Impact of Breast Cancer Grade Discordance on Prediction of Outcome

Emad A Rakha[1,4], Mohammed A Aleskandarany[1,4], Michael S Toss[1], Nigel Mongan[2], Maysa E ElSayed[4], Andrew R Green[1], Ian O Ellis[1] and Leslie W Dalton[3]

[1] Division of Cancer and Stem Cells, School of Medicine, University of Nottingham, Nottingham City Hospital, Hucknall Road, Nottingham NG5 1PB, UK

[2]Faculty of Medicine & Health Sciences, University of Nottingham, Sutton Bonington Campus, Leicestershire, LE12 5RD

[3]Department of Histopathology, South Austin Hospital, Austin, TX USA

[4]Faculty of Medicine, Menoufyia University, Egypt

Correspondence:
Professor Emad Rakha
Department of Histopathology,
Nottingham University Hospital NHS Trust,
City Hospital Campus, Hucknall Road, Nottingham,
NG5 1PB, UK
Tel: (44) 0115-9691169
Fax: (44) 0115- 9627768
Email: Emad.rakha@nuh.nhs.uk
emadrakha@yahoo.com

Keywords: Breast pathology, grade, agreement, discordance, outcome

Running Title: Grade discordance and prognosis

ABSTRACT

BACKGROUND: Histological grade is an independent prognostic variable in breast cancer (BC). Previous concordance studies of BC grade have reported moderate levels of agreement; a typical finding in morphological assessment of biological variables. This study aims at investigating the impact of discordance on the prognostic value of grade and identifying the best reporting approach in borderline cases. METHODS: A large (n=1675) well-characterised annotated cohort of BC originally graded in routine practice using glass slides was re-graded twice, by an expert breast pathologist using virtual microscopy with a three months washout period. Outcome was assessed using breast cancer specific and distant metastasis free survival (median follow-up =135 months). RESULTS: 58% of the cases showed absolute agreement in the three separate grading sessions whereas grade 1/2 and grade 2/3 discordance were observed in 21% and 21% respectively. Absolute intra-observer agreement using virtual microscopy was observed in 77% of the cases whereas 13% and 10% showed grade 1/2 and grade 2/3 discordance respectively. Despite the concordance, outcome analysis revealed significant associations between tumour grade and patients' outcome in the three grading sessions. Grade 1/2 and grade 2/3 discordant cases showed intermediate survival between grade 1 and grade 2 tumours and grade 2 and grade 3 tumours, respectively. Grade 1/2 discordant cases showed a worse outcome when compared with grade 1 tumours ($p=0.008$) but no statistical difference was identified when compared with grade 2 tumours. Similarly, grade 2/3 discordant cases showed a significant difference from grade 2 tumours ($p<0.001$) but no statistical difference was identified when compared with grade 3 tumours. CONCLUSIONS: BC grade discordance is likely a reflection of biologically, and hence morphologically, borderline tumours. Cases with borderline features for grade are more likely to behave similar to the higher grade category. Repeating histological grade of borderline cases or double reporting may improve correlation with

outcome.

INTRODUCTION

Histological grade of breast cancer (BC) is one of the strongest prognostic factors in early stage disease [1-3]. Histological grade, using the Nottingham grading system comprises one of the main components of several management decision tools [4-8] and it has recently been included in the American Joint Committee on Cancer (*AJCC*) TNM staging system as a stage modifier [9]. A concern in regard to BC grading is the subjective nature of the technique with subsequent variation among pathologists in the assignment of all tumours into the same grades [10]. A supposed advantage of modern era techniques, such as molecular biomarkers, is the high objectivity with a corresponding increase in reproducibility. However, in reality this perceived potential has yet to be realised as concordance of modern era molecular assays has not shown any improved agreement compared to human eye histological grading [11, 12].

A distinct advantage of grading in addition to the low cost and short assessment time, is the relative ease in obtaining multiple opinions. From multiple opinions, discordance in grade assignments will most certainly arise. The most likely reflex for the resultant discordance is to be considered as a disadvantage. This is only true if discordance discovery offers no useful information or just reflects poor performance of the reader. However, if a particular case is susceptible to discordance in grade assignment resulting from borderline morphological features it may reflect the biology of the tumour and its eventual behaviour.

Increasing emphasis is being placed upon obtaining second or multiple opinions and with increasing use of digital pathology [13-17], the number of second opinions is likely to further increase. Yet, it is not well understood how discordant grade assignments might impact risk assignment. Knowledge of this might guide the methodology of how to integrate multiple opinions into quality assurance programmes, education, interpretation of research results, and into improved patient care.

In a previous study we assessed the level of inter-observer and intra-observed concordance of BC grading based on virtual microscopy (VM) as compared to the original glass-slides based grading [18] and this showed high concordance levels and demonstrated the reliability of VM in BC grading. In this study, the impact of grade assignment discordance on patients' outcome is investigated along with outlining practical guidelines on how to handle discordance.

PATIENTS AND METHODS

This study has been performed on a large series (n=1675) of early stage invasive primary BC patients presented to Nottingham City Hospital from 1999-2006. This is a well-characterised cohort of breast cancer (BC) with long-term clinical follow-up (median =135 months) and detailed clinicopathological profiles. Data included primary tumour histologic grade and grade components, tumour size and histotype, lymph node stage, lymphovascular invasion, Nottingham Prognostic Index (NPI) and oestrogen receptor (ER) status. Patients' outcome information was collected and prospectively maintained. The latter include BC-specific survival (BCSS), defined as time (in months) from the date the primary surgical treatment to the time of death from BC, and distant metastasis free survival (DMFS) was defined as the time (in months) from primary surgery until the first event of distant metastasis. Patient and tumour demographics are summarised in Table 1.

This tumour cohort was originally graded using the Nottingham grading system during routine pathology reporting using light microscopy (LM) and utilising all available tumour slides for each case (average four slides per case) [2]. For the purpose of this study, data for the final grade as well as the individual grade components (tubule formation, nuclear pleomorphism and mitotic count scores) was retrieved from the patients' records. In addition, 1-3 tumour blocks per case were retrieved and freshly prepared H&E slides were reviewed. A

representative slide per case was selected by a specialised breast pathologist (EA Rakha). The slides were selected based on the presence of adequate invasive tumour sufficient for VM grading regardless of the grade of tumour tissue in the selected slide. Slides with artefacts, which would potentially interfere with image quality or grading, were excluded. Selected slides were scanned into high-resolution (0.19 μm/pixel) digital images at 20x magnification using 3D Histech Panoramic 250 Flash II scanner (3DHISTECH Ltd., Budapest, Hungary). Whole slide digital images (WSI) were generated, stored and viewed using the 3DHistech Pannoramic Viewer (3DHISTECH Ltd., Budapest, Hungary; http://www.3dhistech.com/downloads) on a high-resolution screen. For virtual microscopy (VM) grading, digital images were initially examined at low magnification where tubule formation was assessed. Also, low to intermediate magnification was performed for the identification of potential "hotspots" for mitotic counting. For mitotic counting, the distance measure tool of the software was used. This was important for determining the number of mitotic figures in a given area.

To allow for intra-observer agreement assessment of BC grading using WSI, the whole cohort was graded again using the same criteria by the same observer (L Dalton who is an experienced breast pathologist with special interest in BC grading and digital microscopy). The second grading session was performed after a 3-month washout time without further training. In both WSI grading sessions (V1 and V2), grade components were assigned blinded to the original LM grade as well as other clinicopathological parameters.

This study was approved by Nottingham Research Ethics Committee 2 under the title of "Development of a molecular genetic classification of breast cancer".

Survival analysis

Survival analysis was performed using SPSS 23 (SPSS 23 for Windows, Chicago, IL, USA) using log rank test and Kaplan Meier plots. Much reliance was placed on simple inspection of

survival curves [19]. Survival analysis included separate determinations of BCSS and DMFS. The baseline grade assignment was the originally performed Nottingham grade by LM. From this baseline, two additional reviews generated by VM grading resulted in discordant assessments. Survival curves were constructed which tracked the survival associated with concordance/discordance. Statistical significance in survival stratification was calculated by the log-rank method and univariate cox regression analysis. A p-value of less than 0.05 (two tailed) was considered significant.

RESULTS

In this study, two VM grading sessions were performed by an expert breast pathologist for a large (n=1675) clinically annotated early-stage primary invasive BC with a three-months washout period. 58% of the cases showed absolute agreement in all three grading (original LM grade and 2 VM grade) sessions (13%, 21% and 24% for grades 1, 2 and 3, respectively) whereas grade 1/2 and grade 2/3 discordance were observed in 21% and 21%, respectively (Table 2c). High/low discordance was uncommon and occurred in only 26 cases (1.6%). The intra-observer agreement between the two VM sessions was 77%, whereas 13% and 10% showed grade 1/2 and grade 2/3 discordance, respectively (Table 2a). Only six cases (0.3%) had high/low discordance as assigned by one observer using VM. Figures 1-3 illustrate examples of concordant and discordant grades and example of difficulty in the interpretation of mitotic figures at VM.

Based on the original assessment, grade 2 tumours totalled 683 (41%) of the whole cohort. After VM1, the number of cases remained as grade 2 (i.e. in the intermediate category) was 420 (25%) cases; VM1 has resulted in shifting of some grade 2 tumours into grade 1 (n=215) or 3 (n-48) tumours. Table 2 shows that VM tends to down grade tumours (p=1.0x10^{-13}) with more cases assigned to lower-grade than the higher-grade categories. We assumed that the

experience with digital microscopy is the reasons in the first session. However, the same observation was identified in the second session, which may reflect the relatively reduced ability to identify mitotic counts on the screen. To avoid the confounding effect of the platform on the concordance, we analysed the impact on outcome using the two VM sessions and by one observer as well as the original LM grade assigned by different observer.

Figure 4 shows the survival curve for the originally assigned grade, and for the first, and most naive, of the two VM sessions. The 342 discordant grade 1/2 tumours in the 3 grading sessions (Table 2c) showed a relatively favourable outcome compared to grade 2 tumours over the short-term follow-up. However, long-term outcome analysis revealed survival figures concordant with grade 2 tumours (Table 3). At the opposite end of the spectrum, concordant high grade tumours were associated with the worst patient outcome (Table 3). The 276 discordant grade 2/3 tumours showed relatively better outcome compared to concordant grade 3 BC during the early follow-up times however; this meagre improvement disappeared after longer-term follow-up and the final outcome of these grade 2/3 discordant cases was similar to grade 3 concordant tumours.

To test for how the alteration of the original grade might be impacted by discordance/ concordance of the additional reviews using WSI, concordance of VM1 and VM2 was explored. Comparison of the two VM grading sessions showed a smaller number of discordant assessments reflecting high level of intra-observer concordance. The outcome of discordance as related to the 2 VM grade assignment sessions is outlined in figure 5. Again, the discordant assignments corresponded to interval levels of patients' survival. In addition, it also demonstrated the existence of "solid" or repeatable intermediate grade tumours. Survival curves in figure 6 display these results. A repeatable intermediate grade assignment is more aligned with intermediate survival. Meanwhile grade 2/3 discordant cases in the 2 VM sessions were more aligned with the original high-grade assignment. Figure 7 allows

visualisation at the grade 1/2 end of the spectrum.

When the cohort was stratified into oestrogen receptor positive (ER+) and negative (ER-) subgroups it was in the ER+ subset where interval levels of survival corresponded to discordant grade assignment. In the ER- negative group no statistical significance was found in survival of grade 2, grade 2/3 or grade 3 cases. Table 4 shows the distribution of ER status among the five concordance/discordance levels. Also, the distribution of discordance/ concordance levels seen in patients of younger age is listed. Less than 46 years was chosen given the high probability that patients under this age are pre-menopausal[20].

DISCUSSION

In the seminal paper on assessment of histological grade [21] each tumour was graded independently by two observers. In those tumours having had discordant grade assignment, the observers resolved the matter by joint review at a dual-headed microscope. Therefore, at the outset grading was accomplished by the review of two pathologists. Strictly speaking, the procedure used in the original validation study, should be the procedure used going forward. Of course, since then the practice of single pathologist review is common, and many datasets have shown significance of grade based on a single ~~revision~~ review. With the increasing expectations for outside-institutional second review, and with the advent of digital microscopy, discordance will be encountered, or “discovered” more frequently among different pathologists. Therefore, the current investigation is partly a matter of necessity. Especially since those rendering second opinions may not have any incentive to arrive at a collegial joint decision. Subsequently, questions might arise as to: who might be correct or who might be wrong. There may be no right or wrong if discordant assessments belong into separate and potentially informative categories. The ultimate aim was to test whether discordant assessments can be allocated into separate and potentially informative categories. In other

words, to explore the hypothesis that grade discordance is a biological rather than a pure technical phenomenon.

Based on the findings herein, all opinions may be correct bearing in mind the nature of a cancer itself when it expresses borderline attributes, be they phenotypic, genotypic, or proteomic, therefore raising the susceptibility to discordance in risk assignment. In other terms, discordance may not be resulting from observers' faults, whether it is a man or a machine, but an inherent cancer trait. If tumours with discordant grade assignments are linked to a robust patient outcome data, impact of discordance could be interrogated whether it affects patients' risk stratification and hence management. The ~~more~~ advantage of our approach is having discordance discovery become procedural or, at least, it is worthwhile to expand our knowledge as to the meaning of discordance.

An advantage of modern genomic and molecular techniques is their potential for higher objectivity with a corresponding increase in reproducibility as compared to ~~the known~~ subjective human eye histological grading. However, concordance of modern assays is showing no or marginal agreement [11, 12, 22-24]. Because of technical ease, low cost, and in that grading does not consume additional tissue, grading could be considered to hold a unique advantage to molecular techniques. Furthermore, with grading, discordance/concordance discovery is feasible. It is dubious that discordance in risk assignments, both in morphological parameters or molecular biomarkers, will be completely eliminated. To our knowledge this is the largest study of its kind with the approach followed in this report serving as an illustrative start point.

The findings here, and prior work [25] contravene conventional wisdom. As for two separate opinions, concordance/discordance discovery constructed a risk scale with five categories. The originally assigned grade was, of course, a three-category scheme. A five-category risk scale affords more flexibility in deciding patient treatment strategies. For instance, if those

patients with ER positive tumours show concordance of high grade assignment, it may be deemed reasonable to assume that the patient is located at the definite higher risk end of the spectrum related to tumour grade. In other terms, the agreement between two assignments of high grade can increase pathologists' confidence that the tumour ~~being~~ is a real risk to patient survival. The opposite recommendation is applicable to concordant low-grade ER positive tumours, which in this case are better treated with hormone manipulation without chemotherapy. Although the five-tier system is more reflective of tumour biology and provides detailed representation of BC heterogeneity and more accurate patients' risk stratification, using five categories in routine practice could also be associated with its own disadvantages. There is a tendency to apply prognostic variables in a dichotomised fashion to allow further management of patients in terms of systemic therapy. Regardless of a three-tier or five-tier system, oncologists tend to translate data into a binary variable to decide further management options making the three-tier system more pragmatic. Also, the five-tier system requires that all BC be double graded which has time and cost implications. Importantly, Nottingham grading as a ternary scheme has been so well validated it is not advisable to adjust to a five-tier scheme without further study. Favoured is to simply note in a report that grading in a given case has been based on consensus review. Findings of the current study would suggest, that until proven otherwise, to assign adjacent level discordance into the higher grade. High versus low grade discordance should certainly be subjected to thorough scrutiny.

The results of this study demonstrate the association between grade discordance and outcome, which we interpreted as a reflection of tumour biology and hence the differences in the outcome. Concordant grade 1 cases, the lowest risk group in our five-category risk scale, appears to represent the very well differentiated cancers at one end of the differentiation continuum, while concordant grade 3 cases were the least differentiated at the other end. This study highlights the importance of inter-tumour heterogeneity of BC and that some tumours

show borderline molecular features [26], and hence borderline morphological characteristics, making tumour assignment into a specific grade category subjective and challenging. These tumours comprise the majority of grade discordant cases as demonstrated by the association with distinct outcome in-between the two concordant grade cases. Intra-tumour heterogeneity may also contribute to grade discordance in research studies, including this study; if different slides are used in grading by different observers. This may explain discordance in few cases in this study in which the original grade was assigned based on examination of 4 tumour slides whereas the virtual grade assignment was based on one slide that represents part of the tumour.

In view of outcome analysis in this study linking tumour biology to grading assignment, the impact on pathology practice is twofold. Firstly, BC showing grade discordance between reporting pathologists are likely to eventually behave in a way similar to the higher-grade category and are likely to have high risk than the assigned lower grade. Thus, if more opinion is sought, the higher the grade assignment, the higher the risk the tumour may have. Secondly, some tumours will be assigned to grade 2 category regardless of the number of reviews indicating that grade 2 BC is genuine intermediate grade along the risk scale and not just a basket for lack of assignment of cases. Using molecular assays to assign BC into two grades may not be an optimal approach for risk stratification of individual tumours especially intermediate risk cases [26, 27]. Using other prognostic variables in these cases to determine BC outcome and behaviour is warranted rather than assigning these intermediate grade prognostically borderline tumours into one of the extreme end categories.

One caveat pertains to grade one versus two discordance. As seen by survival curve inspection, in the short term (60 months) low/intermediate track with low grade. It is over a longer term where grade 1 versus grade 2 discordance inclines toward intermediate. The short-term behaviour may influence the decision to avoid treatments which are aimed at short

term response, especially if a treatment cut-point has been set for high sensitivity.

Until there is further validation no formal rules are proposed based on findings of this study. Instead, we would offer two recommendations. Firstly, to maintain the original procedure described by Elston and Ellis, and if discordance is discovered, then resolve discordance by collegial peer review. Should discordance occur beyond a peer setting, it cannot be assumed that the original opinion deserves the label of mistake or error. Instead, the difference may best be attributed to the inherent nature of the tumour itself. Secondly and in view of the time and cost limitations, it is suggested that it is useful to review tumours with borderline features by the same pathologist (after a time interval) or by a different pathologist. Based on practice experience, pathologists are usually aware if they are having some difficulty in deciding between low and intermediate grade (i.e., score five versus score six tumour) or intermediate and high grade (score seven versus eight) tumours. These are the cases that could potentionally be subject to a second opinion. Audits of grade in routine practice can also help in identifying the proportions and features of tumours reported by different pathologists that should be submitted for a second opinion. The findings here can help guide how to resolve the discordance.

An additional recommendation pertains to research studies. If dual (or more) pathologist review is performed, then level of pathologist reproducibility must be assessed at level of consensus before grading can be criticised for lack of reproducibility. Decades ago, it was shown that consensus opinion among groups had higher reproducibility than individual opinions, and consensus opinion corrected for outliers [28].

In this study, whilst a single histopathologist provided the two additional grade assessments and the observations might be strengthened by assessments of additional histopathologists, the large number of cases in this study and the ability to correlate intra-observer concordance with outcome reinforce the value of the current study. Also grade assignments had been

rendered on different microscopy platforms; namely LM and VM, New sentence? in routine practice the additional opinions are increasingly obtained using digital microscopy of WSI.

Moreover, it may seem that attributing interobserver (LM vs VM grading) and intra-observer (VM1 and VM2 grading) to differences to intrinsic biology of the tumour may inadvertently reduce the importance of achieving grading consistency by different observers, However, we would like to emphasise that this phenomenon is a typical feature of biological processes particularly those assessed based on morphological characteristics, such as tumour differentiation by BC histological grading. Importantly discordance was limited to certain tumours whereas the majority of the tumours were consistently assigned to specific grade category. These discordant tumours also showed distinct outcome and their identification can help refining risk stratification of patients. It is also important to highlight that the results of this study refer to discordance of grade between expert pathologists, which is mainly related to intrinsic tumour features and not related to a difference in the application of grade methodology or inability of individual pathologists to consistently assign the “correct” grade.

In this study, we also noticed that mitotic figure recognition is not optimal on VM slides. As VM is a relatively emerging procedure, more practice and comparing the morphology of mitotic figures in LM and VM will help to establish the criteria and experience to identify mitotic figures with reproducible accuracy. A study to improve our ability to identify mitotic figures and differentiate them from apoptotic bodies using high-resolution and high-definition digital images, using Z-stacking image technology and immunohistochemistry for staining of mitotic cells is also proposed.

From a future perspective, the VM grade represents a realistic platform as the use of digital microscopy is currently expanding making the second review accomplishable. Further investigation of the findings of the current study could be achieved by integration of VM grading/second opinion into QA and/or educational programmes. The involvement of

practising pathologists would test, in real practice, the concordance levels between observers/ graders as grading of this cohort has been performed by expert breast pathologists. Moreover, the integration of this VM grading into educational programmes could help accomplish training tasks and to audit trainees' performance compared to experts.

REFERENCES

1. 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.

2. Rakha EA, El-Sayed ME, Lee AH *et al.* Prognostic significance of nottingham histologic grade in invasive breast carcinoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2008;**26**;3153-3158.

3. Rakha EA, Reis-Filho JS, Baehner F *et al.* Breast cancer prognostic classification in the molecular era: The role of histological grade. *Breast Cancer Res* 2010;**12**;207.

4. Ellis IO, Rakha EA, Lee AHS *et al. Pathology reporting of breast disease in surgical excision specimens incorporating the dataset for histological reporting of breast cancer*. London: The Royal College of Pathologists, June 2016;160.

5. Candido Dos Reis FJ, Wishart GC, Dicks EM *et al.* An updated predict breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Res* 2017;**19**;58.

6. Lakhani SR, I.O. E, Schnitt SJ, Tan PH, van de Vijver MJ eds. *Who classification of tumours of the breast*. IARC press, Lyon 2012.

7. Senkus E, Kyriakides S, Ohno S *et al.* Primary breast cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2015;**26 Suppl 5**;v8-v30.

8. *College of american pathologists. Protocol for the examination of specimens from patients with invasive carcinoma of the breast based on ajcc/uicc tnm*. 7th ed, June 2012.

9. Giuliano AE, Connolly JL, Edge SB *et al.* Breast cancer-major changes in the american joint committee on cancer eighth edition cancer staging manual. *CA Cancer J Clin* 2017;**67**;290-303.

10. Rakha EA, Ahmed MA, Aleskandarany MA *et al.* Diagnostic concordance of breast pathologists: Lessons from the national health service breast screening programme pathology external quality assurance scheme. *Histopathology* 2017;**70**;632-642.

11. Bartlett JM, Bayani J, Marshall A *et al.* Comparing breast cancer multiparameter tests in the optima prelim trial: No test is more equal than the others. *J Natl Cancer Inst* 2016;**108**.

12. Fan C, Oh DS, Wessels L *et al.* Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med* 2006;**355**;560-569.

13. Allen TC. Digital pathology and federalism. *Arch Pathol Lab Med* 2014;**138**;162-165.

14. Brachtel E, Yagi Y. Digital imaging in pathology--current applications and challenges. *Journal of biophotonics* 2012;**5**;327-335.

15. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: Current status and future perspectives. *Histopathology* 2012;**61**;1-9.

16. Volynskaya Z, Evans AJ, Asa SL. Clinical applications of whole-slide imaging in anatomic pathology. *Adv Anat Pathol* 2017;**24**;215-221.

17. Morrison AO, Gardner JM. Microscopic image photography techniques of the past, present, and future. *Arch Pathol Lab Med* 2015;**139**;1558-1564.

18. Rakha EA, Aleskandarani M, Toss MS *et al.* Breast cancer histologic grading using digital microscopy: Concordance and outcome association. *Journal of clinical pathology* 2018.

19. Hernan MA. The hazards of hazard ratios. *Epidemiology* 2010;**21**;13-15.

20. McKinlay SM, Brambilla DJ, Posner JG. The normal menopause transition. *Maturitas* 1992;**14**;103-115.

21. 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.

22. Meattini I, Bicchierai G, Saieva C *et al.* Impact of molecular subtypes classification concordance between preoperative core needle biopsy and surgical specimen on early breast cancer management: Single-institution experience and review of published literature. *Eur J Surg Oncol* 2017;**43**;642-648.

23. Polley MY, Leung SC, Gao D *et al.* An international study to increase concordance in ki67 scoring. *Mod Pathol* 2015;**28**;778-786.

24. Stalhammar G, Rosin G, Fredriksson I, Bergh J, Hartman J. Low concordance of biomarkers in histopathological and cytological material from breast cancer. *Histopathology* 2014;**64**;971-980.

25. 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.

26. Sotiriou C, Wirapati P, Loi S *et al.* Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006;**98**;262-272.

27. Metzger Filho O, Ignatiadis M, Sotiriou C. Genomic grade index: An important tool for assessing breast cancer tumor grade and prognosis. *Crit Rev Oncol Hematol* 2011;**77**;20-29.

28. Dalton LW, Page DL, Dupont WD. Histologic grading of breast carcinoma. A reproducibility study. *Cancer* 1994;**73**;2765-2770.

Acknowledgement
We thank the Nottingham Health Science Biobank and Breast Cancer Now Tissue Bank for the provision of tissue samples.

Tables

Table 1: Characteristics of the cohort of invasive breast cancer included in this study

Parameters	**Number of cases (%)**
Age	
>50	1098 (65.6)
≤50	549 (32.8)
Missing	28 (1.7)
Tumour size	
> 2.0cm	588 (35.1)
≤2.0cm	1058 (63.2)
Missing	29 (1.7)
Lympho-vascular Invasion	
Negative	1197 (71.5)
Positive	450 (26.9)
Missing	28 (1.7)
Lymph node status	
Negative	1132 (67.6)
Positive	515 (30.7)
Missing	28 (1.7)
Lymph Node Stage	
1	1027 (62.4)
2	457 (27.3)
3	162 (9.7)
Missing	29 (1.7)
Nottingham Prognostic Index	
Good	568 (33.9)
Moderate	820 (49)
Poor	256 (15.3)
Missing	31 (1.9)
Histologic types	
Ductal NST	1258 (75.1)
Lobular	102 (6.1)
Tubular/Invasive Cribriform	60 (3.6)
Pure Mucinous	22 (1.3)
Invasive Micropapillary	13 (0.8)
Other types including Medullary-like	220 (13.1)

Distant metastasis	
Yes	357 (21.3)
No	1288 (76.9)
Missing	30 (1.6)
Outcome Status at end of follow-up	
Alive	1190 (71)
Died from Breast cancer	297 (17.7)
Died from other causes	156 (9.3)
Missing	32 (1.9)

Table 2: Cross comparison of the two virtual microscopy grading sessions (VM1 and VM2) (Table 2a) and between the light microscopy (LM) grading and both virtual grading sessions (Table 2b) and between the 3 grading sessions (Table 2c)

Table 2a

Grade VM1	**Grade VM2**			**Total percentage**
	Grade 1	**Grade 2**	**Grade 3**	
Grade 1	363	107	2	28.2
Grade 2	101	504	67	40.1
Grade 3	4	106	421	31.7
Total Percentage	27.9	42.8	29.3	100
Percent exact agreement: 77%, Percent adjacent level: 22.7%, Percent high/low: 0.3%				

Table 2b

Grade (VM1 and 2)	**Grade (Light Microscopy)**			Total percentage
	Grade 1	**Grade 2**	**Grade 3**	
Grade 1 VM1	232	215	25	28.2
VM2	233	210	25	27.9
Grade 2 VM1	39	420	213	40.1
VM2	37	440	240	42.8
Grade 3 VM1	1	48	482	37.1
VM2	2	33	455	29.3
Total Percentage	16.2	40.8	43.0	100
Percent exact agreement: 68%, Percent adjacent level: 30.5%, Percent high/low: 1.5%				

Table 2c

Grade (Light Microscope)	Grade VM1	Grade VM2			Total percentage
		Grade 1	**Grade 2**	**Grade 3**	
Grade 1	**Grade 1** **Grade 2** **Grade 3**	212 21 0	19 17 1	1 1 0	28.2
Grade 2	**Grade 1** **Grade 2** **Grade 3**	138 71 1	76 335 29	1 14 18	40.1
Grade 3	**Grade 1** **Grade 2** **Grade 3**	13 9 3	12 152 76	0 52 403	37.1
Total Percentage		27.9	42.8	29.3	100
Percent exact agreement: 58%, Percent adjacent level: 41.2%, Percent high/low: 0.8%					

Table 3: Probability of patient survival (Life table analysis) corresponding to concordance and discordance of originally assigned grade with the two VM additional reviews (VM1 and VM2).

Grade	Interval Start Time (months)	Number Entering Interval	Number Exposed to Risk	Number of Terminal Events	Proportion Surviving	Cumulative Proportion Surviving at End of Interval
Grade 1	0	205	205	0	1.00	1.00
	30	202	200	2	0.99	0.99
	60	197	189	3	0.98	0.97
	90	178	166	2	0.99	0.96
	120	153	129	1	0.99	0.96
	150	104	68	3	0.96	0.91
Grade 1/2	0	334	332	1	1.00	1.00
	30	329	322	5	0.98	0.98
	60	309	296	3	0.99	0.97
	90	279	261	9	0.97	0.94
	120	233	188	8	0.96	0.90
	150	134	95	5	0.95	0.85
Grade 2	0	329	325	5	0.98	0.98
	30	316	310	8	0.97	0.96
	60	296	280	10	0.96	0.93
	90	255	239	5	0.98	0.91
	120	218	175	8	0.95	0.86
	150	125	91	4	0.96	0.83
Grade 2/3	0	335	332	13	0.96	0.96
	30	315	311	24	0.92	0.89
	60	283	270	18	0.93	0.83
	90	239	226	22	0.90	0.75
	120	191	150	9	0.94	0.70
	150	99	73	5	0.93	0.65
Grade 3	0	399	395	34	0.91	0.91
	30	357	353	44	0.88	0.80
	60	304	293	21	0.93	0.74
	90	260	240	7	0.97	0.72
	120	213	171	4	0.98	0.70
	150	125	93	0	1.00	0.70

Table 4: Distribution of concordance/discordance levels corresponding to oestrogen receptor (ER) status, and cancers occurring in younger patients having a high chance for premenopausal status (age < 46).

	ER positive No (%)	ER negative No (%)	Age <46 years No (%)
Concordant low grade	224 (16.6)	2 (0.7)	8 (2.5)
Discordant low/intermediate	245 (18.2)	2 (0.7)	25 (7.8)
Concordant intermediate	395 (29.3)	18 (5.9)	58 (18.0)
Discordant intermediate/high	212 (15.8)	45 (14.9)	65 (20.2)
Concordant high	244 (18.2)	235 (77.6)	158 (49.1)
High/low discordance	24 (1.7)	1 (0.3)	8 (2.5)
Number of Patients	**1344**	**303**	**322**

The above represents a summary with regard to the originally assigned grade and the first of the additional VM reviews; VM1 and VM2.

Figure legends

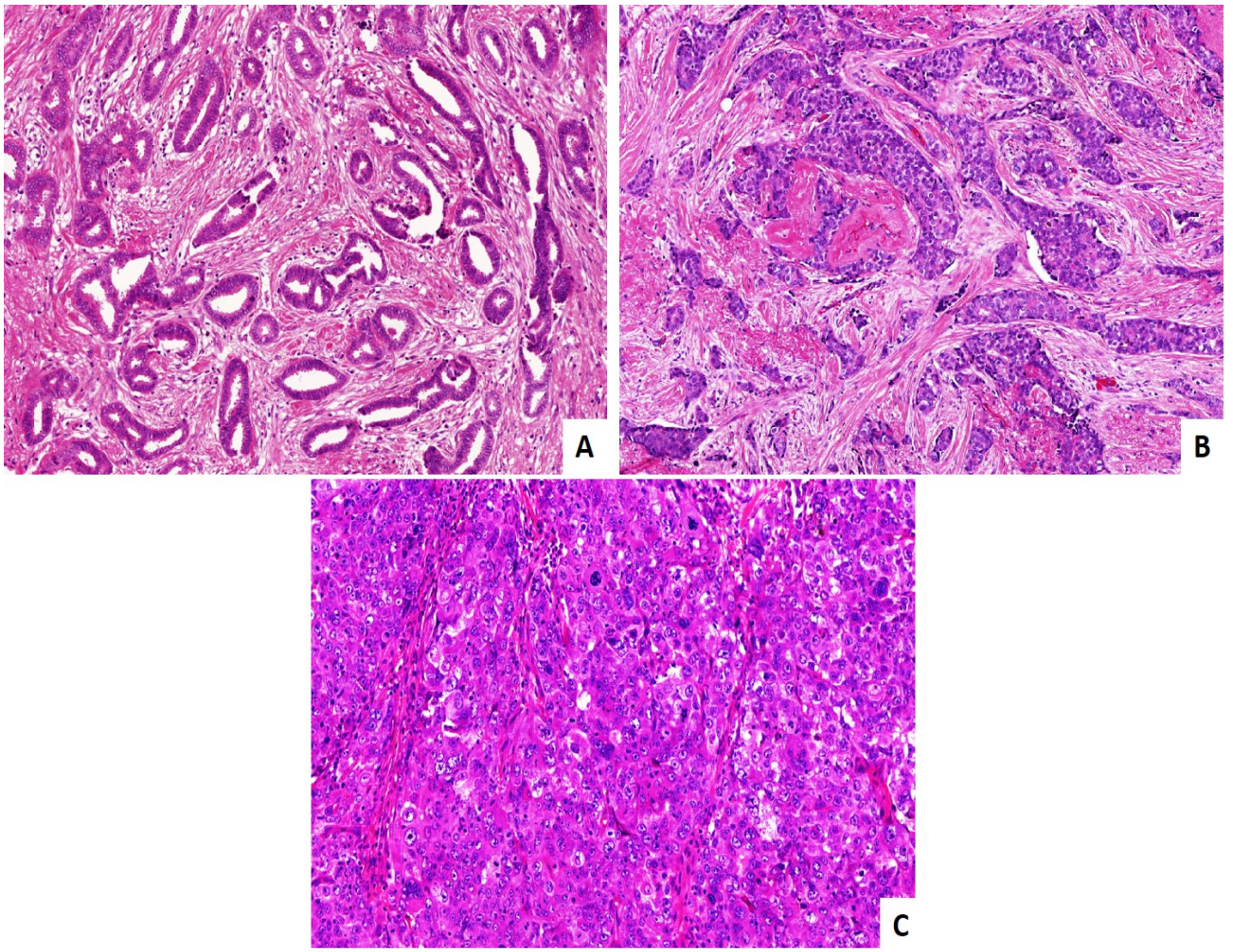

Figure 1: Photomicrographs demonstrating grade concordance between VM1 and VM2; A) A case of concordant Grade 1 tumour, B) A case of concordant Grade 2 tumour, C) An example of concordant Grade 3 tumour.

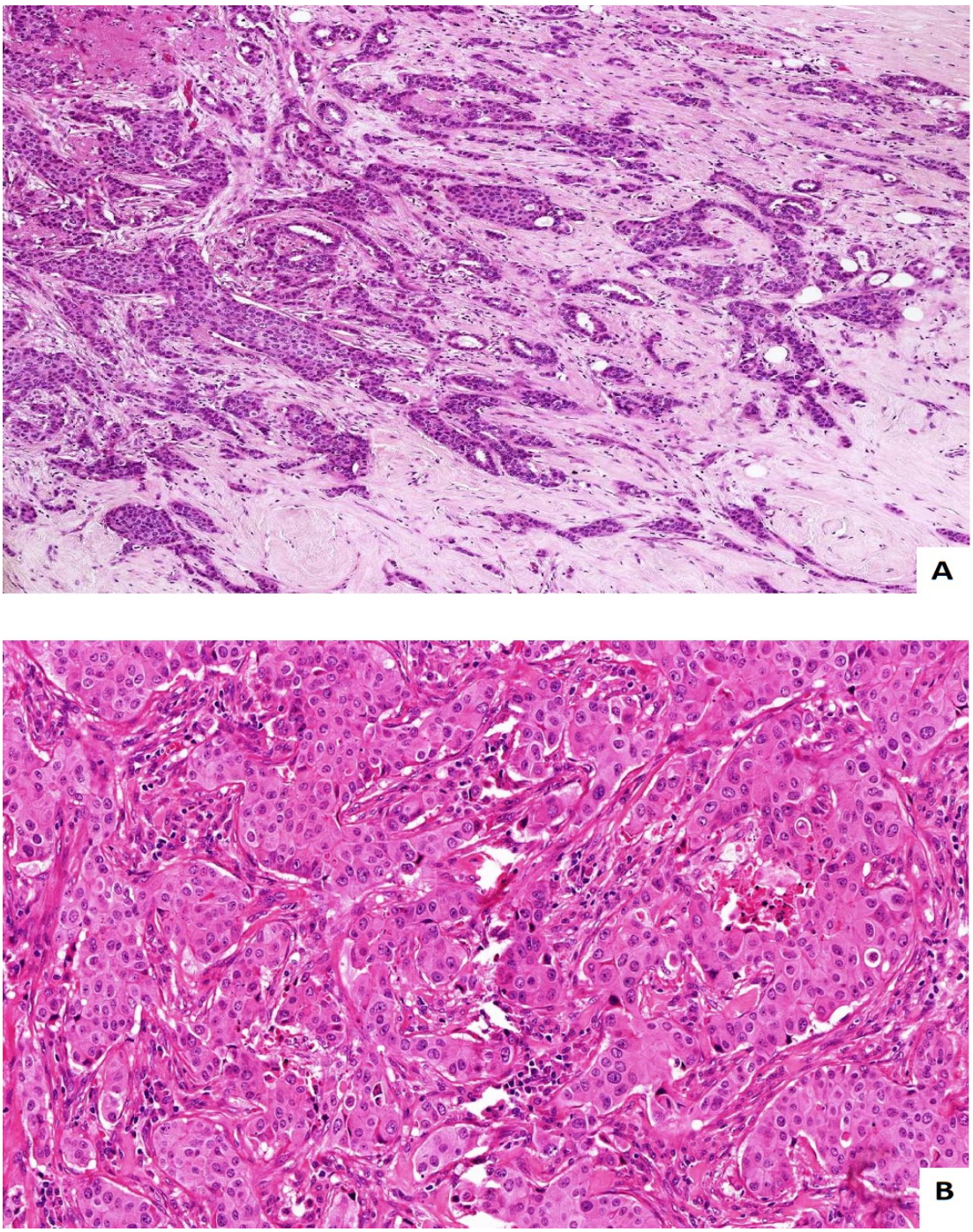

Figure 2: Photomicrographs illustrating discordance in grade (discordance between VM1 and VM2. A) A case of Grade 1/2 discordance; and B) a case of Grade 2/3 discordance) that represent borderline morphological features (tubule formation in A and pleomorphism in B).

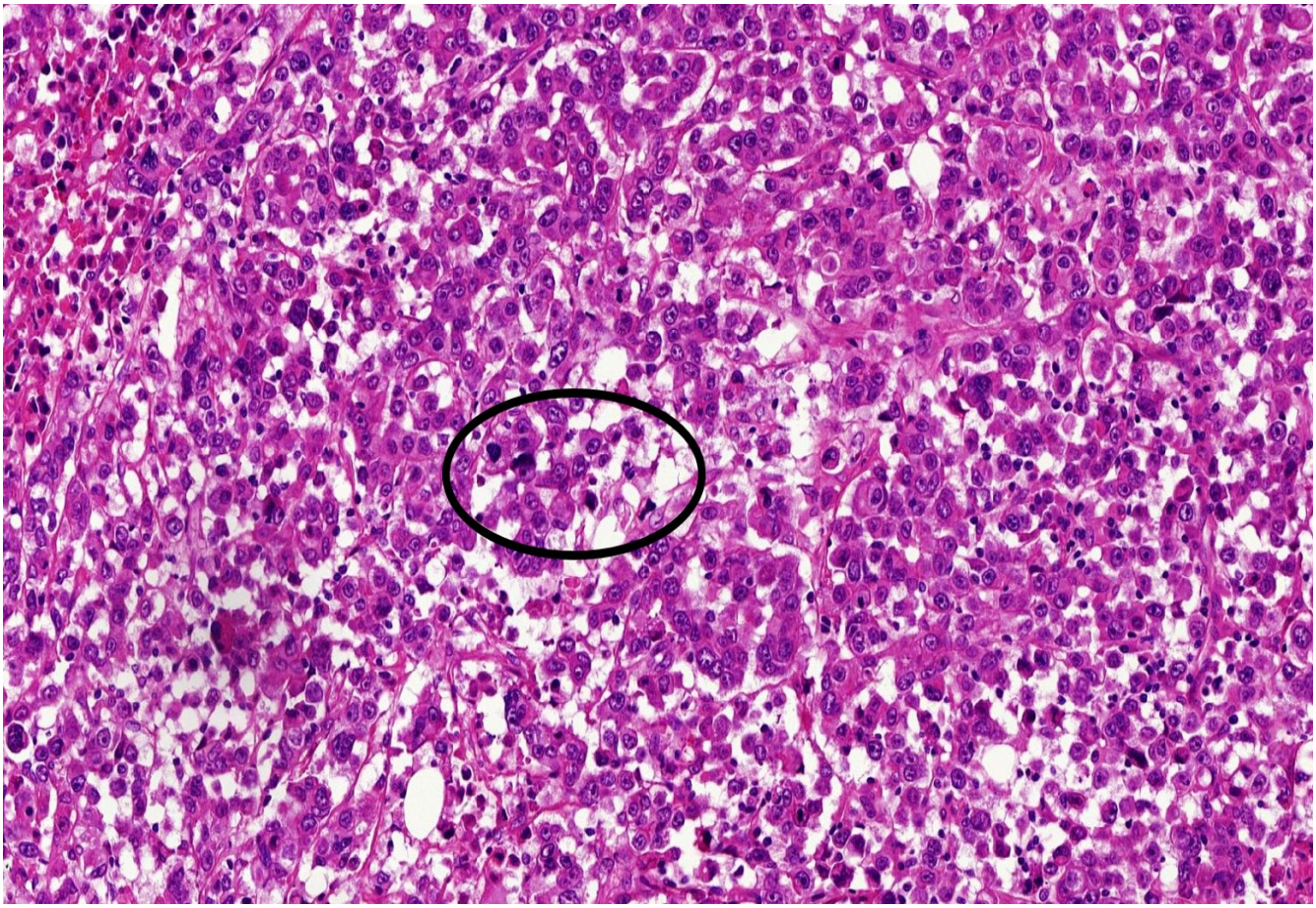

Figure 3: A case illustrating difficulties in differentiating mitotic figures from apoptotic bodies on virtual microscope grading. It is scored 3 for mitotic count using glass slides and scored 1 on the digital image (High power view; 200x).

Figure 4: Kaplan Meier survival curves comparing Nottingham grade originally assigned (A) with the first additional review grade (based on virtual microscopy) (B).

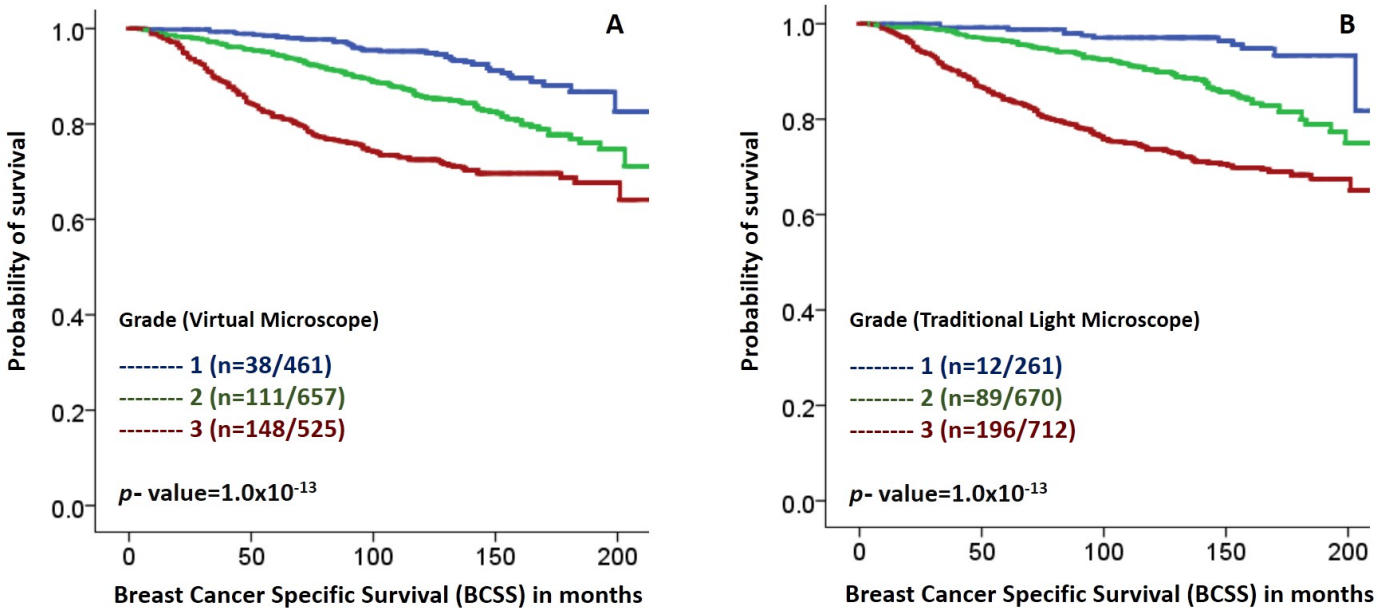

Figure 5: Association between histologic grade concordance and outcome (A- breast cancer specific survival [BCSS] and B- distant metastasis free interval [DMFI]; both $p<0.00001$). Red curves represent tumours with grade concordance (grade 1: upper curve, grade 2: middle curve and grade 3: lower curve) in both VM sessions. Black curve represents cases with grade 1 and 2 discordance whereas grey curve represents cases with grade 2 and 3 discordance.

Figure 6: Association between grade 2/3 concordance and BCSS ($p<0.00001$). Red curves represent tumour with grade concordance in all 3 grading sessions (grade 2: upper curve and grade 3: lower curve). Grey curve represents cases with grade discordance; original grade 3 and then grade 2 in the two sessions whereas purple curve represents cases with original grade 3 and one grade 3 and one grade 2 in the two sessions.

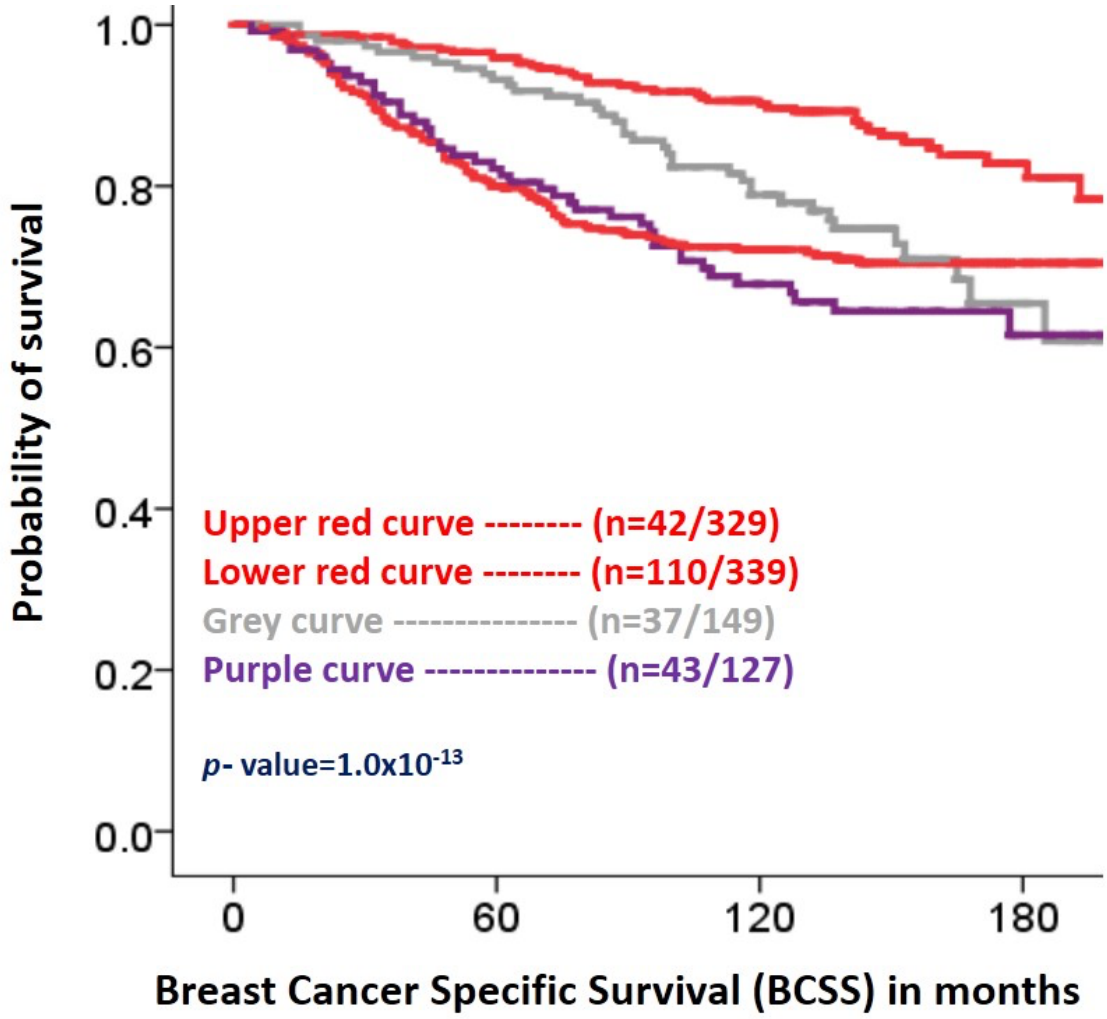

Figure 7: Association between grade 1/2 concordance and BCSS (p=0.001). Red curves represent tumour with grade concordance (grade 1: upper red curve and grade 2: lower red curve) in all 3 grading sessions. Grey curve represents cases with grade discordance; original grade 1 and then one grade 1 and one grade 2 in the two sessions, the blue curve represents cases with original grade 2 and the grade 1 in both sessions whereas the purple curve represents cases originally graded as 2 and then as one grade 2 and one grade 1 in the two sessions.