *et al.* Molecular subclasses of breast cancer: how do we define them? The IMPAKT 2012 Working Group Statement. Ann Oncol 2012 23: 2997–3006.

7 Parker JS, Mullins M, Cheang MC *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160–1167.

8 Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. J Natl Cancer Inst 2015;107:357.

9 Prat A, Cheang MC, Martin M *et al.* Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. J Clin Oncol 2013;31:203–209.

10 Sgroi DC, Sestak I, Cuzick J *et al.* Prediction of late distant recurrence in patients with oestrogen-receptor-positive breast cancer: a prospective comparison of the breast-cancer index (BCI) assay, 21-gene recurrence score, and IHC4 in the TransATAC study population. Lancet Oncol 2013;14:1067–1076.

11 Bahreini F, Soltanian AR, Mehdipour P. A meta-analysis on concordance between immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) to detect HER2 gene overexpression in breast cancer. Breast Cancer 2014;22:615–625.

12 Cheang MC, Martin M, Nielsen TO *et al.* Defining breast cancer intrinsic subtypes by quantitative receptor expression. Oncologist 2015;20:474–482.

13 Polley MY, Leung SC, Gao D *et al.* An international study to increase concordance in Ki67 scoring. Mod Pathol 2015;28:778–786.

14 Stalhammar G, Rosin G, Fredriksson I *et al.* Low concordance of biomarkers in histopathological and cytological material from breast cancer. Histopathology 2014;64:971–980.

15 Ekholm M, Beglerbegovic S, Grabau D *et al.* Immunohistochemical assessment of Ki67 with antibodies SP6 and MIB1 in primary breast cancer: a comparison of prognostic value and reproducibility. Histopathology 2014;65:252–260.

16 Christgen M, von Ahsen S, Christgen H *et al.* The region-of-interest size impacts on Ki67 quantification by computer-assisted image analysis in breast cancer. Hum Pathol 2015;46:1341–1349.

Digital image analysis in breast cancer
G Stålhammar *et al*
329

17 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.

18 Grabau D, Kovács A, Tolockiene E *et al.* Quality- and standardization document for breast tumors. Swedish Quality- and Standardization Committee (KVAST) tSPS 2014;1:1–49.

19 Romero Q, Bendahl PO, Fernö M *et al.* A novel model for Ki67 assessment in breast cancer. Diagn Pathol 2014;9:118.

20 Klintman M, Bendahl PO, Grabau D *et al.* The prognostic value of Ki67 is dependent on estrogen receptor status and histological grade in premenopausal patients with node-negative breast cancer. Mod Pathol 2010;23:251–259.

21 Reyal F, Hajage D, Savignoni A *et al.* Long-term prognostic performance of Ki67 rate in early stage, pT1-pT2, pN0, invasive breast carcinoma. PLoS One 2013;8:e55901.

22 Hammond ME, Hayes DF, Dowsett M *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Clin Oncol 2010;28: 2784–2795.

23 Knutsvik G, Stefansson IM, Aziz S *et al.* Evaluation of Ki67 expression across distinct categories of breast cancer specimens: a population-based study of matched surgical specimens, core needle biopsies and tissue microarrays. PLoS One 2014;9:e112121.

24 Calza S, Hall P, Auer G *et al.* Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. Breast Cancer Res 2006;8:R34.

25 Pawitan Y, Björhle J, Amler L *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 2005;7: R953–R964.

26 Al Sarakbi W, Salhab M, Thomas V *et al.* Is preoperative core biopsy accurate in determining the hormone receptor status in women with invasive breast cancer? Int Semin Surg Oncol 2005;2:15.

27 Nadji M, Gomez-Fernandez C, Ganjei-Azar P *et al.* Immunohistochemistry of estrogen and progesterone receptors reconsidered: experience with 5,993 breast cancers. Am J Clin Pathol 2005;123:21–27.

28 Cottu PH, Asselah J, Lae M *et al.* Intratumoral heterogeneity of HER2/neu expression and its consequences for the management of advanced breast cancer. Ann Oncol 2008;19:595–597.

29 Yi M, Huo L, Koenig KB *et al.* Which threshold for ER positivity? A retrospective study based on 9639 patients. Ann Oncol 2014;25:1004–1011.

30 Wolff AC., Hammond ME, Hicks DG *et al.* Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol 2013;31: 3997–4013.

31 Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166–169.

32 Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11:R25.

33 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–140.

34 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.

35 Eroles P, Bosch A, Pérez-Fidalgo JA *et al.* Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. Cancer Treat Rev 2012;38:698–707.

36 Tibshirani R, Hastie T, Narasimhan B *et al.* Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002;99: 6567–6572.

37 Cleary AS, Leonard TL, Gestl SA *et al.* Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. Nature 2014;508:113–117.

38 Nguyen LV, Pellacani D, Lefort S *et al.* Barcoding reveals complex clonal dynamics of de novo transformed human mammary cells. Nature 2015;528: 267–271.

39 Gerlinger M, Rowan AJ, Horswell S *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012;366: 883–892.

40 Kårsnäs A, Strand R, Doré J *et al.* A histopathological tool for quantification of biomarkers with sub-cellular resolution. Comput Methods Biomech Biomed Eng Imaging Vis 2015;3:25–46.

41 Holten-Rossing H, Møller Talman ML, Kristensson M *et al.* Optimizing HER2 assessment in breast cancer: application of automated image analysis. Breast Cancer Res Treat 2015;152:367–375.

42 Laurinaviciene A, Dasevicius D, Ostapenko V *et al.* Membrane connectivity estimated by digital image analysis of HER2 immunohistochemistry is concordant with visual scoring and fluorescence in situ hybridization results: algorithm evaluation on breast cancer tissue microarrays. Diagn Pathol 2011;6:87.

43 Gavrielides MA, Gallas BD, Lenz P *et al.* Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. Arch Pathol Lab Med 2011;135:233–242.

44 Rizzardi AE, Johnson AT, Vogel RI *et al.* Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. Diagn Pathol 2012;7:42.

45 Nielsen PS, Riber-Hansen R, Jensen TO *et al.* Proliferation indices of phosphohistone H3 and Ki67: strong prognostic markers in a consecutive cohort with stage I/II melanoma. Mod Pathol 2013;26:404–413.

46 Fitzgibbons PL, Dillon DA, Alsabeh R *et al.* Template for reporting results of biomarker testing of specimens from patients with carcinoma of the breast. Arch Pathol Lab Med 2014;138:595–601.

47 Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 1991;19:403–410.