Dichotomous histopathological assessment of ductal carcinoma *in situ* of the breast results in substantial interobserver concordance

Mieke Van Bockstal,[1,2] [iD] Marcella Baldewijns,[3,*] Cécile Colpaert,[4,*] Hélène Dano,[5,*] Giuseppe Floris,[6,7,*] Christine Galant,[5,*] Kathleen Lambein,[8,9,*] Dieter Peeters,[3,*] Sofie Van Renterghem,[2,*] Anne-Sophie Van Rompuy,[6,*] Sofie Verbeke,[2,*] Stephanie Verschuere[10,*] & Jo Van Dorpe[2]

[1]Department of Pathology, Erasmus Medical Centre, Rotterdam, The Netherlands, [2]Department of Pathology, Ghent University Hospital, Ghent, [3]Department of Pathology, Antwerp University Hospital, Antwerp, [4]Department of Pathology, GZA, Antwerp, [5]Department of Pathology, University Clinics St Luc, Brussels, [6]Department of Pathology, University Hospitals Leuven, Leuven, [7]Department of Imaging and Pathology, Laboratory of Translational Cell & Tissue Research, KU Leuven, Leuven, [8]Department of Pathology, AZ St Lucas Hospital, Ghent, [9]Department of Surgical Oncology, University Hospitals Leuven, Leuven, and [10]Department of Pathology, AZ Delta, Roeselare, Belgium

Van Bockstal M, Baldewijns M, Colpaert C, Dano H, Floris G, Galant C, Lambein K, Peeters D, Van Renterghem S, Van Rompuy A-S, Verbeke S, Verschuere S & Van Dorpe J

(2018) *Histopathology* **73**, 923–932. https://doi.org/10.1111/his.13741

Aims: Robust prognostic markers for ductal carcinoma *in situ* (DCIS) of the breast require high reproducibility and thus low interobserver variability. The aim of this study was to compare interobserver variability among 13 pathologists, in order to enable the identification of robust histopathological characteristics.

Methods and results: One representative haematoxylin and eosin-stained slide was selected for 153 DCIS cases. All pathologists independently assessed nuclear grade, intraductal calcifications, necrosis, solid growth, stromal changes, stromal inflammation, and apocrine differentiation. All characteristics were assessed categorically. Krippendorff's alpha was calculated to assess overall interobserver concordance. Cohen's kappa was calculated for every observer duo to further explore interobserver variability. The highest concordance was observed for necrosis, calcifications, and stromal inflammation.

Assessment of solid growth, nuclear grade and stromal changes resulted in lower concordance. Poor concordance was observed for apocrine differentiation. Kappa values for each observer duo identified the 'ideal' cut-off for dichotomisation of multicategory variables. For instance, concordance was higher for 'non-high versus high' nuclear grade than for 'low versus non-low' nuclear grade. 'Absent/mild' versus 'moderate/extensive' stromal inflammation resulted in substantially higher concordance than other dichotomous cut-offs.

Conclusions: Dichotomous assessment of the histopathological features of DCIS resulted in moderate to substantial agreement among pathologists. Future studies on prognostic markers in DCIS should take into account this degree of interobserver variability to define cut-offs for categorically assessed histopathological features, as reproducibility is paramount for robust prognostic markers in daily clinical practice. A new

Address for correspondence: M Van Bockstal, Department of Pathology, Erasmus Medical Centre, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands. e-mails: mieke.vanbockstal@ugent.be; m.vanbockstal@erasmusmc.nl
*These authors contributed equally to this work.

prognostic index for DCIS might be considered, based on two-tier grading of histopathological features. Future research should explore the prognostic potential of such two-tier assessment.

Introduction

Ductal carcinoma *in situ* (DCIS) of the breast represents a heterogeneous group of non-obligate preinvasive precursors of invasive carcinoma of no special type. The ability of DCIS to progress towards invasive cancer is likely to result from a complex interplay between intrinsic and extrinsic factors, such as genetic anomalies within the neoplastic cells, stromal remodelling, and stroma-derived proinvasive signalling.[1] Although DCIS has been the subject of intensive research during the past four decades, no robust prognostic markers have emerged that enable prediction of progression towards invasive cancer. Because surgery has always been the cornerstone of DCIS treatment, the natural biology of DCIS is poorly understood. Ongoing clinical trials such as the COMET, LORD and LORIS trials, which are investigating watchful waiting strategies for low-grade DCIS, might provide crucial information to fill this gap in our knowledge.[2–4] The main motive for such trials is the increasing awareness that significant numbers of DCIS patients are currently overtreated, as not all DCIS cases will become invasive.[1,5]

If these trials can prove that watchful waiting is not inferior to standard surgical treatment in a selected subgroup of patients, the search for proper prognostic markers is likely to be intensified. If DCIS is left untreated, adequate prognosticators are necessary to estimate the risk of upstaging of DCIS to invasive cancer on the initial biopsy, and to evaluate the risk of evolution towards invasion in the future. In addition, these prognosticators could be used to estimate the recurrence risk in patients treated with lumpectomy, thereby providing information on the need for additional radiotherapy. Ideally, adequate prognostic markers should enable us to reduce the number of patients who undergo surgery, as well as the number of patients who receive adjuvant irradiation and hormonal therapy. The current uncertainty about what constitutes 'high-risk DCIS' causes significant variability in treatment.[5] In the past, several histopathological and immunohistochemical characteristics of DCIS have been investigated, often with conflicting results regarding their potential for recurrence risk prediction. We wondered whether interobserver variability might explain such conflicting results. Adequate prognostic markers require robustness of assessment, i.e. high reproducibility and thus low interobserver variability.

The goal of this study was to investigate the levels of interobserver variability among histopathologists in the assessment of DCIS. Investigation of interrater concordance should enable the identification of reproducible robust histopathological features. In addition, this study could compel us to develop new definitions or cut-offs for those characteristics with poor interrater concordance, to enhance the robustness of assessment in the daily routine of histopathology laboratories.

Materials and methods

PATIENTS

All patients were women who consecutively underwent breast-conserving surgery or mastectomy for DCIS between 1 January 2007 and 31 December 2015 at Ghent University Hospital (Ghent, Belgium). Needle biopsies or vacuum-assisted core biopsies were excluded. DCIS cases with associated microinvasive foci were included, but DCIS cases admixed with invasive carcinoma (size of >1 mm) were excluded. All haematoxylin and eosin-stained slides were retrieved from the archives of the Department of Pathology (Ghent University Hospital), and reviewed by one pathologist (M.V.B.). One representative glass slide was selected for each lesion, and this single slide was used for subsequent histopathological assessment. No consecutive sections were cut from the corresponding tissue blocks. All pathologists (coded P1–P13) assessed the selected slides independently by using a light microscope. No digitally scanned slides were used. All 'observers' are routinely involved in breast pathology in academic and non-academic hospitals. This study was approved by the ethics committee of Ghent University Hospital (EC/2018/0014).

### DEFINITIONS FOR CATEGORICAL ASSESSMENT

All histopathological features were assessed categorically. All participants were provided with detailed guidelines regarding the variables within each category. No training

set was used, as this study concerned commonly assessed features. Nuclear grade was assessed as a three-tier variable according to the American Society of Clinical Oncology/College of American Pathologists protocol for examination of DCIS specimens,[6] which is partly based on the 1997 Consensus Conference on DCIS classification.[7] Nuclear grade was categorised as low, intermediate, or high, on the basis of nuclear size, pleomorphism, chromatin distribution, nucleoli, mitoses, and orientation of nuclei.[6,7]

Ductal carcinoma *in situ* architecture included solid, cribriform, papillary and micropapillary growth, regardless of comedonecrosis,[8,9] and was assessed dichotomously: predominantly solid and non-solid architectures were defined as ≥50% and <50% solid growth, respectively. Necrosis was classified into four categories.[6,9] Necrosis was defined as areas of confluent eosinophilic material containing ghost cells and karyorrhectic debris, and categorised as follows: no necrosis, single-cell necrosis, focal necrosis (necrotic debris in <50% of affected ducts), and extensive necrosis (necrotic debris in ≥50% of affected ducts).[9] Calcifications within DCIS were scored as present or absent.

Periductal stromal changes were classified semiquantitatively as previously described.[10] Myxoid stroma was defined as loosely arranged collagen fibres, interspersed with an amorphous, slightly basophilic substance (illustrated in Van Bockstal *et al.*[10–12] and Figure 1A,B). Periductal stromal changes were divided into four categories. DCIS showed either periductal sclerotic stroma without (<1% of ducts) myxoid changes, a mild amount (i.e. ≥1% but <33% of ducts) of myxoid stroma, a moderate amount (i.e. ≥33% but <66% of ducts) of myxoid stroma, or an extensive amount (i.e. ≥66% of ducts) myxoid stroma.

The presence and degree of a chronic inflammatory infiltrate in the periductal stroma were recorded semiquantitatively as previously described.[9,10,13,14] Four categories were discerned. In DCIS with 'no stromal inflammation' (Figure 2A), the periductal stroma was not infiltrated by lymphocytes. In DCIS with 'mild stromal inflammation' (Figure 2B), the periductal stroma surrounding the affected ducts was infiltrated by a few loosely arranged lymphocytes but dense lymphocytic aggregates were absent, so the periductal stroma was easy to perceive. In DCIS with 'moderate stromal inflammation' (Figure 2C), the majority (i.e. ≥50%) of the ducts were surrounded by a moderately dense inflammatory infiltrate: lymphoid aggregates were present but lymphoid follicle formation was absent, and assessment of the periductal stroma was not hampered by the density of the infiltrate. In DCIS with 'extensive stromal inflammation' (Figure 2D), the majority (i.e. ≥50%) of the ducts were surrounded
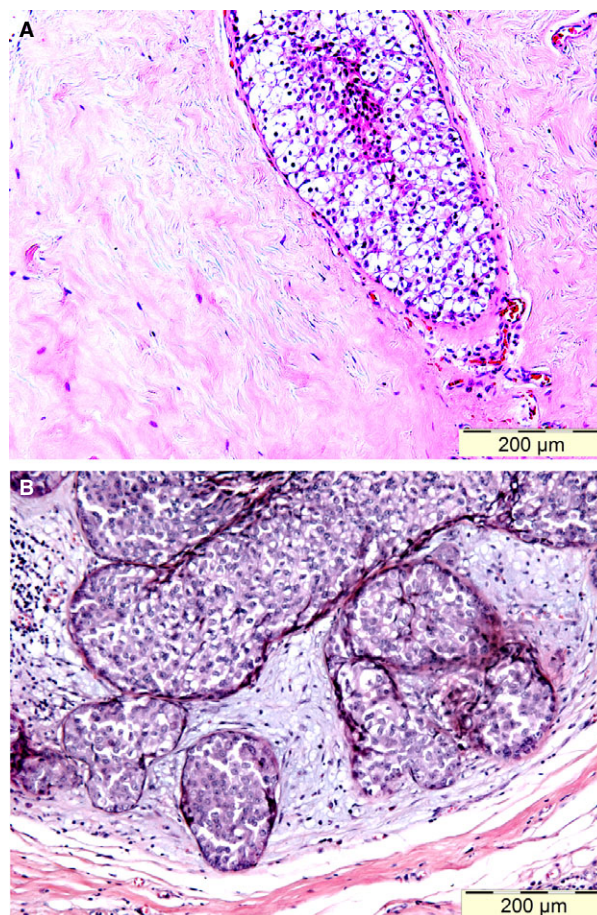
Figure 1. Photomicrographs of haematoxylin and eosin staining illustrating ductal carcinoma *in situ* (DCIS) with predominantly sclerotic periductal stroma (**A**), and DCIS with predominantly myxoid periductal stroma (**B**).

by a dense inflammatory infiltrate, consisting of large aggregates of lymphocytes. The density of this infiltrate hampered assessment of the periductal stroma. Lymphoid follicle formation could be present, but was not a prerequisite.

Apocrine differentiation was defined as previously described,[15] and was assessed regardless of the growth pattern. DCIS was classified as apocrine when >50% of the lesion showed apocrine features, i.e. abundant eosinophilic cytoplasm and large pleomorphic vesicular nuclei with prominent nucleoli.[15] No additional immunohistochemistry was used to determine apocrine differentiation.

STATISTICS

Statistical analyses were performed with IBM SPSS STATISTICS 25.0 (IBM, Chicago, IL, USA). The arithmetic mean was calculated for each DCIS lesion and
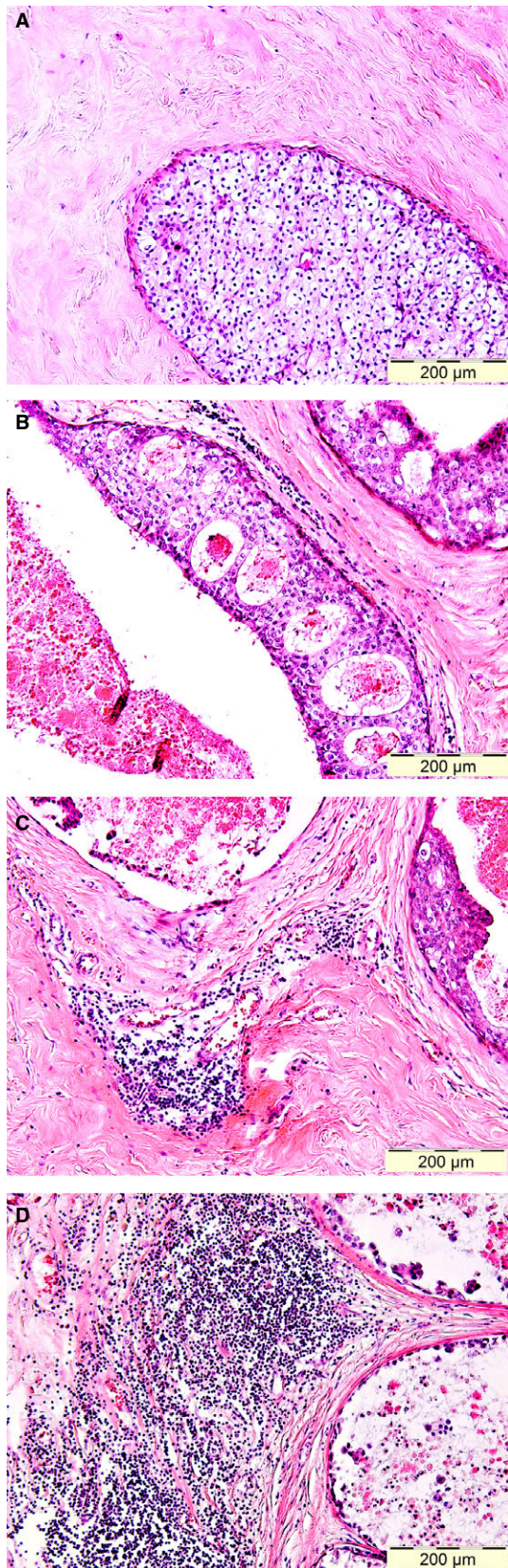
Figure 2. Photomicrographs of haematoxylin and eosin staining illustrating ductal carcinoma *in situ* with no periductal stromal inflammation (**A**), mild periductal stromal inflammation (**B**), moderate periductal stromal inflammation (**C**), and extensive periductal stromal inflammation (**D**).
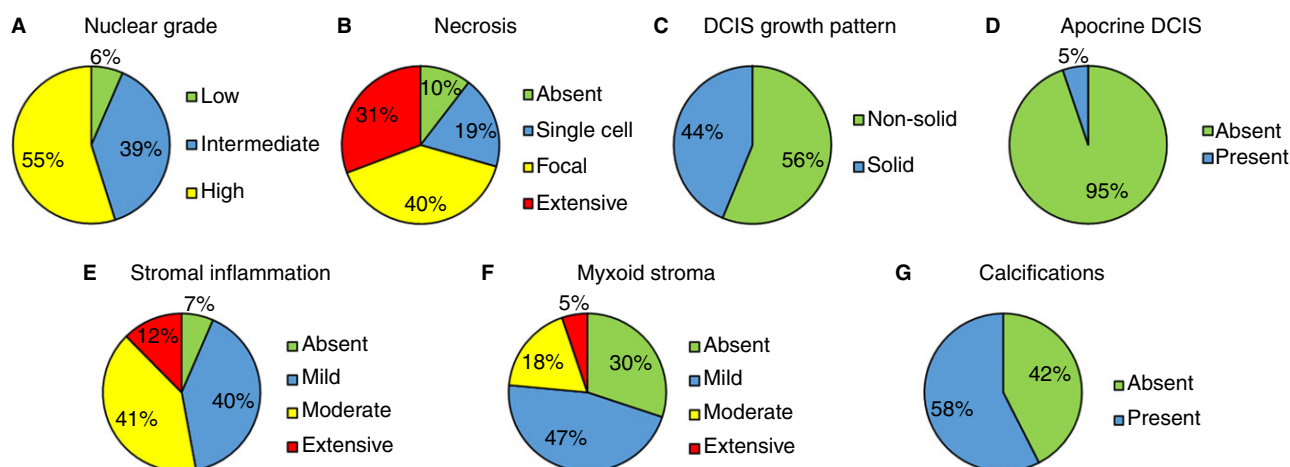
for each histopathological feature, to assess the distribution of each feature within this cohort of DCIS cases. This average score accords with the most commonly addressed category for a specific feature for each lesion. For calculation of Krippendorff's alpha (KA) reliability estimates, the 'Kalpha' macro provided by Hayes and Krippendorff was used (http://afhayes.com/spss-sas-and-mplus-macros-and-code.html). This macro computes KA for categorical data, regardless of the number of observers and categories, and regardless of any missing data.[16,17] KA (with the number of bootstrap samples set at 10 000) was used to investigate overall interrater concordance for each feature. Cohen's kappa values were calculated for each observer duo (i.e. 78 kappa values for each dichotomised histopathological feature), as this allowed detailed investigation of differences in reciprocal interrater concordance. Box-and-whisker plots were constructed in SPSS to visualise distributions of Cohen's kappa for each dichotomised feature. Interpretation was performed according to Landis and Koch.[18] Spearman's rho was determined to investigate possible correlations between the degree of concordance among pathologists and their relative experience, i.e. the number of years in practice.

Results

PERCENTAGE AGREEMENT

All histopathological features were classified in two, three or four categories by 13 observers. The participating pathologists had been in practice for 14 years on average (range 3–31 years). No significant correlation was observed between number of years in practice and degree of interobserver concordance (Spearman's rho ranging from −0.142 to 0.059; $P > 0.05$), although this might have been due to the limited number of participants. All observers rated all cases ($n = 153$). There were four missing values in the dataset. The distribution of all histopathological features is shown in Figure 3, based on the average scores provided by all observers. Percentage agreement signifies the number of cases that were rated identically by all pathologists. Twenty-seven of 153 DCIS cases (17.6%) received an identical nuclear

Figure 3. Pie charts illustrating the distribution of six histopathological features throughout the ductal carcinoma *in situ* (DCIS) cohort (*n* = 153). All features were assessed by 13 pathologists, and the average score was calculated for nuclear grade (**A**), necrosis (**B**), predominantly solid or predominantly non-solid DCIS growth pattern (**C**), apocrine differentiation (**D**), periductal stromal inflammation (**E**), myxoid periductal stromal changes (**F**), and the presence or absence of intraductal calcifications (**G**). [Colour figure can be viewed at wileyonlinelibrary.com]

grade, and these were all considered to be high-grade. Fifty-five of 153 DCIS cases (36.0%) were rated identically for solid growth pattern. Intraductal calcifications were assessed identically in 72 of 153 DCIS cases (47.1%). Ten of 153 DCIS cases (6.5%) received identical scores for necrosis. Periductal stromal changes were rated identically in six of 153 DCIS cases (3.9%). No DCIS cases (0%) were rated similarly for stromal inflammation. All pathologists agreed that apocrine differentiation was lacking in 79 of 153 DCIS cases (51.6%), but disagreement existed for the other 74 cases (48.4%).

OVERALL INTEROBSERVER CONCORDANCE

Krippendorff's alpha was highest for necrosis [0.6885, 95% confidence interval (CI) 0.6315–0.7413], calcifications (0.6456, 95% CI 0.5356–0.7476), and stromal inflammation (0.6355, 95% CI 0.5807–0.6863). KA was lower for solid growth pattern (0.5935, 95% CI 0.4985–0.6815), nuclear grade (0.5629, 95% CI 0.4688–0.6523), and periductal stromal changes (0.4463, 95% CI 0.3635–0.5264). KA was very low for apocrine differentiation (0.2819, 95% CI −0.0841 to 0.5870), indicating poor concordance.

COMPARISON OF PATHOLOGIST DUOS

Krippendorff's alpha does not allow detailed analysis of interobserver variability among different participants. Therefore, Cohen's kappa was calculated for

each pathologist duo, which required 78 calculations for each histopathological feature (Tables 1, 2 and Tables S1–S12). All categorical variables were dichotomised with all possible cut-offs to enable comparison of the kappa distributions for each feature. The 'ideal' cut-off (i.e. the cut-off with the highest median interrater concordance and the most narrow distribution) was identified. No particular patterns of scoring by the participants were observed: no differences were noted between those with greater or fewer years in practice, or between those within the same practice.

Nuclear grade was dichotomised in two ways: first, intermediate grade was combined with low grade to form a 'non-high-grade' group; and second, intermediate grade was combined with high grade to form a 'non-low-grade' group. Non-high-grade versus high-grade dichotomisation resulted in higher concordance (median kappa of 0.526) and a narrower range than low-grade versus non-low-grade dichotomisation (median kappa of 0.394; Figure 4). Assessment of solid DCIS growth and the presence of calcifications resulted in median kappa values of 0.600 and 0.648, respectively, and both histopathological features showed narrow distributions of kappa values (Figure 4).

The four necrosis categories were dichotomised according to three different cut-offs: no necrosis versus any necrosis with a median kappa of 0.417; no necrosis or single-cell necrosis versus focal or extensive necrosis with a median kappa of 0.627; and non-extensive necrosis versus extensive necrosis with a median kappa of 0.587. The first and the last of

Table 1. Cohen's kappa values for each duo of pathologists, assessing nuclear grade as a two-tier feature, i.e. non-high nuclear grade versus high nuclear grade

|	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
P1													
P2	0.545												
P3	0.557	0.583											
P4	0.674	0.465	0.542										
P5	0.539	0.537	0.607	0.568									
P6	0.503	0.501	0.593	0.477	0.613								
P7	0.459	0.410	0.430	0.388	0.360	0.336							
P8	0.543	0.489	0.554	0.490	0.653	0.637	0.352						
P9	0.480	0.403	0.380	0.362	0.293	0.320	0.741	0.312					
P10	0.513	0.619	0.621	0.530	0.571	0.585	0.478	0.572	0.396				
P11	0.578	0.473	0.634	0.543	0.585	0.547	0.515	0.586	0.432	0.541			
P12	0.618	0.566	0.673	0.530	0.571	0.507	0.532	0.572	0.423	0.633	0.594		
P13	0.429	0.478	0.462	0.581	0.505	0.487	0.296	0.500	0.282	0.482	0.522	0.430	
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13

Table 2. Cohen's kappa values for each duo of pathologists, assessing nuclear grade as a two-tier feature, i.e. low nuclear grade versus non-low nuclear grade

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
P1													
P2	0.434												
P3	0.505	0.316											
P4	0.393	0.315	0.386										
P5	0.416	0.376	0.443	0.632									
P6	0.398	0.335	0.329	0.464	0.414								
P7	0.247	0.286	0.383	0.243	0.261	0.152							
P8	0.315	0.147	0.243	0.422	0.514	0.541	0.136						
P9	0.341	0.316	0.401	0.497	0.443	0.329	0.589	0.243					
P10	0.645	0.414	0.303	0.529	0.405	0.446	0.184	0.311	0.396				
P11	0.394	0.233	0.211	0.458	0.579	0.477	0.203	0.479	0.355	0.444			
P12	0.208	0.037	0.316	0.511	0.532	0.279	0.286	0.405	0.565	0.331	0.500		
P13	0.542	0.373	0.302	0.396	0.560	0.461	0.171	0.462	0.366	0.491	0.522	0.373	
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13

these dichotomisations resulted in a substantially broader kappa distribution than dichotomisation as no or single-cell necrosis versus focal or extensive necrosis. Assessment of apocrine differentiation resulted in low concordance (median kappa of 0.277).
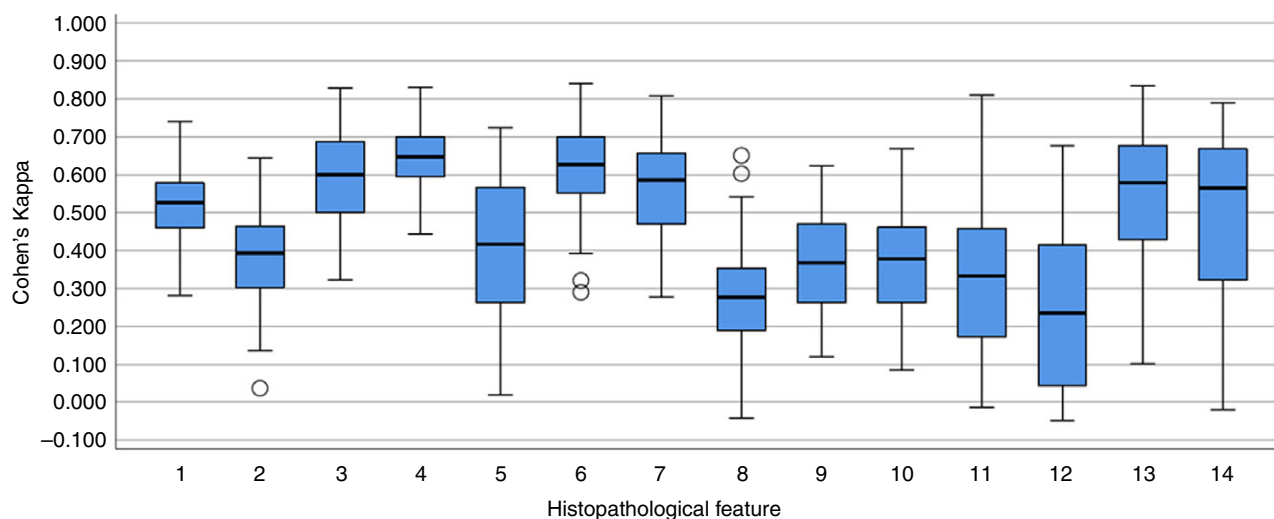
Figure 4. Box-and-whisker plots illustrating the distribution of Cohen's kappa values (*y*-axis) for each dichotomised histopathological feature (*x*-axis). The following histopathological features were assessed as two-tier variables: non-high nuclear grade versus high nuclear grade (1), low nuclear grade versus non-low nuclear grade (2), non-solid ductal carcinoma *in situ* (DCIS) growth pattern versus solid DCIS growth pattern (3), presence or absence of intraductal calcifications (4), no necrosis versus any necrosis (5), no or single-cell necrosis versus focal or extensive necrosis (6), non-extensive necrosis versus extensive necrosis (7), absence or presence of apocrine differentiation (8), <1% periductal myxoid stroma versus ≥1% periductal myxoid stroma (9), <33% periductal myxoid stroma versus ≥33% periductal myxoid stroma (10), <66% periductal myxoid stroma versus ≥66% periductal myxoid stroma (11), no stromal inflammation versus any stromal inflammation (12), no or mild stromal inflammation versus moderate or extensive stromal inflammation (13), and non-extensive stromal inflammation versus extensive stromal inflammation (14). [Colour figure can be viewed at wileyonlinelibrary.com]

The four categories of periductal stromal change were dichotomised according to three different cut-offs of 1%, 33% and 66% of ducts surrounded by myxoid periductal stroma, resulting in median kappa values of 0.368, 0.378, and 0.333, respectively. Similarly, stromal inflammation was dichotomised according to three different cut-offs: absent inflammation versus any inflammation with a median kappa of 0.235; absent or mild inflammation versus moderate to extensive inflammation with a median kappa of 0.579; and non-extensive inflammation versus extensive inflammation with a median kappa of 0.566 (Figure 4).

Discussion

Before the prognostic potential of a particular trait is investigated, the reproducibility of assessment should be examined. If a characteristic is deemed to be significantly associated with prognosis, its assessment should not be prone to high interobserver variability. As studies on prognostic markers in DCIS have often reported conflicting evidence, we wondered whether the different results might be due to substantial interobserver discordance. Therefore, we investigated interobserver variability among a group of 13 pathologists.

Percentage agreement among all observers was generally poor, varying from 0% for stromal inflammation to 51.6% for apocrine differentiation. Percentage agreement is a weak reliability measure, as it gives only approximate estimates of interobserver agreement without taking into account agreement due to chance. It is extremely influenced by the number of categories and the prevalence of a particular trait, which explains why percentage agreement was lower for features that were not assessed dichotomously, such as periductal stromal changes and necrosis. Apocrine differentiation was uncommon in this cohort, and was assessed dichotomously, which resulted in higher percentage agreement. Multicategorically assessed features with a more regular distribution (such as stromal inflammation and necrosis) automatically have lower percentage agreement. To overcome the weaknesses of percentage agreement, additional reliability measures were determined.

Krippendorff's alpha was selected as an overall reliability estimate, as it is used for analyses of subjective judgements, regardless of the number of raters, the number of rated cases, and the presence of missing data.[16] In the social sciences, a cut-off of KA ≥ 0.800 is generally required for data to be considered to be reliable.[19] Data are considered to be unreliable when KA is <0.667. With 0.800 > KA ≥ 0.667, it is

advised to draw only tentative conclusions.[19] In that respect, histopathological assessment would be considered to be unreliable, as KAs for all but one feature were <0.667. We would like to emphasise that such cut-offs are chosen arbitrarily. Interestingly, disagreement itself can provide valuable prognostic information, as discordant nuclear grades were shown to predict outcome in invasive breast cancer.[20,21]

Although KA is the correct reliability measure for categorical data assessed by multiple observers, it does not allow for investigation of mutual relationships among pathologists. Therefore, Cohen's kappa was calculated for all observer duos, and this resulted in 78 kappa values for each dichotomised feature. Degrees of reliability were compared between all features and their applied cut-offs.

So far, (dis)agreement on nuclear grade has been most intensively studied, and reported kappa values were generally low, varying from 0.27 to 0.49, depending on the classification system applied.[22–24] In this study, dichotomisation of nuclear grade as non-high versus high was more reproducible and thus more robust than dichotomisation as low versus non-low. Pathologists seem to have more difficulties with discerning low grade from intermediate grade than in discerning intermediate grade from high grade, which confirms previous findings.[25] By analogy with the two-tier system of low-grade squamous intraepithelial lesions and high-grade squamous intraepithelial lesions that replaced the three-tier system of cervical intraepithelial neoplasia,[26] it may be appropriate to also dichotomise nuclear atypia in DCIS. Two-tier grading systems have also replaced three-tier grading systems in other organ systems, such as the gastrointestinal tract.[27] In these different settings, better agreement among pathologists was obtained by implementing two-tier grading of dysplasia. Moreover, genomic and molecular studies have provided evidence for a low-grade pathway and a high-grade pathway in breast cancer development.[28–30] Interestingly, two-tier morphological grading of DCIS is corroborated by gene expression profiles, as molecular grading indicated a binary grading scheme for DCIS.[31] Unfortunately, it is impossible to investigate the interaction of features with one another, as all histopathological features are present (or absent) in one lesion. For instance, it is known that high-grade DCIS more often presents with necrosis, but it is impossible to investigate whether there is a true correlation or whether pathologists tend to assign a higher nuclear grade to DCIS with necrosis.

Comparison of median kappa values identified cut-offs with the lowest interobserver variability for the presence of necrosis, myxoid stromal changes, and stromal inflammation. Assessment of necrosis was most reliable with dichotomisation as no or single cell-necrosis versus focal or extensive necrosis. Previous studies provided contradictory evidence on the prognostic value of necrosis in DCIS, but methods of assessment differed.[9,32–34] It would be interesting to re-evaluate the prognostic potential of necrosis in a large DCIS cohort on the basis of the definition used in this study, as we have shown that this feature is assessed most reproducibly. Similarly, it would be worthwhile investigating the prognostic potential of dichotomous assessment of stromal inflammation (i.e. absent to mild inflammation versus moderate to extensive inflammation according to the aforementioned definitions), as the currently available reports provide contradictory evidence on the prognostic value of so-called tumour-infiltrating lymphocytes in DCIS.[9,35–37] Likewise, myxoid changes in the periductal stroma cannot be overlooked. We previously reported an association between myxoid stromal changes and recurrence rates in DCIS patients.[10] Here, we have determined a robust cut-off of 33% of ducts surrounded by myxoid changes, and we now aim to further explore the relationship between dichotomised periductal stromal changes and outcome in an independent cohort of DCIS patients. In addition, we aim to investigate whether upfront dichotomous assessment results in similar interobserver concordance as *post-hoc* dichotomisation.

Most studies on interobserver variability among pathologists have focused on the differentiation between atypical ductal hyperplasia (ADH), DCIS, and invasive cancer,[38–40] but only a few studies have been performed on agreement regarding histopathological features within DCIS, mainly focusing on nuclear grade and architectural patterns.[41–43] Here, we show that dichotomous histopathological assessment results in substantial interobserver concordance but depends on the chosen cut-off. We advocate the investigation of interobserver variability before examination of the prognostic potential of a particular trait, because prognostic markers are practically useless when they are not reproducible. In conclusion, it might be worthwhile considering new DCIS grading systems or prognostic indexes, based on two-tier assessment of several histopathological features. A potential disadvantage of two-tier assessment is the loss of information that might be clinically relevant, but we assume that two-tier assessment will be most robust when used in daily practice. First and foremost, this requires further investigation of the prognostic value of each dichotomously assessed

histopathological feature separately, which was beyond the scope of this study.

Acknowledgements

Conflicts of interest

Author contributions

Study design: M. Van Bockstal. Patient selection: M. Van Bockstal. Data collection: all contributors. Statistical analysis: M. Van Bockstal. Writing: M. Van Bockstal. All contributors reviewed and edited the manuscript, and approved its final version.

References

1. Yeong J, Thike AA, Tan PH, Iqbal J. Identifying progression predictors of breast ductal carcinoma in situ. *J. Clin. Pathol.* 2017; **70**; 102–108.
2. Francis A, Fallowfield L, Rea D. The LORIS trial: addressing overtreatment of ductal carcinoma in situ. *Clin. Oncol.* 2015; **27**; 6–8.
3. 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.
4. COMET study. Available at: https://dcisoptions.Org/comet (accessed 16 April 2018).
5. Gorringe KL, Fox SB. Ductal carcinoma in situ biology, biomarkers, and diagnosis. *Front. Oncol.* 2017; **7**; 248.
6. Lester SC, Bose S, Chen YY *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.
7. Consensus Conference on the classification of ductal carcinoma in situ. The Consensus Conference Committee. *Cancer* 1997; **80**; 1798–1802.
8. Schnitt SJ, Collins LC. *Biopsy interpretation of the breast.* Philadelphia, PA: Wolters Kluwer, Lippincott Williams & Wilkins, 2013.
9. 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.
10. Van Bockstal M, Lambein K, Gevaert O *et al.* Stromal architecture and periductal decorin are potential prognostic markers for ipsilateral locoregional recurrence in ductal carcinoma in situ of the breast. *Histopathology* 2013; **63**; 520–533.
11. Van Bockstal M, Libbrecht L, Floris G, Lambein K. The Baader-Meinhof phenomenon in ductal carcinoma in situ of the breast. *Histopathology* 2016; **69**; 522–523.
12. Van Bockstal M, Lambein K, Van Gele M *et al.* Differential regulation of extracellular matrix protein expression in carcinoma-associated fibroblasts by TGF-β1 regulates cancer cell spreading but not adhesion. *Oncoscience* 2014; **1**; 634–648.
13. Van Bockstal M, Lambein K, Denys H *et al.* Histopathological characterization of ductal carcinoma in situ (DCIS) of the breast according to HER2 amplification status and molecular subtype. *Virchows Arch.* 2014; **465**; 275–289.
14. Doebar SC, de Monye C, Stoop H, Rothbarth J, Willemsen SP, van Deurzen CH. Ductal carcinoma in situ diagnosed by breast needle biopsy: predictors of invasion in the excision specimen. *Breast* 2016; **27**; 15–21.
15. Durham JR, Fechner RE. The histologic spectrum of apocrine lesions of the breast. *Am. J. Clin. Pathol.* 2000; **113**; S3–S18.
16. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* 2007; **1**; 77–89.
17. Krippendorff K. Estimating the reliability, systematic error and random error of interval data. *Educ. Psychol. Meas.* 1970; **30**; 61–70.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**; 159–174.
19. Krippendorff K. Reliability in content analysis: some common misconceptions and recommendations. *Hum. Commun. Res.* 2004; **30**; 411–433.
20. Dalton LW, Gerds TA. The advantage of discordance: an example using the highly subjective nuclear grading of breast cancer. *Am. J. Surg. Pathol.* 2017; **41**; 1105–1111.
21. Rakha EA, Aleskandarany MA, Toss MS *et al.* Impact of breast cancer grade discordance on prediction of outcome. *Histopathology* 2018. https://doi.org/10.1111/his.13709
22. Schuh F, Biazus JV, Resetkova E, Benfica CZ, Edelweiss MI. 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.
23. 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.
24. Douglas-Jones AG, Morgan JM, Appleton MAC *et al.* Consistency in the observation of features used to classify duct carcinoma in situ (DCIS) of the breast. *J. Clin. Pathol.* 2000; **53**; 596–602.
25. 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.
26. Darragh TM, Colgan TJ, Cox JT *et al.* The Lower Anogenital Squamous Terminology Standardization Project for HPV-Associated Lesions: background and consensus recommendations from the College of American Pathologists and the American Society for Colposcopy and Cervical Pathology. *Arch. Pathol. Lab. Med.* 2012; **136**; 1266–1297.
27. Schlemper RJ, Riddell RH, Kato Y *et al.* The Vienna classification of gastrointestinal epithelial neoplasia. *Gut* 2000; **47**; 251–255.

28. Lopez-Garcia MA, Geyer FC, Lacroix-Triki M, Marchio C, Reis-Filho JS. Breast cancer precursors revisited: molecular features and progression pathways. *Histopathology* 2010; **57**; 171–192.

29. Pang JM, Gorringe KL, Fox SB. Ductal carcinoma in situ – update on risk assessment and management. *Histopathology* 2016; **68**; 96–109.

30. Hannemann J, Velds A, Halfwerk JB, Kreike B, Peterse JL, van de Vijver MJ. Classification of ductal carcinoma in situ by gene expression profiling. *Breast Cancer Res.* 2006; **8**; R61.

31. Balleine RL, Webster LR, Davis S *et al.* Molecular grading of ductal carcinoma in situ of the breast. *Clin. Cancer Res.* 2008; **14**; 8244–8252.

32. Solin LJ, Gray R, Baehner FL *et al.* A multigene expression assay to predict local recurrence risk for ductal carcinoma in situ of the breast. *J. Natl Cancer Inst.* 2013; **105**; 701–710.

33. Kong I, Narod SA, Taylor C *et al.* Age at diagnosis predicts local recurrence in women treated with breast-conserving surgery and postoperative radiation therapy for ductal carcinoma in situ: a population-based outcomes analysis. *Curr. Oncol.* 2014; **21**; e96–e104.

34. Rakovitch E, Pignol JP, Hanna W *et al.* Significance of multifocality in ductal carcinoma in situ: outcomes of women treated with breast-conserving therapy. *J. Clin. Oncol.* 2007; **25**; 5591–5596.

35. Pruneri G, Lazzeroni M, Bagnardi V *et al.* The prevalence and clinical relevance of tumor-infiltrating lymphocytes (TILs) in ductal carcinoma in situ of the breast. *Ann. Oncol.* 2017; **28**; 321–328.

36. Toss MS, Miligy I, Al-Kawaz A *et al.* Prognostic significance of tumor-infiltrating lymphocytes in ductal carcinoma in situ of the breast. *Mod. Pathol.* 2018; **31**; 1226–1236.

37. Van Bockstal M, Libbrecht L, Floris G, Lambein K, Pinder S. Stromal inflammation, necrosis and HER2 overexpression in ductal carcinoma in situ of the breast: another causality dilemma? *Ann. Oncol.* 2017; **28**; 2317.

38. Elmore JG, Longton GM, Carney PA *et al.* Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* 2015; **313**; 1122–1132.

39. Gomes DS, Porto SS, Balabram D, Gobbi H. Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma in situ of the breast. *Diagn. Pathol.* 2014; **9**; 121.

40. Jain RK, Mehta R, Dimitrov R *et al.* Atypical ductal hyperplasia: interobserver and intraobserver variability. *Mod. Pathol.* 2011; **24**; 917–923.

41. Sloane JP, Amendoeira I, Apostolikas N *et al.* Consistency achieved by 23 European pathologists in categorizing ductal carcinoma in situ of the breast using five classifications. European Commission Working Group on Breast Screening Pathology. *Hum. Pathol.* 1998; **29**; 1056–1062.

42. Sneige N, Lagios MD, Schwarting R *et al.* Interobserver reproducibility of the Lagios nuclear grading system for ductal carcinoma in situ. *Hum. Pathol.* 1999; **30**; 257–262.

43. Bethwaite P, Smith N, Delahunt B, Kenwright D. Reproducibility of new classification schemes for the pathology of ductal carcinoma in situ of the breast. *J. Clin. Pathol.* 1998; **51**; 450–454.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Cohen's kappa values for each duo of pathologists, assessing solid growth of DCIS as a two-tier feature, i.e. predominantly solid versus predominantly non-solid.

Table S2. Cohen's kappa values for each duo of pathologists, assessing the presence of necrosis as a two-tier feature, i.e. no necrosis versus any amount of necrosis (either single-cell, focal or extensive necrosis).

Table S3. Cohen's kappa values for each duo of pathologists, assessing the presence of necrosis as a two-tier feature, i.e. no necrosis or single-cell necrosis versus focal necrosis or extensive necrosis.

Table S4. Cohen's kappa values for each duo of pathologists, assessing the presence of necrosis as a two-tier feature, i.e. non-extensive necrosis (either no necrosis or single-cell necrosis or focal necrosis) versus extensive necrosis.

Table S5. Cohen's kappa values for each duo of pathologists, assessing intraductal calcifications as a two-tier feature, i.e. absent versus present.

Table S6. Cohen's kappa values for each duo of pathologists, assessing the presence or absence of apocrine differentiation as a two-tier feature.

Table S7. Cohen's kappa values for each duo of pathologists, assessing periductal stromal changes as a two-tier feature, i.e. <1% periductal myxoid stroma versus ≥1% periductal myxoid stroma.

Table S8. Cohen's kappa values for each duo of pathologists, assessing periductal stromal changes as a two-tier feature, i.e. <33% of ducts surrounded by myxoid stroma versus ≥33% of ducts surrounded by myxoid stroma.

Table S9. Cohen's kappa values for each duo of pathologists, assessing periductal stromal changes as a two-tier feature, i.e. <66% periductal myxoid stroma versus ≥66% periductal myxoid stroma.

Table S10. Cohen's kappa values for each duo of pathologists, assessing stromal inflammation as a two-tier feature, i.e. no periductal stromal inflammation versus any periductal stromal inflammation (either mild, moderate, or extensive).

Table S11. Cohen's kappa values for each duo of pathologists, assessing periductal stromal inflammation as a two-tier feature, i.e. absent or mild stromal inflammation versus moderate or extensive stromal inflammation.

Table S12. Cohen's kappa values for each duo of pathologists, assessing stromal inflammation as a two-tier feature, i.e. non-extensive periductal stromal inflammation (no inflammation, or mild or moderate inflammation) versus extensive periductal stromal inflammation.