Ductal carcinoma in situ and invasive breast cancer: diagnostic accuracy and prognosis

Seijen, M. van

Chapter 3

Variability in grading of ductal carcinoma in situ among an international group of pathologists

Maartje van Seijen
Katarzyna Jóźwiak
Sarah E. Pinder
Allison Hall
Savitri Krishnamurthy
Jeremy S.J. Thomas
Laura C. Collins
Jonathan Bijron
Joost Bart
Danielle Cohen
Wen Ng
Ihssane Bouybayoune
Hilary Stobart
Jan Hudecek
Michael Schaapveld
Alastair M. Thompson
Esther H. Lips*
Jelle Wesseling*

on behalf of the PRECISION team

*Both senior authors contributed equally

J Pathol Clin Res. 2021 Feb; Online ahead of print.

ABSTRACT

The prognostic value of cytonuclear grade in ductal carcinoma in situ (DCIS) is debated, partly due to high interobserver variability and by the use of multiple guidelines. The aim of this study was to evaluate the interobserver agreement in grading DCIS between Dutch, British and American pathologists. Hematoxylin and eosin-stained slides of 425 women with primary DCIS were independently reviewed by nine breast pathologists, based in the Netherlands, the UK and the USA. Chance corrected kappa (κ_{ma}) for association between pathologists was calculated based on a generalized linear mixed model using the ordinal package in R. Overall κ_{ma} for grade of DCIS (low, intermediate or high) was estimated as 0.50 (95% confidence interval (CI) 0.44-0.56), indicating a moderate association between pathologists. When the model was adjusted for national guidelines, the association for grade did not change (κ_{ma} = 0.53; 95% CI 0.48-0.57); subgroup analysis for pathologists using the UK pathology guidelines only had significantly higher association (κ_{ma} = 0.58; 95% CI 0.56-0.61). To assess if concordance of grading relates to expression of the estrogen receptor (ER) and HER2, archived immunohistochemistry (IHC) was analysed on a subgroup (n=106). This showed that non-high grade according to the majority opinion was associated with ER-positivity and HER2-negativity (100% and 89% of non-high grade cases, respectively). In conclusion, DCIS grade showed only moderate association using whole slide images scored by nine breast pathologists. Since therapeutic decisions and inclusion in ongoing clinical trials are guided by DCIS grade, there is a pressing need to reduce interobserver variability in grading. ER and HER2 might be supportive to prevent accidental and unwanted inclusion of high grade DCIS in such trials.

Introduction

Ductal carcinoma in situ (DCIS) is a non-obligate precursor of invasive breast cancer (IBC) in which the proliferating epithelial cells remain within the boundaries of the ducto-lobular system of the breast. DCIS is graded by pathologists using a three-tier system: well-differentiated (low nuclear grade, grade 1), intermediately differentiated (intermediate nuclear grade, grade 2), and poorly differentiated (high nuclear grade, grade 3). This histological assessment of grade is prognostic, in terms of subsequent ipsilateral in situ and invasive lesion risk, and is used to guide treatment decisions and to determine eligibility for inclusion in clinical trials. Although different guidelines are used to grade DCIS there seems to be a substantial difference in interpretation (interobserver variability) in grading, even using the same guidelines.[1] Consequently, the prognostic and clinical value of DCIS grade is still a subject of debate.[2–4] There are, however, no other histological features or widely tested biomarkers presently available that can be used to predict reliably the progression of DCIS lesions to IBC[5]. Because of this uncertainty, almost all women with DCIS receive similar treatment to that given for invasive breast cancer: i.e. mastectomy or breast conserving surgery often supplemented by radiotherapy and/or endocrine therapy.

To investigate how to distinguish indolent from potentially hazardous DCIS and to be able to stratify DCIS based on risk of progression to invasive disease, we established the international PREvent ductal Carcinoma In Situ Invasive Overtreatment Now (PRECISION) initiative[6]. PRECISION synergizes comprehensive prospective and retrospective DCIS studies[2,7], modelling and prospective clinical trials. Three ongoing prospective trials (COMET[8], LORIS[9], and LORD[10]) randomize patients between standard treatment and active surveillance for low risk DCIS. The identification of low risk DCIS based on morphological features is key for accrual into these trials, but also for international collaborations for conducting research studies of DCIS. We embarked on a DCIS interobserver variability study including whole slide digital images of hematoxylin and eosin (H&E) stained sections of DCIS including cohorts from three countries, namely the United States (USA), the United Kingdom (UK) and the Netherlands (NL) that were reviewed by breast pathologists practicing in these three countries. Our primary goal was to evaluate the extent of interobserver variability in DCIS grading between pathologists from the same and from different health care systems. Subsequently, we aimed to assess possible causes for the variability and then address strategies to establish greater uniformity of grading.

Methods

Slide collection

Four institutions, The Netherlands Cancer Institute (NKI; the Netherlands (NL)), Kings College London (KC; UK), MD Anderson Cancer Center (MDACC; USA) and Duke University Medical Center (DUMC; USA), participated in this study and contributed H&E stained whole slides images of tissue sections of DCIS. The cases were selected to represent the distribution of cytonuclear grade of DCIS (according to the pathology report or from previous review) from the participating countries or individual centers (supplementary table 1). The cases originated from the prospective, population-based Sloane DCIS cohort (KCL; UK)[2], the retrospective nation-wide Dutch DCIS cohort[5] and the retrospective, hospital-based DUMC and MDACC cohorts. Whole slides images of one representative H&E-stained section obtained from a formalin fixed paraffin embedded (FFPE) tissue block of a breast surgical resection were scanned at each centre, anonymized and uploaded to the NKI and evaluated using the web-based software platform Slidescore (supplementary table 1).[11] To assess the number of slides that had to be evaluated, power calculations were performed (see section power calculations, Supplementary file).

Histology & pathologists

To recapitulate pathology reporting in daily clinical practice, the breast pathologists interpreted the whole slide images of H&E tissue sections of DCIS without specific study-related guidelines for all evaluated variables (see supplementary table 2 for detailed information about the used diagnostic guidelines). The following histological variables were assessed (see scoring form, supplemental file): presence of DCIS/ atypical ductal hyperplasia (ADH)/ lobular carcinoma in situ (LCIS), DCIS grade (1, 2 or 3), DCIS grade (low or high), dominant histological architecture (comedo/solid, cribriform, (micro)papillary, flat/clinging, other), presence and semi-quantitative frequency of mitosis (sparse, many), lymphocytic infiltrate (absent, subtle, prominent), presence of calcifications (absent or present), presence of periductal fibrosis (absent, subtle, prominent) and presence and type of necrosis (absent, present - comedo, present – focal, present – comedo and focal).
Three breast pathologists from each country (NL, UK and USA) evaluated all the slides independently. The participating pathologists completed a short questionnaire to collect information about their experience and criteria for DCIS grading that they followed in their clinical practice (see supplementary table 3,).

Data analysis & statistics

The primary aim was the extent of variability between the nine pathologists for histological grade of DCIS based on review of the H&E scanned slides in Slidescore. Tissue slides of insufficient quality, as judged by more than 50% of

the participating pathologists for any histological variable were excluded from analysis (n=12).

As each slide was evaluated by each pathologist, generalized linear mixed models (GLMM) for cross-classified data structure were used to calculate kappa values as chance corrected association between pathologists (κ_{ma})[12,13]. κ_{ma} were obtained by taking into account levels of exact concordance, i.e. where pathologists assigned the exact same grade to a slide, and the level of disagreement among pathologists' classifications. κ_{ma} values were interpreted as measurement of agreement using the criteria suggested by Landis & Koch[14], which are based on the interpretation that 0.00 is pure coincidence and 1.00 is perfect agreement: <0.00 as no, 0.00-0.20 as poor to slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial and 0.81-1.00 as almost perfect agreement.

We modelled the histological variables separately and to analyze the influence of the tissue slides' and pathologists' characteristics on each of the histological variables, GLMM were adjusted for guidelines used, experience, country and using dominant or highest grade in case of heterogeneous DCIS as characteristics of the pathologists and origin of the slide (both country and centre) as characteristics of the slides. Since all the pathologists from the same country used the same guidelines (except in the USA; supplementary table 3), including both 'country of pathologists' and 'guidelines' in the same multivariable model resulted in collinearity. We therefore chose to use the 'guidelines' as covariate instead of country to evaluate variation. The different values of κ_{ma} from the different adjusted models were compared to the results of the intercept only models. The ordinal package within the open source software R was used for all the calculations.

Majority opinion and influence of ER, PR and HER2 expression

For each slide the majority opinion classification, defined as the grade given by most of the pathologists, was assigned. When there was no majority opinion (i.e. equal number of pathologists, for example, four pathologists graded 2, four pathologists graded 3 and one pathologists did not complete the form), the slide was assigned as not applicable (NA). The variable 'number of pathologists' was defined as the number of pathologists that make up the majority opinion and reflects the strength of agreement.

To investigate how to decrease interobserver variability, we retrospectively collected information about the status of estrogen receptor (ER), progesterone receptor (PR) and overexpression of HER2 through immunohistochemical stains (IHC) obtained from whole slides from the NKI, and the ER and PR status of the DUMC whole slides. MDACC had no IHC data available and KCL assessed biomarker IHC on tissue microarrays (TMAs) and was therefore excluded. For the IHC evaluated in NKI,

≥10% ER, ≥10% PR and ≥10% strong membrane expression of HER2 was considered positive, for 2+ HER2 expression (equivocal) silver in situ hybridization (SISH) was performed. The IHC from USA (DUMC) was examined by the Allred method[15] and a score of >2 was considered as positive. See supplementary table 4 for more details about the scoring details, antibodies used and IHC staining procedures.

Results

Cohort information & slide collection

In total, 425 slides were provided by the participating centers (110 by NKI, KCL, and DUMC and 95 by MDACC). All slides were independently evaluated by the international group of nine breast pathologists. Twelve of the 425 slides (2.8%) were excluded from all analyses based on quality issues, as noted by the majority of the participating pathologists. For the histological variables of grade and mitoses, two and five additional cases, respectively, were excluded based on quality issues. supplementary table 3 (Supplementary file) shows both the characteristics of the included cases, and of the participating pathologists.

Differences between pathologists

Figure 1A demonstrates both the individual evaluation and the majority opinion of grading as low (grade 1), intermediate (grade 2) and high (grade 3) per pathologist. It demonstrates substantial variability in grading the same lesion (see supplementary figure 1 for histological examples of concordant and discordant slides). In addition, some pathologists had a tendency for lower grading, while others had a tendency for higher grading; variability diminished only slightly when grade 1 and 2 were grouped together (Figure 1B).

Associations measure between pathologists, κ_{ma}

According to the GLMM model, the probability that an individual H&E section of DCIS was classified into grade 1, 2 or 3 was 8%, 44% and 48%, respectively. The model-based chance corrected measure of association, κ_{ma}, was estimated as 0.50 (95% confidence Interval (CI) 0.44-0.56, table 1), indicating moderate association between the nine pathologists. For dichotomized grade 1 and 2 vs 3, the κ_{ma} also indicated moderate association (0.51; 95% CI 0.43-0.59). When the pathologists had to select between low or high grade as a binary grading system for all cases, the κ_{ma} was 0.52 (95% CI 0.45-0.59). The highest association was achieved for the category of dominant architectural pattern with κ_{ma} of 0.61 (95% CI 0.57-0.64, table 1), indicating substantial association.

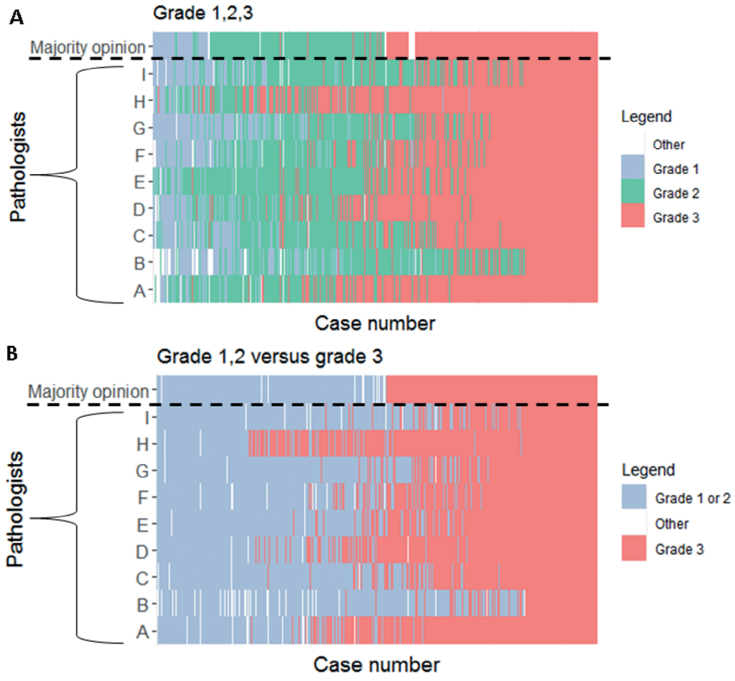


Figure 1. DCIS grades by pathologist (y-axis) and by case (x-axis). The upper row reflects the majority opinion. A) for grade 1 or 2 or 3. B) for grade 1 or 2 vs 3.

Table 1. Model-based measure of association (κ_{ma}) for histological variables.

Histological variable	Model based weighted kappa (κ_{ma})	95% CI
DCIS grade 1, 2, 3 (n=411; intercept only model)	0.50	0.44 - 0.56
DCIS grade 1 and 2 vs 3 (n=411)	0.51	0.43 - 0.59
DCIS grade 1 vs 2 and 3 (n=411)	0.45	0.41 - 0.50
DCIS grade as binary, low vs high (n=411)	0.52	0.45 - 0.59
Necrosis; absent vs present (n=413, manually dichotomized)	0.55	0.51 - 0.59
Calcifications; absent vs present (n=413)	0.51	0.48 - 0.55
Lymphocytic infiltrate; absent vs subtle vs prominent (n=413)	0.47	0.38 - 0.55
Periductal fibrosis; absent vs subtle vs prominent (n=413)	0.35	0.03 - 0.31
Mitoses; sparse vs many (n=408)	0.33	0.24 - 0.42
Architectural pattern; solid and comedo vs cribriform, flat and (micro)papillary (n=413)	0.61	0.58 - 0.64

DCIS grade 1 denotes low grade, 2 intermediate grade, 3 high grade

When incorporating guidelines used as covariate on the pathologist level, the κ_{ma} in the univariable GLMM model for DCIS grade did not change in comparison to the intercept only model (κ_{ma} =0.53; 95% CI 0.48-0.57; p=0.52, table 2). We aimed to investigate whether the κ_{ma} improved when we only included pathologists using the same guideline into the GLMM model. A minimum of three observers was necessary enabling us to analyze the UK and WHO guidelines. Pathologists utilizing the UK pathology guideline had better association between each other (κ_{ma} 0.58, 95% CI 0.56- 0.61) compared to pathologists using the WHO guidance, which showed a κ_{ma} of 0.48 (95% CI 0.36-0.61; p= 0.80), and a model including use of UK pathology guideline shows better association between pathologists compared to the standard model (p=0.02).

For DCIS cytonuclear grading, the associations between pathologists did not change when the following covariates were separately added to the model on pathologist and case level: pathologist's experience (κ_{ma}=0.50; 95% CI 0.44-0.57), country of the pathologist (κ_{ma}=0.51; 95% CI 0.44-0.57) and country of origin of the case (κ_{ma}= 0.49; 95% CI 0.42-0.55). When the model was adjusted for additional histological variables separately, the κ_{ma} for DCIS nuclear grade did not improve (table 2). Multivariable modelling including the variables characterizing the pathologists (i.e. use of guidelines, experience and manner of reporting cases of heterogeneous DCIS) showed an increased but not statistically improved κ_{ma} of 0.57 (95% CI 0.55-0.60; p=0.06). When the model was adjusted for all other histological variables together, the reproducibility for DCIS grading decreased (κ_{ma} = 0.31; 95% CI 0.26-0.36, table 2).

Majority opinion and influence of ER and HER2 expression

Grade 3 DCIS showed less variability than grade 1 or grade 2 disease: 62% of lesions were scored by eight or nine pathologists as grade 3 (see figure 2). We then explored whether ER and/or HER2 expression could help in the identification of grade 3 (high grade) lesions (see figure 3 and supplementary table 5). Figure 3, representing only NKI cases (n=106), shows that lesions categorized as grade 1 DCIS by the majority opinion were all ER positive and HER2 negative, those categorized as grade 2 were predominantly ER positive (100%) and HER2 negative (88%). Grade 3 DCIS cases, determined by the majority opinion, were heterogeneous for ER and HER2 expression, with both positive and negative cases represented. We were able to validate the results of ER expression in the IHC data from DUMC (USA) (see Supplementary table 5); none of the low grade cases of DCIS according to majority opinion were ER negative.

Table 2. Univariable and multivariable analysis with pathological and histological features as covariates to determine the influence on DCIS grade (1 or 2 or 3).

Variable	Model based weighted kappa (κ_{ma})	95% CI	P-value for kappa comparison with the outcome only
DCIS grade 1, 2 or 3	0.50	0.44 - 0.56	
Univariable analysis – adjusted for features of the pathologists			
Experience	0.50	0.44 - 0.57	0.95
Country of pathologist	0.51	0.44 - 0.57	0.91
Heterogeneous DCIS; highest vs most prominent vs other	0.53	0.48 - 0.57	0.54
Guideline used	0.53	0.48 - 0.57	0.52
Split according to guideline used			
1 Consensus Conference	*Only used by one pathologist, not possible*		
2 UK Royal College of Pathologists	0.58	0.56 - 0.61	0.02*
3 College of American Pathologists	*Only used by two pathologists, not possible*		
4 WHO	0.48	0.36 - 0.61	0.80
Univariable analysis - histological features			
Necrosis; absent vs present	0.45	0.39 - 0.52	0.31
Calcification; absent vs present	0.50	0.44 - 0.57	0.97
Lymphocytic infiltrate; absent vs subtle vs prominent	0.46	0.41 - 0.52	0.37
Periductal fibrosis; absent vs subtle vs prominent	0.48	0.43 - 0.54	0.72
Mitoses; sparse vs many	0.46	0.40 - 0.52	0.40
Architectural pattern; solid and comedo vs. cribriform, flat and (micro)papillary	0.45	0.39 - 0.52	0.33
Multivariable analysis – adjusted for features of the pathologists			
Guidelines + experience + solution to heterogeneity of DCIS	0.57	0.54 - 0.59	0.06
Country + experience + solution to heterogeneity DCIS	0.53	0.49 - 0.58	0.41
Multivariable analysis –adjusted for histological features			
Necrosis + calcification + lymphoid infiltrate + periductal fibrosis + mitosis + architectural pattern	0.31	0.26 - 0.36	<0.01*

* p-value showing a significant effect, i.e. p-value<0.05

DCIS grade 1 denotes low grade, 2 intermediate grade, 3 high grade, WHO world health organization

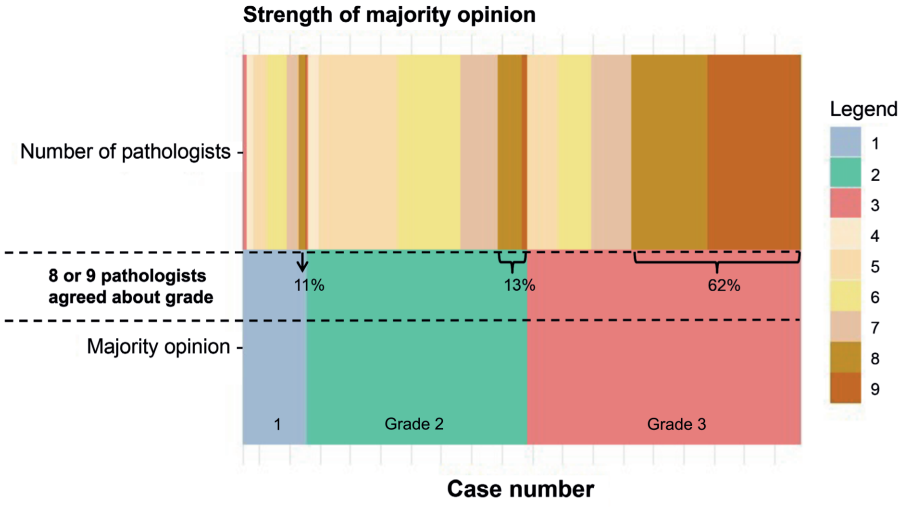


Figure 2. The strength of the majority opinion for low, intermediate and high grade. The bottom row shows the distribution of DCIS grade according to the majority opinion and the upper row the number of pathologists that represent the majority opinion.

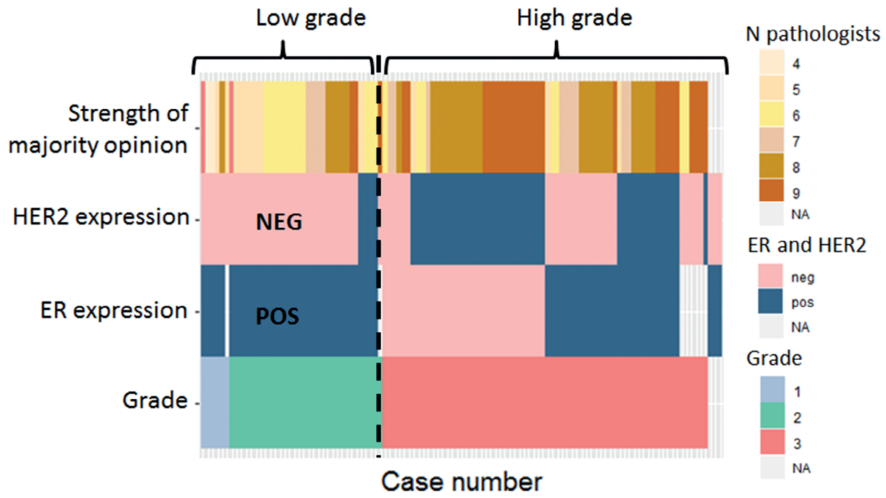


Figure 3. ER and HER2 expression in relation to low (grade 1), intermediate (grade 2) and high (grade 3) grade according to the majority opinion and to the strength of the majority opinion including the NL (NKI)cases (n=110) only.

Discussion

Although reproducibility of the *diagnosis* of DCIS has been demonstrated to have substantial agreement[16], this international study among nine pathologists showed kappa values of 0.5-0.6 for assessment of DCIS *grade*, based on a generalized linear mixed model, indicating only a moderate association between pathologists. Including guidelines as a covariate in to the GLMM model did not improve the association; analyzing the data specifically for the UK pathology guidelines[17] showed a statistically significant improvement in association between pathologists compared to the standard model. Linking the interobserver variability data to immunohistochemical stains demonstrated that almost all non-high grade DCIS lesions according to the majority opinion were ER-positive (100%) and HER2-negative (89%), whereas 55% high grade DCIS were ER-negative and/or HER2-positive (62%). Applying these biomarker stains might be helpful to prevent accidental selection of high grade DCIS, for example in active surveillance protocols.

The significance of cytonuclear grade of DCIS, whilst generally regarded as a predictor of risk of recurrence as subsequent in situ or invasive disease [2,18], is not universally accepted [3,7]. We show here variability in grading DCIS; twenty percent of cases were highly discordant as different pathologists categorized the exact same lesion, on a single identical H&E scanned slide, as grade 1, 2 or 3. This discrepancy might result in a low correlation between prognosis and grade. Multiple studies have shown high interrater variability of DCIS grade and have suggested methods for improvements in consistency, such as dichotomous scoring[19–21], assessing the proportions of DCIS heterogeneity[22], adding uniform e-learning[23] and using second opinions[24]. Our results are based on a GLMM model taking into account the same pathologists examined the same slides[25]. Such variability in grading of DCIS has profound consequences for inclusion of cases of DCIS in active surveillance trials (COMET[8], LORIS[9], LORD[10]), where low or intermediate grade (or low and lower portion of intermediate grade in LORIS) are inclusion criteria. Regarding the COMET and LORD, where no central review is performed, patients are deemed eligible or ineligible based on examination by an individual local pathologist. For all these reasons, it is essential to achieve a globally reproducible scoring system.

As noted, some pathologists tended to score substantially more DCIS lesions as low grade than others while the opposite also occurred. In the case of heterogeneous DCIS, one pathologist categorized the lesion according to the most prominent grade while the majority (7/9) classified the DCIS by the highest cytonuclear grade present, which could explain some of the differences presented. One guideline (UK) clarifies that the highest grade should be recorded when, uncommonly, more than one form is present[17]. Other, previous, guidelines like the 2012 WHO[26] or

1997 Consensus conference[27] have advised that all grades present should be noted. Specifically, in this study we sought to simulate daily clinical practice and therefore did not provide specific guidelines beforehand for grading or for any of the other histological features recorded. Compared to the standard model, pathologists who followed the UK pathology guidelines[17] showed significantly more mutual concordance (κ_{ma} =0.58; p=0.02; table 2) than those who used the 2012 WHO guidance[26] (κ_{ma} =0.48; p=0.80). However, when exploring the details of the various guidelines no major differences were apparent that could explain the better concordance for the UK guideline compared to the others[26–28] (supplementary table 2). It is the case that in the UK, adherence to the breast reporting guidelines is mandated for breast screening pathologists, as is participation in a twice yearly national breast external quality assurance slide review scheme (that includes cases of DCIS) as well as attendance at regional meetings to discuss these. However, two of the three UK breast pathologists are central reviewers in the LORIS trial (through which they have also provided advice and educational webinars for other UK pathologists) and two work in the same department (albeit where cases are reported by the individual). It is therefore difficult to know if the greater concordance of the 3 UK pathologists represents the recent focus on consistency of grading of DCIS in the UK, the overall educational and quality assurance mechanisms in place, or simply that they have had the opportunity to work together, discuss problematic cases and align their approach to DCIS grading. Nevertheless, this supports the use of one international DCIS grading system along with a uniform training program, as also suggested by other studies[1,19–21,29].

To improve guidance for clinical decision making, we explored the use of IHC. In our data on the NKI-series, majority opinion low and intermediate grade DCIS was characterized by ER positivity and HER2 negativity. We were able to validate this in DUMC (USA) slides for ER expression, scored by an alternative (Allred[15]) method (supplementary table 5). This is in line with other studies which also showed that ER was frequently expressed in low and intermediate grade DCIS, whereas HER2 positivity was much more frequent in high grade disease[30,31]. The proportion of pure DCIS that is ER positive is 68%-83%[5,30–33] whilst HER2 positivity ranges from 25%-35%[5,31,32,34]. IHC scoring for ER and HER2 is reported to have high interobserver agreement between pathologists (intra class coefficient >0.8)[5], which is better than the interobserver agreement for grade (presented here and other studies[19–22,35–37]). Globally, the use of IHC within DCIS is variable; no marker is at present included in international DCIS pathology minimum datasets, although in some national datasets (e.g. USA) ER assessment is mandated. In the USA, half of the patients with ER positive DCIS are treated with endocrine therapy[38], but this is still a subject of debate, and is much lower in other countries [2–4]. Positive ER/PR and negative HER2 status is used in the COMET trial as inclusion criteria for the active surveillance regimen[8] in keeping with the data presented

here; when DCIS shows ER negativity and/or HER2 positivity, classification as high grade DCIS should be considered.

The present study has several limitations. Firstly, only limited outcome data was available for many of the cases and therefore the primary outcome was histological interobserver variability, instead of recurrence or progression of disease. Unfortunately we were not able to validate the results of the N=106 NKI cases in another cohort. To our knowledge, only one single centre study has correlated interobserver variability with progression to invasive breast cancer and found that using majority opinion based scores of grade (grade 1+2 versus 3), mitotic activity and growth pattern were associated with outcome in patients treated with breast conserving surgery (BCS) only and not in patients treated with BCS plus radiotherapy. Furthermore, we sought to simulate daily clinical practice and therefore did not require adherence to guidelines assigned specifically for the study. The concordance may have been better if we had provided guidance for assessment of the slides. It should also be noted that most of the study pathologists are not using digital slides to diagnose cases in their daily practice, although digital pathology will become daily practice in the near future. In this study, a DCIS case was represented by one slide, while in daily practice multiple slides are typically examined in evaluating DCIS. Moreover, increasing the number of (international) pathologists would have provided more information about the differences between countries and the guidelines used. Lastly, independent validation of the data on ER and HER2 expression presented is necessary in order to prove the association between low and intermediate grade DCIS with immunohistochemical ER positivity and HER2 negativity.
The strength of this study is the international character of both the cases of DCIS and the participating pathologists. Moreover, the data has been analysed using a method that takes into account the cross-classified data structure.

In conclusion, in this international study we show a moderate concordance for a range of histological features of DCIS between nine specialist breast pathologists. As cytonuclear grade of DCIS plays a role as a prognostic parameter in treatment decisions there is an urgent need for adherence of pathologists to a more objective scoring system. As a first step in improving reproducibility, we suggest that ER negativity and/or HER2 positivity of an individual DCIS lesion is indicative of a high grade lesion, which may be of value in distinguishing this from low and intermediate grade DCIS, although validation is required.

Additional Information

Conflicts of interest

All the authors declare no conflict of interests.

Data availability

The data generated and analysed during this study will be available from the corresponding author upon reasonable request.

Ethics Approval and consent to participate

Local IRB's approved the use of the tissue blocks of NKI, MD Anderson Cancer Center and Duke University with the waiver of informed consent because of the retrospective character of the study.
For the UK slides held at Guy's and St Thomas' Hospitals in the King's Health Partner's Cancer Biobank facility, this is licensed by the Human Tissue Authority (license 12121). Ethics Committee approval was not required for this prospective cohort study originally conducted under the NHS Cancer Screening Program's application to the Patient Information Advisory Group.

Consent for publication

Not applicable

Funding

The PRECISION Team is recipient of a Cancer Research UK Grand Challenge Award 2017, jointly funded by Cancer Research UK and the Dutch Cancer Society.

Acknowledgements

We would like to acknowledge the funders, Cancer Research UK (CRUK) and the Dutch Cancer Society (KWF) of the PRECISION Team.

Authors contributions

Conception and design: MvS, SP, SK, AH, HS, AT, EL, JW
IT-support: JH
Statistical support: KJ, MS
Collection and assembly of data: MvS, SP, IB, SK, AH, AT, JT, WN, LC, DC, JB, JB, JW, EL
Data analysis and interpretation: MvS, KJ, SP, AH, SK, AT, HS, MS, EL, JW
Manuscript writing: all authors
Final approval of manuscript: all authors

References

1 Schuh F, Biazús JV, Resetkova E, *et al.* Reproducibility of three classification systems of ductal carcinoma in situ of the breast using a web-based survey. *Pathol Res Pract* 2010; **206**: 705-711

2 Thompson AM, Clements K, Cheung S, *et al.* Management and 5-year outcomes in 9938 women with screen-detected ductal carcinoma in situ: the UK Sloane Project On behalf of the Sloane Project Steering Group (NHS Prospective Study of Screen-Detected Non-invasive Neoplasias) 1. *Eur J Cancer* 2018; **101**: 210-219

3 Bijker N, Peterse JL, Duchateau L, *et al.* Risk factors for recurrence and metastasis after breast-conserving therapy for ductal carcinoma-in-situ: analysis of European Organization for Research and Treatment of Cancer Trial 10853. *J Clin Oncol* 2001; **19**: 2263-2271

4 Elshof LE, Schaapveld M, Schmidt MK, *et al.* Subsequent risk of ipsilateral and contralateral invasive breast cancer after treatment for ductal carcinoma in situ: incidence and the effect of radiotherapy in a population-based cohort of 10,090 women. *Breast Cancer Res Treat* 2016; **159**: 553-563

5 Visser LL, Elshof LE, Van de Vijver K, *et al.* Discordant Marker Expression Between Invasive Breast Carcinoma and Corresponding Synchronous and Preceding DCIS. *Am J Surg Pathol* 2019; **00**: 1-9

6 van Seijen M, Lips EH, Thompson AM, *et al.* Ductal carcinoma in situ: to treat or not to treat, that is the question. *Br J Cancer* 2019; **121**: 285-292

7 Elshof LE, Schaapveld M, Schmidt MK, *et al.* Subsequent risk of ipsilateral and contralateral invasive breast cancer after treatment for ductal carcinoma in situ: incidence and the effect of radiotherapy in a population-based cohort of 10,090 women. *Breast Cancer Res Treat* 2016; **159**: 553-563

8 Hwang ES, Hyslop T, Lynch T, *et al.* The COMET (Comparison of Operative versus Monitoring and Endocrine Therapy) trial: a phase III randomised controlled clinical trial for low-risk ductal carcinoma in situ (DCIS). *BMJ Open* 2019; **9**: e026797

9 Francis A, Thomas J, Fallowfield L, *et al.* Addressing overtreatment of screen detected DCIS; the LORIS trial. *Eur J Cancer* 2015; **51**: 2296-2303

10 Elshof LE, Tryfonidis K, Slaets L, *et al.* Feasibility of a prospective, randomised, open-label, international multicentre, phase III, non-inferiority trial to assess the safety of active surveillance for low risk ductal carcinoma in situ - The LORD study. *Eur J Cancer* 2015; **51**: 1497-1510

11 Hudecek J. Slide Score. Available from: www.slidescore.com.

12 Nelson KP, Edwards D. Measures of agreement between many raters for ordinal classifications. *Stat Med* 2015; **34**: 3116-3132

13 Nelson KP, Edwards D. A measure of association for ordered categorical data in population-based studies. *Stat Methods Med Res* 2018; **27**: 812-831

14 Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2529310. *Biometrics* 1977; **33**: 159-174

15 Allred DC, Harvey JM, Berardo M, *et al.* Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Mod Pathol* 1998; **11**: 155-168

16 Rakha EA, Bennett RL, Coleman D, *et al.* Review of the national external quality assessment (EQA) scheme for breast pathology in the UK. *J Clin Pathol* 2017; **70**: 51-57
17 NHS Cancer Screening Programmes. Pathology reporting of breast disease in surgical excision specimens incorporating the dataset for histological reporting of breast cancer. *R Coll Pathol* 2016
18 Pinder SE, Duggan C, Ellis IO, *et al.* A new pathological system for grading DCIS with improved prediction of local recurrence: results from the UKCCCR/ANZ DCIS trial. *Br J Cancer* 2010; **103**: 94-100
19 Groen EJ, Hudecek J, Mulder L, *et al.* Prognostic value of histopathological DCIS features in a large-scale international interrater reliability study. *Breast Cancer Res Treat* 2020
20 Alghamdi SA, Krishnamurthy K, Garces Narvaez SA, *et al.* Low-Grade Ductal Carcinoma In Situ. *Am J Clin Pathol* November 2019: 1-8
21 Van Bockstal M, Baldewijns M, Colpaert C, *et al.* Dichotomous histopathological assessment of ductal carcinoma in situ of the breast results in substantial interobserver concordance. *Histopathology* 2018; **73**: 923-932
22 Elston CW, Sloane JP, Amendoeira I, *et al.* Causes of inconsistency in diagnosing and classifying intraductal proliferations of the breast. *Eur J Cancer* 2000; **36**: 1769-1772
23 van Dooijeweert C, Deckers IAG, de Ruiter EJ, *et al.* The effect of an e-learning module on grading variation of (pre)malignant breast lesions. *Mod Pathol* 2020
24 Elmore JG, Tosteson ANA, Pepe MS, *et al.* Evaluation of 12 strategies for obtaining second opinions to improve interpretation of breast histopathology: Simulation study. *BMJ* 2016; **353**: 1-10
25 Mitani AA, Freer PE, Nelson KP. Summary measures of agreement and association between many raters' ordinal classifications. *Ann Epidemiol* 2017; **27**: 677-685.e4
26 Lakhani SR, Ellis IO, Schnitt SJ, *et al.* WHO Classification of Tumours of the Breast. International Agency for Research on Cancer; 2012.
27 Consensus Conference on the classification of ductal carcinoma in situ. The Consensus Conference Committee. *Cancer* 1997; **80**: 1798-1802
28 Lester SC, Bose S, Chen Y-Y, *et al.* Protocol for the examination of specimens from patients with ductal carcinoma in situ of the breast. *Arch Pathol Lab Med* 2009; **133**: 15-25
29 Van Dooijeweert C, Van Diest PJ, Baas IO, *et al.* Variation in breast cancer grading: The effect of creating awareness through laboratory-specific and pathologist-specific feedback reports in 16 734 patients with breast cancer. *J Clin Pathol* 2020: 1-7
30 Meijnen P, Peterse JL, Antonini N, *et al.* Immunohistochemical categorisation of ductal carcinoma in situ of the breast. *Br J Cancer* 2008; **98**: 137-142
31 Tamimi RM, Baer HJ, Marotti J, *et al.* Comparison of molecular phenotypes of ductal carcinoma in situ and invasive breast cancer. *Breast Cancer Res* 2008; **10**: R67
32 Clark SE, Warwick J, Carpenter R, *et al.* Molecular subtyping of DCIS: heterogeneity of breast cancer reflected in pre-invasive disease. *Br J Cancer* 2011; **104**: 120-127
33 Rauch GM, Kuerer HM, Scoggins ME, *et al.* Clinicopathologic, mammographic, and sonographic features in 1,187 patients with pure ductal carcinoma in situ of the breast by estrogen receptor status. *Breast Cancer Res Treat* 2013; **139**: 639-647
34 Siziopikou KP. Ductal carcinoma in situ of the breast: current concepts and future directions. *Arch Pathol Lab Med* 2013; **137**: 462-466

35 Schuh F, Biazús JV, Resetkova E, *et al.* Histopathological grading of breast ductal carcinoma in situ: validation of a web-based survey through intra-observer reproducibility analysis. *Diagn Pathol* 2015; **10**: 93
36 Elmore JG, Longton GM, Carney PA, *et al.* Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *JAMA* 2015; **313**: 1122
37 Onega T, Weaver DL, Frederick PD, *et al.* The diagnostic challenge of low-grade ductal carcinoma in situ. *Eur J Cancer* 2017; **80**: 39-47
38 Ward EM, DeSantis CE, Lin CC, *et al.* Cancer statistics: Breast cancer in situ. *CA Cancer J Clin* 2015; **65**: 481-495

Supplementary files

Power calculation method

Based on a statistical power calculation, we aimed to have at least 379 tissue slides all evaluated by nine pathologists, i.e. three pathologists from each country. This number was obtained taking into consideration the proportion of high-grade DCIS anticipated between three countries and taking correlations within pathologists from the same country into account. We expected overall proportions of high-grade DCIS from the NL, US and UK to be 42%, 52% and 62% respectively, and we assumed that correlations between randomly chosen grades of DCIS within pathologists from the same country was 0.60. Such a design would give us at least 80% power to detect all pairwise comparisons of proportions between different countries using a corrected for multiple testing significance level of 0.016.

Scoring form for evaluating the DCIS slides

☐ **Disease present: DCIS**
☐ **Disease present: ADH**
☐ **Disease present: LCIS**

Dominant architectoral pattern:

- Not assessable
- Comedo
- Solid
- Cribiform
- Flat/Clinging
- (Micro)papillary

Calcification present:

- Not assessable
- Absent
- Present

Necrosis present:

- Not assessable
- Absent
- Present: Comedo
- Present: Focal
- Present: Comedo and focal

Periductal fibrosis present:

- Not assessable
- Absent
- Subtle
- Prominent

Lymphocytic infiltrate (in relation with DCIS) present:

- Not assessable
- Absent
- Subtle
- Prominent

Histological grade DCIS (1/2/3):

- Not assessable
- Low grade
- Intermediate grade
- High grade

Histological grade DCIS (low/high):

- Not assessable
- Low grade
- High grade

Frequency of mitoses:

- Not assessable
- Sparse
- Many

Comments (other diagnosis or otherwise):
open text field

Supplementary table 1. Information regarding included slides

Before inclusion in the study, all slides were evaluated to ensure that they were in focus. Pilot study was performed to check if the quality between the slides from the different centers was similar.

	UK (KCL)	**NL (NKI)**	**USA (Duke)**	**USA (MDACC)**
Type of scanner	NanoZoomer 2.0 HT Slide Scanner (Hamatsu Photonics)	Aperio AT2 Slide Scanner (Leica Biosystems)	Leica Aperio scanner	Aperio AT2 Slide Scanner (Leica Bio systems)
Magnification	40x	20x	20x	20x
Type of slides	Whole breast images	Whole breast images	Whole breast images	Whole breast images
Format	.ndpi	.svs	.svs	.svs
Grade to original pathology reports (N, %)				
Grade 1	12 (11%)	19 (17%)	5 (5%)	14 (16%)
Grade 2	34 (31%)	36 (33%)	41 (45%)	38 (45%)
Grade 3	64 (58%)	55 (50%)	46 (50%)	33 (39%)
Excluded*	0	0	18	10

* Slides originally evaluated as grade 1-2 or 2-3 were excluded

Supplementary table 2. Criteria of the guidelines used.

	WHO 2012	RCPath Guidelines UK	College of American pathologists (Lester et al)	Consensus conference on classification of DCIS
Grade 1 (low)				
Cell appearance	Small monomorphic cells with regular chromatin inconspicuous nucleoli	Monomorphic, evenly spaced cells with rounded, centrally placed nuclei and inconspicuous nucleoli	Monomorphic cells, usually diffuse finely dispersed chromatin, only occasional nucleoli	Monomorphic, usually exhibit diffuse, finely dispersed chromatin, only occasional nucleoli and mitotic figures
Pattern	Arcades, micropapillae, cribiform, or solid	Generally arranged in micropapillary and cribriform patterns		
Orientation	Polarized cells around rosettes	Usually polarisation of cells covering the micropapillae	Polarized toward luminal spaces	Usually associated with polarization of constituent cells
Nuclear Size	Nuclear: Uniform size,	Nuclear: 2x-3x erythrocyte	Nuclear: 1x-2x size of normal RBC or normal duct epithelial cell nucleus	Nuclear: 1.5-2.0 RBC or duct epithelial cell nuclear dimensions
Mitosis	Rare	Few		
Necrosis	uncommon	rarely individual cell necrosis		
Calcifications	Often psammomatous type			

Supplementary table 2. Continued.

	WHO 2012	RCPath Guidelines UK	College of American pathologists (Lester et al)	Consensus conference on classification of DCIS
Grade 2 (intermediate)				
Cell appearance	Mild to moderate variability in shape, variably coarse chromatin, variably prominent nucleoli	Moderate pleomorphism, nuclear to cytoplasmic ratio is often high, and one or two nucleoli may be identified. one or two nucleoli. Clear cell or apocrine types often fall into this category	Intermediate pleomorphism, intermediate chromatin, nucleoli	Nuclei that are neither NG1 or NG2
Pattern		Solid, cribriform or micropapillary.		
Orientation	Cell polarization is not well developed as in low-nuclear grade	some degree of polarization	Intermediate polarization	
Size	Nuclear: variability in size	Nuclear: 2-3x size of an erythrocyte	Nuclear: intermediate	
Mitosis	Maybe present		Intermediate	
Calcifications	Distribution of amorphous of or laminated microcalcifications is generally similar to low-nuclear-grade			
Necrosis	Puncate or comedo necrosis maybe present			

Supplementary table 2. Continued.

	WHO 2012	RCPath Guidelines UK	College of American pathologists (Lester et al)	Consensus conference on classification of DCIS
Grade 3 (high)				
Cell appearance	Highly atypical cells with pleiomorphic nuclei	pleomorphic, irregularly spaced and, nuclei exhibiting marked variation in size with irregular nuclear contours, coarse chromatin and prominent nucleoli	Markedly pleomorphic, usually vesicular with irregular chromatin distribution, prominent nucleoli	Markedly pleiomorphic, usually vesicular and exhibit irregular chromatin distribution and prominent often multiple nucleoli
Pattern	Solid, cribiform or micropapillary patterns	It is often solid with comedo-type central necrosis. Also micropapillary and cribriform patterns frequently associated with central comedo type necrosis		
Orientation	Poorly polarized	rarely any polarization of cells	Usually not polarized toward the luminal space	
Size	Lesion: usually >5mm	Nuclear: >3x the size of erythrocytes	Nuclear : >2.5x size of RBC or normal duct epithelial cell nucleus	Nuclear: usually>2.5 x RBC or duct epithelial cell nuclear dimensions
Mitosis	Usually common (not required)	usually frequent and abnormal forms may be seen		Might be conspicuous
Calcifications	Amorphous microcalcifications are common and usually associated with intraluminal debris			
Necrosis	Frequently presence of comedo necrosis (not obligatory)			

Supplementary table 3. Characteristics of participating pathologists and examined tissue slides

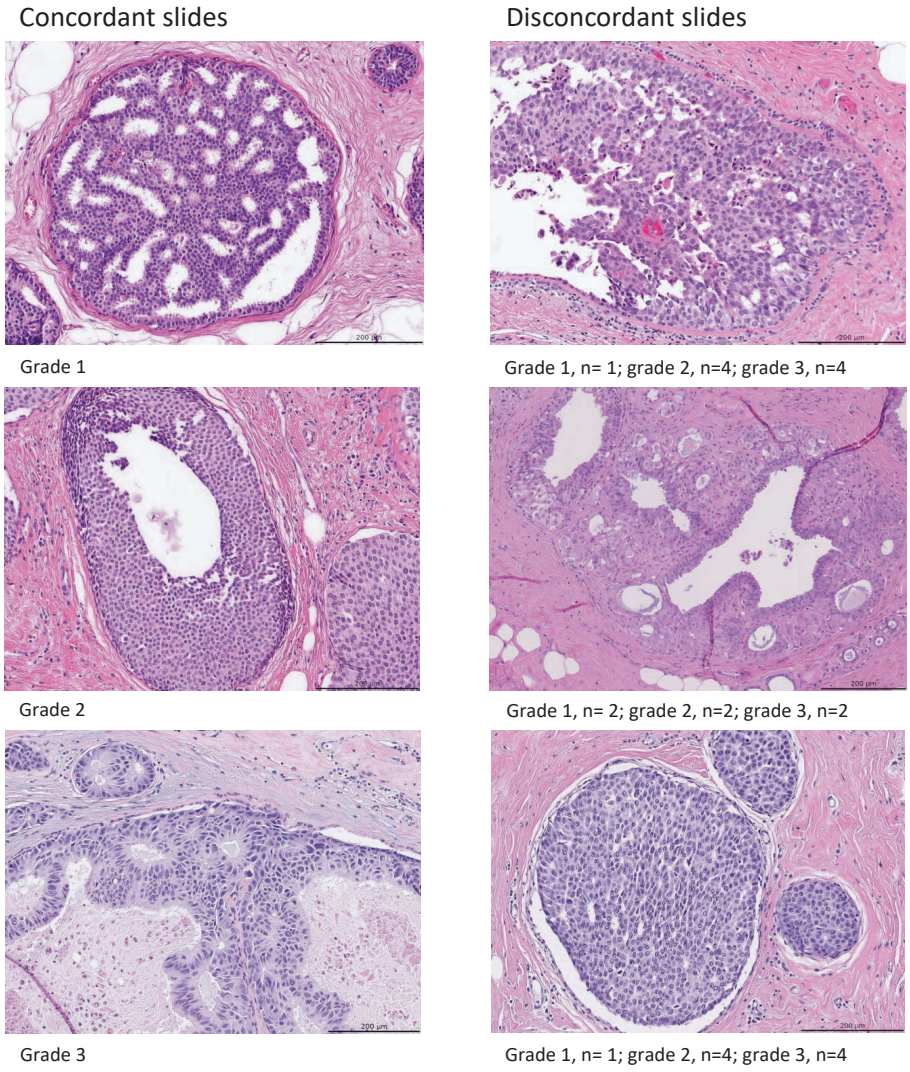
Pathologists (N, %)	**9 (100%)**	**Slides (N, %)**	**425 (100%)**
Country		*Center*	
The Netherlands (NL)	3 (33%)	NKI	110 (26%)
United Kingdom (UK)	3 (33%)	KCL	110 (26%)
United States (US)	3 (33%)	Duke	110 (26%)
		MDACC	95 (22%)
Experience		*Grade according to majority opinion*	399 (100%)
Median	12.0 years	1	45 (11%)
<10 yrs	5 (56%)	2	158 (40%)
>=10 yrs	4 (44%)	3	196 (49%)
Guidelines			
WHO	3 (33%)		
UK RCPath Guidelines	3 (33%)		
College of American pathologists	2 (22%)		
Consensus conference on classification of DCIS	1 (11%)		
In case of heterogeneous DCIS			
Highest grade	7 (78%)		
Most prominent grade	1 (11%)		
Other	1 (11%)		

Supplementary table 4. Details about the scoring of the immunohistochemical stains and characteristics of the used antibodies

	Antigen	NL (NKI)	USA (Duke)
Clone	ER	SP1	1D5 and ER-2-123
	PR	1E2	PgR1294
	HER2	4B5	*Not used in this study*
Dilution	ER	ready-to-use	ready-to use
	PR	ready-to-use	ready-to-use
	HER2	ready-to-use	*Not used in this study*
Manufacturer	ER	Ventana medical systems	Dako / Agilent
	PR	Ventana medical systems	Dako / Agilent
	HER2	Ventana medical systems	*Not used in this study*
Type of slides		Whole slides	Whole slides
Scorings method	ER	% of positive cells; ≥10% is positive	Allred method; >2 is considered as positive
	PR	% of positive cells; ≥10% is positive	Allred method; >2 is considered as positive
	HER2	% membrane staining; ≥10% is positive(3+), if incomplete or weak (2+) SISH was performed	*Not used in this study*
Number of observers		7 (5 pathologists)	1 out of 5 breast pathologists
More details		Supplementary table Visser et al. Clin Can Res 2018	

Supplementary table 5. ER, PR and HER2 expression in relation to interobserver variability in a subset. Grade 1, grade 2, or grade 3 are established according to the majority opinion. For 'certain' cases eight or nine pathologists agreed and 'uncertain' cases <8 pathologists agreed.

NL	**ERneg**	**ERpos**		**PRneg**	**PRpos**		**HER2neg**	**HER2pos**	
Grade 1	**0**	**5**	**5**	**0**	**5**	**5**	**6**	**0**	**6**
Certain (8,9)	0	1		0	1		1	0	
Uncertain (<8)	0	4		11	4		5	0	
Grade 2	**0**	**31**	**31**	**13**	**18**	**31**	**28**	**4**	**32**
Certain (8,9)	0	7		2	5		8	0	
Uncertain (<8)	0	24		11	13		20	4	
Grade 3	**34**	**28**	**62**	**47**	**15**	**62**	**26**	**42**	**68**
Certain (8,9)	27	18		34	11		14	35	
Uncertain (<8)	7	10		13	4		12	7	
Total	**34**	**64**	**98**	**60**	**38**	**98**	**60**	**46**	**106**
USA (Duke)	**ERneg**	**ERpos**		**PRneg**	**PRpos**				
Grade 1	**0**	**8**	**8**	**0**	**8**	**8**			
Certain (8,9)	0	0		0	0				
Uncertain (<8)	0	8		0	8				
Grade 2	**0**	**46**	**46**	**2**	**42**	**44**			
Certain (8,9)	0	5		0	5				
Uncertain (<8)	0	41		2	37				
Grade 3	**16**	**33**	**49**	**21**	**26**	**47**			
Certain (8,9)	14	15		17	12				
Uncertain (<8)	2	18		4	14				
Total	**16**	**87**	**103**	**23**	**76**	**99**			



Supplementary figure 1. Histological examples of concordant and discordant slides.