Quality improvement
of biomarker assessment in breast cancer

Carmen van Dooijeweert

Quality improvement
of biomarker assessment in breast cancer

Kwaliteitsverbetering van de beoordeling van biomarkers in borstkanker
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 29 september 2020 des middags te 4.15 uur

Door

Carmen van Dooijeweert
geboren op 9 mei 1991
te Zeist

Promotoren:
Prof. dr. P.J. van Diest
Prof. dr. E. van der Wall

Copromotoren:
Dr. I.A.G. Deckers
Dr. I.O. Baas

Table of content

1

General introduction
and thesis outline

General introduction

Breast cancer is the most common type of cancer in women, with a worldwide incidence of 2.1 million [1,2]. In the Netherlands, one in seven women will develop invasive breast cancer (IBC) during life [3,4], which accounts for about 15,000 new diagnoses annually [5]. In addition, each year, more than 2,300 women are diagnosed with ductal carcinoma *in situ* (DCIS) of the breast, which is generally considered to be a precursor of IBC [6], although it is unknown which proportion of untreated DCIS will actually progress into IBC [7].

The overall IBC survival rates of 88% after five years, and 79% after ten years, are relatively good, especially in comparison to other cancer types. However, in the Netherlands, each year over 3,000 women still die from breast cancer [5].

Diagnosis
All Dutch women between the ages of 50 and 75 are invited to participate in the national population screening, consisting of a biannual mammography, with the purpose to detect breast cancer at an early stage, as this leads to a better prognosis, less invasive treatment and increased survival [3]. As a result of screening, the detection rate for DCIS and IBC increased markedly for women between 50 and 75 years of age [4,8,9]. Currently, almost half of all IBC cases are screen-detected, while this is the case for over 70% of DCIS cases [10].

Prognostic factors and subtyping
Prognostic factors for IBC consist of patient-related factors, like age and performance status, and tumor-related factors (biomarkers) like tumor histology, tumor size, histologic grade, hormone- and HER2-receptor status, and the number of involved regional lymph nodes [2]. Furthermore, the three major breast cancer subtypes are identified according to the hormone- and HER2-receptor status of the tumor; i.e. estrogen (ER) and/or progesterone (PR)-driven IBC (HER2-) IBC ("luminal"), HER2-driven IBC (ER-/PR-/HER2+) and triple negative IBC (ER-, PR-, HER2-) (TNBC) [11-15].

Treatment
Management of breast cancer consists of many different modalities including surgery, radiotherapy, chemotherapy, endocrine therapy, and targeted therapy, which have considerably improved prognosis of IBC patients over the last decades [4]. Main therapy goals for non-metastatic breast cancer are tumor eradication from the breast and its regional lymph nodes, and preventing distant metastases by systemic therapy [2]. In contrast, for metastatic breast cancer, which is considered an incurable disease [2], main treatment goals are prolonging (quality of) life and symptom control [4].

Which treatment modalities are indicated for individual breast cancer patients is primarily guided by the aforementioned biomarkers. This highlights the crucial role of high-quality assessment of these biomarkers by pathologists. Hence, pathology may be considered as the cornerstone of breast cancer management.

Biomarker assessment: ER, PR, HER

Receptor assessment plays a crucial role in IBC management, as both hormone- and HER2-receptors, next to their prognostic role, form molecular targets for specific therapies, i.e. systemic endocrine- and anti-HER2 therapy [4, 16-18]. For example, the risk of recurrence and mortality is reduced by systemic endocrine therapy for early ER- and/or PR positive IBC patients, even independent of chemotherapy administration [19,20]. In addition, for HER2-positive IBC patients, adjuvant anti-HER2-therapy combined with specific chemotherapy regimens is considered regardless of other tumor characteristics [4].

In daily pathology practice, tumor receptor status for ER, PR, and HER2, is determined by immunohistochemical (IHC) tissue analyses (for ER, PR, HER2) and/or *in situ* hybridization (ISH) (for HER2) [4,21]. Reliable and reproducible receptor assessment is of major clinical importance as false negative results may result in withholding effective treatment, while false-positive results may result in overtreatment with costly and ineffective therapy, as well as exposure of patients to unwanted (direct and long-term) side effects [4, 22-27].

Biomarker assessment: histologic grading

Invasive breast cancer

Histologic grade is a biomarker that is strongly associated with both breast cancer-specific and disease free survival [28,29]. In addition, it has even been suggested that grade can predict tumor behavior more accurately than 'time-dependent' prognostic factors like tumor size [28-32]. Subsequently, histologic grading is widely used to guide management of IBC [4,28,29,33], for example in selecting the right patients for adjuvant systemic chemo- and/or endocrine therapy.

Histologic grade is determined by pathologists according to the modified Bloom and Richardson guideline (Elston-Ellis modification of Scarff-Bloom-Richardson grading system, also known as the Nottingham grading classification) [34,35], which combines pathologists' scores of degree of differentiation (tubule formation), the assessment of cell morphology (nuclear pleomorphism), and proliferation (mitotic count), resulting in an overall score and subsequent grade [35].

Ductal carcinoma in situ of the breast
For DCIS, it is currently believed that a considerable proportion of patients is treated for lesions that may never become invasive (i.e. IBC) [36-38]. Yet, standard treatment consists of either breast conserving surgery followed by radiotherapy, or a mastectomy, both sometimes even followed by endocrine therapy [39].

To counteract this presumed overtreatment [40], four clinical trials now aim to identify a group of patients at low risk of invasive progression who, under active surveillance, may safely forgo surgical treatment [37,38,41-43]. Under the hypothesis that progression risk, or at least speed of progression is higher for high-grade lesions [40,44], and if a low-grade DCIS eventually does become invasive, it will be a low-grade invasive carcinoma with favorable characteristics and excellent survival rates after treatment [37,45], all trials aim to identify their 'low-risk' subgroup based on histologic grade. Hence, grade may become the single biomarker that is used to determine whether treatment of DCIS is required.

Reproducibility of biomarker assessment
Accurate and reproducible biomarker assessment is of key-importance for identification of individual breast cancer patients who may benefit from specific treatment modalities. However, many studies, in which sets of IBC slides were reviewed by multiple pathologists, found substantial differences in receptor-assessment between testing laboratories [18, 46-56], while reproducibility of histologic grading was no more than moderate [57-60]. This raises the question whether this also holds true for biomarker assessment in daily clinical practice.

The nationwide network and registry of histo- and cytopathology in the Netherlands (PALGA)
PALGA plays an important role in daily pathology practice in the Netherlands. Besides governing all pathology reports in the Netherlands (since 1991) [61], thereby providing an incredible source for research, PALGA also facilitates synoptic (protocolled) reporting in daily pathology practice. In contrast to narrative reporting, synoptic reporting results in an increased overall completeness of pathology reports [62], and it enables easy data extraction, as all variables are stored in a standardized manner. Moreover, it has been shown that synoptic reporting significantly improves patient care in colorectal cancer [63]. For breast cancer, synoptic reporting is adapted well in the field, as approximately 80% of IBC resection specimens are currently reported via the synoptic protocol [64].

Thesis outline

Part 1 - Biomarker assessment in daily pathology practice

The first part of this thesis evaluates variation in biomarker assessment in daily pathology practice by using real-world nationwide (synoptic) PALGA data. In **chapter 2** we compare ER-, PR-, and HER2-receptor positivity rates between Dutch pathology laboratories, thereby controlling for patient and tumor characteristics (case-mix). This study design enables laboratories and pathologists to compare their (case-mix adjusted) positivity rates with other laboratories, which may be crucial to create awareness as pathologists may feel best addressed by their own data. In **chapter 3** we assess the inter- and intra-laboratory differences in histologic grading of DCIS, which may have important implications for the ongoing clinical trials and/or their implementation in daily pathology practice. Furthermore, a questionnaire on DCIS grading practices of pathologists may gain further insight in the difficulties of DCIS grading. In **chapter 4** we assess the inter- and intra-laboratory variation in histologic grading of IBC, which guides adjuvant systemic therapy, radiotherapy after a mastectomy, and the use of gene-expression profiling (GEP). In addition, we examine variation in the three components of the modified Bloom and Richardson grading classification: tubule formation, nuclear pleomorphism and mitotic count. In **chapter 5** we evaluate the increasing relevance of histologic grading in tailoring adjuvant systemic therapy in breast cancer patients in the Netherlands, thereby taking into account the past and present Dutch breast cancer guidelines.

Part 2 - Initiatives to decrease variation in biomarker assessment

The second part of this thesis focusses on two initiatives that were launched to minimize grading variation in DCIS and IBC; laboratory- and pathologist-specific feedback reports, and an e-learning module. In **chapter 6** we evaluate the effect of laboratory- and pathologist-specific feedback reports on grading variation of DCIS. The effect of these reports, in which laboratory-specific case-mix adjusted proportions per grade are benchmarked against other laboratories, was studied up to one year after the reports were sent. Similarly, in **chapter 7** we evaluate the effect of the laboratory- and pathologist-specific feedback reports on grading variation of IBC. In **chapter 8** we assess whether an e-learning module, in which pathologists and residents are trained in histologic grading of both DCIS and IBC, may be a feasible tool to decrease grading variation. The results of this thesis and future perspectives are discussed in **chapter 9**.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.

2. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019;321(3):288-300.

3. National Institute for Health and Environment (RIVM) Ministry of Health, Wellness and Sport. Chance of breast cancer. Avalaible from: https://www.rivm.nl/bevolkingsonderzoek-borstkanker/wat-is-borstkanker/kans-op-borstkanker. Accessed October 10 2019.

4. The Netherlands Comprehensive Cancer Organization (IKNL). Breast Cancer Guideline 2017 [updated 2012-07-02]. Available from: https://www.oncoline.nl/borstkanker. Accessed 17 January 2019.

5. Netherlands Cancer Registry supplied by IKNL, Cijfers over kanker. Available from: www.cijfersoverkanker.nl. Accessed March 20, 2019.

6. Pinder SE, Ellis IO. The diagnosis and management of pre-invasive breast disease: ductal carcinoma in situ (DCIS) and atypical ductal hyperplasia (ADH)--current definitions and classification. Breast Cancer Res. 2003;5(5):254-7.

7. Welch HG, Black WC. Using autopsy series to estimate the disease "reservoir" for ductal carcinoma in situ of the breast: how much more breast cancer can we find? Ann Internal Med. 1997;127(11):1023-8.

8. van Steenbergen LN, Voogd AC, Roukema JA, et al. Screening caused rising incidence rates of ductal carcinoma in situ of the breast. Breast Cancer Res, Treat. 2009;115(1): 181-183..

9. Louwman WJ, Voogd AC, van Dijck JAAM, et al. On the rising trends of incidence and prognosis for breast cancer patients diagnosed 1975-2004: a long-term population-based study in southeastern Netherlands. Cancer Causes Control. 2008;19(1):97-106.

10. The Netherlands Comprehensive Cancer Organization (IKNL). Monitor Breast Cancer Sreening 2017. Available from: https://www.iknl.nl/getmedia/e2abd7e5-c6e5-4398-842f-e91facd3d6bd/iknl_rapport-borstkanker-in-nederland-311018-int.pdf. Accessed October 10, 2019.

11. Goetz MP, Gradishar WJ, Anderson BO, et al. NCCN Guidelines Insights: Breast Cancer, Version 3.2018. 2019;17(2):118.

12. The Netherlands Comprehensive Cancer Organization (IKNL). Oncoline: Cancer clinical practice guidelines (the Netherlands) 2019. Available from: https://www.oncoline.nl/. Accessed May 8, 2020.

13. Burstein HJ, Curigliano G, Loibl S, et al. Estimating the Benefits of Therapy for Early Stage Breast Cancer The St Gallen International Consensus Guidelines for the Primary Therapy of Early Breast Cancer 2019. Ann Oncol. 2019;30(10):1541-1557.

14. Henry NL, Somerfield MR, Abramson VG, et al. Role of Patient and Disease Factors in Adjuvant Systemic Therapy Decision Making for Early-Stage, Operable Breast Cancer: Update of the ASCO Endorsement of the Cancer Care Ontario Guideline. J Clin Oncol. 2019;37(22):1965-77.

15. Eisen A, Fletcher GG, Gandhi S, et al. Optimal systemic therapy for early breast cancer in women: a clinical practice guideline. Curr Oncol 2015;22(Suppl 1):S67-81.

16. Coates AS, Winer EP, Goldhirsch A, et al. Tailoring therapies - improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer

2015. Ann Oncol: 2015;26(8):1533-46.

17. Duffy MJ, Harbeck N, Nap M, et al. Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). Eur J Cancer. 2017;75:284-98.

18. McCullough AE, Dell'orto P, Reinholz MM, et al. Central pathology laboratory review of HER2 and ER in early breast cancer: an ALTTO trial [BIG 2-06/NCCTG N063D (Alliance)] ring study. Breast Cancer Res Treat. 2014;143(3):485-92.

19. Davies C, Godwin J, Gray R, et al. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet. 2011;378(9793):771-84.

20. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet. 2005;365(9472):1687-717.

21. Choritz H, Busche G, Kreipe H. Quality assessment of HER2 testing by monitoring of positivity rates. Virchows Arch. 2011;459(3):283-9.

22. Bowles EJA, Wellman R, Feigelson HS, et al. Risk of Heart Failure in Breast Cancer Patients After Anthracycline and Trastuzumab Treatment: A Retrospective Cohort Study. J Natl Cancer Inst. 2012;104(17):1293-305.

23. Perez EA, Suman VJ, Davidson NE, et al. Cardiac Safety Analysis of Doxorubicin and Cyclophosphamide Followed by Paclitaxel With or Without Trastuzumab in the North Central Cancer Treatment Group N9831 Adjuvant Breast Cancer Trial. J Clin Oncol. 2008;26(8):1231-8.

24. Romond EH, Jeong JH, Rastogi P, et al. Seven-year follow-up assessment of cardiac function in NSABP B-31, a randomized trial comparing doxorubicin and cyclophosphamide followed by

paclitaxel (ACP) with ACP plus trastuzumab as adjuvant therapy for patients with node-positive, human epidermal growth factor receptor 2-positive breast cancer. J Clin Oncol. 2012;30(31):3792-9.

25. Rüschoff J, Lebeau A, Kreipe H, et al. Assessing HER2 testing quality in breast cancer: variables that influence HER2 positivity rate from a large, multicenter, observational study in Germany. Mod Pathol. 2016;30:217.

26. Russell SD, Blackwell KL, Lawrence J, et al. Independent Adjudication of Symptomatic Heart Failure With the Use of Doxorubicin and Cyclophosphamide Followed by Trastuzumab Adjuvant Therapy: A Combined Review of Cardiac Data From the National Surgical Adjuvant Breast and Bowel Project B-31 and the North Central Cancer Treatment Group N9831 Clinical Trials. J Clin Oncol. 2010;28(21):3416-21.

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

28. Rakha EA, El-Sayed ME, Lee AH, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol. 2008;26(19):3153-8.

29. 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(4):207.

30. Sundquist M, Thorstenson S, Brudin L, Nordenskjold B. Applying the Nottingham Prognostic Index to a Swedish breast cancer population. South East Swedish Breast Cancer Study Group. Breast Cancer Res Treat. 1999;53(1):1-8.

31. Frkovic-Grazio S, Bracko M. Long term prognostic value of Nottingham histological grade and its components in early

(pT1N0M0) breast carcinoma. J Clin Pathol. 2002;55(2):88-92.

32. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat. 1992;22(3):207-19.

33. Curigliano G, Burstein HJ, Winer EP, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. Ann Oncol. 2017;28(8):1700-12.

34. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer. 1957;11(3):359-77.

35. 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(5):403-10.

36. Benson JR, Jatoi I, Toi M. Treatment of low-risk ductal carcinoma in situ: is nothing better than something? The Lancet Oncology. 2016;17(10):e442-e51.

37. 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(12):1497-510.

38. Francis A, Thomas J, Fallowfield L, et al. Addressing overtreatment of screen detected DCIS; the LORIS trial. Eur J Cancer. 2015;51(16):2296-303.

39. Gorringe KL, Fox SB. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. Front Oncol. 2017;7:248.

40. Toss M, Miligy I, Thompson AM, et al. Current trials to reduce surgical intervention in ductal carcinoma in situ of the breast: Critical review. Breast. 2017;35:151-6.

41. 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(3):e026797.

42. Groen EJ, Elshof LE, Visser LL, et al. Finding the balance between over- and under-treatment of ductal carcinoma in situ (DCIS). Breast. 2017;31:274-83.

43. Lippey J, Spillane A, Saunders C. Not all ductal carcinoma in situ is created equal: can we avoid surgery for low-risk ductal carcinoma in situ? ANZ J Surg. 2016;86(11):859-60.

44. Sanders ME, Schuyler PA, Dupont WD, Page DL. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. Cancer. 2005;103(12):2481-4.

45. Lopez-Garcia MA, Geyer FC, Lacroix-Triki M, Marchió C, Reis-Filho JS. Breast cancer precursors revisited: Molecular features and progression pathways. Histopathology. 2010;57(2):171-92.

46. Cuadros M, Cano C, López FJ, et al. HER2 status in breast cancer: experience of a Spanish National Reference Centre. Clin Transl Oncol. 2011;13(5)335-340.

47. Denkert C, Huober J, Loibl S, et al. HER2 and ESR1 mRNA expression levels and response to neoadjuvant trastuzumab plus chemotherapy in patients with primary breast cancer. Breast Cancer Res. 2013;15(1):R11.

48. Orlando L, Viale G, Bria E, et al. Discordance in pathology report after central pathology review: Implications for breast cancer adjuvant treatment. The Breast. 2016;30:151-5.

49. Paik S, Bryant J, Tan-Chiu E, et al. Real-world performance of HER2 testing -National Surgical Adjuvant Breast and Bowel Project experience. J Natl Cancer

Inst. 2002;94(11):852-4.

50. Perez EA, Suman VJ, Davidson NE, et al. HER2 testing by local, central, and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial. J Clin Oncol. 2006;24(19):3032-8.

51. Roche PC, Suman VJ, Jenkins RB, et al. Concordance between local and central laboratory HER2 testing in the breast intergroup trial N9831. J Natl Cancer Inst. 2002;94(11):855-7.

52. Rosa M, Khazai L. Comparison of HER2 testing among laboratories: Our experience with review cases retested at Moffitt Cancer Center in a two-year period. The Breast Journal. 2017;24(2):139-47.

53. Viale G, Regan MM, Maiorano E, et al. Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1-98. J Clin Oncol. 2007;25(25):3846-52.

54. Douglas-Jones AG, Morgan JM, Appleton MA, et al. Consistency in the observation of features used to classify duct carcinoma in situ (DCIS) of the breast. J Clin Pathol. 2000;53(8):596-602.

55. Elston CW, Sloane JP, Amendoeira I, et al. Causes of inconsistency in diagnosing and classifying intraductal proliferations of the breast. European Commission Working Group on Breast Screening Pathology. Eur J Cancer. 2000;36(14):1769-72.

56. 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(10):705-11.

57. Boiesen P, Bendahl PO, Anagnostaki L, et al. Histologic grading in breast cancer--

reproducibility between seven pathologic departments. South Sweden Breast Cancer Group. Acta Oncol. 39;(1):41-45.

58. Frierson HF Jr., Wolber RA, Berean KW, et al. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. Am J Clin Pathol. 103(2):195-198.

59. Italian Network for Quality Assurance of Tumour Biomarkers (INQAT) Group. Quality control for histological grading in breast cancer: an Italian experience. Pathologica. 2005;97(1):1-6.

60. Meyer JS, Alvarez C, Milikowski C, et al. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. Mod Pathol. 2005;18(8):1067-78.

61. Casparie M, Tiebosch AT, Burger G, et al. Pathology databanking and biobanking in The Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. Cell Oncol. 2007;29(1):19-24.

62. Sluijter CE, van Lonkhuijzen LR, van Slooten HJ, Nagtegaal ID, Overbeek LI. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows Arch. 2016;468(6):639-49.

63. Sluijter CE, van Workum F, Wiggers T, et al. Improvement of Care in Patients With Colorectal Cancer: Influence of the Introduction of Standardized Structured Reporting for Pathology. JCO Clin Cancer Inform. 2019;3:1-12.

64. Foundation PALGA. Annual report 2018. Available from: https://www.palga.nl/assets/uploads/Jaarverslag/2018%20Palga%20web.pdf. Accessed October 10, 2019.

PART ONE

Biomarker assessment
in daily pathology practice

2

Hormone- and HER2-receptor assessment in 33,046 breast cancer patients: a nationwide comparison of positivity rates between pathology laboratories in the Netherlands

Carmen van Dooijeweert
Ivette A. G. Deckers
Inge O. Baas
Elsken van der Wall
Paul J. van Diest

Abstract

Purpose
Patient management of invasive breast cancer (IBC) is to a large extent based on hormone- and HER2-receptor assessment. High-quality, reliable receptor assessment is of key-importance as false results may lead to under- or overtreatment of patients. Surveillance of case-mix adjusted positivity rates has been suggested as a tool to identify laboratories with insufficient testing assays, as this covers the whole process of receptor assessment and enables laboratories to benchmark their positivity rates against other laboratories. We studied laboratory-specific variation in hormone- and HER2-positivity rates of 33,046 breast cancer patients using real-life nationwide data.

Methods
All synoptic pathology reports of IBC resection-specimens, obtained between 2013-2016, were retrieved from the nationwide Dutch Pathology registry (PALGA). Absolute and case-mix adjusted receptor positivity rates were compared to the mean national proportion and presented in funnel plots in separate analyses for estrogen (ER), progesterone (PR) and HER2. Case-mix adjustment was performed by multivariable logistic regression.

Results
33,794 IBC lesions from 33,046 patients of 39 pathology laboratories were included. After case-mix adjustment, mean positivity rates were 87.2% for ER (range: 80.4-94.3), 71.3% for PR (62.5-77.5%), and 9.9% for HER2 (5.5-12.7%). Overall, 14 (35.9%), 17 (43.6%) and 11 (28.2%) laboratories showed positivity rates outside the 95% confidence interval for ER, PR and HER2, respectively.

Conclusion
This nationwide study shows that absolute variation in hormone- and HER2-receptor positivity rates between Dutch pathology laboratories is limited. Yet, the considerable number of outlying laboratories shows that there is still need for improvement. Continuous monitoring and benchmarking of positivity rates may help to realize this.

Introduction

Patient management of invasive breast cancer (IBC) is to a large extent based on estrogen- (ER), progesterone- (PR), and HER2-receptor assessment as they determine whether targeted anti-hormonal, anti-HER2 therapy and/or chemotherapy are indicated [1-4]. For early ER- and/or PR positive breast cancer, the risk of recurrence and mortality is reduced by anti-endocrine therapy, independent of the administration of chemotherapy [5, 6]. In addition, for HER2 positive breast cancer, adjuvant anti-HER2-therapy combined with chemotherapy is considered, regardless of other characteristics like tumor grade [1]. In addition, different chemotherapy regimens are considered for HER2-positive breast cancer patients [1].

ER-, PR- or HER2-receptor status of a tumor is established by pathological analysis of tumor tissue by immunohistochemistry (IHC) (ER, PR and HER2) and/or *in situ* hybridization (ISH) (HER2) [1,7], which, according to global guidelines, is mandatory for all newly diagnosed primary IBC cases [1, 2, 4, 8-10]. High-quality, reliable receptor assessment is of key-importance as false-negative results may result in withholding effective treatment, whilst false-positive results could result in overtreatment with costly and ineffective therapy at the same time resulting in unwanted direct and long-term side effects [1, 11-16].

The quality of ER-, PR- and HER2 testing has been extensively studied over the past two decades. Central review of trial cases or cases from local pathology laboratories mainly showed that substantial differences between testing laboratories occurred [3, 17-24], which was confirmed by reversed studies in which samples or tissue microarrays were send to different laboratories [25-30]. Proficiency testing programs were launched as a promising remedy [31-35], but it has been argued that they render only a temporary and incomplete assessment of testing performance, which does not necessarily reflect reliability of testing over time [7]. For example, crucial steps like tissue fixation and processing are not covered by these tests [36].

Recently, surveillance of positivity rates has been suggested as a tool to identify laboratories with insufficient testing assays and a high yield of false-positive- or false-negative results [7, 14, 16, 37]. However, as test-accuracy is not the only potential factor in receptor positivity rates, it is important to also take patient- and tumor characteristics into account [14, 38]. Such a study-design would enable laboratories and pathologists to compare their receptor positivity rates with other laboratories, while controlling for differences in population characteristics ("case-mix") [38]. This may be crucial to create awareness, as pathologists and their laboratories may feel addressed by their own case-mix adjusted "mirror"-data. Previous studies using such a design found significant

variation between pathology laboratories in Germany with a range of HER2-positivity rates varying from 7.6 to 31.6% [7, 14] with significant outliers even after case-mix correction [14]. To the best of our knowledge such studies have not been performed for ER- and PR-receptor positivity rates.

To create insight and awareness in the Netherlands, we compared ER, PR and HER2-receptor positivity rates from daily clinical practice between pathology laboratories using real-life data from synoptic (structured) pathology reports of 33,046 IBC patients from the Dutch nationwide pathology registry (PALGA).

Methods

Data source and study population
We extracted data from PALGA, the nationwide network and registry of histo- and cytopathology in the Netherlands, which contains pathology reports from all Dutch Pathology laboratories since 1991 [39]. Data from the PALGA database are pseudonymized by a trusted third party (ZorgTTP, Houten, the Netherlands). As all pathology laboratories were initially anonymized, we obtained further written consent for the additional analysis of inter-pathologist variation within individual laboratories (n=7). This study was approved by the scientific and privacy committee of PALGA and all data were retrieved and handled in compliance with the General Data Protection Regulation act.

All synoptic pathology reports of patients with IBC resection specimens between January 1 2013 and December 31 2016 in the Netherlands (n=48,665) were extracted. Synchronous IBC was defined as an ipsilateral lesion within six months of the previous IBC resection during the study period. As these lesions were considered paired measurements, we only included the first lesion. Reports of resection specimens without a primary tumor were excluded. Likewise, pathology reports of patients who received neoadjuvant treatment were excluded as tumor receptor status may be converted by neo-adjuvant treatment [40-42] (Fig. 1).

Overall, 40 out of 46 Dutch pathology laboratories reported on breast resection specimens using the synoptic (PALGA) pathology protocol. Of these laboratories, we only included those that synoptically reported ≥250 IBC resection specimens during the study period (n=39). For inter-pathologist variation within individual laboratories, we only analyzed data from pathologists from the consenting laboratories who synoptically reported ≥20 IBC during the study period.

From each pathology report we extracted patient characteristics (sex, age, type of surgery) and tumor characteristics (tumor size, histologic subtype, histologic grade, ER- and PR-receptor status, and HER2-receptor status). ER- and PR-status were determined by IHC, whereas HER2-status was determined either by IHC and/or ISH. Lastly, reports of IBC with any missing data (histologic grade, ER, PR- or HER2-receptor status) were excluded from further data-analysis (Fig 1).

Analysis of ER- and PR-receptor status

Within the synoptic protocol, and according to the Dutch guideline [1], the ER- and PR-receptor status is considered positive when ≥10% of tumor cells show ER- and PR-specific staining by IHC. Overall, both ER- and PR-receptor status were taken into account as a binary variable, either positive (≥10%) or negative (<10%), since the percentage of stained tumor nuclei (not an obligatory item) was not known for ~20% of cases. When one or both receptors were missing, the status on biopsy was considered the true receptor status (~7.5%), as this is common practice in clinical management.

Surrogate intrinsic subtype

Surrogate intrinsic subtype was established as before by Perou et al [43] as Luminal A = ER +, PR +/-, HER2-, Luminal B = ER +, PR +/-, HER 2 +, HER2-driven = ER- , PR-, HER2 +, Basal-like = ER-, PR-, HER2-.

Analysis of HER2-receptor status

HER2-receptor status was taken into account as a binary variable, either positive or negative, regardless of which techniques were used (IHC and/or ISH). In general, and as recommended by the Dutch guideline [1], IHC is performed first, followed by amplification testing in case of a 2+ IHC score. As described for ER- and PR, when HER2-receptor status was missing on resection specimen, biopsy HER2-receptor status was considered the true receptor status (5.5%).

Survey among laboratories

A survey was sent to all 46 Dutch pathology laboratories to gain insight into their processes and interpretation of receptor assessment in daily clinical practice. The survey included questions on whether receptor status was assessed on biopsy and/or resection specimen, the cutoff percentages used for receptor-positivity (ER/PR), the interpretation of IHC scores for HER2 (0, 1+, 2+, 3+), techniques used for HER2-assessment and the order in which they were executed.

Synoptic pathology reports of 48,665 invasive breast tumors (IBC) from PALGA database 2013-2016 (42,621 patients)

2,104 reports →

No primary tymor
- complete regression (1,432 reports)
- after biopsy (363 reports)
- re-excision (309 reports)

46,561 IBC
(40,869 patients)

1 report →

Small laboratories
(total IBC <250)

46,560 IBC
(40,868 patients)

5,828 reports →

Neoadjuvant treatment

40,732 IBC
(35,684 patients)

4,170 reports →

Synchronous IBC
- synchronous (4,155 reports)
- synchronous unknown (15 reports)

36,562 IBC
(35,677 patients)

2,462 reports →

Unknown receptor status
(ER, PR and/or HER2)

306 reports →

Unknown tumor grade

33,794 IBC
(33,046 patients)

Figure 1. Flowchart of included lesions of invasive breast cancer (IBC) to assess variation in receptor (ER, PR, HER2) positivity rates between laboratories.

Statistical analysis

Separate analyses were performed for ER-, PR, and HER2-receptor status as outcome measure. Patient and tumor characteristics were summarized and differences between receptor-positive and –negative status (ER, PR, and HER2) were tested by means of χ^2-test for categorical variables and by a non-parametric Kruskal-Wallis test for continuous variables.

Overall positivity rates per receptor (ER, PR, HER2) were determined and considered the national proportion. Absolute differences in positivity rates between laboratories were presented in funnel plots per receptor, in which the positive-receptor proportions per laboratory were plotted against the number of included IBC reports per laboratory, with the overall national proportion with its 95% confidence limits as target [44].

For case-mix correction, all available clinicopathological risk factors were selected *a priori* based on literature [14, 38, 45-47] and on pathologists' experience. These factors included age, sex, tumor size, type of surgery, histologic subtype, tumor grade, and either the combined hormone-receptor status (for HER2-analysis) or HER2-receptor status (for ER- and PR-analysis). The combined hormone-receptor status (ER/PR) was considered positive when either or both the ER- and PR-receptor were reported as positive. Sex was excluded in the final multivariable logistic regression model, as the number of males was too low. However, males did not cluster in specific laboratories. To calculate case-mix adjusted percentages, the observed percentage (O) per laboratory was divided by the expected percentage (E), based on the multivariate logistic regression model, and multiplied by the overall mean positive percentage per receptor (O/E * mean). Similar to the crude percentages, case-mix adjusted percentages were presented in funnel plots.

For analysis of the inter-pathologist variation within the laboratories we merely compared the proportions per receptor (ER, PR and HER2) between pathologists by Fisher exact test (Monte Carlo option).

Survey results were summarized by frequencies and percentages. P-values below 0.05 were considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics version 25.

Results

Characteristics of patients, tumors, and laboratories
In total, 33,794 unique IBC lesions of 33,046 patients from 39 laboratories were included. Characteristics of all included patients and corresponding invasive breast tumors are listed in Table 1.

Nearly all patients were female (99.2%) and the overall mean (+/- standard deviation (SD)) age was 62.2 (+/- 12.1) years. The majority of patients underwent breast conserving surgery (63.9%) for tumors with a mean (+/- SD) of 1.9 (+/- 1.3) cm. HER2-positivity of tumors was associated with higher histologic tumor grade, whereas ER- and PR-positivity of tumors was associated with lower tumor grade. HER2-positive tumors were less often of lobular subtype and were of larger size than HER2-negative tumors. Furthermore, HER2-positive tumors were less often hormone-receptor positive and vice versa.

The number of synoptically reported IBC lesions per laboratory ranged from 80 to 2,224 (median 794). Overall observed positive proportions were 87.2% for ER, 71.3% for PR and only 9.9% for HER2. Regarding the intrinsic breast cancer subtypes, luminal A, luminal B, HER2-driven, and basal-like subtypes were observed in 80.5%, 6.7%, 3.2% and 9.7%, respectively (Table 1).

Inter-laboratory variation in ER-, PR- and HER2-positivity rates
Positivity-rates between laboratories varied most for PR (60.0-78.8%), followed by ER (77.5-92.7%) and HER2 (5.3-13.0%). After case-mix adjustment, the inter-laboratory range slightly decreased for all receptors; PR (62.5-77.5%), ER (80.4-94.3%), HER2 (5.5-12.7%) (Fig. 2). Overall, 17 laboratories (43.6%) showed positivity rates outside the 95% CI for PR, followed by 14 laboratories (35.9%) for ER and 11 laboratories (28.2%) for HER2 (Fig. 2).

Intra-laboratory variation in ER-, PR- and HER2-positivity rates
Sixty-two pathologists from the seven laboratories that participated in the intra-laboratory variation analysis synoptically reported ≥20 IBC during the study period. Per laboratory the number of analyzed pathologists ranged from 3 to 15 (median 9). The number of analyzed IBC reports per pathologist ranged from 20-257 (median 81). Overall, positivity ranges for ER, PR and HER2 did not significantly differ between pathologists within individual laboratories, except for the ER-positivity rates of the three pathologists from one laboratory (laboratory 10, positivity rates 90.1/98.8/92.9%, p= 0.032, data not shown).

Table 1. Characteristics of the 33,794 included invasive breast cancers from the Dutch national PALGA database 2013-2016

	Total (n=33,794)	ER-negative (n=4,337)	ER-positive (n=29,457)	PR-negative (9,698)	PR-positive (n=24,069)	HER2-negative (n=30,454)	HER2-positive (n=3,340)
Age (y)*	62.2 (12.1)	60.9 (14.2)	62.4 (11.7)	62.9 (12.5)	61.9 (11.9)	62.5 (11.9)	59.6 (13.1)
Sex, n (%)							
Female	33,540 (99.2%)	4,335 (100.0%)	29,205 (99.1%)	9,666 (99.7%)	23,874 (99.1%)	30,219 (99.2%)	3,321 (99.4%)
Male	254 (0.8%)	2 (0.0%)	252 (0.9%)	32 (0.3%)	222 (0.9%)	235 (0.8%)	19 (0.6%)
Tumour size (cm)*	1.9 (1.3)	2.2 (1.6)	1.8 (1.3)	2.0 (1.5)	1.8 (1.3)	1.8 (1.3)	2.1 (1.4)
Type of surgery, n (%)							
Mastectomy	12,208 (36.1%)	1,902 (43.9%)	10,306 (35.0%)	3,961 (40.8%)	8,247 (34.2%)	10,641 (34.9%)	1,567 (46.9%)
Breast conserving	21,586 (63.9%)	2,435 (56.1%)	19,151 (65.0%)	5,737 (59.2%)	15,849 (65.8%)	19,813 (65.1%)	1,773 (53.1%)
Histologic subtype, n (%)							
Ductal	28,549 (84.5%)	3,763 (86.8%)	24,786 (84.1%)	8,137 (83.9%)	20,412 (84.7%)	25,416 (83.5%)	3,133 (93.8%)
Lobular	4,429 (13.1%)	95 (2.1%)	4,334 (14.7%)	1,012 (10.4%)	3,417 (14.2%)	4,291 (14.1%)	138 (4.1%)
Other	816 (2.4%)	479 (11.0%)	337 (1.1%)	549 (5.7%)	267 (1.1%)	747 (2.5%)	69 (2.1%)
Histologic grade, n (%)							
Grade 1	9,494 (28.1%)	130 (3.0%)	9,364 (31.8%)	1,487 (15.3%)	8,007 (33.2%)	9,283 (30.5%)	211 (6.3%)
Grade 2	16,103 (47.6%)	964 (22.2%)	15,139 (51.4%)	3,696 (38.1%)	12,407 (51.5%)	14,767 (48.5%)	1,336 (40.0%)
Grade 3	8,197 (24.3%)	3,243 (74.8%)	4,954 (16.8%)	4,515 (46.6%)	3,682 (15.3%)	6,404 (21.0%)	1,793 (53.7%)
Combined ER/PR status, n (%)							
Negative	4,216 (12.5%)	-	-	-	-	3,179 (10.4%)	1,037 (31.0%)
Positive	29,578 (87.5%)	-	-	-	-	27,275 (89.6%)	2,303 (69.0%)
HER2-receptor status, n (%)							
Negative	30,454 (90.1%)	3,265 (75.3%)	27,189 (92.3%)	7,866 (81.1%)	22,588 (93.7%)	-	-
Positive	3,340 (9.9%)	1,072 (24.7%)	2,268 (7.7%)	1,832 (18.9%)	1,508 (6.3%)	-	-
Intrinsic subtypes**							
Luminal A	27,189 (80.5%)	-	-	-	-	-	-
Luminal B	2.268 (6.7%)	-	-	-	-	-	-
HER2-driven	1.072 (3.2%)	-	-	-	-	-	-
Basal-like	3.265 (9.7%)	-	-	-	-	-	-

* Mean (SD) ** Luminal A = ER +, PR +/-, HER2 -, Luminal B = ER +, PR +/-, HER 2 +, HER2-driven = ER- , PR-, HER2 +, Basal-like = ER-, PR-, HER2 –

***p-values for all variables (positive versus negative receptor-status) <0.0005, except for males and HER2 (p=0.198)
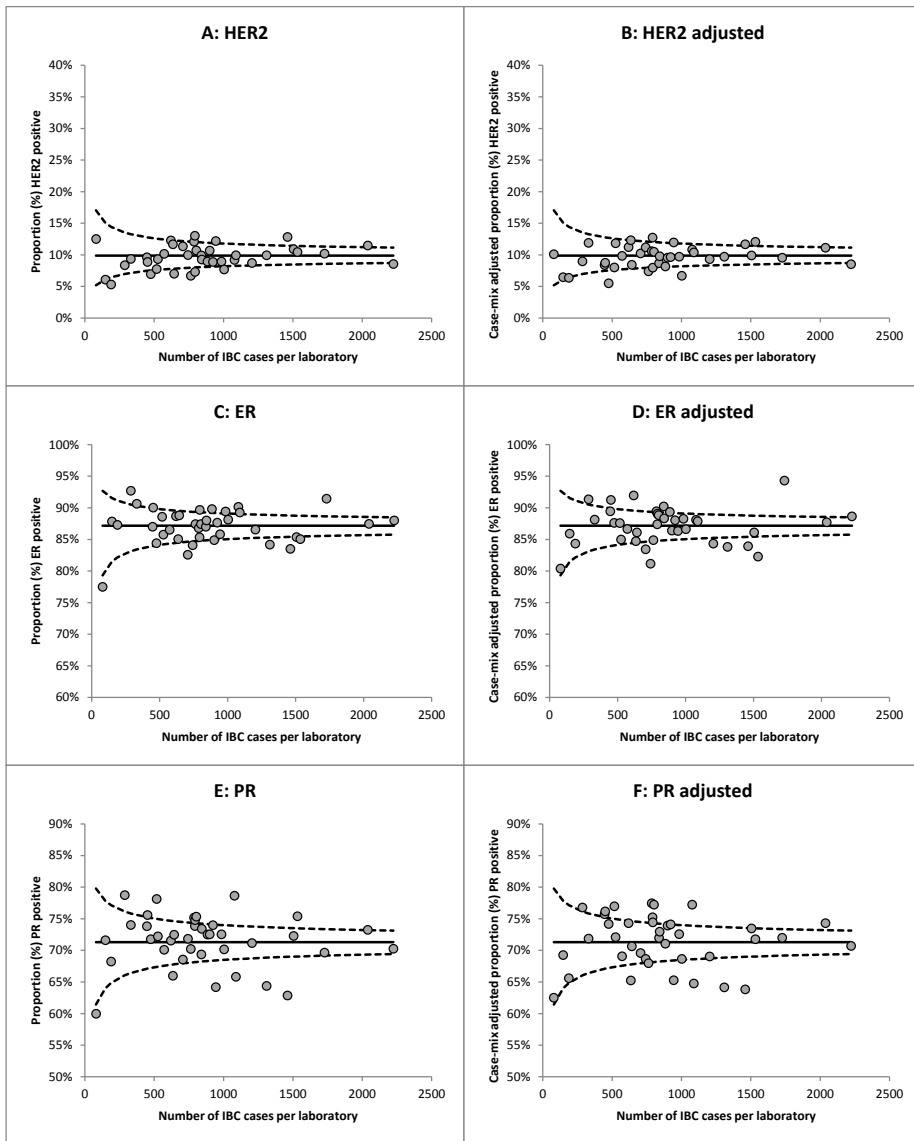
Figure 2. Funnel plots showing the observed (A, C, E) and case-mix adjusted positivity rates (B, D, F) per laboratory (dots) relative to the mean national proportion and its 95% confidence intervals for, for HER2 (A-B), estrogen (C-D) and progesterone (E-F) (2013-2016).

Table 2. Responses of 13 laboratories to our survey on receptor assessment of invasive breast cancer

	n (%)		
	Total (n=13)	Academic (n=6)	Non-academic (n=7)
Testing on biopsy or resection specimen?			
Biopsy*	12 (92.3%)	6 (100.0%)	6 (85.7%)
Resection specimen**	1 (7.7%)	0 (0.0%)	1 (14.3%)
Both	3 (23.1%)	3 (50.0%)	0 (0.0%)
Used cutoff for ER- and PR- receptor positivity?			
≥1%	1 (7.7%)	1 (16.7%)	0 (0.0%)
≥10%	9 (69.2%)	4 (66.7%)	5 (71.4%)
≥11%	3 (23.1%)	1 (16.7%)	2 (28.6%)
Used techniques for HER2-receptor assessment**			
Immunohistochemistry (IHC)	13 (100.0%)	6 (100.0%)	7 (100.0%)
Fluorescence In Situ Hybridization (FISH)	7 (53.8%)	4 (66.7%)	3 (42.9%)
Silver In Situ Hybridization (SISH)	5 (38.5%)	3 (50.0%)	2 (28.6%)
Chromogenic In Situ Hybridization (CISH)	1 (7.7%)	0 (0.0%)	1 (14.3%)
Multiplex ligation-dependent probe amplification	1 (7.7%)	1 (16.7%)	0 (0.0%)
Next generation sequencing	1 (7.7%)	1 (16.7%)	0 (0.0%)
HER2-receptor: order of testing techniques			
IHC followed by amplification testing	10 (76.9%)	4 (66.7%)	6 (85.7%)
FISH followed by IHC when indicated	3 (23.1%)	2 (33.3%)	1 (14.3%)
Primary IHC HER2-test (n=10) scores:			
Score 0			
Reported as negative	10 (100.0%)	4 (100.0%)	6 (100.0%)
Score 1+			
Reported as negative	9 (90.0%)	4 (100.0%)	6 (100.0%)
Additional amplification test	1 (10.0%)	1 (25.0%)	0 (0.0%)
Score 2+			
Additional amplification test	10 (100.0%)	4 (100.0%)	6 (100.0%)
Score 3+			
Reported as positive	8 (80.0%)	3 (75.0%)	5 (83.3%)
Additional amplification test	2 (20.0%)	1 (25.0%)	1 (16.7%)

* Most laboratories repeat receptor assessment on resection specimen in case of a negative receptor status on biopsy

** Receptor assessment on biopsy only when requested by clinician

*** ≥1 answer possible

Results of survey

Thirteen of the 46 Dutch pathology laboratories (28.3%) responded to our online survey, of which six were academic laboratories (Table 2). All responding laboratories participated in mandatory external audits (SKML, NordiQC and/or UK-Neqas) and all IHC stainings were executed mechanically. The vast majority of responding laboratories (92.3%) currently performs receptor assessment on biopsy, which is usually only repeated on resection specimen in case of a negative staining. In accordance with the Dutch national guideline (1), all laboratories, except for one, use 10% as a cutoff for ER- and PR-positivity. The laboratory that uses a different cutoff percentage, i.e. 1%, was not included in our dataset, as they did not synoptically report on breast cancer during the study period. For HER2 testing, 23.1% of the responding laboratories uses an amplification test (FISH), possibly followed by IHC, as primary test. In addition, one academic laboratory performed amplification testing after any plus-score (i.e. 1+, 2+, 3+).

Discussion

We studied inter-laboratory variation in ER-, PR- and HER2-positivity rates in a nationwide cohort of 33.046 invasive breast cancer patients, using real-life data from synoptic pathology reports of the Dutch nationwide pathology registry (PALGA). The results of this study show that absolute differences of ER, PR and HER2-positivity rates between laboratories were reassuringly limited. However, the number of outlying laboratories after case-mix adjustment for ER (14/39), PR (17/39) and HER2 (11/39) clearly shows that there is still room for improvement.

Overall positivity rates were 87.2% for ER, 71.3% for PR and 9.9% for HER2, which, for ER- and PR, is in line with previous studies [48-50], whereas for HER2 this is somewhat lower than the percentages of 15%-25% that are often referred to [7, 14, 16, 26, 51-53]. Although we only included synoptic pathology reports, there is no reason to assume that our synoptic dataset may have been selective, since data from the Dutch Breast Cancer Audit (NBCA), which also holds data from narrative pathology reports, show similar receptor positivity rates [49]. Moreover, over 80% of (pre)malignant breast lesions are currently reported via the synoptic PALGA protocol by Dutch pathologists [54], which results in an increased overall completeness of reports [55] and it enables easy and error-free data extraction. This study stresses the potential of using a population-based registry as it provides information on the actual situation in daily clinical practice, which may differ from data derived from clinical trials, from smaller cohorts or even from neighboring countries.

It could be argued that positivity rates in this study may have been biased for several reasons. First, in case of a missing receptor status on resection specimen, the receptor status of the biopsy, when known from the resection pathology report, was included in the analysis. As, however, discrepancies of receptor status between biopsy and resection specimen are uncommon and, according to literature, results from the core biopsy can be used with confidence [56-60], there is no reason to assume that this has influenced our positivity rates. Secondly, we excluded pathology reports with a missing ER, PR or HER2-receptor status (n=2,462). For the majority of these missing values (~70-75%), the tumor receptor status was reported as '*in progress*'. As IHC staining usually takes overnight, the definitive receptor status may have been added as a narrative addendum to the pathology report afterwards, yet not to the synoptic PALGA protocol, and therefore it is unknown in this dataset. However, it is unlikely that this happens more often to receptor-positive than to receptor-negative tumors. For the remaining 735 reports with missing values the reason remained unknown.

Overall, receptor positivity rates of individual laboratories were compared to the mean national positivity rates, with and without correction for case-mix. Case-mix adjustment only slightly narrowed the range of positivity rates between laboratories, which indicates that there is either little variation in case-mix per laboratory in the Netherlands, or there is little effect of the included case-mix variables. Either way, case-mix does not explain the inter-laboratory differences in this study. In addition, as laboratories with both few and many reports showed positivity rates outside the 95% CI (Fig. 2), laboratory sample size also does not explain the inter-laboratory variation that was found in this study. Furthermore, variation between individual pathologists within laboratories was minimal, which suggest that factors other than pathologists' interpretation of the fixed and immunohistochemically stained tissue slides, may explain the inter-laboratory differences in receptor positivity rates. One could for example think of different ways of tissue fixation or the use of different antibodies between laboratories.

Despite the low response-rate of our survey (13/46 laboratories), it did show that, in spite of a clear national guideline, one of the thirteen responding laboratories uses a different positivity threshold for ER and PR, which is undesirable as this would result in different therapy advice in our country, even if two laboratories would estimate the same percentage of ER or PR stained nuclei. As all laboratories in this study are anonymous, the results of our survey could not be linked to the laboratories in the dataset. Therefore it remains unknown whether the use of different positivity thresholds (1% versus 10%) may (partially) cause the found inter-laboratory variation. However, we do know that only a fairly small proportion of patients shows 'arguable' staining percentages between 1% and 10%. For both ER and PR, the percentage of staining was known in approximately 80% of

reports and of those reports, 1.3% showed ER percentages between 1% and 10%, whereas this was the case for 7.5% for PR. Therefore, the overall influence of (possible) different cutoff percentages is probably be limited.

A nationwide multidisciplinary breast cancer audit (NBCA) has already been implemented in breast cancer care in the Netherlands [49], yet currently there is only one pathology indicator, i.e. whether the PALGA protocol is used for reporting on (pre)malignant breast lesions [61]. We believe that it is important to use this synoptic PALGA protocol to monitor and benchmark the major pathology breast cancer biomarkers, namely ER, PR, HER2 and histologic grade, as these are crucial in decision making in current clinical practice [1]. Although molecular or genetic measures of prognosis may become increasingly important in IBC risk stratification in the near future, the only three mandatory breast cancer biomarkers are still ER-, PR- and HER2-receptor status, despite the massive investment of time and money into development of new biomarkers [4]. What is more, Groenendijk et al. [62] showed that the distribution of genomic risk is mainly influenced by histologic grade and ER- and HER2-status, which shows that these classic biomarkers remain very relevant. Given their prominent role in clinical practice, it seems worthwhile to invest in better and more uniform assessment of these classic biomarkers.

We believe that creating insight and awareness in variation of clinically relevant biomarkers through annual individual pathology "mirror" reports is an important step towards improvement of breast cancer care. Monitoring of receptor positivity rates may help to identify laboratories with a high number of false-positive or false-negative results [7, 14, 16, 37, 38] that are not picked up by the external audits, since crucial steps like tissue fixation and processing are not covered by these tests [36]. Furthermore, pathologists and their laboratories may feel best addressed by their own, case-mix adjusted, "mirror" data visualized against other national laboratories. Indeed, in a previous nationwide breast cancer audit, a HER2-outlier hospital critically evaluated their lab process and found that they used a different approach to HER2-positivity [49].

In conclusion, this nationwide study shows that there is limited absolute variation in ER, PR and HER2-receptor positivity rates between Dutch pathology laboratories in daily clinical practice. Yet, the considerable number of outlying laboratories shows that there is still room for improvement. Continuous monitoring and benchmarking of positivity rates may help to realize this and has been implemented in the Netherlands.

Acknowledgements

2

References

1. The Netherlands Comprehensive Cancer Organization (IKNL). Breast Cancer Guideline 2017 [updated 2012-07-02]. Available from: https://www.oncoline.nl/borstkanker. Accessed 17 January 2019.

2. Coates AS, Winer EP, Goldhirsch A, et al. Tailoring therapies - improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. Ann Oncol: 2015;26(8):1533-46.

3. McCullough AE, Dell'orto P, Reinholz MM, et al. Central pathology laboratory review of HER2 and ER in early breast cancer: an ALTTO trial [BIG 2-06/NCCTG N063D (Alliance)] ring study. Breast Cancer Res Treat. 2014;143(3):485-492.

4. Duffy MJ, Harbeck N, Nap M, et al. Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). Eur J Cancer. 2017;75:284-98.

5. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet. 2005;365(9472):1687-717.

6. Davies C, Godwin J, Gray R, et al. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet. 2011;378(9793):771-84.

7. Choritz H, Busche G, Kreipe H. Quality assessment of HER2 testing by monitoring of positivity rates. Virchows Arch. 2011;459(3):283-9.

8. Burstein HJ, Temin S, Anderson H, et al. Adjuvant Endocrine Therapy for Women With Hormone Receptor–Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. J Clin Oncol 2014;32(21):2255-69.

9. Senkus E, Kyriakides S, Ohno S, et al on behalf of the ESMO Guidelines Committee. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. Ann Oncol. 2015; 26(suppl_5):v8-v30.

10. Hammond MEH, Hayes DF, Dowsett M, et al. American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer. J Clin Oncol. 2010;28(16):2784-95.

11. Bowles EJA, Wellman R, Feigelson HS, et al. Risk of Heart Failure in Breast Cancer Patients After Anthracycline and Trastuzumab Treatment: A Retrospective Cohort Study. J Natl Cancer Inst. 2012 104(17):1293-305.

12. Perez EA, Suman VJ, Davidson NE, et al. Cardiac Safety Analysis of Doxorubicin and Cyclophosphamide Followed by Paclitaxel With or Without Trastuzumab in the North Central Cancer Treatment Group N9831 Adjuvant Breast Cancer Trial. J Clin Oncol. 2008;26(8):1231-8.

13. Romond EH, Jeong JH, Rastogi P, et al. Seven-year follow-up assessment of cardiac function in NSABP B-31, a randomized trial comparing doxorubicin and cyclophosphamide followed by paclitaxel (ACP) with ACP plus trastuzumab as adjuvant therapy for patients with node-positive, human epidermal growth factor receptor 2-positive breast cancer. J Clin Oncol 2012;30(31):3792-9.

14. Rüschoff J, Lebeau A, Kreipe H, et al. Assessing HER2 testing quality in breast cancer: variables that influence HER2 positivity rate from a large, multicenter, observational study in Germany. Mod

Pathol. 2016;30:217.

15. Russell SD, Blackwell KL, Lawrence J, et al. Independent Adjudication of Symptomatic Heart Failure With the Use of Doxorubicin and Cyclophosphamide Followed by Trastuzumab Adjuvant Therapy: A Combined Review of Cardiac Data From the National Surgical Adjuvant Breast and Bowel Project B-31 and the North Central Cancer Treatment Group N9831 Clinical Trials. J Clin Oncol. 2010; 28(21):3416-21.

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

17. Cuadros M, Cano C, Lopez FJ, et al. HER2 status in breast cancer: experience of a Spanish National Reference Centre. Clin Transl Oncol. 2011;13(5):335-40.

18. Denkert C, Huober J, Loibl S, et al. HER2 and ESR1 mRNA expression levels and response to neoadjuvant trastuzumab plus chemotherapy in patients with primary breast cancer. Breast Cancer Res. 2013;15(1):R11.

19. Orlando L, Viale G, Bria E, et al. Discordance in pathology report after central pathology review: Implications for breast cancer adjuvant treatment. Breast. 2016; 30:151-5.

20. Paik S, Bryant J, Tan-Chiu E, et al. Real-world performance of HER2 testing--National Surgical Adjuvant Breast and Bowel Project experience. J Natl Cancer Inst. 2002; 94(11):852-4.

21. Perez EA, Suman VJ, Davidson NE, et al. HER2 testing by local, central, and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial. J Clin Oncol. 2006;24(19):3032-8.

22. Roche PC, Suman VJ, Jenkins RB, et al. Concordance between local and central laboratory HER2 testing in the breast intergroup trial N9831. J Natl Cancer Inst. 2002; 94(11):855-7.

23. Rosa M, Khazai L. Comparison of HER2 testing among laboratories: Our experience with review cases retested at Moffitt Cancer Center in a two-year period. Breast J. 2017;24(2):139-47.

24. Viale G, Regan MM, Maiorano E, et al. Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1-98. J Clin Oncol. 2007;25(25):3846-52.

25. Bianchi S, Caini S, Paglierani M, et al. Accuracy and Reproducibility of HER2 Status in Breast Cancer Using Immunohistochemistry: A Quality Control Study in Tuscany Evaluating the Impact of Updated 2013 ASCO/CAP Recommendations. Pathol Oncol Res. 2015;21(2):477-85.

26. Dowsett M, Hanna WM, Kockx M, et al. Standardization of HER2 testing: results of an international proficiency-testing ring study. Mod Pathol. 2007;20:584.

27. Layfield LJ, Goldstein N, Perkinson KR, Proia AD. Interlaboratory Variation in Results from Immunohistochemical Assessment of Estrogen Receptor Status. Breast J. 2003;9(3):257-9.

28. Parker RL, Huntsman DG, Lesack DW, et al. Assessment of interlaboratory variation in the immunohistochemical determination of estrogen receptor status using a breast cancer tissue microarray. Am J Clin Pathol. 2002;117(5):723-8.

29. Regitnig P, Reiner A, Dinges HP, et al Quality assurance for detection of estrogen and progesterone receptors by immunohistochemistry in Austrian pathology laboratories. Virchows Archiv 2002; 441(4):328-34.

30. Rhodes A, Jasani B, Balaton AJ, et al.

Study of interlaboratory reliability and reproducibility of estrogen and progesterone receptor assays in Europe. Documentation of poor reliability and identification of insufficient microwave antigen retrieval time as a major contributory element of unreliable assays. Am J Clin Pathol. 2001; 115(1):44-58

31. Fitzgibbons PL, Murphy DA, Dorfman DM, et al. Interlaboratory comparison of immunohistochemical testing for HER2: results of the 2004 and 2005 College of American Pathologists HER2 Immunohistochemistry Tissue Microarray Survey. Arch Pathol Lab Med. 2006;130(10):1440-5

32. Rhodes A, Jasani B, Anderson E, et al. Evaluation of HER-2/neu immunohistochemical assay sensitivity and scoring on formalin-fixed and paraffin-processed cell lines and breast tumors: a comparative study involving results from laboratories in 21 countries. Am J Clin Pathol. 2002;118(3):408-17

33. Rudiger T, Hofler H, Kreipe HH, et al Quality assurance in immunohistochemistry: results of an interlaboratory trial involving 172 pathologists. Am J Surg Pathol. 2002;26(7):873-82.

34. von Wasielewski R, Mengel M, Wiese B, et al. Tissue Array Technology for Testing Interlaboratory and Interobserver Reproducibility of Immunohistochemical Estrogen Receptor Analysis in a Large Multicenter Trial. Am J Clin Pathol. 2002;118(5):675-82

35. von Wasielewski R, Hasselmann S, Ruschoff J, et al. Proficiency testing of immunohistochemical biomarker assays in breast cancer. Virchows Archiv. 2008; 453(6):537-43.

36. Tong LC, Nelson N, Tsourigiannis J, Mulligan AM. The effect of prolonged fixation on the immunohistochemical evaluation of estrogen receptor, progesterone receptor, and HER2 expression in invasive breast cancer: a prospective study. Am J Surg Pathol. 2011;35(4):545-52.

37. Rakha EA, Pinder SE, Bartlett JM, et al. Updated UK Recommendations for HER2 assessment in breast cancer. J Clin Pathol. 2105; 68(2):93-9

38. Lin CY, Carneal EE, Lichtensztajn DY, et al. Regional Variability in Percentage of Breast Cancers Reported as Positive for HER2 in California: Implications of Patient Demographics on Laboratory Benchmarks. Am J Clin Pathol. 2017;148(3):199-207

39. Casparie M, Tiebosch ATMG, Burger G, et al. Pathology Databanking and Biobanking in The Netherlands, a Central Role for PALGA, the Nationwide Histopathology and Cytopathology Data Network and Archive. Cell Oncol. 2007;29(1):19-24

40. Xian Z, Quinones AK, Tozbikian G, Zynger DL. Breast cancer biomarkers before and after neoadjuvant chemotherapy: does repeat testing impact therapeutic management? Hum Pathol. 2017;62:215-21.

41. Yang YF, Liao YY, Li LQ, et al. Changes in ER, PR and HER2 receptors status after neoadjuvant chemotherapy in breast cancer. Pathol Res Pract. 2013;209(12):797-802.

42. Yang L, Zhong X, Pu T, et al. Clinical significance and prognostic value of receptor conversion in hormone receptor positive breast cancers after neoadjuvant chemotherapy. World J Surg Oncol. 2018;16(1):51.

43. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.

44. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med. 2005;24(8):1185-202.

45. Huang HJ, Neven P, Drijkoningen M, et al. Association between tumour characteristics and HER-2/neu by immunohistochemistry in 1362 women with primary operable breast cancer. J Clin Pathol. 2005; 58(6):611-6.

46. Sullivan PS, Apple SK. Should histologic type be taken into account when considering neoadjuvant chemotherapy in breast carcinoma? Breast J. 2009;15(2):146-54.

47. Tubiana-Hulin M, Stevens D, Lasry S, et al. Response to neoadjuvant chemotherapy in lobular and ductal breast carcinomas: a retrospective study on 860 patients from one institution. Ann Oncol. 2006; 17(8):1228-33.

48. Parise CA, Caggiano V. Breast Cancer Survival Defined by the ER/PR/HER2 Subtypes and a Surrogate Classification according to Tumor Grade and Immunohistochemical Biomarkers. J Cancer Epidemiol. 2014; 469251.

49. van Bommel AC, Spronk PE, Vrancken Peeters MT, et al. Clinical auditing as an instrument for quality improvement in breast cancer care in the Netherlands: The national NABON Breast Cancer Audit. J Surg Oncol. 2017;115(3):243-9.

50. van Deurzen CH. Predictors of Surgical Margin Following Breast-Conserving Surgery: A Large Population-Based Cohort Study. Ann Surg Oncol. 2016;23(Suppl 5):627-33.

51. Hynes NE, Stern DF. The biology of erbB-2/ neu/HER-2 and its role in cancer. Biochim Biophys Acta. 1994;1198(2-3):165-84.

52. Owens MA, Horten BC, Da Silva M. HER2 amplification ratios by fluorescence in situ hybridization and correlation with immunohistochemistry in a cohort of 6556 breast cancer tissues. Clinical Breast Cancer. 2004 5(1):63-9.

53. Ross JS. Breast cancer biomarkers and HER2 testing after 10 years of anti-HER2 therapy. Dru News Perspect. 2009; 22(2):93-106.

54. Foundation PALGA. Annual report 2017. Available from: https://www. palga.nl/assets/uploads/Jaarverslag/ Jaarverslag_2017_websiteversie.pdf. Accessed January 17 2019

55. Sluijter CE, van Lonkhuijzen LR, van Slooten HJ, et al. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows Archiv. 2016;468(6):639-49.

56. Arnedos M, Nerurkar A, Osin P, et al. Discordance between core needle biopsy (CNB) and excisional biopsy (EB) for estrogen receptor (ER), progesterone receptor (PgR) and HER2 status in early breast cancer (EBC). Ann Oncol. 2009;20(12):1948-52.

57. Clark BZ, Onisko A, Assylbekova B, et al. Breast cancer global tumor biomarkers: a quality assurance study of intratumoral heterogeneity. Mod Pathol 2018; doi: 10.1038/s41379-018-0153-0. [Epub ahead of print]

58. Chen X, Yuan Y, Gu Z, Shen K. Accuracy of estrogen receptor, progesterone receptor, and HER2 status between core needle and open excision biopsy in breast cancer: a meta-analysis. Breast Cancer Res Treat. 2012;134(3):957-67.

59. Dekker TJ, Smit VT, Hooijer GK, et al. Reliability of core needle biopsy for determining ER and HER2 status in breast cancer. Ann Oncol 2013;24(4):931-7.

60. Tsuda H, Kurosumi M, Umemura S, et al. HER2 testing on core needle biopsy specimens from primary breast cancers: interobserver reproducibility and concordance with surgically resected specimens. BMC Cancer. 2010;10:534.

61. Dutch Institute for Clinical Auditing (DICA)/ Netherlands Comprehensive Cancer Organisation/ NABON Breast Cancer Audit (NBCA) (2018) Factsheet Indicators NABON Breast Cancer Audit (NBCA) 2018, version 2018.3.

62. Groenendijk FH, Jager A, Cardoso F, van Deurzen CHM. A nationwide registry-based cohort study of the MammaPrint genomic risk classifier in invasive breast cancer. Breast. 2018;38:125-31.

3

Significant inter- and intra-laboratory variation in grading of ductal carcinoma in situ of the breast: a nationwide study of 4,901 patients in the Netherlands

Carmen van Dooijeweert
Paul J. van Diest
Stefan M. Willems
Chantal. C. H. J. Kuijpers
Lucy I. H. Overbeek
Ivette A. G. Deckers

Abstract

Purpose

A considerable part of ductal carcinoma *in situ* (DCIS) lesions may never progress into invasive breast cancer. However, standard treatment consists of surgical excision. Trials aim to identify a subgroup of low-risk DCIS patients that can safely forgo surgical treatment based on histologic grade, which highlights the importance of accurate grading. Using real-life nationwide data, we aimed to create insight and awareness in grading variation of DCIS in daily clinical practice.

Methods

All synoptic pathology reports of pure DCIS resection specimens between 2013-2016 were retrieved from PALGA, the nationwide Dutch Pathology Registry. Absolute differences in proportions of grade I-III were visualized using funnel plots. Multivariable analysis was performed by logistic regression to correct for case-mix, providing odds ratios (ORs) and 95% confidence intervals (CI) for high-grade (III) versus low-grade (I-II) DCIS.

Results

4,952 DCIS reports from 36 laboratories were included, of which 12.5% were reported as grade I (range 6.1-24.4%), 39.5% as grade II (18.2-57.6%), and 48.0% as grade III (30.2-72.7%). After correction for case-mix, 14 laboratories (38.9%) reported a significantly lower (n=4) or higher (n=10) proportion of high-grade DCIS than the reference laboratory. Adjusted ORs (95%CI) ranged from 0.52 (0.31-0.87) to 3.83 (1.42-10.39). Significant grading differences were also observed among pathologists within laboratories.

Conclusion

In this cohort of 4,901 patients we observed substantial inter- and intra-laboratory variation in DCIS grading, not explained by differences in case-mix. Therefore, there is an urgent need for nationwide standardization of grading practices, especially since the future management of DCIS may alter significantly depending on histologic grade.

Introduction

Ductal carcinoma *in situ* (DCIS) of the breast is generally considered a precursor of invasive ductal carcinoma [1], although it is unknown which proportion of untreated DCIS lesions will progress into invasive breast cancer [2]. In fact, it is believed that a considerable part of DCIS patients is treated for lesions that may never progress into invasive breast cancer [3-5]. Nevertheless, standard treatment of DCIS currently consists of either mastectomy or breast conserving surgery followed by radiotherapy and/or followed by endocrine therapy [6]. In addition, standard treatment decisions are made regardless of histologic grade, while progression risk or at least speed of progression is higher for high grade lesions [7,8].

To counteract this presumed overtreatment, four clinical trials currently aim to identify a subgroup of low-risk DCIS patients that, under active surveillance, could safely forgo surgical treatment [3,4,9,10]. All these trials aim to identify this subgroup solely based on histologic grade, hence, histologic grading of DCIS might be of great clinical importance in the near future. This perspective highlights the key importance of accurate, consistent, and reproducible grading by pathologists. However, current evidence suggests that there is considerable variation in the grading of DCIS in daily clinical practice. Previous studies in which several pathologists reviewed a set of DCIS cased show that the classification of DCIS is associated with significant inter-observer variation [11-13]. Moreover, various classifications are used to subdivide DCIS into lesions of good (grade I), moderate (grade II) and poor (grade III) differentiation. Although previous studies showed that there is considerable variation in grading, individual practicing pathologists may not have been influenced by these data as it did not provide them insight into their own grading practice. Moreover, grading was performed in a study setting, which may not resemble grading in real-life clinical practice. In this context, studies in a nationwide cohort of (pre)malignant colorectal lesions in real-life daily clinical practice by Kuijpers et al showed considerable inter-laboratory differences in the histologic grading of both colorectal adenomas and adenocarcinomas [14,15](23, 24). Therefore, we expected substantial variation between pathology laboratories, as well as individual pathologists in the grading of breast DCIS in daily clinical practice. To create insight and awareness in grading variation of DCIS, especially with the potential future treatment consequences in mind, we studied the laboratory-specific variation in histologic grading of DCIS in a nationwide daily clinical practice study.

Using the Dutch nationwide pathology registry (PALGA), we assessed the variation in histologic grading of nearly 5,000 patients with DCIS between Dutch pathology laboratories and between individual pathologists using data from synoptic (structured) pathology reports from real-life daily pathology practice. Furthermore, we conducted a questionnaire among pathologists to gain insight into their histologic grading practices.

Creating insight into these laboratory-specific differences may help design an intervention to improve standardization among laboratories and pathologists. This may ultimately enable more accurate risk stratification of patients with low-risk DCIS, which is highly relevant since the future management of DCIS may alter significantly depending on histologic grade.

Materials and Methods

Data source and study population
Data were extracted from the PALGA database, the nationwide registry of histo- and cytopathology in The Netherlands, which contains excerpts of all pathology reports from Dutch pathology laboratories since 1991 [16]. Data from the PALGA database are pseudonymised by a trusted third party (ZorgTTP, Houten, the Netherlands). All Dutch laboratories gave consent for the storage of their data in the PALGA database, including scientific use of these data. Additional consent was obtained for analysis of inter-pathologist variation within the individual laboratories. All laboratories were anonymized. The scientific and privacy committee of PALGA approved this study.

We retrieved all synoptic pathology reports of DCIS resection specimens without the presence of any coexistent invasive component between January 1, 2013 and December 31, 2016 in The Netherlands. Synchronous DCIS was defined as an ipsilateral DCIS lesion within six months of the first DCIS diagnosis. These lesions were considered paired measurements of which only the first was included. We excluded DCIS pathology reports of re-excisions and reports of residual *in situ* lesions after neoadjuvant treatment of primary invasive tumors.

In total, 39 out of 46 Dutch laboratories synoptically reported DCIS on breast resection specimens. Of these, we included those that synoptically reported ≥75 DCIS during the study period. Similarly, for inter-pathologist variation within individual laboratories, we analyzed only data from pathologists who synoptically reported ≥10 DCIS during the study period. Lastly, reports of DCIS with unknown tumor grade or unknown tumor size were excluded from further data-analysis.

From each report we extracted patients' sex and age, DCIS size, histologic grade, type of surgery, and date of diagnosis.

Histologic grading

The primary outcome measure of this study was the inter-laboratory variation in histologic grading of DCIS. DCIS grade III was considered as high-grade, whereas DCIS grades I and II were considered low-grade in our multivariate inter-laboratory analysis. Secondary outcome measure was the inter-pathologist variation in histologic grading within the individual laboratories.

Questionnaire among pathologists

A questionnaire was sent to all pathology laboratories in The Netherlands to identify how pathologists determine the histologic grade of DCIS in daily practice. The questionnaire contained questions on whether pathologists consider themselves as specialized breast pathologists, the number of years of experience as a pathologist, which classification system they use for grading DCIS, and how they deal with heterogeneity of histologic grade within one specimen.

Statistical analysis

The overall proportions of histologic grades I, II, and III DCIS were determined and considered the national proportion. Absolute differences in proportion of histologic grades between laboratories are presented in funnel plots per grade, in which the proportions per laboratory were plotted against the number of included DCIS per laboratory. The target of these funnel plots was set at the national proportions with 95% confidence intervals (CI) as limits [17].

To compare relative differences among laboratories, odds ratios (OR) and 95% CIs per laboratory were calculated by logistic regression. For the choice of the reference laboratory, the sum-score of absolute deviations from the grade-specific national proportions was calculated to compare the absolute deviation for all three grades at once. The laboratory with the lowest sum-score was deemed best for resembling the national distribution and was therefore chosen as the reference laboratory.

Patient and tumor characteristics were summarized and differences between histologic grades were tested by means of a χ^2-test for categorical variables and by a non-parametric Kruskal Wallis test for continuous variables.

A multivariate logistic regression analysis was performed to correct for differences in case-mix. To identify these potential confounding factors, we selected clinicopathological variables *a priori* based on literature [18-20] and on pathologists' experience, namely age, sex, tumor size, type of surgery, and year of diagnosis. Only tumor size and type of surgery appeared to be significantly associated with grade and were therefore

included in the final multivariate model. Since differences between the univariate and multivariate model were limited, only the adjusted ORs (95% CI) are presented in a forest plot.

For analysis of the inter-pathologist variation within the individual laboratories, we merely compared the proportions per histologic grade between pathologists by Fisher exact test. Results of the questionnaire were summarized by frequencies and percentages. P-values below 0.05 were considered statistically significant. All statistical analyses were performed by using IBM SPSS Statistics version 21.

Results

### *Characteristics of patients, DCIS lesions, and laboratories*
A total of 4,952 DCIS lesions of 4,901 patients from 36 laboratories were included in our final data-analysis (Fig. 1). Characteristics of these included patients and corresponding DCIS are listed in Table 1. Mean (SD) age at diagnosis was 59.5 (10.1) years and patients were predominantly female (99.8%). The majority of patients underwent breast conserving surgery (67.0%). Both tumor size and mastectomy rate increased with histologic grade.

The number of synoptically reported DCIS lesions per laboratory ranged from 22 to 324 (median 109). Overall national proportions for DCIS grades I, II and III were 12.5%, 39.5% and 48.0%, respectively.

Inter-laboratory differences in histologic grading
Laboratories varied mostly in DCIS grade III (range: 30.2%-72.7%), followed by DCIS grade II (range: 18.2%-57.6%) and DCIS grade I (range: 6.1%-24.4%). Overall, half of the laboratories (18/36) showed proportions outside the 95% CI for grade III, indicating that these laboratories graded significantly different from the national proportion, which was similar for grade II (17 laboratories). In contrast, for grade I, only eight laboratories graded significantly different from the national proportion (Fig. 2).
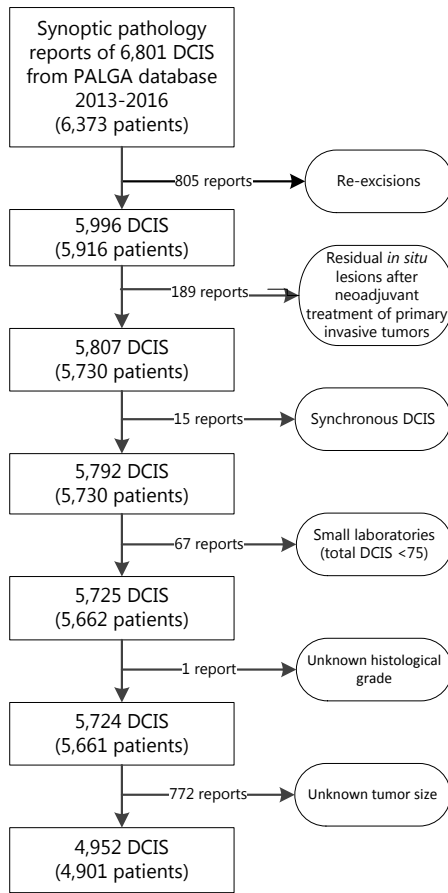
Figure 1. Flowchart of included cases of ductal carcinoma *in situ* (DCIS) of the breast to assess histopathologic grading variation between laboratories.

Table 1. Characteristics of the 4,952 included cases of ductal carcinoma *in situ* (DCIS) of the breast

	Total (n=4,592)	Grade 1 (n=618)	Grade 2 (n=1,956)	Grade 3 (n=2,378)	p-value
Age (y)*	59.5 (10.1)	58.5 (10.2)	59.9 (10.1)	59.3 (10.1)	0.001
Sex, n (%)					**
Female	4,944 (99.8%)	616 (99.7%)	1,950 (99.7%)	2,378 (100.0%)	
Male	8 (0.2%)	2 (0.3%)	6 (0.3%)	0 (0.0%)	
Tumor size (cm)*	2.4 (2.2)	1.6 (1.8)	2.1 (2.0)	2.9 (2.3)	0.000
Type of surgery, n (%)					0.000
Mastectomy	1,636 (33.0%)	130 (21.0%)	538 (27.5%)	968 (40.7%)	
Breast conserving	3,316 (67.0%)	488 (79.0%)	1,418 (72.5%)	1,410 (59.3%)	
Year of diagnosis, n (%)					0.518
2013	1,051 (21.2%)	125 (20.2%)	438 (22.4%)	488 (20.5%)	
2014	1,157 (23.4%)	147 (23.8%)	447 (22.9%)	563 (23.7%)	
2015	1,298 (26.2%)	161 (26.1%)	489 (25.0%)	648 (27.2%)	
2016	1,446 (29.2%)	185 (29.9%)	582 (29.8%)	679 (28.6%)	

* Mean (SD) ** Number of males too low for statistical testing

The sum-score was lowest and only 0.95% for laboratory 13, which was therefore chosen as reference laboratory. The maximum sum-score, in contrast, was 49.4%. Multivariate logistic regression showed that 14 laboratories (38.9%) reported a significantly higher (n=10) or lower (n=4) proportion of high grade DCIS than the reference laboratory (Fig. 3). For 2 laboratories (laboratories 27 and 29) the conclusion on multivariate analysis differed from the conclusion on univariate logistic regression analysis. The ORs of these two laboratories became significantly deviant. Adjusted ORs (95% CI) of individual laboratories for high- versus low-grade DCIS ranged from 0.52 (0.31-0.87) to 3.83 (1.42-10.39).

Intra-laboratory differences in histologic grading

Forty-six pathologists from eight analyzed laboratories synoptically reported ≥10 DCIS during the study period. Per laboratory, the number of analyzed pathologists ranged from 2 to 10 (median 6). In addition, the number of analyzed DCIS per pathologist ranged from 10 to 88 (median 15.5). Overall, 14 pathologists (28.3%) graded significantly deviant compared to the national proportions for grade II and III DCIS, while this was the case for ten pathologists (21.7%) for grade I (Fig. 4). Together, pathologists within individual laboratories differed mostly in the reporting of grade II (range: 15.4-76.9%) and grade III (11.8-69.2%), both in laboratory 2. In contrast, the maximum variation for DCIS grade I was considerably smaller (range: 6.3-39.4%) and was found in laboratory 26.

The differences in the distribution of histologic grade of DCIS among pathologists within the laboratories were significant in two laboratories (laboratories 1 and 2; Supplementary 1).

Results of questionnaire

Seventy-nine (25%) pathologists (out of the approximately 320 practicing pathologists in The Netherlands [14]) responded to our online questionnaire, of which 37 (46.8%) were reportedly specialized breast pathologists. Grading practice of general- and specialized breast pathologists did not seem to differ. Pathologists reported numerous different guidelines, articles and books as a reference for the histologic grading of DCIS, with most pathologists (35.4%) using the guideline of Holland et al [21]. Sixteen pathologists (20.3%) stated that they (partially) grade DCIS based on intuition. The majority of pathologists (76.0%) graded DCIS of heterogeneous differentiation based on the highest grade (Table 2).
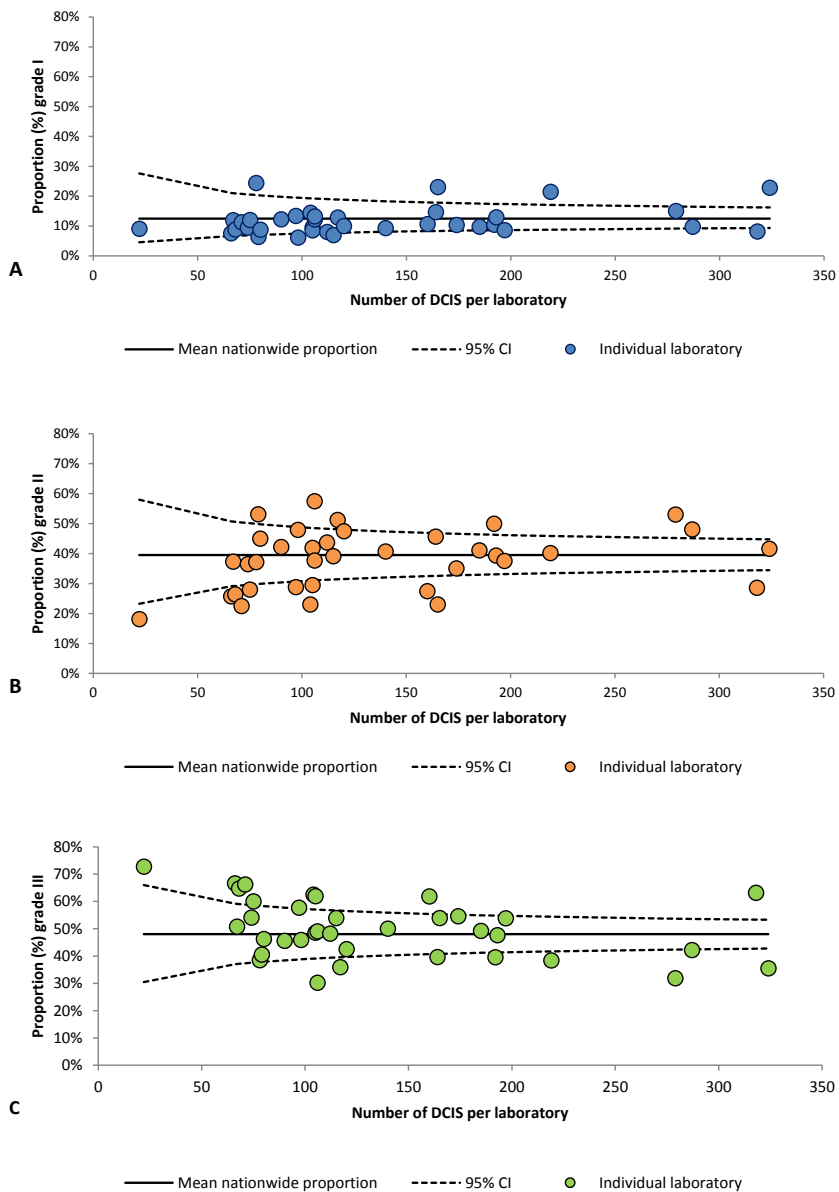
Figure 2. Funnel plots showing the observed proportion of ductal carcinoma *in situ* (DCIS) of the breast per grade per laboratory (dots) relative to the national proportion and its 95% confidence intervals for DCIS grades I (A), II (B) and III (C) (2013-2016).

3
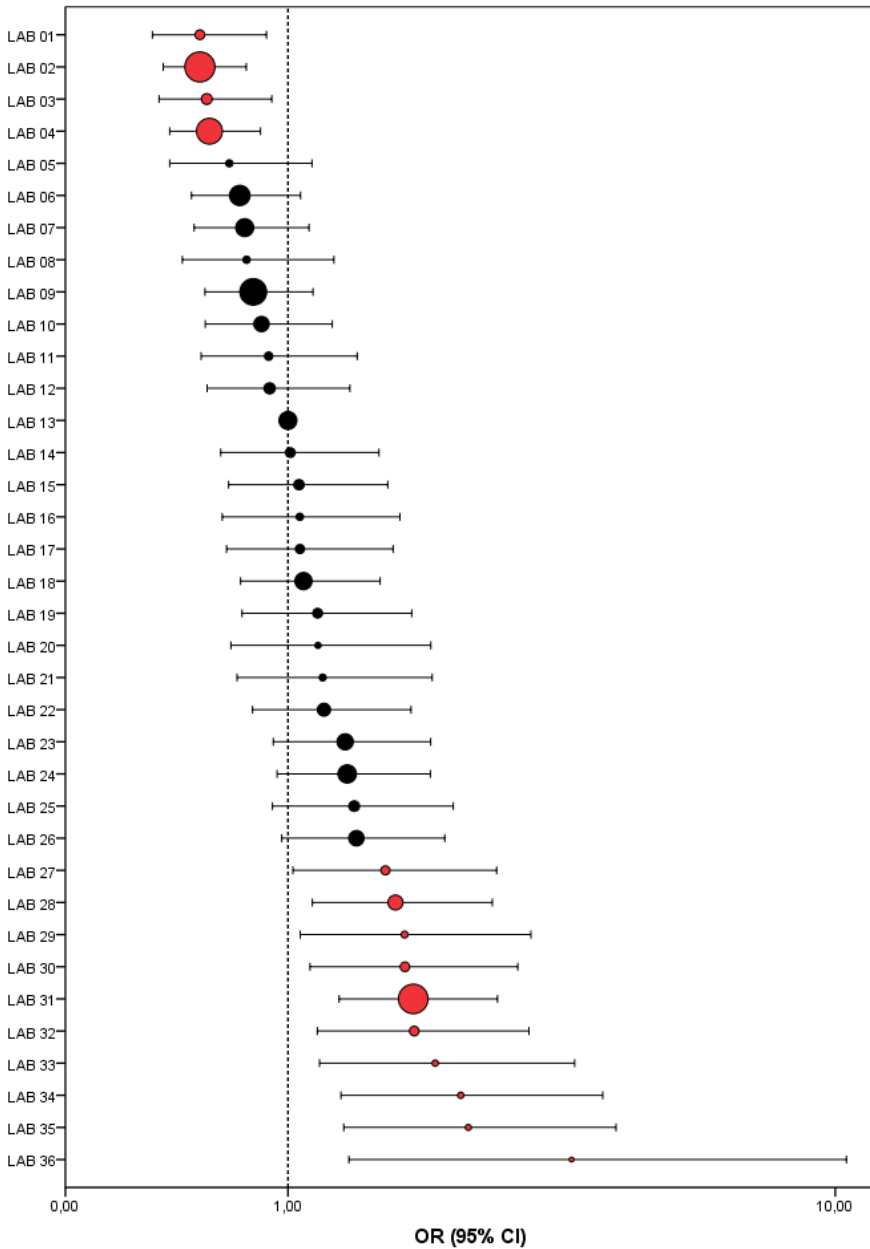
Figure 3. Forest plot showing the adjusted odds ratios (OR) and 95% confidence intervals (CI) of high grade ductal carcinoma *in situ* (DCIS) of the breast (grade III) versus low grade DCIS (grade I-II) per laboratory in comparison to the reference laboratory (#13). Dot size indicates the total number of analyzed synoptically reported DCIS cases per laboratory. Red dots indicate laboratories with a significantly deviant OR as compared to the reference laboratory.
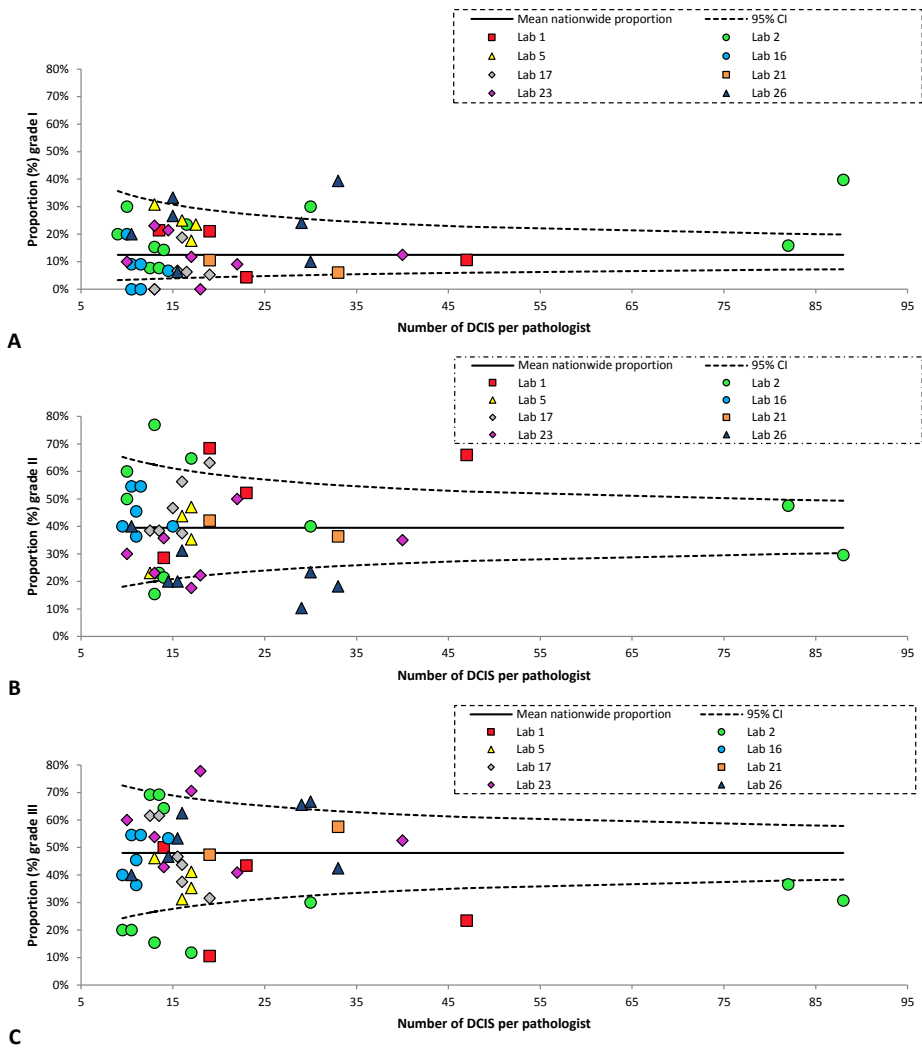
Figure 4. Funnel plots showing the observed proportion of ductal carcinoma *in situ* (DCIS) of the breast per grade per pathologist (dots) of eight laboratories relative to the national proportion for DCIS grades I (A), II (B) and III (C) (2013-2016).

Table 2. Results of 79 pathologists responding to our questionnaire on histologic grading of ductal carcinoma *in situ* (DCIS) of the breast

	n (%)		
	Total (n=79)	**Breast pathologist (n=37)**	**General pathologist (n=42)**
Laboratory			
Academic	15 (19.0%)	11 (29.7%)	4 (9.5%)
Peripheral	64 (81.0%)	26 (70.3%)	38 (90.5%)
Years of experience			
0-5	28 (35.4%)	11 (29.7%)	17 (40.5%)
6-10	15 (19.0%)	6 (16.2%)	9 (21.4%)
11-20	17 (21.5%)	9 (24.3%)	8 (19.0%)
>20	19 (24.1%)	11 (29.7%)	8 (19.0%)
Based on which guideline or reference do you grade DCIS?*			
Holland et al (1994) [25]	28 (35.4%)	14 (37.8%)	14 (33.3%)
Pinder et al (2010) [29]	16 (20.3%)	9 (24.3%)	7 (16.7%)
Intuition	16 (20.3%)	8 (21.6%)	8 (19.0%)
WHO [30]	11 (13.9%)	5 (13.5%)	6 (14.3%)
I don't know	4 (5.1%)	0 (0.0%)	4 (9.5%)
Tavassoli et al (Pathology of the Breast, 1999) [31]	3 (3.8%)	1 (2.7%)	2 (4.8%)
Van Nuys (Silverstein et al, 1995) [32]	2 (2.5%)	1 (2.7%)	1 (2.4%)
College of American Pathologists Guidelines [33]	1 (1.3%)	0 (0.0%)	1 (2.4%)
Combination (n.o.s.)	1 (1.3%)	1 (2.7%)	0 (0.0%)
How do you grade a DCIS of heterogeneous differentiation?			
Based on the highest grade	60 (76.0%)	28 (75.7%)	32 (76.2%)
I report the percentages of each grade	9 (11.4%)	4 (10.8%)	5 (11.9%)
Based on the predominant grade	7 (8.9%)	4 (10.8%)	3 (7.1%)
Not within the protocol	3 (3.8%)	1 (2.7%)	2 (4.8%)

* Multiple answers possible (n=82)
n.o.s. indicates not otherwise specified

3

Discussion

Using real-life daily clinical practice data from the nationwide pathology database PALGA, we studied the inter-laboratory variation in histopathologic grading of DCIS in daily clinical practice to create insight and awareness in grading variation, which is highly relevant as the future management of DCIS may alter significantly depending on histologic grade.

Approximately half of the lesions of this nationwide cohort of 4,952 DCIS were reported as grade III (48.0%), whereas 39.5% were reported as grade II and only 12.5% as grade I. The observed overall proportions per grade in this study are in line with previous studies of smaller cohorts of DCIS patients (n=853-1430), which showed similar distribution percentages for DCIS grades I (15-18%), II (32-39%) and III (42-54%) [19].

Laboratory-specific data was analyzed in an absolute and relative manner, comparing individual laboratories to both the national proportion and the reference laboratory. This indicates that inter-laboratory differences in histologic grading were substantial. This was highlighted by the substantial range of absolute proportions of histologic grade between laboratories, by the sum-score that varied up to 48.5%, by the number of laboratories with significantly deviant proportions from the national distribution per grade (~50% for grade II and III), and by the number of laboratories with significantly deviant case-mix adjusted ORs from the reference laboratory (~40%). Case-mix adjusted ORs (95%CI) ranged between 0.52 (0.31-0.87) and 3.83 (1.42-10.39), indicating that the chance of a DCIS being graded as high-grade in the 'lowest' laboratory is approximately two times lower than in the reference laboratory, and nearly four times higher in the 'highest' laboratory. Consequently, the difference between the 'lowest' and the 'highest' laboratory is even larger.

Although unlikely, since breast case-mix is not known to show regional differences in the Netherlands, we could not *a priori* exclude the possibility that grading practices of DCIS lesions may be influenced by patient- and tumor characteristics. Comparison of the results of univariate and multivariate regression analysis, however, showed that the substantial inter-laboratory variation is not influenced by the most important clinicopathologic variables. Nonetheless, other variables, like imaging or how the lesions were diagnosed could play a role. Unfortunately, these factors are not, or very rarely, documented in pathology reports. However, the Dutch breast cancer screening program refers patients randomly to all local Dutch hospitals and it is therefore unlikely to be an important factor in case-mix correction. This is further supported by the small effect of the case-mix variables that could be taken into account in the current study, which implies that regional differences in breast case mix are limited in the Netherlands. Another possible confounder that could not be analyzed is comedo necrosis, since this is not an obligatory variable in the synoptic reporting module

and was only available in a minority of cases. We have, however, no reason to assume that this feature is not evenly distributed between laboratories. Lastly, all laboratories were anonymized to the researchers, but we do know that five of the eight Dutch academic laboratories are represented in our dataset. There were no striking differences between academic and regional laboratories, but the incomplete dataset does not allow to draw firm conclusions.

For multivariable data analysis, histologic grade was dichotomized into low-grade DCIS (grade I-II) and high-grade DCIS (grade III), based on the definition of low-risk DCIS in the majority of current clinical trials [3,4,9,10]. Moreover, given the low proportion of grade I DCIS, most variation between laboratories was expected between grade II and III. In a sensitivity analysis, we were able to validate the main results of our logistic regression model in a multinomial regression model, which allows a multinomial endpoint (data not shown).

Data included in this study were merely from patients synoptically reported DCIS lesions, because synoptic reporting, compared to narrative reporting, results in improved reporting of relevant clinical data and an increased overall completeness of pathology reports [22]. In addition, all variables are stored in a standardized manner, which enables reliable and easy data extraction. Over 80 percent of (pre)malignant breast lesions is currently reported synoptically by pathologists via the PALGA protocol [23]. To check whether our case selection was likely to be representative for all DCIS patients, we compared our data with aggregated data from the Netherlands Cancer Registry, which also holds narrative reports, and observed a similar distribution of histologic grade (data not shown), indicating that our case selection based on synoptic reporting is likely to be representative.

Because we used data from a nationwide pathology database, we were able to extract reports of pure DCIS lesions, indicating that reports did not indicate the presence of any coexistent invasive component. This is important because grading of DCIS lesions may not be independent of the invasive component. For example, Farabegoli et al. showed that pure DCIS and DCIS associated with invasive ductal carcinoma may be genetically distinct [24]. It is unclear whether a previous history of invasive breast cancer might influence the grading of a later DCIS lesion, but it is unlikely that a pathologist being aware of the breast pathology history would interpret DCIS morphology differently. Nevertheless, the influence on our results would be limited as the proportion of patients with such history in our dataset was small (~5%).

In addition to the substantial inter-laboratory variation, significant differences were also observed between pathologists within two out of eight analyzed laboratories. These results emphasize that even within the laboratories analyzed, histologic grading is not performed

in a standardized manner. Moreover, this implies that the observed inter-laboratory variation may predominantly be the result of different grading practices of individual pathologists and not of differences in case-mix, as shown in the present study. The findings from our questionnaire, where numerous reference classifications were mentioned as a guideline by pathologists and 20% of pathologists even stated that they grade DCIS based on intuition, further illustrate that histologic grading is currently insufficiently standardized. This calls for (inter)national consensus on the grading system and criteria to improve reproducibility in view of the therapeutic consequences. About half the intuitive graders regarded themselves as breast pathologists, which calls for better criteria for this status.

The results of this study may raise awareness among pathologists, emphasizing that histologic grading of DCIS is currently not meeting high enough standards, which is an important first step to improvement. The fact that we needed a threshold of 10 to allow analysis of inter-pathologist variation indicates that perhaps too many pathologist engage in DCIS diagnosis. In addition, pathologists are enabled to discuss and reflect on their grading practices, as these "mirror" data were also send to the laboratories by PALGA. In this context, annual benchmarking of histologic grading of DCIS based on "mirror" PALGA data is already being considered by the Dutch Society for Pathology and may be adopted much broader in the field. Future research might focus on the development of an e-learning module to train pathologists in determining the histologic grade of DCIS, thereby aiming to attribute to the synchronization and better reproducibility of DCIS grading.

This improvement is especially relevant since the decision to manage low-risk DCIS trough active surveillance may be solely dependent on histologic grade in the near future. However, the risk stratification of all four clinical trials [3,4,9,10] based merely on histological grade is criticized by Toss et al [7], who emphasize that trial outcomes will be influenced by the inherent subjectivity of the current simple grading system and that DCIS risk stratification should be a combination of histologic grading and more objective biomarkers such as ER and HER2 [7], molecular markers, or deep learning strategies on digital images. In this context, standardization and synchronization of histologic grading will not only improve health care in general, but it might also improve risk stratification of low-risk DCIS and subsequently might improve (clinical) studies that take histologic grading into account.

In conclusion, both inter- and intra-laboratory results of this nationwide cohort of nearly 5,000 patients show that there is substantial variation in the histologic grading of DCIS by pathologists in routine daily clinical practice. This implies that there is an urgent need for improvement and better standardization of DCIS grading, especially since the future management of DCIS may alter significantly depending on histologic grade.
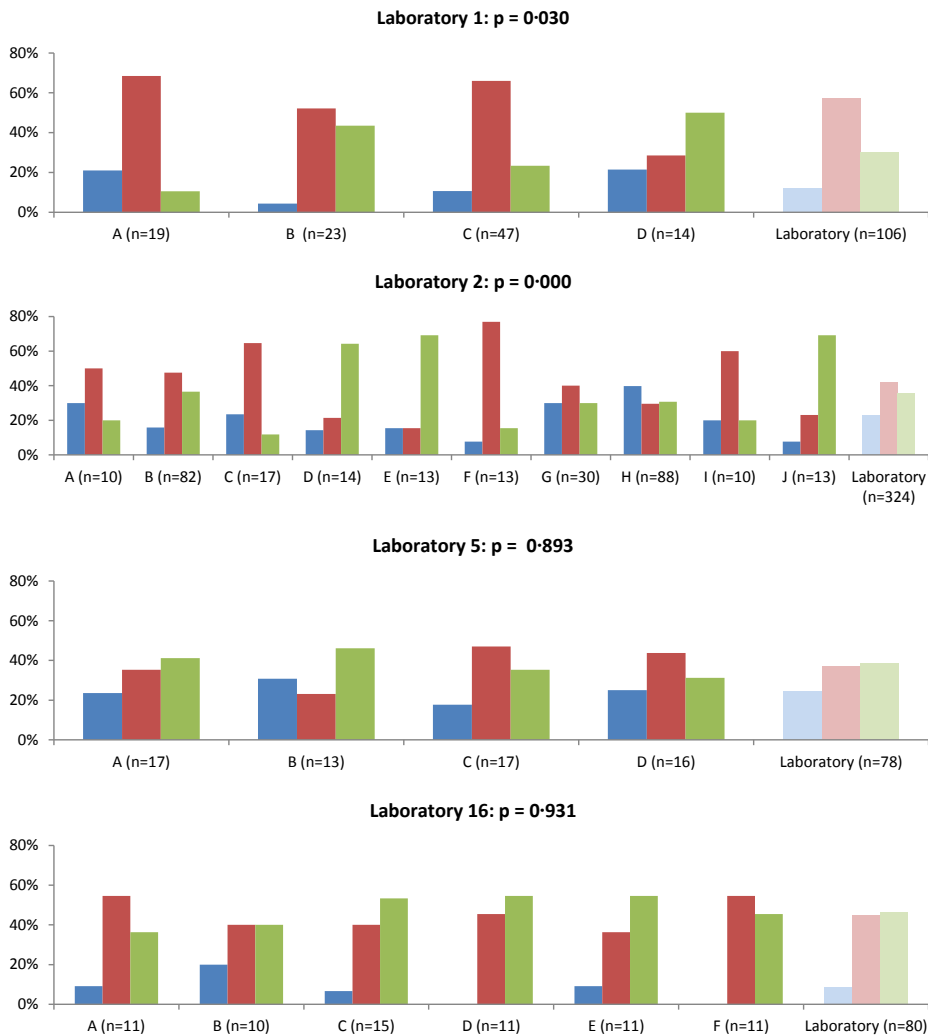
Acknowledgements

3

References

1. Pinder SE, Ellis IO. The diagnosis and management of pre-invasive breast disease: ductal carcinoma in situ (DCIS) and atypical ductal hyperplasia (ADH)- current definitions and classification. Breast Cancer Res. 2003; 5(5): 254-7.

2. Welch HG, Black WC. Using autopsy series to estimate the disease "reservoir" for ductal carcinoma in situ of the breast: how much more breast cancer can we find? Ann of Intern Med. 1997; 127(11): 1023-8.

3. Francis A, Thomas J, Fallowfield L, et al. Addressing overtreatment of screen detected DCIS; the LORIS trial. Eur J Cancer. 2015; 51(16): 2296-303.

4. 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(12): 1497-510.

5. Benson JR, Jatoi I, Toi M. Treatment of low-risk ductal carcinoma in situ: is nothing better than something? Lancet Oncol. 2016; 17(10): e442-e451.

6. Gorringe KL, Fox SB. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. Front Oncol. 2017; 7: 248.

7. Toss M, Miligy I, Thompson AM, et al. Current trials to reduce surgical intervention in ductal carcinoma in situ of the breast: Critical review. Breast. 2017; 35: 151-6.

8. Sanders ME, Schuyler PA, Dupont WD, Page DL. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. Cancer. 2005; 103(12): 2481-4.

9. Groen EJ, Elshof LE, Visser LL, et al. Finding the balance between over- and under-treatment of ductal carcinoma in situ

(DCIS). Breast. 2017; 31: 274-83.

10. Youngwirth LM, Boughey JC, Hwang ES. Surgery versus monitoring and endocrine therapy for low-risk DCIS: The COMET Trial. Bull Am Coll Surg. 2017; 102(1): 62-3.

11. 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(10): 705-11.

12. Elston CW, Sloane JP, Amendoeira I, et al. Causes of inconsistency in diagnosing and classifying intraductal proliferations of the breast. European Commission Working Group on Breast Screening Pathology. Eur J Cancer. 2000; 36(14): 1769-72.

13. Douglas-Jones AG, Morgan JM, Appleton MA, et al. Consistency in the observation of features used to classify duct carcinoma in situ (DCIS) of the breast. J Clin Pathol. 2000; 53(8): 596-602.

14. Kuijpers CC, Sluijter CE, von der Thüsen JH, et al. Interlaboratory Variability in the Histologic Grading of Colorectal Adenocarcinomas in a Nationwide Cohort. Am J Surg Pathol. 2016; 40(8): 1100-8.

15. Kuijpers CC, Sluijter CE, von der Thüsen JH, et al. Interlaboratory variability in the grading of dysplasia in a nationwide cohort of colorectal adenomas. Histopathology. 2016; 69(2): 187-97.

16. Casparie M, Tiebosch AT, Burger G, et al. Pathology databanking and biobanking in The Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. Cell Oncol. 2007; 29(1): 19-24.

17. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med. 2005; 24(8): 1185-202.

18. Liu H, Peng W. MRI morphological classification of ductal carcinoma in situ

(DCIS) correlating with different biological behavior. Eur J Radiol. 2012; 81(2): 214-7.

19. van Luijt PA, Heijnsdijk EA, Fracheboud J, et al. The distribution of ductal carcinoma in situ (DCIS) grade in 4232 women and its impact on overdiagnosis in breast cancer screening. Breast Cancer Res. 2016; 18(1): 47.

20. Weigel S, Hense HW, Heidrich J, Berkemeyer S, Heindel W, Heidinger O. Digital Mammography Screening: Does Age Influence the Detection Rates of Low-, Intermediate-, and High-Grade Ductal Carcinoma in Situ? Radiology. 2016; 278(3): 707-13.

21. Holland R, Peterse JL, Millis RR, et al. Ductal carcinoma in situ: a proposal for a new classification. Semin Diagn Pathol. 1994; 11(3): 167-80.

22. Sluijter CE, van Lonkhuijzen LR, van Slooten HJ, Nagtegaal ID, Overbeek LI. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows Archiv. 2016; 468(6): 639-49.

23. Foundation PALGA. Annual report 2017. Available from: https://www.palga.nl/assets/uploads/Jaarverslag/Jaarverslag_2017_websiteversie.pdf Accessed 1 September 2018.

24. Farabegoli F, Champeme MH, Bieche I, et al. Genetic pathways in the evolution of breast ductal carcinoma in situ. J Pathol. 2002; 196(3): 280-6.

25. 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. British Journal of Cancer. 2010; 103(1): 94-100.

26. Lakhani SR. WHO classification of tumours of the breast. 4th ed. Lyon: International Agency for Research on Cancer; 2012.

27. Tavasolli FA. Pathology of the Breast. 2nd ed. Maidenhead. McGraw-Hill Professional - Europe; 1999.

28. Silverstein MJ, Poller DN, Waisman JR, et al. Prognostic classification of breast ductal carcinoma-in-situ. Lancet. 1995; 345(8958): 1154-7.

29. College of American Pathologists. Protocol for the Examination of Specimens From Patients With Ductal Carcinoma in situ (DCIS) of the Breast 2017. Available from: http://www.cap.org/ShowProperty?nodePath=/UCMCon/Contribution%20Folders/WebContent/pdf/cp-breast-dcis-18protocol-4100.pdf. Accessed 1 September 2018.

3

Supplementary figure 1a. Intra-laboratory variation between pathologists in laboratories 1, 2, 5 and 16. Blue, red and green bars indicate the proportions of ductal carcinoma *in situ* (DCIS) grades I, II and III, respectively. Capital letters on the X-axis indicate pathologists within one laboratory, followed by the overall proportion of the laboratory per differentiation grade. Intra-laboratory differences are calculated by Fisher exact test (Monte Carlo option in laboratory 2).
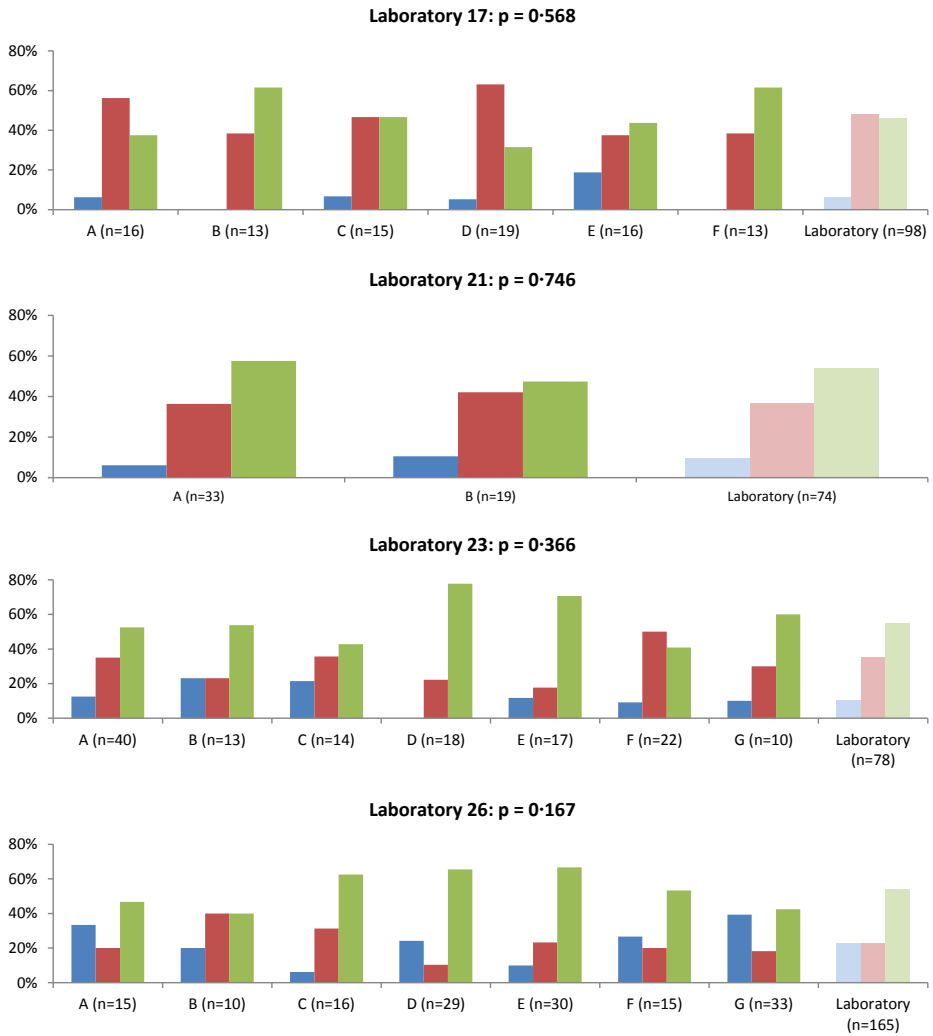
Supplementary figure 1b. Intra-laboratory variation between pathologists in laboratories 17, 21, 23 and 26. Blue, red and green bars indicate the proportions of ductal carcinoma *in situ* (DCIS) grades I, II and III, respectively. Capital letters on the X-axis indicate pathologists within one laboratory, followed by the overall proportion of the laboratory per differentiation grade. Intra-laboratory differences are calculated by Fisher exact test (Monte Carlo option in laboratory 26).

4

Significant inter- and intra-laboratory variation in grading of invasive breast cancer: a nationwide study of 33,043 patients in the Netherlands

Carmen van Dooijeweert

Paul J. van Diest

Stefan M. Willems

Chantal. C. H. J. Kuijpers

Elsken van der Wall

Lucy I. H. Overbeek

Ivette A. G. Deckers

Abstract

Accurate, consistent, and reproducible grading by pathologists is of key-importance for identification of individual patients with invasive breast cancer (IBC) that will or will not benefit from adjuvant systemic treatment. We studied the laboratory-specific grading variation using nationwide real-life data to create insight and awareness in grading variation.

Synoptic pathology reports of all IBC resection-specimens, obtained between 2013-2016, were retrieved from the nationwide Dutch Pathology Registry (PALGA). Absolute differences in laboratory-proportions of grade I-III were compared to the national reference. Multivariable logistic regression provided laboratory-specific odds ratios (ORs) for high- versus low-grade IBC.

33,792 IBC pathology reports of 33,043 patients from 39 laboratories were included, of which 28.1% were reported as grade I (range between laboratories 16.3-43.3%), 47.6% as grade II (38.4-57.8%), and 24.3% as grade III (15.5-34.3%). Based on national guidelines, the indication for adjuvant chemotherapy was dependent on histologic grade in 29.9% of patients. After case-mix correction, 20 laboratories (51.3%) showed a significantly deviant OR. Significant grading differences were also observed among pathologists within laboratories.

In this cohort of 33,043 breast cancer patients we observed substantial inter- and intra-laboratory variation in histologic grading. It can be anticipated that this has influenced outcome including exposure to unnecessary toxicity, since choice of adjuvant chemotherapy was dependent on grade in nearly a third of patients. Better standardization and training seems warranted.

Introduction

About one in seven women in the Netherlands will develop breast cancer during her life [1], which makes breast cancer the most common type of cancer in Dutch women with approximately 15,000 new diagnoses per year [2]. Histologic grade is one of the best established prognostic factors in breast cancer, and is strongly and independently associated with both breast cancer-specific and disease-free survival [3, 4]. Studies even suggest that histologic grade can predict tumour behaviour more accurately than other 'time-dependent' prognostic factors like tumour size [3-7]. Hence, histologic grade is an important clinical contributor and is widely used to guide therapeutic breast cancer management [3, 4, 8]. Furthermore, since breast cancer screening programs resulted in earlier detection and thereby a greater proportion of both smaller tumours and lymph node negative tumours [9-11], histologic grade is determinative in patient management in a substantial number of cases, including the use of genetic profiling tests [8, 12-15]. Additionally, in contrast to prognostic genetic profiling tests, the evaluation of histologic grade is cheap and can in principle be performed in all cases of breast cancer [16].

The most widely used grading system for invasive breast cancer (IBC) is the modified Bloom and Richardson guideline (Elston-Ellis modification of Scarff-Bloom-Richardson grading system, also known as the Nottingham grading system) [17, 18], which combines the assessment of cell morphology (nuclear polymorphism), measurement of differentiation (tubule formation) and assessment of proliferation (mitotic count), resulting in a total score and derived grade[18]. This system is considered suitable for evaluating IBC in routine clinical setting and is globally incorporated in breast cancer guidelines [3]. Furthermore, Lundin *et al* concluded that, even when assessed by pathologists who have had no special training in breast cancer pathology, histologic grading in breast cancer is of substantial and independent prognostic value [19].

While grading has systematically been proven to be prognostically very important, accurate, consistent, and reproducible grading by pathologists is of key-importance for identification of individual patients who, based on their prognosis, may or may not benefit from adjuvant treatment. However, current evidence suggests that there is considerable and clinically relevant variation in the grading of IBC. Previous studies, in which a set of IBC was reviewed by several pathologists, mostly showed an overall reproducibility that was no more than moderate [16, 20-22]. Yet, these conclusions are derived from smaller studies where grading was performed in study setting, and thus this may not resemble real-life grading in daily clinical practice. Moreover, individual practicing pathologists may not have felt addressed by these data, as it did not provide them insight into their own grading practice.

4

In this context, nationwide daily clinical practice studies did show that there is considerable variation between laboratories and individual pathologists in the grading of e.g. colorectal adenomas [23] and colorectal adenocarcinomas [24]. In addition, we previously reported substantial nationwide inter- and intra-laboratory variation in grading of ductal carcinoma *in situ* of the breast (DCIS) [25]. Grading of DCIS is methodologically different from and less standardized than grading of IBC, and DCIS treatment is currently independent of any histopathologic features, whereas treatment of IBC is widely guided by histologic grading. Therefore, it is unclear whether the same conclusions can be drawn for IBC. In light of its current, and important, clinical consequences, it is also particularly important to create insight and awareness in grading variation of IBC.

We studied the laboratory-specific variation in histologic grading of IBC in a nationwide study in the Netherlands. Using the Dutch nationwide pathology registry (PALGA), we assessed the variation in histologic grading of over 33,000 patients with IBC, between Dutch pathology laboratories and between individual pathologists using real-life data from synoptic (structured) pathology reports from daily pathology practice. In addition, we also analysed the variation of the three components of grading (according to the modified Bloom and Richardson classification) between laboratories. Furthermore, we conducted a questionnaire among pathologists to gain insight into their grading practices. As grade is an important decision-tool in adjuvant treatment, inter- and intra-laboratory variation in grading may lead to under- and overtreatment of a substantial percentage of primary breast cancer patients. Creating insight into these laboratory-specific differences may help to design an intervention to improve standardization among laboratories and pathologists.

Materials and Methods

Data source and study population
Data were extracted from PALGA, the nationwide network and registry of histo- and cytopathology in the Netherlands, which contains excerpts of all pathology reports from Dutch Pathology laboratories since 1991 [26]. All data from the PALGA database are pseudonymized by a trusted third party (ZorgTTP, Houten, The Netherlands). Consent was given by all Dutch laboratories for the storage of their data by PALGA, and for scientific use of these data. Pathology laboratories were initially anonymized and further consent was obtained for additional analysis of inter-pathologist variation within the individual laboratories (n=8 laboratories). The scientific and privacy committee of PALGA approved this study. All data were retrieved and handled in compliance with the General Data Protection Regulation Act (GDPR).

We retrieved all synoptic pathology reports of patients with resection specimens of IBC between January 1 2013 and December 31 2016 in the Netherlands (n=48,667). Synchronous IBC was defined as an ipsilateral lesion within six months of the first IBC resection. These lesions were considered paired measurements of which only the first was included. We solely included patients with primary tumours, thereby excluding resection specimens with complete regression of the tumour, specimens without a tumour after biopsy, and specimens of re-excisions. As neoadjuvant therapy may influence grading [27-29], pathology reports of patients who received neoadjuvant treatment were excluded (Supplementary 1).

In total, 40 out of 46 Dutch laboratories synoptically reported IBC on breast resection specimens. Of these, we included those that synoptically reported ≥250 IBC during the study period. For inter-pathologist variation within individual laboratories, we only analysed data from pathologists from the eight participating laboratories who synoptically reported ≥20 IBC during the study period.

For each patient we extracted sex and age, type of surgery, IBC tumour size, histologic subtype, histologic grade, ER/PR-receptor status (immunohistochemistry (IHC)) and HER2-receptor status (IHC and/or in situ hybridisation (ISH)). Reports of IBC with any missing data were excluded from further data analysis (Supplementary 1).

Analysis of histologic grading
In the PALGA synoptic reporting module, histologic grade was determined according to the modified Bloom and Richardson guideline (Elston-Ellis modification of Scarff-Bloom-Richardson grading system, also known as the Nottingham grading system). According to this guideline the three components (tubule formation, nuclear polymorphism and mitotic count) are scored from 1-3, which results in a total score and derived overall histologic grade (score 3-5 = grade I, score 6-7 = grade II, score 8-9 = grade III) [17-18]. The primary outcome measure of this study was the inter-laboratory variation in total histologic grading of IBC and separate for its three components. Secondary outcome measure was the inter-pathologist variation in histologic grading within a single laboratory.

Histologic grading in relation to clinical management
To gain insight into the influence of histologic grading on therapeutic patient management, we identified a subgroup of patients who, in view of current national guidelines[8], were eligible for adjuvant systemic chemotherapy solely on the basis of histologic grade. This concerned patients ≥35 years of age with a negative HER2-receptor status and a tumour size of 1.1-2cm , or, in patients < 35 years, those with a negative HER2-receptor status and a tumour of ≤1cm, or a positive HER2-receptor status and a tumour of <0.5cm. In these patients, grade II-III tumours qualify for adjuvant chemotherapy, whereas this is not

recommended for patients with grade I tumours. In addition, for this group, it was checked whether the total (modified Bloom and Richardson) score was on a switch point of grades, (i.e., scores 5 (grade I) or 6 (grade II) and scores 7 (grade II) or 8 (grade III)), where the difference of only one point on the total Bloom and Richardson score could already alter the overall histologic grade and thereby chemotherapy indication.

Questionnaire survey among pathologists

A questionnaire survey was sent to all 46 pathology laboratories in the Netherlands to identify how pathologists determine the histologic grade of IBC in daily clinical practice. The survey contained questions on whether pathologists consider themselves specialized breast pathologists, the number of years of experience as a pathologist, how they count mitoses and how they deal with heterogeneity of histologic grade within one specimen (Supplementary 2).

Statistical analysis

Patient and tumour characteristics were summarized and differences between histologic grades were tested by means of a χ^2-test for categorical variables and by a non-parametric Kruskal-Wallis test for continuous variables.

The overall proportions of histologic grade I, II, and III were determined and considered the national proportion. Absolute differences in proportion of histologic grades between laboratories are presented in funnel plots per grade, in which the proportions per laboratory are plotted against the number of included IBC per laboratory (Fig. 1). The target of these funnel plots was set at the national proportions with their 95% confidence intervals (CI) as limits[30]. Absolute inter-laboratory differences in proportions of the three components of grading were also analysed.

To compare relative differences among laboratories, odds ratios (OR) and 95% CIs per laboratory were calculated by logistic regression. As there is no clear binary cut-off for low-grade and high-grade in clinical practice, we performed two logistic regression analyses, with different definitions of low- and high-grade IBC. In our first logistic regression analysis we defined low-grade IBC as grade I and high grade IBC as grade II-III. In our second logistic regression analysis we defined low-grade IBC as grade I-II and high grade IBC as grade III. Both analyses resulted in ORs and 95% CI for high- versus low-grade IBC.

For the choice of the reference laboratory of the logistic regression models, the sum of absolute deviations from the grade specific national proportions was calculated to compare the absolute deviation for all three grades at once. The laboratory with the lowest sum-score was deemed best resembling the national distribution and was thereupon chosen as reference laboratory.

Two multivariable logistic regression analyses for high- versus low-grade IBC were performed to correct for differences in case-mix. To identify potential confounding factors, we selected clinicopathological variables *a priori* based on literature [18, 31-35] and on pathologists' experience. These factors included age, sex, tumour size, type of surgery, histologic subtype, hormone-receptor status and HER2-receptor status. Hormone-receptor status (ER/PR) was considered positive when either or both the ER- or PR-receptor were positive. According to the Dutch guideline, and within the synoptic PALGA protocol module, the ER- and PR-receptor status is considered positive when ≥10% of the tumour cells show ER- and PR-specific staining on IHC [8]. Overall, hormone-receptor status was taken into account as a binary variable (either positive (≥10%) or negative (<10%)) and not as a continuous variable (percentage of stained tumour nuclei). All variables, except for sex, as the number of males was too low, appeared to be significantly associated with grade and were therefore included in both final multivariate models. It was checked whether males clustered in specific laboratories, but this was not the case. The adjusted ORs (95% CI) are presented in a forest plot (Fig. 2).

For analysis of the inter-pathologist variation within the laboratories we merely compared the proportions per histologic grade between pathologists by Fisher exact test (Monte Carlo option) (Fig 3., Supplementary 3). Results of the questionnaire were summarized by frequencies and percentages.

Values of p < 0.05 were considered statistically significant. All statistical analyses were performed by using IBM SPSS Statistics version 21.

Results

Characteristics of IBC lesions and laboratories
A total of 33,792 IBC lesions from 33,043 patients were included in our data analysis. For some patients we included more than one pathology report as this concerned either a bilateral tumour or an ipsilateral tumour more than six months after the first IBC resection. All patients originate from a total of 39/46 Dutch pathology laboratories as one laboratory graded less than 250 IBC lesions within the synoptic PALGA protocol module and six laboratories had not yet implemented synoptic reporting at the time of the study (Supplementary 1). Characteristics of these included patients and corresponding invasive breast tumours are listed in Table 1.

Table 1. Characteristics of the 33,792 included invasive breast cancer (IBC) lesions from the PALGA database 2013-2016

	Total (n=33,792)	Grade 1 (n=9,495)	Grade 2 (n=16,098)	Grade 3 (n=8,199)	p-value
Age (y)*	62.2 (12.1)	62.4 (10.8)	62.8 (11.8)	60.7 (13.8)	0.000
Sex, n (%)					
Female	33,537 (99.2%)	9,441 (99.4%)	15,967 (99.2%)	8,129 (99.1%)	0.045
Male	255 (0.8%)	54 (0.6%)	131 (0.8%)	70 (0.9%)	
Tumour size (cm)*	1.9 (1.3)	1.4 (0.9)	1.9 (1.4)	2.3 (1.5)	0.000
Type of surgery, n (%)					
Mastectomy	12,209 (36.1%)	2,548 (26.8%)	6,172 (38.3%)	3,489 (42.6%)	0.000
Breast conserving	21,583 (63.9%)	6,947 (73.2%)	9,926 (61.7%)	4,710 (57.4%)	
Histologic subtype, n (%)					
Ductal	28,547 (84.5%)	8,727 (91.9%)	12,382 (76.9%)	7,438 (90.7%)	0.000
Lobular	4,432 (13.1%)	647 (6.8%)	3,465 (21.5%)	320 (3.9%)	
Other	813 (2.4%)	121 (1.3%)	251 (1.6%)	441 (5.4%)	
ER/PR receptor status, n (%)					
Positive	29,576 (87.5%)	9,373 (98.7%)	15,162 (94.2%)	5,041 (61.5%)	0.000
Negative	4,216 (12.5%)	122 (1.3%)	936 (5.8%)	3,158 (38.5%)	
HER2-receptor status, n (%)					
Positive	3,340 (9.9%)	212 (2.2%)	1,335 (8.3%)	1,793 (21.9%)	0.000
Negative	30,452 (90.1%)	9,283 (97.8%)	14,763 (91.7%)	6,406 (78.1%)	

* Mean (SD)
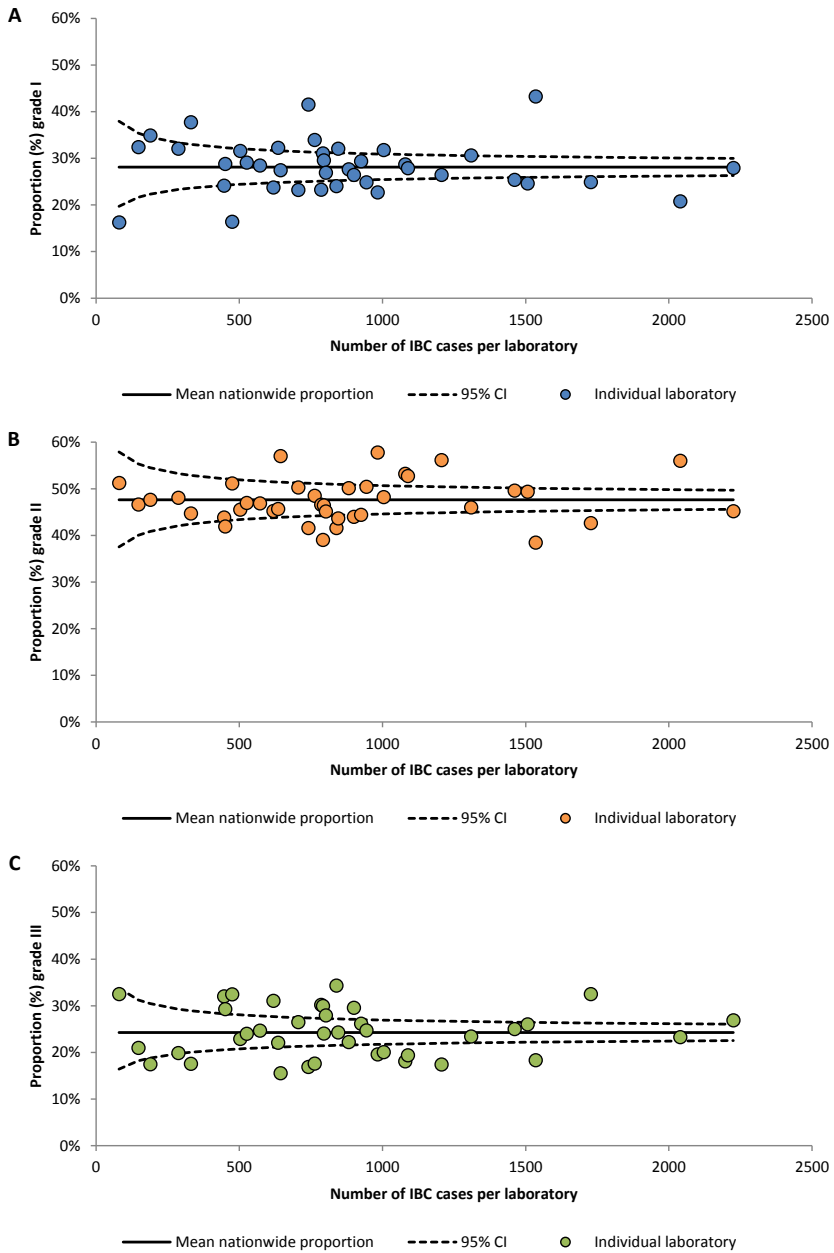
Figure 1. Funnel plots showing the observed proportion per IBC grade per laboratory (dots) relative to the mean national proportion and its 95% confidence intervals for IBC grades I (A), II (B) and III (C) (2013-2016).

Mean (SD) age at diagnosis was 62.2 (12.1) years and patients were predominantly female (99.2%). Breast conserving surgery was performed in the majority of patients (63.9%). Higher histologic grade was positively associated with mastectomy rate and tumour size, and with a negative ER/PR-receptor status and a positive HER2-receptor status. The number of synoptically reported IBC lesions per laboratory ranged from 80 to 2 225 (median 795). Overall national proportions for IBC grade I, II, and III were 28.1%, 47.6% and 24.3%.

Inter-laboratory differences in histologic grading

Laboratories varied mostly in the reporting of IBC grade I (16.3-43.3%), followed by IBC grade II (38.4-57.8%) and IBC grade III (15.5-34.3%). Overall, more than half of the laboratories (22/39) showed proportions outside the 95% CI for both grade I and grade III (56.4%), whereas this was the case for 41.0% of the laboratories for grade II (Fig. 1).

The sum-score was lowest and only 1.6% for laboratory 22, which was thereupon chosen as reference laboratory. The maximum sum-score, in contrast, was 30.3% (laboratory 1). Using the first definition of high-grade IBC (grade II-III), multivariate logistic regression showed that 11 laboratories (28.2%) reported a significantly higher (n=4) or lower (n=7) proportion of high-grade IBC (grade II-III) than the reference laboratory (Fig. 2*a*). Adjusted ORs of individual laboratories ranged from 0.43 (95% CI: 0.35-0.54) to 1.98 (95% CI: 1.03-3.79).

Using the second, alternative, definition of high-grade IBC (grade III), multivariate logistic regression analyses showed that 13 laboratories (33.3%) reported a significantly higher (n=7) or lower (n=6) proportion of high-grade IBC (III) than the reference laboratory (Fig. 2*b*). Adjusted ORs of individual laboratories ranged from 0.47 (95% CI: 0.34-0.65) to 1.95 (95% CI: 1.52-2.50).

After correction for case-mix in both analyses, using different definitions of high-grade IBC, 20 laboratories (51.3%) had at least one significantly higher or lower OR than the reference laboratory. Four laboratories (10.3%) had significantly deviant ORs on both analyses (Fig. 2).

Inter-laboratory differences in components of histologic grading

Regarding the three components of grading, most variation between laboratories was observed for nuclear polymorphism (broadest range in category 3 (severe) 14.2-55.0%), followed by mitotic count (broadest range in category 1 (≥13 mitoses per 2mm^2): 47.2-75.2%) and tubular formation (broadest range for category 3 (<10% of cells with tubular differentiation): 52.7-74.2%).

Overall, the majority of tumours (76.8%) had a total grading score on a switch point of grades, i.e. scores 5 or 6 (49.4%) and scores 7 or 8 (27.4%) (Table 2), for which the difference of only one point on the total Bloom and Richardson score could alter the overall histologic grade and thereby the indication for chemotherapy.

Intra-laboratory differences in histologic grading within laboratories

Sixty-eight pathologists from the eight participating laboratories synoptically reported ≥20 tumours during the study period. Per laboratory, the number of analysed pathologists ranged from 3 to 15 (median 8). In addition, the number of analysed IBC lesions per pathologist ranged from 20-257 (median 82.5). Overall, 22 pathologists (32.4%) graded significantly deviant compared to the national proportions for IBC grade I, while this was the case for 16 pathologists (23.5%) for grade II and for 14 pathologists (20.6%) for grade III (Fig. 3).

Most variation between pathologists within the individual laboratories was observed within laboratory 7 for grade I (range 8.3-50.0%) and grade II (range 23.7-62.8%), whereas most variation for grade III (range 14.7%-45.6%) was observed in laboratory 5. For five laboratories (62.5%) the distribution of histologic grade (i.e. the proportions of grades I-III) significantly differed between pathologists within that laboratory (Supplementary 3*a*, 3*b*).

Indication for adjuvant systemic chemotherapy

In 19,461 of the 33,792 IBC lesions (57.6%) the pathology reports held complete information on all relevant variables that in current clinical practice are used to establish the indication for adjuvant chemotherapy in primary breast cancer (i.e. lymph node status, HER2-status, age, tumour size and histologic grade).

Histologic grade determined the indication for adjuvant chemotherapy in 5,821 patients (29.9%) (Fig. 4). Of this group, 1,801 tumours (30.9%) were reported as grade I and thus, according to current guidelines [8], would not have had an indication for adjuvant chemotherapy. In 4,020 tumours (69.1%), solely based on histologic grade, adjuvant chemotherapy has likely been advised, as they were reported as grade II or III tumours. In total, of the tumours in which the indication for adjuvant chemotherapy was dependent on histologic grade (n=5,821), 3,187 (54.8%) even had a total score on the switch point of grades I and II (i.e. an overall score of five (grade I) or six (grade II)).
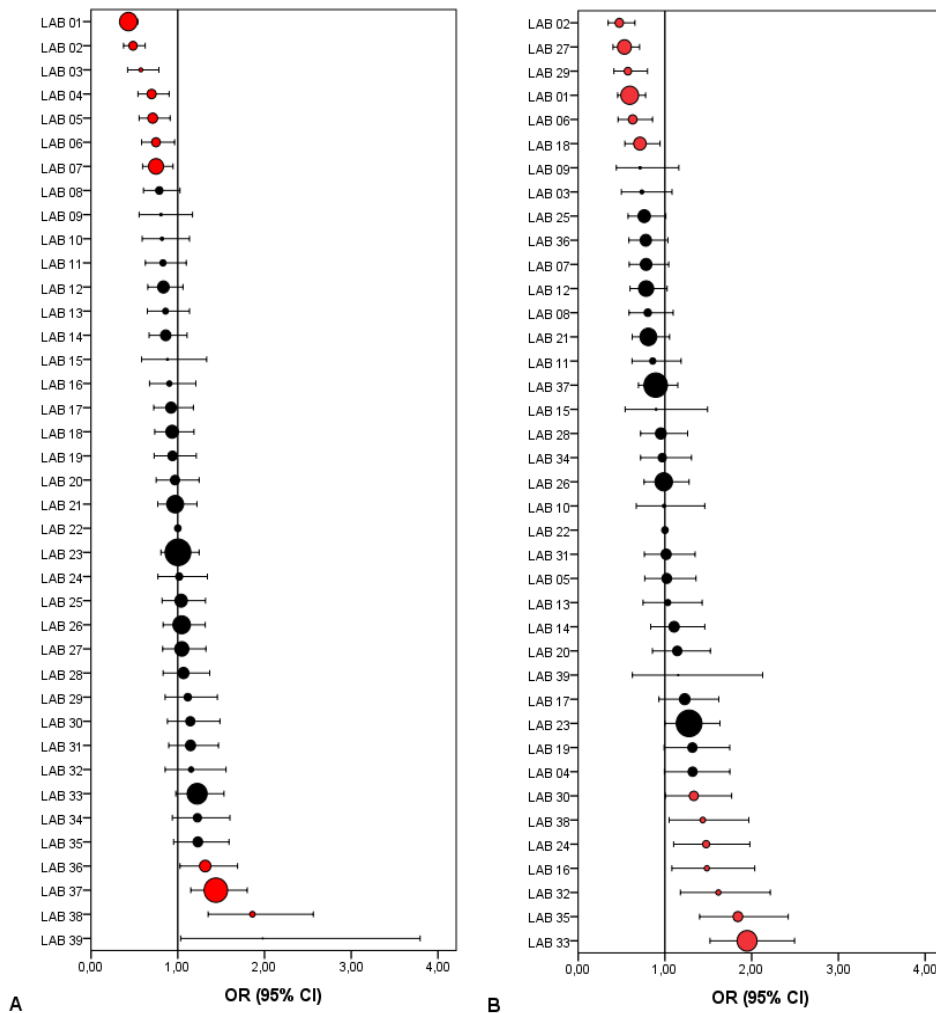
Figure 2. Forest plots showing the adjusted odds ratios (OR) and 95% confidence intervals (CI) of invasive breast cancer (IBC) grade II-III versus IBC grade I (A) and of IBC grade III versus IBC grade I-II (B) in comparison to the reference laboratory (#22). Dot size indicates the total number of analysed synoptically reported IBC lesions per laboratory. Red dots indicate laboratories with a significantly deviant OR as compared to the reference laboratory. ORs are adjusted for age, tumour size, type of surgery, histologic subtype, hormone receptor status and HER2 receptor status.

Table 2. Scores of the three components of the modified Bloom and Richardson classification and overall score for the 33,972 included invasive breast cancer (IBC) lesions from the PALGA database 2013-2016

Characteristics	Total (n = 33,792)	Grade I (n = 9,495)	Grade II (n = 16,098)	Grade III (n = 8,199)
Tubular differentiation, n (%)				
1 > 75% of cells	3,895 (11.5%)	3,698(38.9%)	197 (1.2%)	0 (0.0%)
2 10-75% of cells	8,724 (25.8%)	4,694 (49.4%)	3,371 (20.9%)	659 (8.0%)
3 <10% of cells	21,173 (62.7%)	1,103 (11.6%)	12,530 (77.8%)	7,540 (92.0%)
Nuclear polymorphism, n (%)				
1 Mild 1*	2,942 (8.7%)	2,818 (29.7%)	124 (0.8%)	0 (0.0%)
2 Moderate 2**	20,741 (61.4%)	6,545 (68.9%)	12,258 (76.1%)	1,938 (23.6%)
3 Severe 3***	10,109 (29.9%)	132 (1.4%)	3,716 (23.1%)	6,261 (76.4%)
Mitotic count, n (%)				
1 < 7 per 2mm^2	21,164 (62.6%)	9,273 (97.7%)	11,891 (73.9%)	0 (0.0%)
2 ≥ 8 ≤12 per 2mm^2	5,163 (15.3%)	213 (2.2%)	3,270 (20.3%)	1,680 (20.5%)
3 ≥ 13 per 2mm^2	7,465 (22.1%)	9,(0.1%)	937 (5.8%)	6,519 (79.5%)
Total score				
3 grade I	1,127 (3,3%)	1,127 (11.9%)	-	-
4 grade I	2,796 (8.3%)	2,796 (29.5%)	-	-
5 grade I	5,572 (16.5%)	5,572 (58.7%)	-	-
6 grade II	11,127 (32.9%)	-	11,127 (69.1%)	-
7 grade II	4,971 (14.7%)	-	4,971 (30.9%)	-
8 grade III	4,277 (12.7%)	-	-	4,277 (52.2%)
9 grade III	3,922 (11.6%)	-	-	3,922 (47.8%)

* Nuclei small with little increase in size in comparison with normal breast epithelial cells, regular outlines, uniform nuclear chromatin, little variation in size.
** Cells larger than normal with open vesicular nuclei, visible nucleoli, and moderate variability in both size and shape
*** Vesicular nuclei, often with prominent nucleoli, exhibiting marked variation in size and shape, occasionally with very large and bizarre forms

4
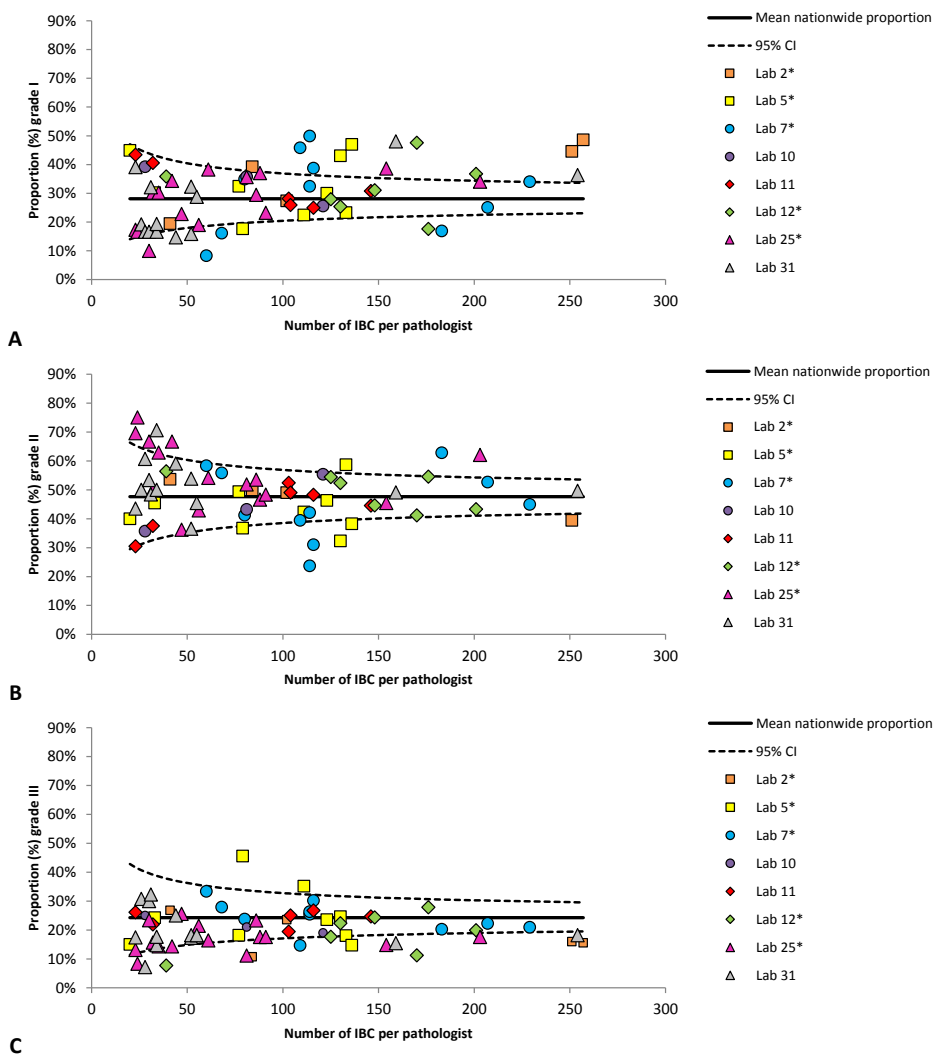
Figure 3. Funnel plots showing the observed proportion of invasive breast cancer (IBC) lesions per grade per pathologist (dots) of eight laboratories relative to the mean national proportion for IBC grades I (A), II (B) and III (C) (2013-2016).

* Indicates that the distribution of grade I-III significantly differed between pathologists within the individual laboratory (calculated by Fishers Exact test; Monte Carlo option).
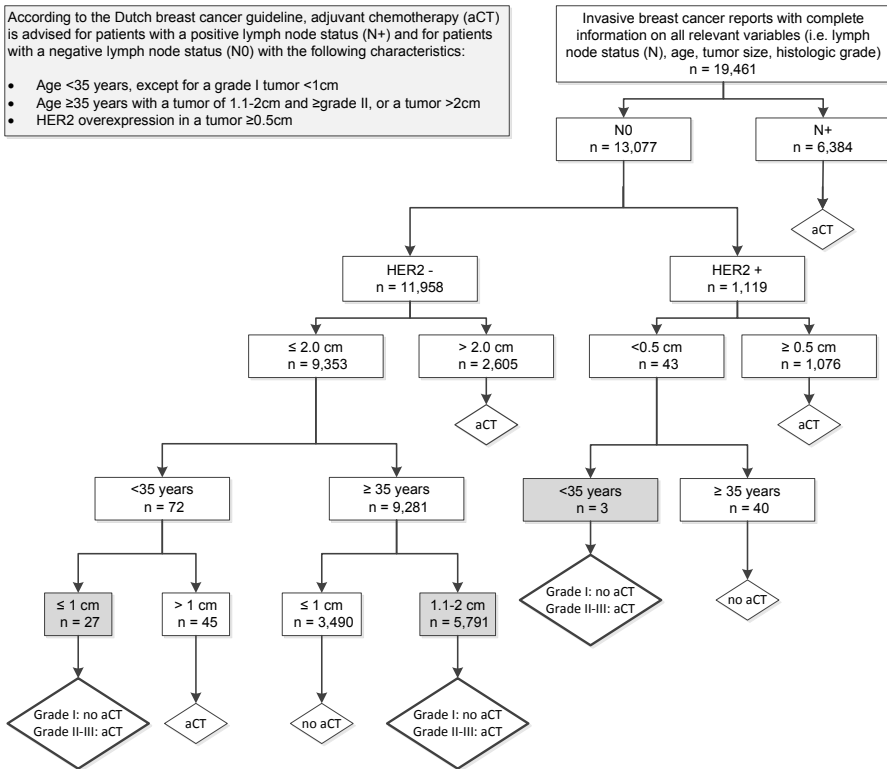
Figure 4. Flowchart showing the decision tree for adjuvant chemotherapy (aCT) in breast cancer patients according to the current Dutch guideline for the 19,461 tumors that held complete information on all relevant variables (i.e. lymph node status (N), age, tumor size and histologic grade). Grey squares indicate tumors of patients in which the aCT indication is dependent on histologic grade (n=5,821, 29.9%).

Results of questionnaire survey

Seventy-nine pathologists out of the approximately 320 practicing pathologists in the Netherlands (25%) [24], responded to our online questionnaire, of which 19.0% worked in an academic hospital at the time. Thirty-seven (46.8%) pathologists denoted themselves as experts in breast pathology. Grading practice of general- and specialised breast pathologists did not seem to differ (data not shown). All pathologists reported the modified Bloom and Richardson grading guideline as a reference for the histologic grading of IBC, however, eleven pathologists (13.9%) also responded that in their opinion specific histologic subtypes per definition have a specific grade. In case of heterogeneity of histologic grade within one specimen, the majority of pathologists (76.0%) report the highest grade as overall histologic grade.

Discussion

In this nationwide cohort of 33,043 invasive breast cancer patients, approximately half of the lesions were reported as grade II (47.6%), whereas grade I and grade III were scored in about a quarter of all lesions (28.1% and 24.3% respectively). The observed overall proportions per grade are in line with previous cohort studies that showed a similar distribution pattern for IBC grades I (15-30%), II (41-62%), and III (22-33%), although specific percentages vary [36-41].

As synoptic reporting, compared to narrative reporting, results in an increased overall completeness of pathology reports [42], and as it enables easy data extraction because all variables are stored in a standardized manner, data included in this study were solely from synoptically reported IBC lesions. Currently, over 80 percent of (pre)malignant breast lesions is reported via the synoptic PALGA protocol by pathologists in the Netherlands [43]. As a control, we compared our data with aggregated data from the Netherlands Cancer Registry, which also holds narrative reports, and observed a similar distribution (data not shown), indicating that the distribution of histologic grade in our population, based on synoptic reporting, is likely to be representative for all IBC patients in the Netherlands.

Laboratory-specific data was analysed in an absolute and relative manner, in which individual laboratories were compared to both the national proportion and a reference laboratory, all indicating that inter-laboratory differences in histologic grading of IBC are substantial. This was illustrated by the large range of proportions per grade, by the sum score with variation of up to 30.3% and by the fact that the reported proportions per grade and the adjusted ORs were significantly deviant from the national distribution in approximately half of the laboratories.

In addition to the substantial inter-laboratory variation, significant intra-laboratory differences were also observed between pathologists within five of eight analysed laboratories (62.5%). These findings emphasize that, even within the laboratories, histologic grading is not performed in a similar manner among pathologists, although the same guidelines are used in all laboratories and by all pathologists (modified Bloom and Richardson guideline). In the eight participating laboratories, 38 of 68 pathologists (55.9%) synoptically graded <100 IBC in the study period of four years. Although there is no external standard or benchmark to indicate whether a pathologist is an expert in IBC grading, this may imply that there are too many pathologists with too little experience in grading. However in absolute manner, pathologists may grade more tumours than it seems from our results, as approximately 20% of IBC cases are still graded outside the synoptic protocol. Furthermore, our results show that both pathologists who grade few and pathologists who

grade many IBC show significantly deviant proportions. Nevertheless, the results of this study raise the question whether it is desirable that some pathologists may actually only grade a few IBC cases per year. This may be the subject of future research.

In line with previous studies [16,20-22], most variation between laboratories was observed for nuclear polymorphism, which might be explained by the fact that scoring of this category is least quantitative, when compared to mitosis counting and to scoring the percentage of tubular differentiation. In addition, more than three-quarters of all patients had a tumour with a total score on a switch point of grades (i.e. scores 5 or 6, and scores 7 or 8), which shows that the variation of only one point in the total score of the three components may already change its subsequent histologic grade, and thus may influence patient management.

The results of this study hopefully raise awareness among pathologists and clinical oncologists, emphasizing that treatment decisions depend on histologic grade in a substantial number of patients and that, for individual patients the difference of only one point on the total score could mean the difference between adjuvant chemotherapy or not. Therefore, accurate, consistent, and reproducible grading is of utmost importance. However, this study also shows that histopathologic grading may currently not meet high enough clinical standards for individual patients, which is a crucial first step to improvement. Furthermore, pathologists are enabled to discuss and reflect on their grading practices as these "mirror" data were also sent to the laboratories by PALGA, which may lead to regression to the mean. Furthermore, these data should not only be discussed by pathologists, but also in multidisciplinary meetings with clinical oncologists. In this context, the Dutch Society of Pathology is already considering annual benchmarking of histologic grading of IBC based on "mirror" PALGA data, which may be adopted much broader in the field. In addition, future research might focus on specifically training pathologists in the assessment of histologic grade, which is underlined by Elston and Ellis, who emphasize that grading IBC should only be undertaken by trained pathologists[44]. Pathologists might for example be trained by an e-learning, which could attribute to better synchronization of histologic grading.

Despite the indisputable need to improve histologic grading practices, it should be noted that other variables guiding breast cancer patient management have limitations as well. For example, HER2 and ER scoring, and the assessment of small nodal metastases are also subject to inter-observer variation [3,45]. In addition, in contrast to other prognostic parameters, like genetic profiling tests, the evaluation of histologic grade is cheap and can in principle be performed in all breast cancer cases [12,16]. Furthermore, although molecular or genetic measures of prognosis may become increasingly important in the

risk stratification of IBC, it is believed by Elston and Ellis that the future clinical application of molecular measures will be in combination with-, and analogous to histologic grade, which is underlined by current (international) guidelines [8, 13, 14, 44]. What is more, the decision to apply expression profiling of IBC is to a great extent based on histologic grade of the tumour [8, 12, 14, 15, 46]. Thus, the assessment of histologic grade may remain of great clinical importance as one of the best established prognostic factors for patients with breast cancer.

The impact of histologic grading is further underlined by our findings that treatment decisions on adjuvant therapy, according to the current guidelines, are solely dependent on histologic grade in almost one in every three patients, which highlights that histologic grading is of great clinical importance, as it influences treatment decisions and may subsequently influence outcome in a substantial part of patients. More than half of this group of patients (54.8%), for whom the indication for adjuvant chemotherapy was dependent on histologic grade (n=5,821), even had a score on the switch point of grades (i.e. score 5 or score 6), indicating that a difference of only one point on the total score would already alter their indication for adjuvant chemotherapy. With the observed substantial grading variation in this study, it is very likely that this may have influenced treatment decisions. Whether this subsequently influenced outcome of these patients should be the subject of future research. Overall, variation in grading may very easily lead to different treatment indications in a substantial part of patients.

Conclusion
In conclusion, the results of this large nationwide study show that there is substantial variation in the histopathologic grading of IBC, both between and within pathology laboratories. Reducing variation in grading is highly clinically relevant, as, for almost one in every three patients, the decision on adjuvant systemic chemotherapy solely depends on histologic grade. Hence, it is very likely that variation in grading influences treatment decisions and subsequently may influence outcome and exposure to unnecessary toxicity of individual patients. Interventions to improve nationwide histologic grading, for example by e-learning, may especially focus on the assessment of nuclear polymorphism, as most variation was observed in this category.

Acknowledgements

4

References

1. National Institute for Health and Environment (RIVM) Ministry of Health, Wellness and Sport. Chance of breast cancer. Avalaible from: https://www.rivm.nl/bevolkingsonderzoek-borstkanker/wat-is-borstkanker/kans-op-borstkanker. Accessed March 20 2019.

2. Netherlands Cancer Registry supplied by IKNL, Cijfers over kanker. Available from: www.cijfersoverkanker.nl. Accessed March 20, 2019.

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(4):207.

4. Rakha EA, El-Sayed ME, Lee AH, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol. 2008;26(19):3153-8.

5. Sundquist M, Thorstenson S, Brudin L, Nordenskjold B. Applying the Nottingham Prognostic Index to a Swedish breast cancer population. South East Swedish Breast Cancer Study Group. Breast Cancer Res Treat. 1999;53(1):1-8.

6. Frkovic-Grazio S, Bracko M. Long term prognostic value of Nottingham histological grade and its components in early (pT1N0M0) breast carcinoma. J Clin Pathol. 2002;55(2):88-92.

7. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat. 1992;22(3):207-19.

8. The Netherlands Comprehensive Cancer Organisation. Breast Cancer Guideline 2017 [updated 2012-07-02]. Available from: https://www.oncoline.nl/borstkanker. Accessed October 30, 2018.

9. Anttinen J, Kautiainen H, Kuopio T. Role of mammography screening as a predictor of survival in postmenopausal breast cancer patients. Br J Cancer. 2006;94(1):147-51.

10. Shen Y, Yang Y, Inoue LY, Munsell MF, Miller AB, Berry DA. Role of detection method in predicting breast cancer survival: analysis of randomized screening trials. J Natl Cancer Inst. 2005;97(16):1195-203.

11. Hanrahan EO, Valero V, Gonzalez-Angulo AM, Hortobagyi GN. Prognosis and management of patients with node-negative invasive breast carcinoma that is 1 cm or smaller in size (stage 1; T1a,bN0M0): a review of the literature. J Clin Oncol. 2006;24(13):2113-22.

12. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med. 2016;375(8):717-29.

13. Senkus E, Kyriakides S, Ohno S, et al, on behalf of the ESMO Guidelines Committee. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. Ann Oncol. 2015;26(suppl_5):v8-v30.

14. Curigliano G, Burstein HJ, Winer EP, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. Ann Oncol. 2017;28(8):1700-12.

15. Krop I, Ismaila N, Andre F, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. J Clin Oncol. 2017;35(24):2838-47.

16. Boiesen P, Bendahl PO, Anagnostaki L, er al. Histologic grading in breast cancer - reproducibility between seven pathologic departments. South Sweden Breast Cancer Group. Acta Oncol. 2000;39(1):41-5.

17. Bloom HJ, Richardson WW. Histological

grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer. 1957;11(3):359-77.

18. 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(5):403-10.

19. Lundin J, Lundin M, Holli K, et al. Omission of histologic grading from clinical decision making may result in overuse of adjuvant therapies in breast cancer: results from a nationwide study. J Clin Oncol. 2001;19(1):28-36.

20. Frierson HF Jr., Wolber RA, Berean KW er al. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. Am J Clin Pathol. 1995;103(2):195-8.

21. Italian Network for Quality Assurance of Tumour Biomarkers (INQAT) Group. Quality control for histological grading in breast cancer: an Italian experience. Pathologica. 2005;97(1):1-6.

22. Meyer JS, Alvarez C, Milikowski C, et al. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. Mod Pathol. 2005;18(8):1067-78.

23. Kuijpers CC, Sluijter CE, von der Thusen JH, et al. Interlaboratory variability in the grading of dysplasia in a nationwide cohort of colorectal adenomas. Histopathology. 2016;69(2):187-97.

24. Kuijpers CC, Sluijter CE, von der Thusen JH, et al. Interlaboratory Variability in the Histologic Grading of Colorectal Adenocarcinomas in a Nationwide Cohort. Am J Surg Pathol. 2016;40(8):1100-8.

25. van Dooijeweert C, van Diest PJ, Willems SM, Kuijpers C, Overbeek LIH, Deckers IAG. Significant inter- and intra-laboratory variation in grading of ductal carcinoma in situ of the breast: a nationwide study of 4901 patients in the Netherlands. Breast Cancer Res Treat. 2019;174(2):479-488.

26. Casparie M, Tiebosch AT, Burger G, et al. Pathology databanking and biobanking in The Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. Cell Oncol. 2007;29(1):19-24.

27. Adams AL, Eltoum I, Krontiras H, Wang W, Chhieng DC. The effect of neoadjuvant chemotherapy on histologic grade, hormone receptor status, and HER2/neu status in breast carcinoma. Breast J. 2008;14(2):141-6.

28. Zheng S, Zhang BL, Xiao T, et al. [Comparison of histopathologic changes and expression of biomarkers in breast carcinoma before and after neoadjuvant chemotherapy]. Zhonghua bing li xue za zhi (= Chinese journal of pathology). 2011;40(7):465-70.

29. Yin HF, Wang YH, Qin XQ, et al. [Effect of neoadjuvant chemotherapy on histologic grade and expression of biological markers in breast cancer]. Zhonghua zhong liu za zhi [Chinese journal of oncology]. 2009;31(11):858-62.

30. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med. 2005;24(8):1185-202.

31. Veneroso C, Siegel R, Levine PH. Early age at first childbirth associated with advanced tumor grade in breast cancer. Cancer Detect Prev. 2008;32(3):215-23.

32. Huang HJ, Neven P, Drijkoningen M, et al. Association between tumour characteristics and HER-2/neu by immunohistochemistry in 1362 women with primary operable breast cancer. J Clin Pathol. 2005;58(6):611-6.

33. Rosenberg J, Chia YL, Plevritis S. The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the U.S. SEER database. Breast Cancer Res Treat. 2005;89(1):47-54.

4

34. Sullivan PS, Apple SK. Should histologic type be taken into account when considering neoadjuvant chemotherapy in breast carcinoma? The breast journal. 2009;15(2):146-54.

35. Tubiana-Hulin M, Stevens D, Lasry S, et al. Response to neoadjuvant chemotherapy in lobular and ductal breast carcinomas: a retrospective study on 860 patients from one institution Ann Oncol. 2006;17(8):1228-33.

36. Blamey RW, Hornmark-Stenstam B, Ball G, et al. ONCOPOOL- a European database for 16,944 cases of breast cancer. Eur J Cancer. 2010;46(1):56-71.

37. Kiaer HW, Laenkholm AV, Nielsen BB, Bjerre KD. Classical pathological variables recorded in the Danish Breast Cancer Cooperative Group's register 1978-2006. Acta Oncol. 2008;47(4):778-83.

38. Moller H, Henson K, Luchtenborg M,. Short-term breast cancer survival in relation to ethnicity, stage, grade and receptor status: national cohort study in England. Br J Cancer. 2016;115(11):1408-15.

39. Puig-Vives M, Sanchez MJ, Sanchez-Cantalejo J, et al. Distribution and prognosis of molecular breast cancer subtypes defined by immunohistochemical biomarkers in a Spanish population-based study. Gynecol Oncol. 2013;130(3):609-14.

40. Seneviratne S, Lawrenson R, Scott N, Kim B, Shirley R, Campbell I. Breast cancer biology and ethnic disparities in breast cancer mortality in new zealand: a cohort study.

41. Sun J, Chen C, Wei W, et al. Associations and indications of Ki67 expression with clinicopathological parameters and molecular subtypes in invasive breast cancer: A population-based study. Oncol Lett. 2015;10(3):1741-8.

42. Sluijter CE, van Lonkhuijzen LR, van Slooten HJ, Nagtegaal ID, Overbeek LI. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows Archiv. 2016;468(6):639-49.

43. Foundation PALGA. [Accessed October 20, 2018]: Available from: https://www.palga.nl/assets/uploads/Jaarverslag/Jaarverslag_2017_websiteversie.pdf. Accessed 1 September 2018.

44. 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. 2002;41(3a):154-61.

45. Orlando L, Viale G, Bria E, et al. Discordance in pathology report after central pathology review: Implications for breast cancer adjuvant treatment. The Breast. 2016;30:151-5.

46. Groenendijk FH, Jager A, Cardoso F, van Deurzen CHM. A nationwide registry-based cohort study of the MammaPrint genomic risk classifier in invasive breast cancer. Breast. 2018;38:125-31.

PLoS One. 2015;10(4):e0123523.

```
┌─────────────────────────────┐
│  Synoptic pathology reports of │
│  48 667 invasive breast tumors │
│  (IBC) from PALGA database     │
│  2013-2016                     │
│  (42 705 patients)             │
└─────────────────────────────┘
```

No primary tumour
- complete regression (1 432 reports)
- after biopsy (363 reports)
- re-excision (309 reports)

— 2 104 reports →

```
┌─────────────────────────────┐
│  46 563 IBC                    │
│  (40 871 patients)             │
└─────────────────────────────┘
```

Small laboratories
(total IBC <250)

— 1 report →

```
┌─────────────────────────────┐
│  46 562 IBC                    │
│  (40 870 patients)             │
└─────────────────────────────┘
```

Neoadjuvant treatment

— 5 828 reports →

```
┌─────────────────────────────┐
│  40 734 IBC                    │
│  (35 686 patients)             │
└─────────────────────────────┘
```

Synchronous IBC
- synchronous (4 155 reports)
- synchronous unknown (15 reports)

— 4 170 reports →

```
┌─────────────────────────────┐
│  36 564 IBC                    │
│  (35 679 patients)             │
└─────────────────────────────┘
```

Missing tumor grade

— 409 reports →

Missing ER/PR-receptor status

— 606 reports →

Missing HER2-receptor status

— 1 757 reports →

```
┌─────────────────────────────┐
│  33 792 IBC                    │
│  (33 043 patients)             │
└─────────────────────────────┘
```

Supplementary 1. Flowchart of included lesions of invasive breast cancer (IBC) to assess histopathologic grading variation between laboratories.

Supplementary 2. Questionnaire to pathologists on their invasive breast cancer grading practice.

Questionnaire: grading practice of invasive breast cancer

1. In which pathology laboratory do you work?
2. Is this an academic or a non-academic pathology laboratory?
    a. Academic
    b. Non-academic
3. Where (institution, region, country) were you educated as a pathologist?
4. How many years have you worked as a pathologist?
    a. 0-5 years
    b. 6-10 years
    c. 11-20 years
    d. >20 years
5. Do you consider yourself a specialized breast pathologist?
    a. Yes
    b. No
6. Can you give an estimate of the percentage of your daily activities that you spend on breast pathology?
    a. 0-25%
    b. >25%
7. How many pathologists work in your laboratory?
8. How many pathologists within your laboratory perform breast diagnostics?
9. How many breast pathologists work in your laboratory?
10. How often do you use the synoptic breast cancer protocol module for the reporting on breast resections?
    a. >75% of cases
    b. 51-75% of cases
    c. 26-5% of cases
    d. <25% of cases
    e. Never
    f. Other
11. How do you determine histologic grade of invasive breast cancer
    a. According to the modified Bloom and Richardson guideline (Rakha 2008)
    b. Certain histological subtypes have a certain degree by definition
    c. Based on intuition
    d. Other
12. How do you count mitosis?
    a. In the area of the tumour with most cells and most mitosis
    b. In cell-rich peripheral area of the tumour
    c. In the most solid part of the tumour
    d. Spread over the tumour
    e. Other

Supplementary 2. Continued.

Questionnaire: grading practice of invasive breast cancer

13. How do you grade a lesion of heterogeneous grade?
    a. Based on the predominant grade
    b. Based on the highest grade
    c. I mention that there is heterogeneity and mention the percentages of each grade
    d. Other

14. Do you have discussions/meetings between pathologists in your clinic with the aim of uniformly assessing (pre) malignant breast lesions?

4

Laboratory 2, p = 0.001

Laboratory 5, p = 0.000

Laboratory 7, p = 0.000

Laboratory 10, p = 0.238
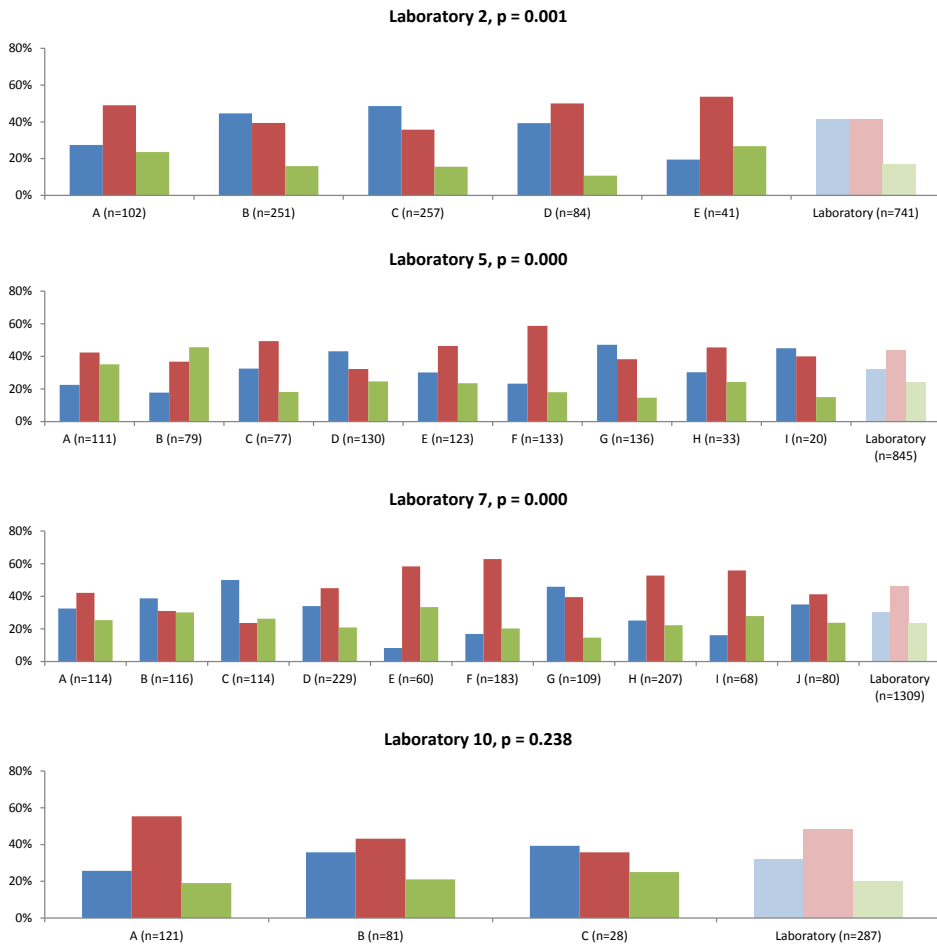
Supplementary 3a. Intra-laboratory variation between pathologists in laboratories 2, 5, 7 and 10. Blue, red and green bars indicate the proportions of invasive breast cancer (IBC) grades I, II and III, respectively. Capital letters on the X-axis indicate pathologists within one laboratory, followed by the overall proportion of the laboratory per differentiation grade. Intra-laboratory differences were calculated by Fisher exact test (Monte Carlo option).

Laboratory 11, p = 0.566

Laboratory 12, p = 0.000

Laboratory 25, p = 0.003

Laboratory 31, p = 0.066
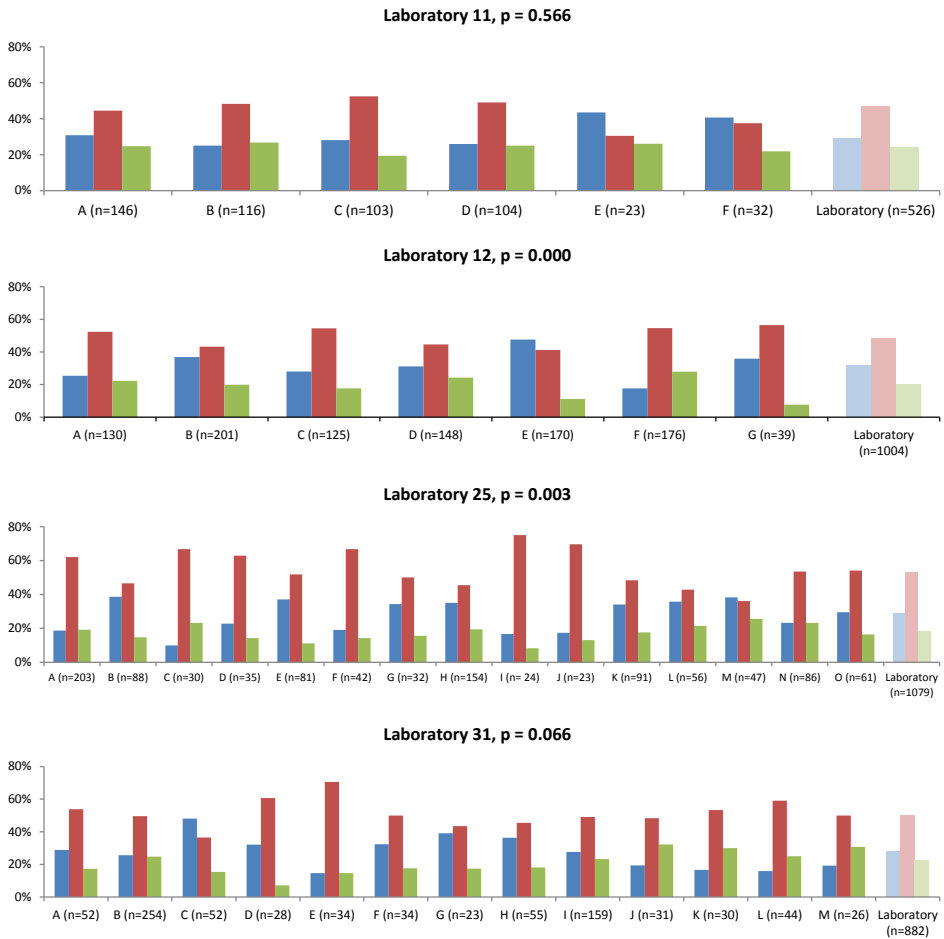
Supplementary 3b. Intra-laboratory variation between pathologists in laboratories 11, 12, 25 and 31 Blue, red and green bars indicate the proportions of invasive breast cancer (IBC) grades I, II and III, respectively. Capital letters on the X-axis indicate pathologists within one laboratory, followed by the overall proportion of the laboratory per differentiation grade. Intra-laboratory differences were calculated by Fisher exact test (Monte Carlo option).

5

The increasing relevance of histologic grading in tailoring adjuvant systemic therapy in 30,843 breast cancer patients in the Netherlands

Carmen van Dooijeweert
Inge O. Baas
Ivette A. G. Deckers
Sabine Siesling
Paul J. van Diest
Elsken van der Wall

Abstract

Background

Large variation between pathology laboratories and pathologists in histologic grading of invasive breast cancer (IBC) has been revealed, which might influence tailoring adjuvant systemic therapy. This study was conducted to evaluate the role of grading in clinical decision-making, thereby taking into account the past and present Dutch national guidelines for IBC management.

Methods

All synoptic pathology reports of IBC resection-specimens, obtained between 2013 and 2016, were extracted from the nationwide Dutch Pathology Registry (PALGA), and linked to treatment data from the Netherlands Cancer Registry (NCR). The relevance of grading with regard to adjuvant chemotherapy (aCT) was quantified by identifying patients for whom grade was the determinative factor. In addition, the relation between grade and aCT-administration was evaluated by multivariate logistic regression for patients with a guideline aCT indication.

Findings

Overall, 30,843 IBC patients were included for data analyses. We identified 7,744 (25.1%) patients for whom grade was the determinative factor for their aCT indication when strictly applying the Dutch guideline that was valid during the study period (2013-2016). For the current guideline, we identified 10,869 (35.2%) patients for whom the aCT indication would be determined by grade. In addition, the indication for adjuvant endocrine therapy (aET), which was added to the 2019 guideline, would solely depend on grade in 9,173 (29.7%) patients. Grade also played a significant role in tailoring aCT de-escalation, as patients with lower grade tumors received aCT significantly less often.

Interpretation

These real-world data illustrate the increasing relevance of histologic grade in tailoring adjuvant systemic therapy. Next to playing a key-role in aCT de-escalation, the role of histologic grading has even expanded to de-escalation of aET. This stresses the need for optimizing grading by pathologists as much as possible, thereby diminishing the risk of worse patient outcome due to non-optimal treatment.

Introduction

Breast cancer is the most common type of cancer in women worldwide (other than non-melanoma skin cancer), with an incidence of 2.1 million [1]. In the Netherlands, invasive breast cancer (IBC) accounts for approximately 15,000 new diagnoses annually, which translates into development of breast cancer during life in about one in every seven women [2, 3].

Preventing distant metastases is one of the main goals of systemic therapy in non-metastatic breast cancer [4]. Primary systemic treatment modalities used in this so-called (neo-) adjuvant setting are chemotherapy, endocrine therapy, and targeted anti-HER2-therapy. Aiming to provide individual breast cancer management, clinicopathologic biomarkers are used to guide treatment decision making in daily practice [4-12]. For example, the indication for adjuvant systemic chemotherapy (aCT), is based on patient factors like age and performance status, and classic pathology features like tumor size, estrogen- (ER), progesterone- (PR), and HER2-receptor status, lymph node status and histologic grade [5-9, 11, 13, 14]. Of these factors, histologic grade is a biomarker that has consistently been found to be associated with both breast cancer-specific and disease free-survival [13, 14].

Within the Dutch national breast cancer guideline, histologic grade plays an important role in the selection of individual patients considered to experience benefit from aCT as higher grade patients are deemed at high risk of occult metastases [11]. This is of particular relevance since we have previously shown that substantial grading variation exists between Dutch pathology laboratories in daily clinical practice [15]. It is therefore highly likely that tumors are under- and over-graded in specific pathology laboratories and/or by specific pathologists, which may lead to under- or over-treatment of IBC patients.

In general, for selecting the right patients for adjuvant systemic chemotherapy (aCT), most clinical guidelines worldwide roughly distinguish three groups, i.e. estrogen (ER)- and or progesterone (PR)-driven HER2 negative IBC, HER2-driven (ER-/PR-) IBC and triple negative breast cancer (TNBC) (ER-, PR-, HER2-) [4, 6, 7, 10-12]. For HER2-driven IBC and TNBC, these guidelines state that practically all patients should receive some form of aCT [6, 7, 10-12]. In contrast, for ER+/PR+ and ER+/PR- driven breast cancer, a group with generally an excellent prognosis [4, 16], focus has shifted to tailoring aCT by de-escalation, while mostly maintaining adjuvant endocrine therapy (aET).

More recently, gene-expression profiling (GEP) by multigene assays (like the 21-gene recurrence score assay (21-GS/Oncotype DX) and the 70-gene signature (70-GS/Mammaprint) [5-9, 11, 13, 14] has been introduced as a supportive tool in clinical-decision

making for ER+/HER2- IBC patients. Although both the 21-GS and 70-GS are commercially available in the Netherlands, the 21-GS is barely used [17]. According to the Dutch breast cancer guideline [11], the 70-GS is encouraged in specific subgroups of patients, based upon clinicopathologic variables, for whom aCT is considered, i.e. ER+, HER2-, pT-2 node negative (N0) or pT1N1 (1 positive lymph node) tumors of ductal/no special histologic type (NST). For these patients, the GEP-indication further depends on tumor size and histologic grade (pN0: grade I/size 3-5cm and grade II/size 2-5cm; pN1: grade I/size 2-5cm, grade II/size 0-5cm) [11]. Thus, the 70-GS is used to tailor aCT de-escalation in a group of well-defined patients, for whom in node-positive and younger breast cancer patients (<50 years of age) its use has been associated with a decreased likelihood of receiving aCT [17].

Overall, histologic grade thus plays an important role in tailoring aCT, both directly and indirectly (i.e. 70-GS indication). In view of its clearly additive role in clinical-decision making, while bearing in mind that substantial variation in grading exists in daily pathology practice, we evaluated the role of histologic grading in tailoring adjuvant systemic therapy in a nationwide clinical setting, in light of both the past and current Dutch breast cancer guidelines, thereby evaluating whether accurate grading remains relevant. Therefore, real world pathology data from PALGA, the nationwide network and registry of histo- and cytopathology in the Netherlands, were linked to nationwide clinical data on treatment from the Netherlands Cancer Registry (NCR).

Methods

Data source
Data were primarily extracted from the Dutch nationwide network and registry of histo- and cytopathology (PALGA), which contains excerpts of all pathology reports in the Netherlands since 1991 [18]. All personal data in this database are pseudonymized by a trusted third party (ZorgTTP, Houten, the Netherlands), and pathology laboratories are anonymized to the researchers. All data were retrieved and handled according to the General Data Protection Regulation Act (GDPR).

Study population and data linking
All synoptic reports of resection specimens with IBC in the Netherlands, between January 1, 2013, and December 1, 2016, were extracted from the PALGA database, including n=48,667 reports from 42,705 patients. From each pathology report we extracted patient characteristics (sex, age, neoadjuvant treatment) and tumor characteristics (tumor size, histologic grade, histologic subtype, and ER, PR and HER2 status).

Pathology reports of patients without a primary tumor (n=2,104) were excluded. Furthermore, pathology reports from patients who received neoadjuvant treatment (n=5,829) were excluded, because treatment for these patients is based upon biopsies, while biomarkers like histologic grade, and receptor-status (ER-, PR- and HER2) may alter after systemic treatment [19-24], hence they may differ from the initial biopsy. For synchronous breast cancer, defined as ipsilateral breast cancer within six months after the first IBC resection specimen report, only the first report was included, as this tumor is usually the index-tumor on which treatment decisions are based. Remaining patients with more than one pathology report (n=1,495), which concerned bilateral tumors in the majority (>90%), were excluded, as in this case, it could not be determined which tumor was considered the index-tumor. Lastly, we excluded reports with any missing data (which concerned missing values on histologic grade, ER/PR-, and HER2-receptor status) (Supplementary fig. 1).

To evaluate the role of histologic grading in tailoring adjuvant systemic therapy in daily clinical practice, thereby taking into account the past and present Dutch breast cancer guidelines, the pathology-specific PALGA data were linked to clinical treatment data from the Netherlands Cancer Registry (NCR) adding information on adjuvant systemic treatment and information on N-status in all patients. This study was therefore approved by the scientific and privacy committee of PALGA and the monitory board of the NCR. In the final step of linking, less than five percent of patients (1,454) could not be matched, leaving a total of 30,843 breast cancer patients for data analysis (Supplementary fig. 1).

Histologic grade and other determinants

Histologic grade was assessed according to the modified Bloom and Richardson guideline, which combines sub-scores on the three grading components (nuclear pleomorphism, tubule formation and mitotic count), resulting in a total score and derived grade [25, 26]. The combined hormone (ER/PR) receptor status was considered positive when ≥10% of tumor cells showed nuclear staining by immunohistochemistry (IHC) for either, or both ER and PR, which is the cutoff used for ER- and PR-receptor positivity in the Netherlands [11]. Based on guideline cutoffs, tumor size was categorized (≤1cm, 1.1-2cm, >2cm). For age, we used different cutoffs based on the guideline indication, as this was deemed clinically most relevant. For GEPs, we only presented numbers on 70-GS, as 21-GS was barely used in the Netherlands during the study period (<1%) [17].

Adjuvant systemic therapy and histologic grading within the past and present guideline

According to the guideline that was valid during the study period (2013-2016) [27], overall four specific patient subgroups had an aCT indication. First this included all N+ patients <70 years (subgroup 1) and, second, N0 patients <70 years but unfavorable characteristics.

These unfavorable characteristics were defined as age <35 years (with exception of grade I tumors ≤1cm) (subgroup 2), or, within patients aged 35-70, grade II-III tumors of 1.1-2cm (subgroup 3), and all tumors >2 cm (subgroup 4).

When specifically focusing on histologic grade, it can be concluded that the aCT indication according to the 2013-2016 guideline specifically depended on histologic grade (i.e. grade was the determinative factor) in patients <35 years with a N0 tumor ≤1cm, and in patients aged 35-70 with a N0 tumor between 1.1-2cm as aCT was indicated when their tumor was graded as grade II or higher.

To evaluate whether the dependency on histologic grade changed, and thus whether accurate grading by pathologists remains relevant, we evaluated grading within the current Dutch breast cancer guideline, which was published in 2019 [11]. Currently, aCT is indicated in N+ patients <70 years (except in grade I tumors <2cm) (subgroup 1), and patients <70 years with a N0-status (or isolated tumor cells (N0(i+)), or micrometastases (N1(mi)) with unfavorable characteristics. These unfavorable characteristics are defined as age <35 years with a grade I tumor >2cm, or a grade II/III tumor >1cm (subgroup 2), age 35-70 with a grade I tumor >3cm, a grade II tumor >2cm, or a grade III tumor >1cm (subgroup 3). Lastly, aCT is indicated in all HER2-overexpressing tumors (subgroup 4).

It can be concluded that the aCT indication within the current guideline depends on histologic grade in three specific patient subgroups. First this concerns patients <35 years, N0/N0(i+)/N1(mi) with a HER2-negative tumor between 1.1 and 2cm. Second, this concerns patients aged 35-70, with macrometastases in their lymph nodes (N1-3), and a hormone-positive, HER2-negative tumor ≤2cm. Lastly, this concerns patients aged 35-70, N0/N0(i+)/N1(mi) with a hormone receptor positive, HER2-negative tumor between 1.1 and 3cm [11].

In addition, histologic grading within the current guideline also plays a role in tailoring (aET). Overall, aET is indicated in patients of all ages with an ER- and/or PR-positive tumor receptor status and N1-status (including (N1(mi)), while in case of an N0-status (including N0 (i+)), aET is only indicated in grade I tumors >2cm, and grade II or grade III tumors >1cm. Thus, for patients of all ages, N0/N0(i+) with a hormone positive tumor between 1.1 and 2cm, the aET indication is determined by histologic grade.

Statistical analysis
Patient-, tumor- and treatment characteristics were summarized using counts and proportions, and means and standard deviations.

First, we identified patients for whom the aCT indication, when strictly following the 2013-2016 guideline, depended on histologic grade. We then performed the same analyses for the 2019 guideline, and we also identified patients for whom the aET indication would depend on histologic grade, as this was added in the 2019 guideline. Second, we calculated which proportion of patients with an aCT indication according to the 2013-2016 guideline actually received aCT. Third, the relation of grade and other clinicopathologic variables (age, ER/PR- and HER2-receptor status, tumor size) and 70-GS use with aCT administration was evaluated by multivariate logistic regression for the overall population with a 2013-2016 guideline aCT-indication (i.e. all four subgroups). Adjusted odds ratios (AORs) and 95% confidence intervals (CIs) were provided for aCT versus no aCT. Grade I was taken into account as reference category for histologic grade. Lastly, to identify factors that may have influenced guideline (non)adherence with regard to aCT, within the four subgroups with a guideline aCT indication, we compared clinicopathologic variables (histologic grade, ER/PR- and HER2-receptor status, age, tumor size) and 70-GS use between patients who did and did not receive aCT, using counts and proportions, which were tested by means of a χ^2 test.

P-values <0.05 were considered statistically significant. All statistical analyses were performed by using IBM SPSS Statistics version 15.0.0.2.

Results

Patient-, tumor-, and treatment characteristics of the 30,843 included patients are summarized in Table 1. The overall mean age was 62.0 years, while about a quarter of patients were 70 years or older. Overall, only 245 males were included (0.8%). Patients were N0 in roughly two thirds of cases (68.6%) and a similar percentage of patients underwent breast conserving surgery (66.1%). Mean tumor size was just under 2 cm and the vast majority of tumors (84.7%) were of ductal/no special type (NST) histologic subtype. Hormone-receptor positivity was reported in 87.6% of tumors and 10.0% were HER2-receptor positive.

Overall, aCT was administered in nearly one third of patients (32.3%), more than half of all patients received aET (55.3%), and 71.9% underwent adjuvant radiotherapy. Targeted therapy was administered in 7.4% of patients, and consisted of the anti-HER2 agent Trastuzumab in nearly all cases. Little over ten percent of patients had a GEP performed, which concerned 70-GS in virtually all cases. The majority of patients with a 70-GS performed were assigned to the 70-GS low-risk category (n=2,026, 61.6%).

Histologic grading

Overall, we identified 7,744 (25.1%) patients for whom the indication for aCT would depend on histologic grade (i.e. is the determinative factor) when strictly applying the guideline that was valid during the study period (2013-2016) (Supplementary table 1). When applying the Dutch breast cancer guideline that was published in 2019 [11] to our cohort of breast cancer patients, we identified an increase in the percentage of patients for whom the aCT indication would depend on histologic grade, being 10,869 (35.2%) patients (Supplementary table 2). In addition, when strictly adhering to the current guideline, the indication for aET would depend on histologic grade in 9,173 (29.7%) patients (Supplementary table 3).

Table 1. Characteristics of a Dutch nationwide cohort of breast cancer patients (n=30,843), without neo-adjuvant treatment, with a single synoptic resection specimen pathology report between 2013 and 2016

Characteristics	Total (n=30,843)
Age (y (SD))	62.0 (12.0)
<60 years	12,572 (40.8%)
≥60 years	18,271 (59.2%)
<70 years	22,338 (72.4%)
≥70 years	8,505 (27.6%)
Sex (n (%))	
Female	30.598 (99.2%)
Male	245 (0.8%)
Lymph-node status (n (%))	
N0	21,148 (68.6%)
N1	7,295 (23.7%)
N2	996 (3.2%)
N3	606 (2.0%)
N unknown	798 (2.6%)
Type of surgery (n (%))	
Mastectomy	10,463 (33.9%)
Breast conserving	20,380 (66.1%)
Tumor size (cm (SD))	1.9 (1.3)
Histologic subtype (n (%))	
Ductal	26,121 (84.7%)
Lobular	3,982 (12.9%)
Other	740 (2.4%)

Table 1. Continued

Characteristics	Total (n=30,843)
Histologic grade(n (%))	
Grade I	8,633 (28.0%)
Grade II	14,682 (47.6%)
Grade III	7,528 (24.4%)
ER/PR receptor status (n (%))	
Positive	27,008 (87.6%)
Negative	3,835 (12.4%)
HER2-receptor status (n (%))	
Positive	3,070 (10.0%)
Negative	27,773 (90.0%)
Chemotherapy (n (%))	9,950 (32.3%)
Endocrine therapy (n (%))	17,041 (55.3%)
Radiotherapy (n (%))	22,186 (71.9%)
Targeted therapy (n (%))	2,281 (7.4%)
70-GS (n (%))	3,291 (10.7%)

Applying the Dutch breast cancer guideline that was valid between 2013 and 2016, we identified 14,954 patients (48.5%) with an indication for aCT (Table 2a). Of this group, only 9,010 patients (60.3%) actually received aCT. Overall, of all patients receiving aCT (n=9,950), 940 patients (9.4%) did so without having an indication according to the guideline. Administration of aCT decreased slightly between 2013 and 2015, while a considerable decrease of 12.8% was observed in 2016 (Fig. 1).

In general, for all patients with an aCT indication, aCT administration was significantly more apparent in patients with tumors of higher grade (AOR grade II: 1.4 (95%CI: 1.3-1.7), AOR grade III: 4.7 (95%CI: 4.1-5.4)), HER2-receptor positive tumors (AOR: 4.9 (95%CI: 4.1-5.8)), and tumors of larger size (AOR per cm increase: 1.4 (95%CI: 1.3-1.4)). In contrast, patients with hormone-receptor positive tumors (AOR: 0.4 (95%CI: 0.4-0.5), an age ≥60 years (AOR: 0.3 (95% CI: 0.3-0.3), or patients with a 70-GS performed (irrespective of 70-GS result category) (AOR: 0.4 (95%CI: 0.4-0.4) had significantly lower odds of receiving aCT (Table 2b).

Table 2a. Number of breast cancer patients from our dataset **(2013-2016)** with an indication for adjuvant chemotherapy (aCT) according to the Dutch breast cancer guideline between 2013 and 2016 vs. number of breast cancer patients who actually received aCT

Characteristics	Number of patients with an indication for aCT	Number of patients who received aCT
All patients <70 years and N+	6,378	4,562
Patients up to 70 years and N0		
Age <35 years: grade I tumors >1cm, grade II-III all tumor sizes	202	180
Age ≥35 years: tumor 1.1-2cm, grade II-III tumor	5,459	2,476
Age ≥35: tumor >2cm	2,915	1,792
All other patients	-	940
Total number of patients (n=30.843)	**14,954 (48.5%)**	**9,950 (32.3%)***

Table 2b: Indicators of guideline adherence (aCT-administration) in 14,954 IBC patients with a guideline indication for adjuvant chemotherapy (aCT) according to the Dutch guideline **(2013-2016)**

Patient- and tumor characteristics	Number of patients	Adjusted OR (95% CI)*
Histologic grade		
Grade I	1,732 (11.6%)	1
Grade II	8,333 (55.7%)	1.5 (1.3-1.7)
Grade III	4,889 (32.7%)	4.7 (4.1-5.4)
ER/PR positive receptor status	12,704 (85.0%)	0.4 (0.4-0.5)
HER2 positive receptor status	1,921 (12.8%)	4.9 (4.1-5.8)
Age ≥60 years	6,084 (40.7%)	0.3 (0.3-0.3)
Tumor size (cm)	2.1 (1.3)**	1.4 (1.3-1.4)
70-GS	2,950 (19.7%)	0.4 (0.4-0.4)

*calculated by multivariate logistic regression; adjusted for ER/PR-receptor status, HER2-receptor status, histologic grade, age, tumor size and 70-GS use.
** mean (SD)

Patients with a guideline aCT indication
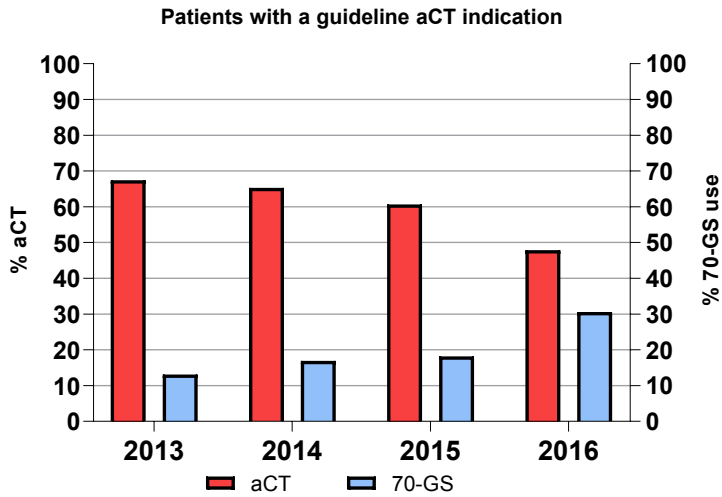
Figure 1. Patients with a strict aCT guideline indication according to the **2013-2016 guideline**

Guideline adherence was most apparent for patients <35 years of age with an aCT indication (89.1%), while it was lowest for the group of N0 patients aged 35-69 and tumors 1.1-2cm in size (45.4%) (Table 2c). Notably, within the latter group, only 20.9% of patients with grade II tumors received aCT, whereas 78.8% of patients with grade III tumors received aCT (p=0.000). In addition, 70-GS use was highest within this group (30.2%) and aCT was only administered in 34.6% of these patients.

As to the role of the 70-GS, this was used in 2,950 patients with a strict guideline aCT indication (19.7%), and only about a third of these patients received aCT (Table 2c). We also observed an only marginal increase in the use of 70-GS in the years between 2013 and 2015, followed by a considerable increase of 12.4% in 2016 (Fig 1).

Table 2c. Characteristics of IBC patients with an indication for adjuvant chemotherapy (aCT) according to the Dutch guideline **(2013-2016)** who received aCT versus patients with an aCT indication who did not receive aCT

Characteristics	Total	aCT received	no aCT received	p-value*
Patients <70 years, N+ tumor	6,378	4,562 (71.5%)	1,816 (28.5%)	**0.000**
Histologic grade				
Grade I	1,316	628 (47.7%)	688 (52.3%)	**0.000**
Grade II	3,245	2,363 (72.8%)	882 (27.2%)	
Grade III	1,817	1,571 (86.5%)	246 (13.5%)	
Receptor status				
ER and/or PR positive	5,657	3,899 (68.9%)	1,758 (31.1%)	**0.000**
HER2 positive	858	794 (92.5%)	64 (7.5%)	**0.000**
Age ≥60 years	2,403	1,394 (58.0%)	1,009 (42.0%)	**0.000**
Tumor size				
Tumor size ≤1.0 cm	623	321 (51.5%)	302 (48.5%)	**0.000**
Tumor size 1.1-2cm	2,826	1,869 (66.1%)	957 (33.9%)	
Tumor size >2 cm	2,929	2,372 (81.0%)	557 (19.0%)	
70-GS	721	240 (33.3%)	481 (66.7%)	**0.000**
Patients <35 years, N0 tumor: grade I >1cm, grade II-III	202	180 (89.1%)	22 (10.9%)	**0.000**
Histologic grade				
Grade I	12	9 (75.0%)	3 (25.0%)	**0.000**
Grade II	61	46 (75.4%)	15 (24.6%)	
Grade III	129	125 (96.9%)	4 (3.1%)	
Receptor status				
ER and/or PR positive	112	93 (83.0%)	19 (17.0%)	**0.002**
HER2 positive	43	40 (93.0%)	3 (7.0%)	0.353
Tumor size				
Tumor size ≤1.0 cm	37	26 (70.3%)	11 (29.7%)	**0.000**
Tumor size 1.1-2cm	99	91 (91.9%)	8 (8.1%)	
Tumor size >2 cm	66	63 (95.5%)	3 (4.5%)	
70-GS	11	3 (27.3%)	8 (72.7%)	0.073
Patients 35-69 years, N0 tumor: 1.1-2cm, grade II-III	5,459	2,476 (45.4%)	2,983 (54.6%)	**0.000**
Histologic grade				
Grade II	3,686	1,078 (29.2%)	2,608 (70.8%)	**0.000**
Grade III	1,773	1,398 (78.8%)	375 (21.2%)	

Table 2c. Continued

Characteristics	Total	aCT received	no aCT received	p-value*
Receptor status				
ER and/or PR positive	4,636	1,782 (38.4%)	2,854 (61.6%)	**0.000**
HER2 positive	690	622 (90.1%)	68 (9.9%)	**0.000**
Age ≥60 years	2,442	774 (31.7%)	1,668 (68.3%)	**0.000**
70-GS	1,648	571 (34.6%)	1,077 (65.4%)	**0.000**
Patients 35-69 years, N0 tumor: >2cm	2,915	1,792 (61.5%)	1,123 (38.5%)	**0.000**
Histologic grade				
Grade I	404	83 (20.5%)	321 (79.5%)	**0.000**
Grade II	1,341	723 (53.9%)	618 (46.1%)	
Grade III	1,170	986 (84.3%)	184 (15.7%)	
Receptor status				
ER and/or PR positive	2,299	1,248 (56.0%)	1,051 (47.2%)	**0.000**
HER2 positive	330	294 (89.1%)	36 (10.9%)	**0.000**
Age ≥60 years	1,239	578 (46.7%)	661 (53.3%)	**0.000**
70-GS	570	204 (35.8%)	366 (64.2%)	**0.000**

* Calculated by means of a χ^2 test, bold numbers indicate statistical significance

Discussion

Using real-world data from PALGA and the NCR, the role of histologic grading in daily clinical practice, in light of both the past and present Dutch breast cancer guidelines, was evaluated in 30,843 breast cancer patients.

These data illustrate that histologic grade remains a key-player in tailoring adjuvant systemic therapy, and that the importance of accurate histologic grading has, when considering the most recent guideline, only increased. As to the role of histologic grading within the guideline that was valid during the study years, our data illustrate that grade likely played an important and significant role in tailoring aCT. Whereas the aCT indication theoretically depended on histologic grade in a quarter of all breast cancer patients (Supplementary table 1), grade also played an important role in tailoring aCT de-escalation, as patients with lower grade tumors received aCT significantly less often. As to the role of histologic grading within the current guideline, the importance of histologic grading in tailoring aCT has increased, grade being decisive for aCT in a third of breast cancer patients (Supplementary

5

table 2) compared to a quarter in the previous guideline. Moreover, tailoring aET depends on histologic grade in a similar percentage of breast cancer patients as well (Supplementary table 3), while grade did not play any role in tailoring aET in the past guideline.

In general, these data illustrate a trend of aCT de-escalation in breast cancer patients whose risk of (developing) distant metastases was deemed low (enough) to withhold adjuvant systemic treatment. Important supportive clinicopathologic biomarkers tailoring aCT were patients' age, N status, hormone- and HER2-receptor status, tumor size and histologic grade. A particularly useful tool in aCT de-escalation in the Netherlands has been the application of the 70-GS, whose use increased considerably in 2016, most likely after publication of the MINDACT-trial [28]. In patients with a strict aCT guideline indication, this increase in 70-GS use was accompanied by a decrease of aCT administration. Overall, only a third of patients who had a 70-GS performed, actually received aCT. It is however important to emphasize that the observed aCT de-escalation can only partially be contributed to 70-GS use, since, overall, 5,944 patients with a strict guideline aCT indication (40.7%) did not receive aCT. Of these patients, only 1,932 (32.5%) had a 70-GS test, leaving 4,012 (67.5%) patients, whose guideline non-adherence was probably based on clinicopathologic biomarkers (i.e. smaller tumor size, lower tumor grade, N0, ER/PR+/HER2-). In addition, guideline (non)adherence may have also been influenced by other patient factors like comorbidity and/or patient preference. However, we did not have data on comorbidities or patient preference in case of guideline non-adherence, but we did observe that patients of older age were less likely to receive aCT, which may have been related to existing comorbidities.

Considering the 2019 Dutch national breast cancer guideline [11], one could conclude that aCT de-escalation, which was already increasingly implemented in daily clinical practice throughout the years 2013 and 2016 despite the guideline ("guideline update anticipation"), has now been translated into official guidelines. Within this current guideline, clear distinctions are made between the three large subgroups based on receptor-status (ER/PR-driven, HER2-driven, triple-negative), where aCT de-escalation is only indicated in ER/PR-driven ("luminal") breast cancer patients. Next to age and receptor status, aCT indication is now primarily based on histologic grade, tumor size, and the degree of lymph node involvement [11]. In addition, there remains a role for the 70-GS as a supportive tool in aCT de-escalation decisions. However, it is important to realize that the indication for 70-GS is guided by histologic grade, histologic subtype, tumor size, N-status and receptor status [11].

Bearing the major clinical implications of histologic grading in mind, it is clear that accurate and reproducible grading by pathologists is of utmost importance to provide high-quality individualized breast cancer care. In that respect, we have previously shown that there is substantial variation in grading between Dutch pathology laboratories and between

pathologists of the same laboratory [15] in daily clinical practice. For example >50% of Dutch pathology laboratories showed significantly deviant ORs in a multivariate logistic regression model, in which the average laboratory was taken into account as a reference [15]. However, from these data, we cannot conclude which laboratories may grade inaccurately, as it can be questioned whether the nationwide average laboratory is the one grading most accurately. Nonetheless, considering the fact that grade is the determinative factor in >10.000 patients, while >50% of Dutch laboratories graded significantly different from the nationwide average, it may be clear that for many thousands of breast cancer patients, treatment may have been based on inaccurate histologic grading, leading to under- or over-treatment.

It may be clear that maximum effort to minimize differences in grading is required. Therefore, two encouraging initiatives, i.e. laboratory-specific feedback reports and training of pathologists by e-learning were successfully launched [29, 30]. This is an ongoing process, yet, and grading variation remains substantial even after training and feedback. Since histologic grade is of major influence in clinical decision making, additional avenues should be explored to ensure high quality oncological care [31]. It would for example also be interesting to see whether and how artificial intelligence could support pathologists in accurate biomarker assessment [32, 33]. In addition, Ki67 is a prognostic proliferation marker that may be of added value, but there is controversy with regard to its clinical utility in routine clinical management due to variations in analytical practice, lack of a standardized procedure for Ki67 assessment, and absence of consensus on cutoff values [34, 35].

Although this study formally only applies to the Dutch situation, we would like to emphasize that breast cancer guidelines in the Netherlands are generally in accordance with international breast cancer guidelines [5-7, 9, 10, 12, 36, 37], which all take histologic grade into account albeit with slight variations (Supplementary table 4). While for some guidelines primary focus has shifted to GEP-use as a first determinant in deciding whether aCT should be administered (mainly the 21-GS) [7, 10, 36], these guidelines also state that, when GEPs are not available, classic clinicopathologic biomarkers, including histologic grade should be considered. In this light, it is highly relevant to underline that GEPs are not accessible (due to the associated costs, i.e. $4.000 per 21-GS, $2.750 per 70-GS) nor applicable for every breast cancer patient [38-45]. For example, probably only up to half of the eligible women in the US actually [43, 44] receive GEP-testing. Hence, for many thousands, and maybe even the majority, of breast cancer patients, classic pathology biomarkers remain the most important indicator(s) for the need of adjuvant systemic treatment. Moreover, recent evidence shows that, in specific subgroups, the 21-GS score may be accurately predicted based on histologic grade and PR-receptor status, perhaps saving the need for expensive tests [46, 50]. Furthermore, the 70-GS is currently not covered by basic health care

insurance in the Netherlands, upon advice of the Dutch National Health Care institute, which states that the clinical utility of the 70-GS has not been sufficiently demonstrated, and that refraining from aCT may lead to a considerable increase in long-term mortality [51].

The short follow-up time unfortunately precluded outcome analyses. Thus, whether aCT de-escalation in almost 40% of breast cancer patients with a strict guideline aCT-indication did not compromise their outcome remains to be elucidated. This will be an important analysis that should be performed after longer follow-up.

An important limitation of this study is that patients who received neoadjuvant treatment were excluded from our data analyses. As neoadjuvant systemic therapy is the preferred initial approach in HER2-driven and TNBC [7], both with a poorer prognosis in general, these groups are most likely underrepresented in our study. Hence, we would like to emphasize that the 30,843 IBC patients within our study do not represent the overall population of breast cancer patients in the Netherlands. However, they do represent the nationwide population of breast cancer patients who were potentially eligible for adjuvant systemic therapy between 2013 and 2016. A final limitation of our study may be that we used synoptic pathology reports only, which excludes about 20% of Dutch breast cancer patients, as their reports are narrative [52]. However, previous comparisons with aggregated data, including all narrative pathology reports, showed a similar distribution of clinicopathologic variables [15]. Furthermore, synoptic reporting, in contrast to narrative reporting, not only results in an increased overall completeness of pathology reports, but has also been shown to improve patient care in colorectal cancer [53]. Lastly, synoptic reports enable the assembly of the laboratory-specific feedback reports [29], thereby creating insight and awareness in variation in biomarker assessment in daily pathology practice.

Conclusion
This study with nationwide real-world data on 30,843 Dutch primary breast cancer patients illustrate the increasing relevance of histologic grading with regard to tailoring adjuvant systemic therapy, as the indications for aCT and aET depend on histologic grade in roughly a third of breast cancer patients. This stresses the need for optimizing grading by pathologists as much as possible, to diminish the risk of worse patient outcome due to non-optimal treatment decisions.

Acknowledgements

5

References

1. World Health Organization (WHO). Breast Cancer; Early diagnosis and screening 2019. Available from: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. Accessed February 14 2020.

2. Netherlands Cancer Registry supplied by IKNL. Cijfers over kanker. Available from: www.cijfersoverkanker.nl. Accessed: Accessed February 14 2020.

3. National Institute for Health and Environment (RIVM) Ministry of Health, Wellness and Sport. Chance of breast cancer. Avalaible from: https://www.rivm.nl/bevolkingsonderzoek-borstkanker/wat-is-borstkanker/kans-op-borstkanker. Accessed October 10 2019.

4. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019;321(3):288-300.

5. Henry NL, Somerfield MR, Abramson VG, et al. Role of Patient and Disease Factors in Adjuvant Systemic Therapy Decision Making for Early-Stage, Operable Breast Cancer: American Society of Clinical Oncology Endorsement of Cancer Care Ontario Guideline Recommendations. J Clin Oncol. 2016;34(19):2303-11.

6. Henry NL, Somerfield MR, Abramson VG, et al. Role of Patient and Disease Factors in Adjuvant Systemic Therapy Decision Making for Early-Stage, Operable Breast Cancer: Update of the ASCO Endorsement of the Cancer Care Ontario Guideline. J Clin Oncol. 2019;37(22):1965-77.

7. Burstein HJ, Curigliano G, Loibl S, et al. Estimating the Benefits of Therapy for Early Stage Breast Cancer The St Gallen International Consensus Guidelines for the Primary Therapy of Early Breast Cancer 2019. Ann Oncol. 2019. 2019 Oct 1;30(10):1541-1557. doi: 10.1093/annonc/mdz235

8. Curigliano G, Burstein HJ, Winer EP., et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. Ann Oncol. 2017;28(8):1700-12.

9. Andre F, Ismaila N, Henry NL, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: ASCO Clinical Practice Guideline Update-Integration of Results From TAILORx. J Clin Oncol. 2019;37(22):1956-64.

10. Goetz MP, Gradishar WJ, Anderson BO, et al. NCCN Guidelines Insights: Breast Cancer, Version 3.2018. 2019;17(2):118.

11. The Netherlands Comprehensive Cancer Organization (IKNL). Oncoline: Cancer clinical practice guidelines (the Netherlands) 2019. Available from: https://www.oncoline.nl/. Accessed May 8, 2020.

12. Eisen A, Fletcher GG, Gandhi S, et al. Optimal systemic therapy for early breast cancer in women: a clinical practice guideline. Curr Oncol. 2015;22(Suppl 1):S67-81.

13. Rakha EA, El-Sayed ME, Lee AH, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol. 2008;26(19):3153-8.

14. 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(4):207.

15. van Dooijeweert C, van Diest PJ, Willems SM, et al. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: A nationwide study of 33,043 patients in the Netherlands. Int J Cancer. 2020;146)3):769-780.

16. Chavez-MacGregor M, Mittendorf EA, Clarke

CA, Lichtensztajn DY, Hunt KK, Giordano SH. Incorporating Tumor Characteristics to the American Joint Committee on Cancer Breast Cancer Staging System. Oncologist. 2017;22(11):1292-300.

17. van Steenhoven JEC, Kuijer A, Schreuder K, et al. The Changing Role of Gene-Expression Profiling in the Era of De-escalating Adjuvant Chemotherapy in Early-Stage Breast Cancer. Ann Surg Oncol. 2019;26(11):3495-501.

18. Casparie M, Tiebosch ATMG, Burger G, Blauwgeers, et al. Pathology Databanking and Biobanking in The Netherlands, a Central Role for PALGA, the Nationwide Histopathology and Cytopathology Data Network and Archive. Cellular Oncology : the Official Journal of the International Society for Cell Oncol. 2007;29(1):19-24.

19. Adams AL, Eltoum I, Krontiras H, Wang W, Chhieng DC. The effect of neoadjuvant chemotherapy on histologic grade, hormone receptor status, and HER2/neu status in breast carcinoma. Breast J. 2008;14(2):141-6.

20. Yin HF, Wang YH, Qin XQ, Zhang H, Li T, Ye JM, et al. [Effect of neoadjuvant chemotherapy on histologic grade and expression of biological markers in breast cancer]. Zhonghua zhong liu za zhi [Chinese journal of oncology]. 2009;31(11):858-62.

21. Zheng S, Zhang BL, Xiao T, Zou SM, Xue LY, Luo W, et al. [Comparison of histopathologic changes and expression of biomarkers in breast carcinoma before and after neoadjuvant chemotherapy]. Zhonghua bing li xue za zhi = Chinese journal of pathology. 2011;40(7):465-70.

22. Xian Z, Quinones AK, Tozbikian G, Zynger DL. Breast cancer biomarkers before and after neoadjuvant chemotherapy: does repeat testing impact therapeutic management? Hum Pathol. 2017;62:215-21.

23. Yang L, Zhong X, Pu T, Qiu Y, Ye F, Bu H. Clinical significance and prognostic value of receptor conversion in hormone receptor positive breast cancers after neoadjuvant chemotherapy. World J Surg Oncol. 2018;16(1):51.

24. Yang YF, Liao YY, Li LQ, Xie SR, Xie YF, Peng NF. Changes in ER, PR and HER2 receptors status after neoadjuvant chemotherapy in breast cancer. Pathol Res Pract. 2013;209(12):797-802.

25. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer. 1957;11(3):359-77.

26. 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(5):403-10.

27. The Netherlands Comprehensive Cancer Organisation (IKNL) (2017). Oncoline: Breast Cancer Guideline. Available from: https://www.oncoline.nl/borstkanker. Accessed March 4 2019.

28. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med. 2016;375(8):717-29.

29. van Dooijeweert C, van Diest PJ, Baas IO, van der Wall E, Deckers IA. 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: jclinpath-2019-206362.

30. van Dooijeweert C, Deckers I.A.G., de Ruiter E.J., et al. The effect of an e-learning module on grading variation of (pre) malignant breast lesions. Mod Pathol. 2020. Epub ahead of print: doi: 10.1038/s41379-020-0556-6

31. Nass SJ, Cohen MB, Nayar R, et al. Improving Cancer Diagnosis and Care: Patient Access to High-Quality Oncologic Pathology. Oncologist. 2019;24(10):1287-90.

5

32. Veta M, van Diest PJ, Jiwa M, Al-Janabi S, Pluim JP. Mitosis Counting in Breast Cancer: Object-Level Interobserver Agreement and Comparison to an Automatic Method. PloS One. 2016;11(8):e0161286.

33. Bejnordi EB, Veta M, van Diest PJ, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017;318(22):2199-210.

34. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. J Natl Cancer Inst. 2011;103(22):1656-64.

35. Penault-Llorca F, Radosevic-Robin N. Ki67 assessment in breast cancer: an update. Pathology. 2017;49(2):166-71.

36. Cardoso F, Kyriakides S, Ohno S, P, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. Ann Oncol. 2019;30(8):1194-220.

37. Duffy MJ, Harbeck N, Nap M, et al. Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). Eur J Cancer. 2017;75:284-98.

38. Orucevic A, Heidel RE, Bell JL. Utilization and impact of 21-gene recurrence score assay for breast cancer in clinical practice across the United States: lessons learned from the 2010 to 2012 National Cancer Data Base analysis. Breast Cancer Res Treat. 2016;157(3):427-35.

39. Roberts MC, Kurian AW, Petkov VI. Uptake of the 21-Gene Assay Among Women With Node-Positive, Hormone Receptor-Positive Breast Cancer. J Natl Compr Canc Netw. 2019;17(6):662-8.

40. Jasem J, Amini A, Rabinovitch R, et al. 21-Gene Recurrence Score Assay As a Predictor of Adjuvant Chemotherapy Administration for Early-Stage Breast Cancer: An Analysis of Use, Therapeutic Implications, and Disparity Profile. J Clin Oncol. 2016;34(17):1995-2002.

41. Press DJ, Ibraheem A, Dolan ME, Goss KH, Conzen S, Huo D. Racial disparities in omission of oncotype DX but no racial disparities in chemotherapy receipt following completed oncotype DX test results. Breast Cancer Res Treat. 2018;168(1):207-20.

42. Weldon CB, Trosman JR, Gradishar WJ, Benson AB, Schink JC. Barriers to the Use of Personalized Medicine in Breast Cancer. J Oncol Pract. 2012;8(4):e24-e31.

43. Davis BA, Aminawung JA, Abu-Khalaf MM, et al. Racial and Ethnic Disparities in Oncotype DX Test Receipt in a Statewide Population-Based Study. J Natl Compr Cancer Netw. 2017;15(3):346-54.

44. Zhang L, Hsieh MC, Petkov V, Yu Q, Chiu YW, Wu XC. Trend and survival benefit of Oncotype DX use among female hormone receptor-positive breast cancer patients in 17 SEER registries, 2004-2015. Breast Cancer Res Treat. 2020;180(2):491-501.

45. Ricks-Santi LJ, McDonald JT. Low utility of Oncotype DX(R) in the clinic. Cancer Med. 2017;6(3):501-7.

46. Wu SG, Zhang WW, Wang J, Lian CL, Sun JY, Chen YX, et al. Progesterone receptor status and tumor grade predict the 21-gene recurrence score of invasive lobular breast cancer. Biomark Med. 2019;13(12):1005-12.

47. Allison KH, Kandalaft PL, Sitlani CM, Dintzis SM, Gown AM. Routine pathologic parameters can predict Oncotype DXTM recurrence scores in subsets of ER positive patients: who does not always need testing? Breast Cancer Res Treat. 2012;131(2):413-24.

48. Huang JL, Kizy S, Marmor S, Altman A, Blaes A, Beckwith H, et al. Tumor grade and progesterone receptor status predict 21-gene recurrence score in early stage invasive breast carcinoma. Breast Cancer Res Treat. 2018;172(3):671-7.

49. Yoo SH, Kim T-Y, Kim M, et al. Development

of a Nomogram to Predict the Recurrence Score of 21-Gene Prediction Assay in Hormone Receptor–Positive Early Breast Cancer. Clin Breast Cancer. 2020;20(2):98-107.

50. Orucevic A, Bell JL, King M, McNabb AP, Heidel RE. Nomogram update based on TAILORx clinical trial results - Oncotype DX breast cancer recurrence score can be predicted using clinicopathologic data. Breast. 2019;46:116-25.

51. National Healthcare Institute of the Netherlands. Point of view: MammaPrint® in women with early stage breast cancer (reassessment) 2018. Available from: https://www.zorginstituutnederland.nl/binaries/zinl/documenten/standpunten/2018/09/27/standpunt-mammaprint-bij-vrouwen-met-vroeg-stadium-borstkanker-herbeoordeling/Standpunt+MammaPrint%C2%AE+bij+vrouwen+met+vroeg+stadium+borstkanker.pdf. Accessed: April 17 2020.

52. Foundation PALGA. Annual report 2018. Available from: https://www.palga.nl/assets/uploads/Jaarverslag/2018%20Palga%20web.pdf. Accessed October 10, 2019.

53. Sluijter CE, van Workum F, Wiggers T, van de Water C, Visser O, van Slooten HJ, et al. Improvement of Care in Patients With Colorectal Cancer: Influence of the Introduction of Standardized Structured Reporting for Pathology. JCO Clin Cancer Inform. 2019;3:1-12.

5

Excerpts of synoptic resection specimen pathology reports with invasive breast cancer (IBC) from the PALGA database between 2013 and 2016
n = 48,667 IBC reports
n = 42,705 IBC patients

Reports without a primary tumor (n=2,104)

n = 48,563 IBC reports
n = 40,871 IBC patients

Reports of patients who received neoadjuvant systemic therapy (n=5,829)

n = 40,734 IBC reports
n = 35,686 IBC patients

Second report of patients with synchronous tumors (n=4,170)

Reports of patients with >1 non-synchronous tumor (n=1,495)

Reports of patients with missings on tumor grade (n=409), ER/PR-receptor status (n=606) or HER2-receptor status (n=1,757)

n = 32,297 IBC reports
n = 32,297 IBC patients

Patients who could not be matched to the Dutch Cancer Registry (n=1,454)

n = 30,843 IBC reports
n = 30,843 IBC patients

Supplementary figure 1. Flowchart of included patients with invasive breast cancer to assess the role of histologic grading in tailoring adjuvant systemic therapy in the Netherlands

Supplementary table 1. Number of breast cancer patients for whom the indication for adjuvant chemotherapy (aCT) depends on histologic grade according to the Dutch breast cancer guideline between **2013 and 2016**

Patient- and tumor characteristics	Number of patients
All patients aged <35 years, with a N0 tumor ≤1cm*	48
All patients aged 35-70 years, with a N0 tumor between 1.1-2cm*	7,696
Total	**7,744 (25.1%)**

* grade I: no indication for chemotherapy, grade II-III: indication for chemotherapy

Supplementary table 2. Number of breast cancer patients from our dataset (2013-2016) for whom the indication for adjuvant chemotherapy (aCT) would depend on histologic grade according to the current Dutch breast cancer guideline **(2019)**

Patient- and tumor characteristics	Number of patients with an indication for aCT solely based on grade
Age <35 years, HER2 negative, N0/N0(i+)/N1(mi), tumor 1.1-2cm*	84
Age 35-70 years, ER/PR positive, HER2 negative, N1-3, tumor ≤2cm*	1,790
Age 35-70 years, ER/PR positive, HER2 negative, N0/N0(i+)/N1(mi), tumor 1.1-3cm**	8,995
Total number of patients (n=30.843)	**10,869 (35.2%)**

* chemotherapy for grade II-III tumors
** tumor 1.1-2cm: chemotherapy in case of grade III, tumor 2.1-3cm: chemotherapy in case of grade II-III

Supplementary table 3. Number of breast cancer patients from our dataset (2013-2016) for whom the indication for adjuvant endocrine therapy (aET) would depend on histologic grade according to the current Dutch breast cancer guideline **(2019**)

Patient- and tumor characteristics	Number of patients with an indication for aET based on grade
All ages, ER and/or PR positive (≥10% of the tumor cells show ER- and/or PR-specific staining on IHC),N0/N0(i+)/N1(mi), tumor 1.1-2cm	9,173
Total number of patients (n=30.843)	**9,173 (29.7%)**

* endocrine therapy for grade II-III tumors

5

Supplementary table 4. Indication for adjuvant chemotherapy (aCT) in early breast cancer according to the different international guidelines

Specific (inter)national guidelines

Cancer Care Ontario (CCO) guideline (12)

- High risk N0 tumor
  - Tumor >0.5cm with ≥1 of the following risk-factors
    - **Grade III**
    - Triple negative (ER-, PR-, HER2-)
    - Lymphovascular invasion
    - HER2 + receptor status
    - Oncotype DX recurrence score associated with an estimated 10-year relapse risk of ≥15%
  - Adjuvant!Online risk >10-15%

American Society of Clinical Oncology guideline (ASCO) (CCO guideline endorsement) (5, 6, 9)

- High risk N0 tumor (exception: tubular and mucinous carcinomas: possibly no aCT)
  - Tumor >0.5cm with ≥1 of the following risk-factors
    - **Grade III***
    - Triple negative (ER-, PR-, HER2-)
    - Lymphovascular invasion*
    - HER2 + receptor status
    - Oncotype DX recurrence score associated with an estimated 10-year relapse risk of >20%
  - Adjuvant!Online risk >10-15%

National Comprehensive Cancer Network (NCCN) guideline (10)

- Ductal, lobular, mixed, metaplastic; pT1, pT2 or pT3; pN0; >0.5cm
  - Strongly consider Oncotype DX
    - aCT <70 years for Recurrence score 26-30 or ≥31
  - No Oncotype DX
    - Take into account factors like N-status, tumor size, **tumor grade**, lymphovascular invasion, age, comorbid conditions

European Society for Medical Oncology (ESMO) (36)

- Luminal A-like (ER+, HER2-, Ki67 low, PR high, low-risk molecular signature (if available))
  - High disease burden (≥4LN, ≥T3)
- Luminal B-like (HER2-) (ER+, HER2- and either Ki67 high or PR low, high-risk molecular signature (if available))
  - Majority of cases aCT, dependent on individual risk score, expected response to aET and patient preference
- Luminal B-like (HER2+) (ER+, HER2+, Ki67, any PR)
  - All patients

Supplementary table 4. Continued

Specific (inter)national guidelines

St. Gallen Expert Panel 2019 (7)

• Genomic signatures preferred for basing the critical yes/no chemotherapy decision

• If genomic signatures are not available, integrate traditional pathology (tumor size, grade, ER/
PR and proliferation) to recommend aCT or not

* Grade 3 and presence of lymphovascular should generally not be used to drive decision
making when considered in isolation and must be interpreted in the overall clinical context,
also: no chemo for well differentiated luminal A like tumors

5

PART TWO

Initiatives to decrease variation
in biomarker assessment

6

Grading variation in 2,934 patients with ductal carcinoma in situ of the breast:
the effect of laboratory- and pathologist-specific feedback reports

Carmen van Dooijeweert
Paul J. van Diest
Inge O. Baas
Elsken van der Wall
Ivette A. G. Deckers

Abstract

Background

Histologic grade of ductal carcinoma *in situ* of the breast (DCIS) may become the single biomarker that decides whether patients will be treated. Yet, evidence shows that grading variation in daily practice is substantial. To facilitate quality improvement, feedback reports, in which laboratory-specific case-mix adjusted proportions per grade were benchmarked against other laboratories, were sent to the individual laboratories by March 1 2018. One year later, the effect of these feedback reports on inter-laboratory variation was studied.

Methods

Synoptic pathology reports of all pure DCIS resection specimens between March 1 2017 and March 1 2019 were retrieved from PALGA (the nationwide Dutch pathology registry). Laboratory-specific proportions per grade were compared to the overall proportion in the year before and after feedback. The absolute deviation for all three grades at once, represented by the overall deviation score (ODS), was calculated as the sum of deviations from the grade-specific overall proportions. Case-mix adjusted, laboratory-specific odds ratios (ORs) for high- (grade III) versus low-grade (grade I-II) DCIS were obtained by multivariable logistic regression.

Results

Overall, 2,954 DCIS reports from 31 laboratories were included. After feedback, the range between laboratories decreased by 22% and 6.5% for grades II and III, while an increase of 6.2% was observed for grade I. Both the mean ODS (27.2% to 24.1%) and maximum ODS (87.7% to 59.6%) decreased considerably. However, the range of case-mix adjusted ORs remained fairly stable and substantial (0.39 (95% CI: 0.18-0.86) to 3.69 (95% CI: 1.30-10.51)).

Conclusion

A promising decrease in grading variation was observed after laboratory-specific feedback for DCIS grades II-III, while this was not observed for DCIS grade I. Overall, grading variation remained substantial which needs to be addressed considering its clinical implications. Nationwide consensus on a classification, and training of (expert breast) pathologists, for example by e-learning, may help to further improve grading standardization.

Introduction

Treatment of ductal carcinoma *in situ* of the breast (DCIS) currently consists of surgery [1,2], radiotherapy [2-6] and sometimes even (low-dose) tamoxifen [2-4, 7-9]. However, it is believed that an unknown number of DCIS patients are treated for lesions that may never progress into invasive breast cancer [10-13]. Therefore, four randomized controlled clinical trials aim to identify a group of low-risk DCIS patients that, under active surveillance, may safely forgo surgical treatment [11, 12, 14-16].

For all trials, the main biomarker which identifies DCIS as being low-risk, is histologic (nuclear) grade, although different classifications are used by the LORD-, LORIS-, COMET-, and LARRIKIN trials [11, 12, 14-16]. The general hypothesis of these trials is that progression risk, or at least speed of progression is higher for high-grade lesions [17, 18], and if a low-grade DCIS does become invasive, it will be a low-grade invasive carcinoma with favorable characteristics and excellent survival rates after treatment [11, 19].

Besides the fact that grade may become the single biomarker that is used to decide whether patients are treated for their DCIS, grade already plays an important role in clinical patient management. For example, grade influences radiotherapy decisions (omitting a boost, considering partial breast irradiation) [1,6] and indicates (on biopsy) whether a sentinel lymph node procedure is required [1]. Thus, accurate, consistent, and reproducible grading is of key importance. However, we previously showed that variation in grading, between pathology laboratories and between pathologists within laboratories, is substantial in daily clinical practice on a nationwide scale in the Netherlands [20]. Furthermore, studies in which a set of DCIS was assessed by several pathologists showed significant inter-observer variation, regardless of the used classification, as well [21-23].

As studies have shown that quality of breast cancer care can be improved by auditing and benchmarking [24-29], the results of our previous study were sent to all participating Dutch pathology laboratories as feedback reports, in which their proportions per grade were benchmarked against other laboratories. This enabled pathologists to discuss and reflect upon their grading practices. The present study was conducted to investigate the effect of these feedback reports on grading variation between laboratories on a nationwide scale.

6

Methods

Data source

All data were retrieved from PALGA, the Dutch nationwide network and registry of histo- and cytopathology, which contains excerpts of all pathology reports from laboratories in the Netherlands [30]. Data within this database are pseudonymized in the laboratories themselves and by a trusted third party (ZorgTTP, Houten, the Netherlands). In addition, as data on the reporting pathologist was not available in PALGA, laboratories provided the pathologist information directly to the PALGA data-analyst. In a final step, the PALGA data-analyst anonymized all laboratories and pathologists to the researchers. This study was approved by the scientific and privacy committee of PALGA and data were retrieved and handled in compliance with the General Data Protection Regulation Act (GDPR).

Study population

We retrieved all synoptic pathology reports of patients with pure DCIS resection specimens (i.e. without a report of a known adjacent invasive breast cancer) in the Netherlands between March 1 2017 and March 1 2019 from PALGA (n=3,336). Per pathology report, patient- and tumor characteristics were extracted (sex, age, type of surgery, histologic grade and tumor size). Reports with missing data on histologic grade and/or tumor size were excluded (Supplementary 1). Pathology reports of residual *in situ* lesions after neoadjuvant treatment of primary invasive breast carcinomas were excluded. Furthermore, ipsilateral DCIS reports within six months of the previous DCIS resection specimen report were considered paired measurements of which we only included the first report.

In total, 38 out of 42 pathology laboratories in the Netherlands reported resection specimens via the synoptic (PALGA) protocol from March 1 2017 and onwards. Of these 38 laboratories, we included those that annually reported a minimum of 15 DCIS resection specimens within the protocol.

Feedback reports

Laboratory-specific feedback reports, with proportions per histologic grade of individual laboratories benchmarked against other anonymized laboratories (20), were sent out by March 1 2018. The general feedback report is available on the PALGA website (in Dutch only) [31], while laboratory-specific reports are only available to the individual laboratories themselves.

Laboratory-specific reports consisted of funnel plots, in which absolute differences in proportion of histologic grade, are plotted against the number of included IBC per laboratory. Within these funnel plots, the national proportions per grade with their 95% confidence intervals as limits were set as targets.

In addition, all laboratories were asked to provide coded information of the reporting pathologist per pathology report, to compare grading practices of the different pathologists within their laboratory. These data on pathologist' level were provided by ten laboratories, which as a result received feedback on both laboratory- and pathologist' level. Thus, pathologists within these ten laboratories were enabled to discuss and reflect upon both their personal- and overall laboratory grading practices.

As feedback reports were sent to the laboratories by March 1 2018, we considered the period from March 1 2017 up to and including March 1 2018 as pre-feedback period, while the period from March 2 2018 up to and including March 1 2019 is considered the post-feedback period.

DCIS grading classification

Histologic grade was defined in the pathology reports as grade I, II or III, without a specification of which classification was used. The Dutch guideline[1] recommends to use the classification of Holland [32]. However, we know from our previous research that numerous different guidelines are used by Dutch pathologists in daily practice [20].

Statistical analysis

Patient and tumor characteristics from the pathology reports were summarized in Table 1. Differences of these characteristics between pre- and post-feedback pathology reports were tested by means of a χ^2-test for categorical variables and by means of a Mann-Whitney U test for continuous variables.

Overall proportions for DCIS grades I-III were determined pre- and post-feedback. Absolute differences from the overall proportions per laboratory are presented in bar charts per feedback period for grades I-III. We calculated an overall deviation score (ODS), consisting of the sum of absolute deviations from the grade-specific overall proportions per period, to compare the absolute deviation for all grades at once. Differences in ODS of the individual laboratories between the pre- and post-feedback period were compared by means of a Wilcoxon signed-rank test.

We used several, arbitrary, definitions of change as a possible way to interpret the type of change in laboratories. Laboratories who showed an absolute change of ≤2% were defined as 'not shifting'. Among laboratories that showed an absolute change of >2%, laboratories moving closer to the overall mean after feedback were defined as 'less deviant', while laboratories moving further from the overall mean after feedback were defined as 'more deviant'.

A logistic regression analyses, providing odds ratios (ORs) and 95% confidence intervals (CI) per laboratory, was used to compare relative differences among laboratories. For the logistic regression model, grade was dichotomized into low-grade DCIS (grade I-II) and high-grade DCIS (grade III), as this is the same definition that is used by the majority of the clinical trials [11, 12, 14, 16, 20]. The laboratory best resembling the national distribution on low- (grade I-II) and high-grade (grade III) was chosen as reference laboratory. We performed a multivariate logistic regression analysis to correct for differences in case-mix. Variables were selected based on our previous research (20) and consisted of tumor size and type of surgery.

To compare differences in case-mix adjusted ORs of the individual laboratories, the positive OR difference, i.e. the difference of the laboratory OR to the reference OR of 1.00, was calculated. These positive OR differences of the individual laboratories were compared pre- and post-feedback by Wilcoxon signed rank test.

All analyses were performed by using IBM SPSS Statistics version 15.0.0.2.

Results

Patient-, tumor- and laboratory characteristics
In total, 2,954 DCIS resection specimen reports from 2,934 patients were included. For 20 patients, two pathology reports were included, as this either concerned a bilateral tumor (n=18) or an ipsilateral tumor >6 months after the first diagnosis (n=2). Of the included reports, 1,493 were defined as pre-feedback, while 1,461 were defined as post-feedback.

Characteristics of the included DCIS resection specimen reports are summarized in Table 1. All included pathology reports originate from 31/42 pathology laboratories. Four laboratories did not implement synoptic reporting between March 1 2017 and March 1 2019. The remaining seven laboratories were excluded as they synoptically graded less than 15 DCIS lesions within the PALGA protocol pre- and/or post-feedback. The overall number of DCIS reports before feedback ranged from 19 to 94 (median 48), whereas after feedback the number of synoptic pathology reports ranged from 19 to 86 (median 39).

Mean age (SD) at diagnosis was 59.8 (10.3) years and mean tumor size (SD) was 2.5 (2.3) cm. Overall, only 11 men were included (0.4%). Breast conserving surgery was performed in almost 70% of DCIS patients, although a small decrease was observed after feedback

(69.5% versus 66.6%, p=0.096). A significant change in distribution of histologic grade was observed after feedback (p=0.016), as the proportion of grade I increased (from 11.3% to 14.6%), whereas the proportion of grade III decreased (from 49.2% to 45.6%), while grade II remained fairly stable (39.5% versus 39.8%) (Table 1).

Table 1. Characteristics of the 2,954 included ductal carcinoma *in situ* (DCIS) lesions from the PALGA database between March 1 2017 and March 1 2018 (PRE feedback), and March 1 2018 to March 1 2019 (POST feedback).

	Total (n=2,954)	PRE (n= 1,493)	POST (n =1,461)	p-value**
Histologic grade, n (%)				
Grade 1	383 (13.0%)	169 (11.3%)	214 (14.6%)	**0.016**
Grade 2	1,171 (39.6%)	590 (39.5%)	581 (39.8%)	
Grade 3	1,400 (47.4%)	734 (49.2%)	666 (45.6%)	
Age (y)*	59.8 (10.3)	59.8 (10.5)	59.8 (10.1)	0.922
Sex, n (%)				0.790
Female	2,943 (99.6%)	1,487 (99.6%)	1,456 (99.7%)	
Male	11 (0.4%)	6 (0.4%)	5 (0.3%)	
Tumor size (cm)*	2.5 (2.3)	2.5 (2.3)	2.5 (2.3)	0.321
Type of surgery, n (%)				0.096
Mastectomy	944 (32.0%)	456 (30.5%)	488 (33.4%)	
Breast conserving	2,010 (68.0%)	1,037 (69.5%)	973 (66.6%)	

* Mean (SD)

** PRE vs POST categorical variables by Chi-square test, continuous variables by Mann Whitney U Test
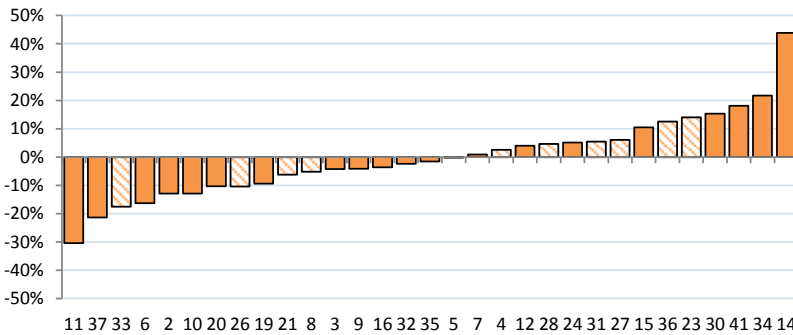
Inter-laboratory differences in histologic grading

For grade II, the total range between laboratories decreased markedly by 22% (from 9.1-83.3% to 10.0-62.2%) and, for grade III, to a lesser extent by 6.5% (from 16.7-75.8% to 21.3-73.9%) (Fig. 1). In contrast, the range between laboratories increased by 6.2% for grade I (from 0.0-21.1% to 2.7-30.0%) (Fig. 1).

After feedback, the maximum ODS decreased considerably from 87.7% to 59.6%, while the mean ODS also decreased from 27.2% to 24.1% (Fig. 2). The ODS of the individual laboratories did not differ significantly (Wilcoxon signed rank test, p=0.456).
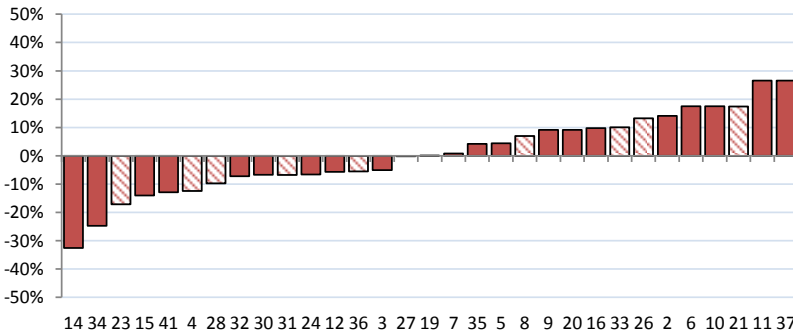
A: Before feedback reports: grade I
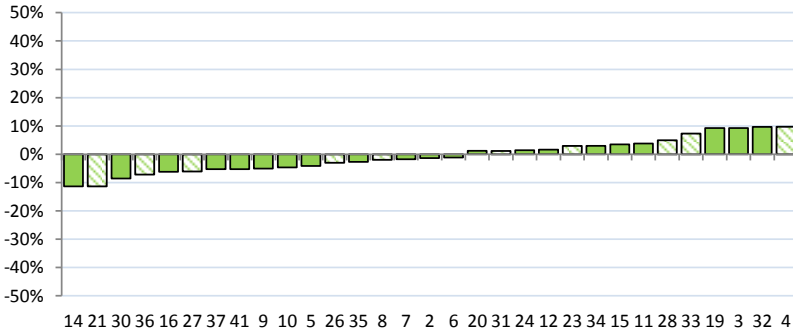
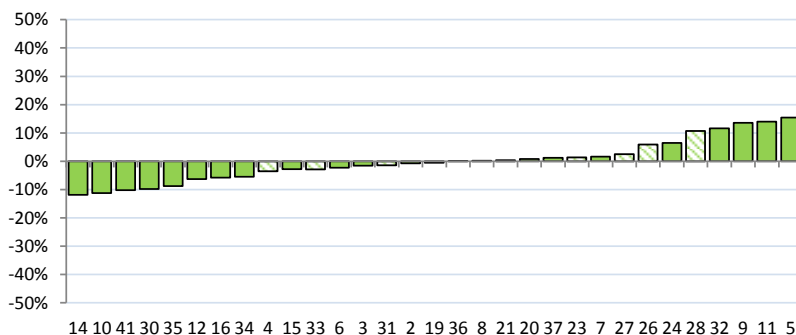
B: Before feedback reports: grade II

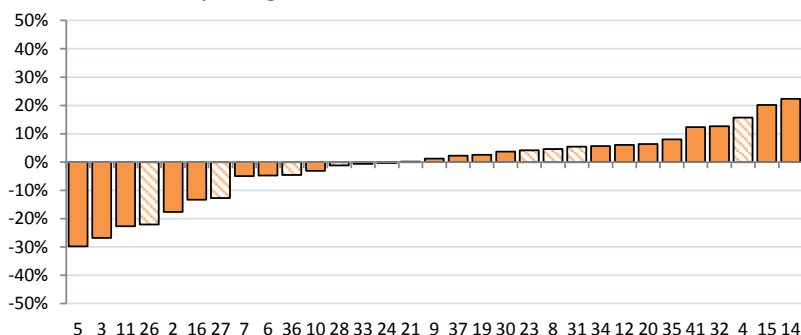**C: Before feedback reports: grade III**

Figure 1. Inter-laboratory (n=31) variation in histologic grading of ductal carcinoma *in situ* (DCIS) of the breast before (A-C) (n=1,493) and after feedback reports (D-F) (n=1,461). Percentages per laboratory show the absolute deviation from the national proportion per histologic grade for grade I (A+D), grade II (B+E) and grade III (C+F). Laboratory numbers for all sub-figures (A-F) correspond. All laboratories are ranged from lower (negative values) to higher proportions (positive values). A decrease in absolute range was observed for grades II and III; grade II (-22.0%), grade III (-6.2%). An increase was observed for grade I (+6.5%). Striped bars indicate laboratories who received feedback on pathologist' level (n=10).
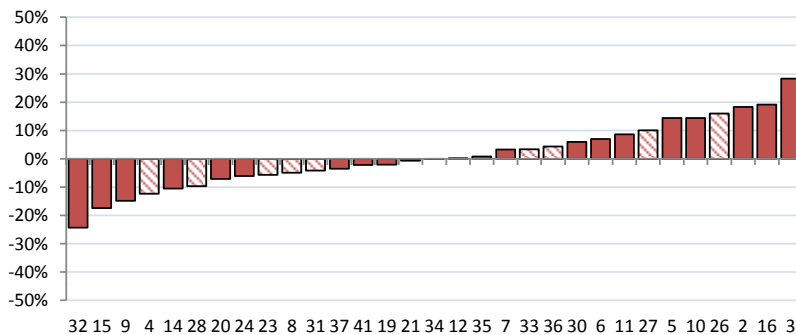
D: after feedback reports: grade I

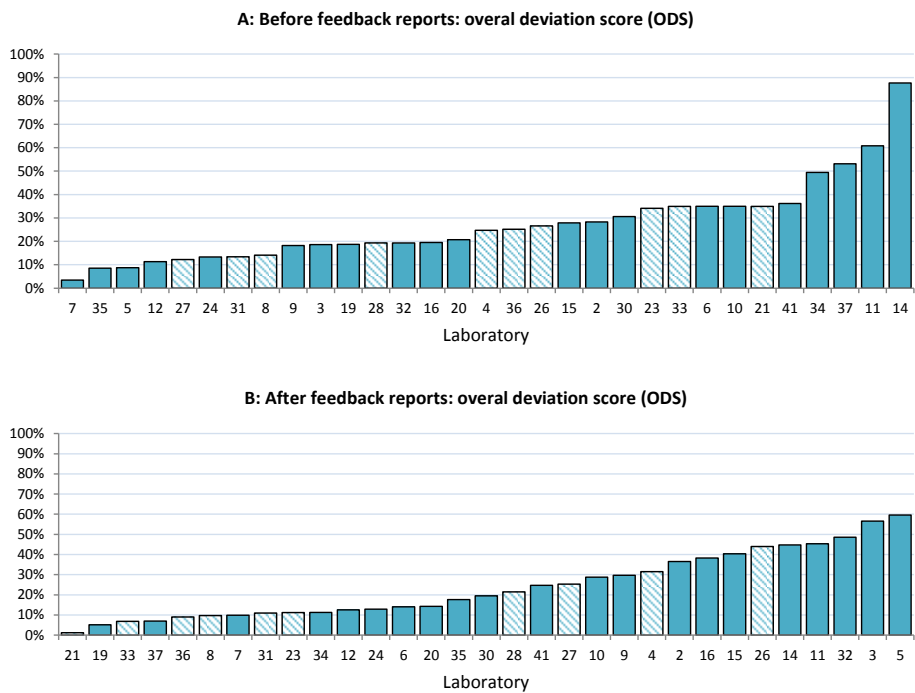**E: After feedback reports: grade II**

F: After feedback reports: grade III

Figure 1. Continued.

The majority of laboratories became less deviant after feedback for grades II and III (18 laboratories (58.1%) and 17 laboratories (54.8%) respectively), while 11 laboratories (35.5%) became more deviant for both grades. For grade I, the number of laboratories that became less deviant was similar to the number of laboratories that became more deviant; 11 (35.5%) versus 12 (38.8%) laboratories, while 8 (25.8%) laboratories showed a shift of ≤2% (Table 2).

A: Before feedback reports: overal deviation score (ODS)

B: After feedback reports: overal deviation score (ODS)

Figure 2. Inter-laboratory (n=31) variation in histologic grading of ductal carcinoma *in situ* (DCIS) of the breast (A) (n=1,493) and after feedback reports (B) (n=1,461). Each bar represents the overall deviation score (ODS) per laboratory. Laboratory numbers for A & B correspond. All laboratories are ranged from lower to higher ODS. The maximum ODS decreased from 87.7% to 59.6% after feedback. Striped bars indicate laboratories who received feedback on pathologist' level (n=10).

Laboratory 19 was chosen as reference laboratory in the multivariate logistic regression model, as it best resembled the national distribution for grade III versus grade I-II (mean deviation for grade III before and after feedback 1.1%). Laboratory-specific ORs ranged from 0.20 (95% CI: 0.06-0.65) to 3.39 (95% CI: 1.34-8.57) before feedback, resulting in an overall OR range of 3.19. After feedback, laboratory-specific ORs ranged from 0.39 (95% CI: 0.18-0.86) to 3.69 (95% CI: 1.30-10.51), with a corresponding OR range of 3.30 (Fig. 3). Subsequently, the overall OR range increased by 3.5% after feedback. Positive OR-differences within the laboratories did not significantly differ before and after feedback (Wilcoxon signed-rank test, p=0.886).

**Table 2.** Type of change in the laboratories after feedback per histologic grade

Type of change	Total number of laboratories (n=31)	Laboratories with feedback on pathologist' level (n=10)	Laboratories without feedback on pathologist' level (n=21)	p-value
Grade I				
No shift (≤2%)	8 (25.8%)	1 (10.0%)	7 (33.3%)	0.165*
Shift(>2%)				
Less deviant	11 (35.5%)	6 (60.0%)	5 (23.8%)	0.147**
More deviant	12 (38.7%)	3 (30.0%)	9 (42.9%)	
Grade II				
No shift (≤2%)	2 (6.5%)	1 (10.0%)	1 (4.8%)	0.579*
Shift (>2%)				
Less deviant	18 (58.1%)	6 (60.0%)	12 (57.1%)	0.732**
More deviant	11 (35.5%)	3 (30.0%)	8 (38.1%)	
Grade III				
No shift (≤2%)	3 (9.7%)	2 (20.0%)	1 (4.8%)	0.180*
Shift (>2%)				
Less deviant	17 (54.8%)	6 (60.0%)	11 (52.4%)	0.328**
More deviant	11 (35.5%)	2 (20.0%)	9 (42.9%)	

* p-value for shift vs. no shift between laboratories with and without feedback on pathologist' level
** p-value for type of change when laboratories shifted >2% after feedback between laboratories with and without feedback on pathologist' level

Feedback on pathologist' level
Compared to the slight decrease in mean ODS (from 28.8% to 27.5%) of laboratories who only received feedback on laboratory-level, the mean ODS of the ten laboratories who did also receive feedback on pathologist' level, showed a larger decrease from 24.0% to 17.1%. As for the type of change in laboratories, it seems that laboratories who also received feedback on pathologist' level showed more improvement, however, these differences were not statistically significant (Table 2).
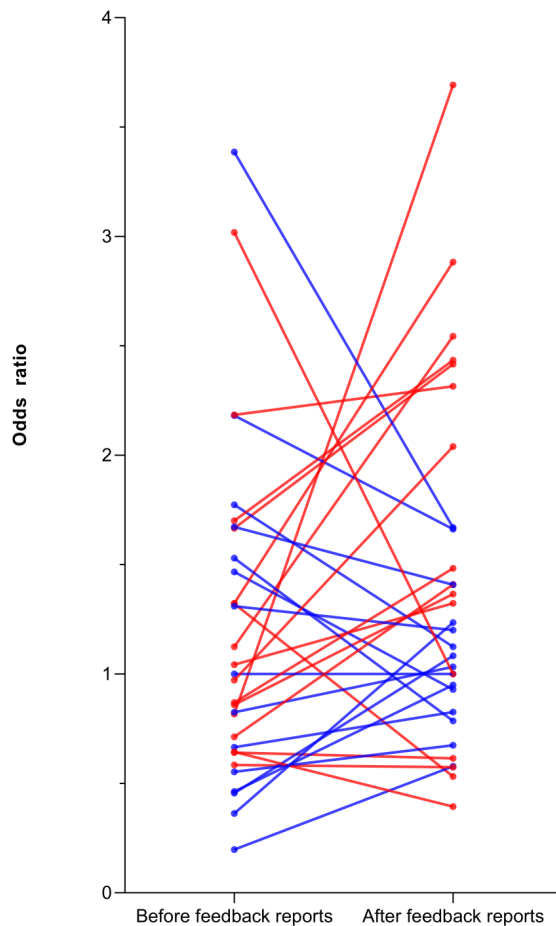
6

Figure 3. Case-mix adjusted odds ratios per laboratory before and after the feedback reports were calculated by multivariate logistic regression analyses for grade III versus grade I-II ductal carcinoma *in situ* (DCIS). ORs are adjusted for tumor size and type of surgery. Each laboratory is represented by two dots (before and after feedback) connected by one line. The color of the line and dots shows whether the OR after feedback shifted towards an OR of 1.00, and thus became less deviant (blue) or the OR after feedback shifted away from an OR of 1.00, and thus became more deviant (red) from the reference laboratory.

Discussion

Using data from structured pathology reports, we investigated the effect of feedback reports on nationwide inter-laboratory grading variation of DCIS. The decrease in absolute variation for grades II (6.5%) and III (22.0%), as well as the overall majority of laboratories becoming less deviant after feedback, and the decrease of both the mean and maximum ODS seem to indicate a promising decrease in DCIS grading variation. However, absolute variation increased for grade I (6.2%), and the range of case-mix adjusted ORs remained fairly stable and large after feedback. Furthermore, the absolute range between laboratories remains substantial, and maybe even clinically unacceptable, for all grades.

We hypothesize that the lack of consensus on a grading classification [13, 20], which is also reflected by the use of different classifications by the trials [11, 12, 14-16], may be the explanation for these mixed results as we believe that uniform grading starts with the use of single grading classification by all pathologists. Furthermore, in comparison to grading of invasive breast cancer, which is performed according to the modified Bloom and Richardson grading classification (Elston-Ellis modification) [33, 34], with scoring of the three subcategories (tubular differentiation, nuclear polymorphism and mitotic count) grading of DCIS is less standardized. Recently, dichotomous histopathological assessment of ductal carcinoma was studied as an alternative with (better) acceptable degrees of interobserver variability, however, interobserver variation remained considerable and at best acceptable [35]. In addition, other recent data showed that even among 35 expert breast pathologists, the threshold for comedonecrosis is highly variable [36]. This highlights the complexity of histologic grading of DCIS and the need for clear and uniform guidelines [36]. This not only important to the possible implementation of trials results into daily clinical practice, but it may also benefit the quality of the trials itself as central pathology review is not always carried out [37].

Interestingly, 7 out of 38 laboratories that used the synoptic PALGA protocol, were excluded for grading less than 15 DCIS resection specimens (via the protocol), which practically means they grade little over one (pure) DCIS specimen per month. For two laboratories this was because they likely only started using the protocol while another laboratory, stopped using the protocol for unknown reasons. Yet, for the remaining four laboratories, numbers per year (pre- or post-feedback) were fairly stable and low. We would like to emphasize, however, that pathologists within these laboratories may still report DCIS resection specimens outside the protocol, and therefore, they may grade more DCIS than our data would suggest. Nevertheless, if these are the actual numbers of DCIS' that are assessed in specific pathology laboratories per year, it may be questioned whether this is desirable, especially with regard to the complexity of histologic grading.

6

Easy data extraction on a large (nationwide) scale and an increased overall completeness of reports [38] were the main reasons to only include synoptic pathology reports. Furthermore, over 80% of breast resection specimens is reported via the synoptic protocol [39].

The mean numbers per histologic grade in this study (13.0% for grade I, 39.6% for grade II and 47.4% for grade III) are in line with other studies [20, 40, 41]. Interestingly, we observed a significant change in grade distribution after feedback as the proportion of grade I tumors increased by over 3%, while the proportion of grade III tumors decreased by a similar percentage. Whether this is initiated by the feedback reports or whether it reflects a true change in the population of DCIS patients remains unknown. Nonetheless, it did make it more difficult to interpret the results regarding deviations from the mean.

Overall, after analyzing the data in an absolute and relative manner, we did observe promising and positive changes, reflected by the decrease in absolute variation for DCIS grade II and III and the decrease of both the mean and maximum ODS. Furthermore, the majority of laboratories became less deviant after feedback for grades II and III. Hence, for grades II and III most deviant laboratories became less deviant, indicating that there is a decrease of the extremes, while changes of individual laboratories (ODS, positive OR-differences) were not significant. In contrast, the results for DCIS grade I showed an increase in the absolute range between laboratories, while the overall range of ORs remained stable and substantially large, ranging from 0.39 (95% CI: 0.18-0.86) to 3.69 (95% CI: 1.30-10.51). This shows that variation in histologic grading is still far from clinically acceptable levels.

Our results confirmed that feedback on the individual level (i.e. the pathologist) may be more effective than feedback on the general level (i.e. laboratories) [42-45]. We observed a larger absolute decrease of the mean ODS of laboratories who also received individual feedback. Furthermore, these laboratories also showed more improvement as compared to laboratories who only received feedback on laboratory-level. Due to the low number of laboratories who received pathologist-specific feedback reports, differences were not statistically significant.

The observed effects on grading variation may not exclusively be attributed to the feedback reports. However, we would like to emphasize that between 2016 and March 1 2019, besides the feedback reports, no other interventions or guideline changes took place. Furthermore, our previous paper [20] was only first published after feedback reports were sent to the individual pathology laboratories. We believe that these feedback reports may be a useful tool to, at least, monitor grading variation in daily clinical practice.

Conclusion

We observed a promising decrease in grading variation for DCIS grades II and III, while this was not observed for DCIS grade I. The overall variation for all grades remains substantial, and it seems that histologic grade is far from being a clinically acceptable biomarker for DCIS, let alone be the single biomarker that decides whether patients may be treated. In light of the current ongoing trials, improvement and standardization of DCIS grading is adamant. Continuing with feedback reports, especially on pathologist' level, helps to create awareness and may open the discussion about nationwide consensus on a single grading classification. In addition, adequate training of (expert breast) pathologists, according to a single classification system, for example by e-learning, may help to establish uniform grading in clinical practice over time.

Acknowledgements

6
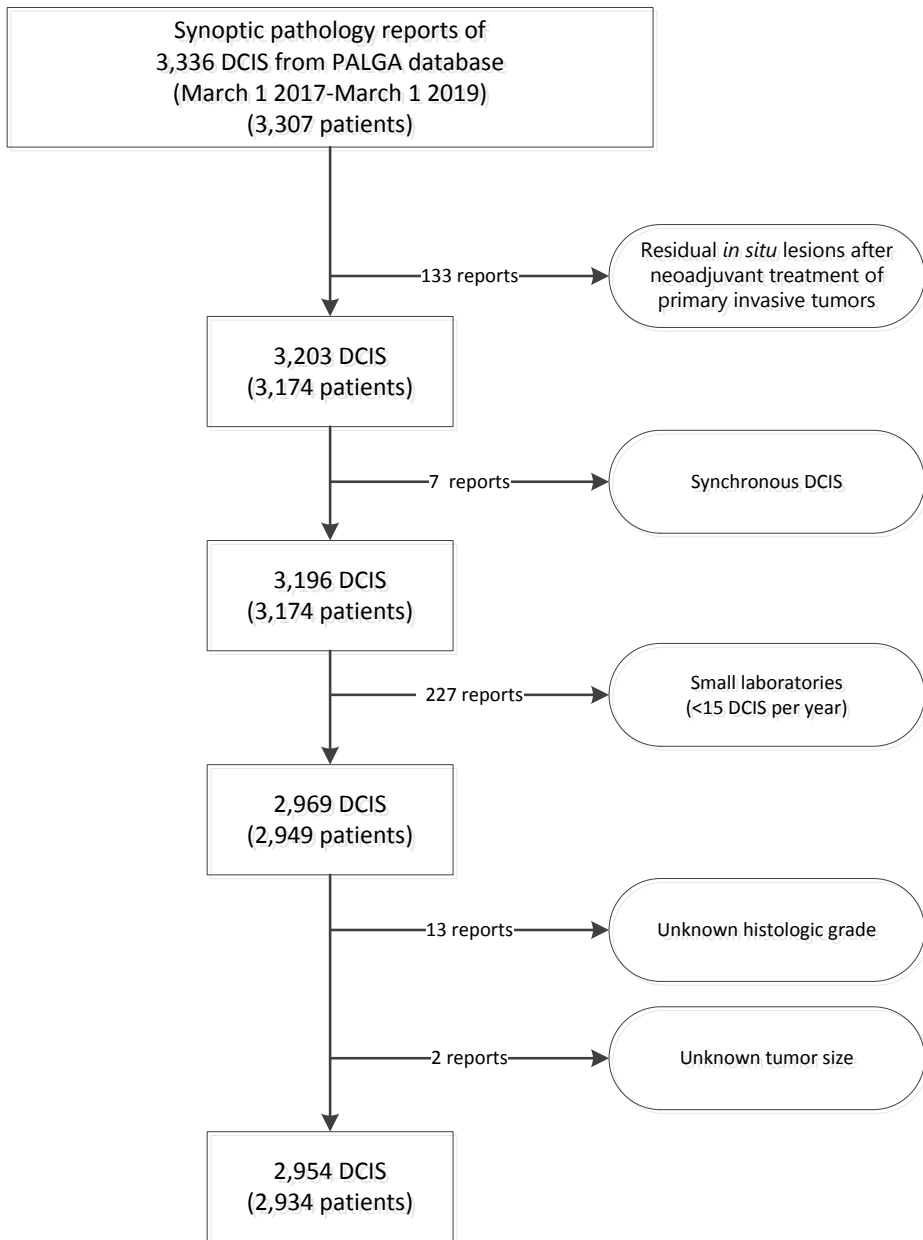
References

1. The Netherlands Comprehensive Cancer Organisation (IKNL) (2017). Oncoline: Breast Cancer Guideline. Available from: https://www.oncoline.nl/borstkanker. Accessed December 4 2019.

2. Curigliano G, Burstein HJ, Winer EP, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer. 2017. Ann Oncol. 2017;28(8):1700-12.

3. Correa C, McGale P, Taylor C, et al. Overview of the randomized trials of radiotherapy in ductal carcinoma in situ of the breast. J Natl Cancer Inst Monogr. 2010;2010(41):162-77.

4. Gorringe KL, Fox SB. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. Front Oncol. 2017;7:248.

5. Morrow M, Van Zee KJ, Solin LJ, et al. Society of Surgical Oncology-American Society for Radiation Oncology-American Society of Clinical Oncology Consensus Guideline on Margins for Breast-Conserving Surgery With Whole-Breast Irradiation in Ductal Carcinoma In Situ. J Clin Oncol. 2016;34(33):4040-6.

6. Smith BD, Bellon JR, Blitzblau R, et al. Radiation therapy for the whole breast: Executive summary of an American Society for Radiation Oncology (ASTRO) evidence-based guideline. Pract Radiat Oncol. 2018;8(3):145-52.

7. Staley H, McCallum I, Bruce J. Postoperative Tamoxifen for ductal carcinoma in situ: Cochrane systematic review and meta-analysis. Breast. 2014;23(5):546-51.

8. Cuzick J, Sestak I, Pinder SE, et al. Effect of tamoxifen and radiotherapy in women with locally excised ductal carcinoma in situ: Long-term results from the UK/ANZ DCIS trial. Lancet Oncol. 2011;12(1):21-9.

9. DeCensi A, Puntoni M, Guerrieri-Gonzaga A, et al. Randomized Placebo Controlled Trial of Low-Dose Tamoxifen to Prevent Local and Contralateral Recurrence in Breast Intraepithelial Neoplasia. J Clin Oncol. 2019;37(19):1629-37.

10. Benson JR, Jatoi I, Toi M. Treatment of low-risk ductal carcinoma in situ: is nothing better than something? Lancet Oncol. 2016;17(10):e442-e51.

11. 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(12):1497-510.

12. Francis A, Thomas J, Fallowfield L, et al. Addressing overtreatment of screen detected DCIS; the LORIS trial. Eur J Cancer. 2015;51(16):2296-303.

13. 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(4):285-92.

14. 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(3):e026797.

15. Groen EJ, Elshof LE, Visser LL, et al. Finding the balance between over- and under-treatment of ductal carcinoma in situ (DCIS). Breast. 2017;31:274-83.

16. Lippey J, Spillane A, Saunders C. Not all ductal carcinoma in situ is created equal: can we avoid surgery for low-risk ductal carcinoma in situ? ANZ J Surg. 2016;86(11):859-60.

17. Sanders ME, Schuyler PA, Dupont WD, Page DL. The natural history of low-grade ductal carcinoma in situ of the breast in women

treated by biopsy only revealed over 30 years of long-term follow-up. Cancer. 2005;103(12):2481-4.

18. Toss M, Miligy I, Thompson AM, et al. Current trials to reduce surgical intervention in ductal carcinoma in situ of the breast: Critical review. Breast. 2017;35:151-6.

19. Lopez-Garcia MA, Geyer FC, Lacroix-Triki M, Marchió C, Reis-Filho JS. Breast cancer precursors revisited: Molecular features and progression pathways. Histopathology. 2010;57(2):171-92.

20. van Dooijeweert C, van Diest PJ, Willems SM, Kuijpers C, Overbeek LIH, Deckers IAG. Significant inter- and intra-laboratory variation in grading of ductal carcinoma in situ of the breast: a nationwide study of 4901 patients in the Netherlands. Breast Cancer Res Treat. 2019;174(2):479-88.

21. Douglas-Jones AG, Morgan JM, Appleton MA, et al. Consistency in the observation of features used to classify duct carcinoma in situ (DCIS) of the breast. J Clin Pathol. 2000;53(8):596-602.

22. Elston CW, Sloane JP, Amendoeira I, et al. Causes of inconsistency in diagnosing and classifying intraductal proliferations of the breast. European Commission Working Group on Breast Screening Pathology. Eur J Cancer. 2000;36(14):1769-72.

23. Schuh F, Biazus 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(10):705-11.

24. Laronga C, Gray JE, Siegel EM, et al. Florida Initiative for Quality Cancer Care: improvements in breast cancer quality indicators during a 3-year interval. J Am Coll Surg. 2014;219(4):638-45.e1.

25. Sacerdote C, Bordon R, Pitarella S, et al. Compliance with clinical practice guidelines for breast cancer treatment: a population-based study of quality-of-care indicators in Italy. BMC Health Serv Res. 2013;13:28.

26. Srigley J, Lankshear S, Brierley J, et al. Closing the quality loop: facilitating improvement in oncology practice through timely access to clinical performance indicators. J Oncol Pract. 2013;9(5):e255-61.

27. van Bommel AC, Spronk PE, Vrancken Peeters MT, et al. Clinical auditing as an instrument for quality improvement in breast cancer care in the Netherlands: The national NABON Breast Cancer Audit. J Surg Oncol. 2017;115(3):243-9.

28. van Dam PA, Verkinderen L, Hauspy J, et al. Benchmarking and audit of breast units improves quality of care. Facts Views Vis Obgyn. 2013;5(1):26-32.

29. Veerbeek L, van der Geest L, Wouters M, et al. Enhancing the quality of care for patients with breast cancer: seven years of experience with a Dutch auditing system. Eur J Surg Oncol. 2011;37(8):714-8.

30. Casparie M, Tiebosch ATMG, Burger G, et al. Pathology Databanking and Biobanking in The Netherlands, a Central Role for PALGA, the Nationwide Histopathology and Cytopathology Data Network and Archive. Cell Oncol. 2007;29(1):19-24.

31. Deckers IAG (2018) Feedback variation in grading of invasive breast cancer and DCIS (in Dutch). PALGA, the nationwide network and registry of histo- and cytopathology in the Netherlands, Houten, The Netherlands. https://www.palga.nl/assets/uploads/Spiegelinfo/PALGA%20rapport%20-%20Terugkoppeling%20variatie%20in%20gradering%20mammacarcinoom_algemeen.pdf. Accessed December 4 2019.

32. Holland R, Peterse JL, Millis RR, et al. Ductal carcinoma in situ: a proposal for a new classification. Sem Diagn Pathol. 1994;11(3):167-80.

33. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer. 1957;11(3):359-77.

34. 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(5):403-10.

35. Dano H, Altinay S, Arnould L, et al. Interobserver variability in upfront dichotomous histopathological assessment of ductal carcinoma in situ of the breast: the DCISion study. Mod Pathol. 2019;33(3):354-366.

36. Harrison BT, Hwang ES, Partridge AH, Thompson AM, Schnitt SJ. Variability in diagnostic threshold for comedo necrosis among breast pathologists: implications for patient eligibility for active surveillance trials of ductal carcinoma in situ. Mod Pathol. 2019;32(9):1257-1262.

37. Van Bockstal MR, Agahozo MC, Koppert LB, van Deurzen CHM. A retrospective alternative for active surveillance trials for ductal carcinoma in situ of the breast. Int J Cancer. 2020;146(5):1189-1197.

38. Sluijter CE, van Lonkhuijzen LR, van Slooten HJ, Nagtegaal ID, Overbeek LI. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows Archiv. 2016;468(6):639-49.

39. PALGA Foundation. Annual report 2018. https://www.palga.nl/assets/uploads/ Jaarverslag/2018%20Palga%20web.pdf. Accessed December 4 2019.

40. van Luijt PA, Heijnsdijk EA, Fracheboud J, et al. The distribution of ductal carcinoma in situ (DCIS) grade in 4232 women and its impact on overdiagnosis in breast cancer screening. Breast Cancer Res. 2016;18(1):47.

41. Cutuli B, Lemanski C, De Lafontan B, et al. Ductal carcinoma in situ (DCIS): a French National Survey. Analysis of 2,125 patients. Clinical Breast Cancer. 2019.

42. Brehaut JC, Colquhoun HL, Eva KW, et al. Practice Feedback Interventions: 15 Suggestions for Optimizing Effectiveness. Ann Int Med. 2016;164(6):435-41.

43. Kluger AN, DeNisi A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin. 1996;119(2):254-84.

44. Ilgen DR, Fisher CD, Taylor MS. Consequences of individual feedback on behavior in organizations. Journal of Applied Psychology. 1979;64(4):349-71.

45. Archer-Kath J, Johnson DW, Johnson RT. Individual versus Group Feedback in Cooperative Groups. The Journal of Social Psychology. 1994;134(5):681-94.

```
                    ┌─────────────────────────────────────┐
                    │   Synoptic pathology reports of     │
                    │  3,336 DCIS from PALGA database      │
                    │   (March 1 2017-March 1 2019)        │
                    │          (3,307 patients)            │
                    └─────────────────────────────────────┘
                                     │
                                     │                    ╭─────────────────────────────╮
                                     ├──133 reports──────▶│ Residual in situ lesions after│
                                     │                    │  neoadjuvant treatment of    │
                                     ▼                    │   primary invasive tumors    │
                    ┌─────────────────────────────┐       ╰─────────────────────────────╯
                    │      3,203 DCIS             │
                    │      (3,174 patients)      │
                    └─────────────────────────────┘
                                     │                    ╭─────────────────────────────╮
                                     ├──7  reports──────▶│     Synchronous DCIS         │
                                     ▼                    ╰─────────────────────────────╯
                    ┌─────────────────────────────┐
                    │      3,196 DCIS             │
                    │      (3,174 patients)      │
                    └─────────────────────────────┘
                                     │                    ╭─────────────────────────────╮
                                     ├──227 reports─────▶│    Small laboratories        │
                                     │                    │   (<15 DCIS per year)        │
                                     ▼                    ╰─────────────────────────────╯
                    ┌─────────────────────────────┐
                    │      2,969 DCIS             │
                    │      (2,949 patients)      │
                    └─────────────────────────────┘
                                     │                    ╭─────────────────────────────╮
                                     ├──13 reports──────▶│  Unknown histologic grade    │
                                     │                    ╰─────────────────────────────╯
                                     │                    ╭─────────────────────────────╮
                                     ├──2 reports───────▶│    Unknown tumor size        │
                                     ▼                    ╰─────────────────────────────╯
                    ┌─────────────────────────────┐
                    │      2,954 DCIS             │
                    │      (2,934 patients)      │
                    └─────────────────────────────┘
```

Supplementary figure 1. Flowchart of included pathology reports of ductal carcinoma *in situ* of the breast (DCIS) resection specimens to assess the effect of feedback reports on variation in histologic grading of DCIS.

7

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

Carmen van Dooijeweert
Paul J. van Diest
Inge O. Baas
Elsken van der Wall
Ivette A. G. Deckers

Abstract

Aims

Histologic grade is widely used to guide management of invasive breast cancer (IBC). Yet, substantial inter- and intra-laboratory grading variation exists in daily pathology practice. To create awareness and to facilitate quality improvement, feedback reports, containing case-mix adjusted laboratory-specific grades benchmarked against other laboratories, were sent to the individual laboratories by March 1 2018. We studied the effect of these feedback reports on inter-laboratory grading variation up till one year later.

Methods

Overall, 17,102 synoptic pathology reports of IBC resection-specimens from 33 laboratories, obtained between March 1 2017 and March 1 2019 were retrieved from the Dutch pathology registry (PALGA). An overall deviation score (ODS), representing the sum of deviations from the grade-specific overall proportions, was calculated to compare the absolute deviation for all grades at once. Case-mix correction was performed by two multivariable logistic regression analyses, providing laboratory-specific odds ratios (ORs) for high versus low-grade IBC.

Results

After feedback, the overall range between laboratories decreased by 3.8%, 6.4%, 6.6% for grades I, II, and III, respectively. Though the mean ODS remained similar (13.8% versus 13.7%), the maximum ODS decreased from 34.1% to 29.4%. The range of laboratory-specific ORs decreased by 21.9% for grade III versus grades I-II.

Conclusions

An encouraging decrease in grading variation of invasive breast cancer was observed after laboratory-specific feedback. Nevertheless, overall grading variation remains substantial. In view of the important role of grading in patient management, it is adamant that not only feedback should be provided on a regular basis but also other interventions, such as additional training, are required.

Introduction

To date, histologic grade is widely used to guide therapeutic management of invasive breast cancer [1-4] as it remains one of the most well established prognostic factors [2, 3, 5]. When strictly adhering to the current Dutch guideline [1], which is similar to the rest of Europe [4], grade even indicates the need for adjuvant chemotherapy in approximately one third of breast cancer patients in the Netherlands [6]. Furthermore, grading is used to guide radiotherapy decisions [1, 4, 7] and the use of genetic profiling tests [1, 4, 8-10].

Despite its important role in patient management, we previously showed that substantial, and clinically relevant, variation in grading of IBC exists on a nationwide scale in daily clinical practice in the Netherlands [6]. Studies in which multiple IBC lesions were graded by several pathologists also showed that reproducibility was no more than moderate [11-14]. This suggests that patients may be under- and over-graded in specific pathology laboratories and or by specific pathologists, which may subsequently result in under-and over-treatment of a substantial number of breast cancer patients [6]. As this may influence outcome, including exposure to unnecessary toxicity, it is clear that standardized histologic grading is of key-importance.

The results of our previous study [6] were sent to the individual laboratories as feedback reports by the nationwide Dutch Pathology Registry (PALGA) to facilitate quality improvement, as auditing and benchmarking improves quality of breast cancer care [15-20]. By benchmarking their laboratory-specific proportions per histologic grade against other laboratories, pathologists in individual laboratories were enabled to discuss and reflect on their grading practices, and could conclude that adaptations were necessary.

This study was conducted to examine the effect of the case-mix adjusted, laboratory-specific feedback reports on the inter-laboratory variation in histologic grading of IBC using real-life data from synoptic (structured) pathology reports in the Netherlands.

Methods

Data source

Data were retrieved from PALGA, the nationwide network and registry of histo- and cytopathology in the Netherlands, which contains excerpts of all pathology reports from Dutch laboratories, with nationwide coverage since 1991 [21]. All data within the PALGA database are pseudonymized both in the laboratories and by a trusted third party (ZorgTTP, Houten, the Netherlands). All pathology laboratories were anonymized to the researchers by PALGA in a

final step. Laboratories who wanted to receive feedback on pathologist' level (in addition to the overall laboratory-feedback), were asked to send their local pathologist information to PALGA, since the PALGA-database did not contain pathologist' information before 2019. All data were retrieved and handled in compliance with the General Data Protection Regulation Act (GDPR) and this study was approved by the scientific and privacy committee of PALGA.

Study population
All synoptic pathology reports of patients with IBC resection specimens in the Netherlands between March 1 2017 and March 1 2019 were retrieved from PALGA (n=25,420) (Supplementary 1).

Overall, 38 out of 42 Dutch pathology laboratories used the synoptic (PALGA) protocol from March 1 2017 and onwards. Of these laboratories, we included those that reported at least 50 IBC resection specimens per year.

We excluded all resection specimen reports of patients with neoadjuvant treatment as grading may be influenced by chemotherapy [22-24]. Furthermore, synchronous IBC, defined as an ipsilateral lesion within six months of the previous IBC resection report, were considered paired measurements of which we only included the first report (Supplementary 1).

Per pathology report, we extracted patient characteristics (sex, age, type of surgery) and tumour characteristics (tumour size, histologic subtype, histologic grade, estrogen- (ER) and progesterone- (PR) receptor status, and HER2-receptor status). Reports with any missing data on one of the patient- or tumour characteristics were excluded from further data-analysis (Supplementary 1).

Feedback reports
Laboratory-specific feedback reports, regarding the variation in grading of IBC between January 1 2013 and December 31 2016, were sent to the laboratories by PALGA by March 1 2018. These feedback reports showed the laboratory specific proportions per histologic grade, benchmarked against the overall national proportions and the proportions of the other anonymized laboratories. Thereby, laboratories were enabled to discuss and reflect on their grading practice, and perhaps conclude that adaptations were necessary, The general feedback report is available on the PALGA website (in Dutch only) [25].

Ten laboratories provided (coded) pathologist information for their data, which gave these pathologists the advantage to benchmark their own grading practice against other pathologists in their laboratory and to the national mean. According to literature, this type of individual feedback is more effective than providing general (laboratory-level) data only [26-29].

Feedback reports were sent to the laboratories by March 1 2018, which resulted in a pre-feedback group of synoptic pathology reports from March 1 2017 to March 1 2018, and a post-feedback group of synoptic pathology reports from March 2 2018 to March 1 2019.

Histologic grading

Histologic grading of IBC was determined according to the modified Bloom and Richardson guideline (Elston-Ellis modification) [30,31], with a score of 1-3 on its three components (tubule formation, nuclear polymorphism and mitotic count). This results in a total score and subsequent grade (3-5 = grade I, 6-7 = grade II, 8-9 = grade III).

Statistical analysis

Patient and tumour characteristics were summarized and differences between the pre- and post-feedback group were tested by means of a χ^2-test for categorical variables and by a non-parametric Mann-Whitney U test for continuous variables.

Overall proportions per grade (I, II, and III) were determined before and after the feedback reports and considered the national proportions. The absolute differences from the national proportion per laboratory are presented in bar charts per grade for the pre- and post-feedback period. Laboratories who also received feedback on pathologist' level are indicated by striped bars.

An overall deviation score (ODS) was computed to compare the absolute deviation for all three grades at once. The ODS was calculated by the sum of absolute deviations from the grade-specific national proportions per period (pre- and post-feedback). Differences in ODS of individual laboratories before and after feedback were compared using a Wilcoxon signed rank test.

As a possible way to interpret the type of change in laboratories after feedback, we used multiple definitions of change. First, we arbitrarily defined laboratories with an absolute change of two percent or less as 'not shifting'. Second, in case of an absolute change of more than two percent, we defined two types of change. Laboratories with a smaller deviation from the overall mean were defined as 'less deviant'. Similarly, laboratories who became more deviant from the overall mean were defined as 'more deviant'.

To compare relative differences among laboratories, we used a logistic regression model, providing odds ratios (OR) and 95% confidence intervals (CI) per laboratory. We performed two logistic regression analyses, with different definitions of low- and high-grade IBC, as there is no clear binary cut-off in clinical practice. For example, grade III is considered a risk factor (high-grade) according to radiotherapy guidelines [1, 4, 7], whereas according to

chemotherapy guidelines grades II-III is considered a risk factor (high-grade) with possible subsequent therapy consequences [1, 4]. Therefore, in one logistic regression analysis, low-grade IBC was defined as grades I-II and high-grade IBC as grade III, whereas in the other logistic regression analysis low-grade IBC was defined as grade I and high-grade IBC as grades II-III.

For the choice of reference laboratory, we arbitrarily chose the laboratory best resembling the national distribution with regard to the specific logistic regression analysis. Multivariable logistic regression analyses were performed to correct for differences in case-mix. Case-mix variables were selected based on our previous research [6] and included age, tumour size, type of surgery, histologic subtype, HER2-receptor status, and hormone-receptor status. Hormone-receptor status was considered positive when either or both the estrogen (ER) or progesterone (PR) receptors were positive, and was taken into account as a binary variable (either positive or negative). According to the current Dutch guideline [1], the receptor status for ER and PR is considered positive when ≥10% of the tumour cells show ER- and PR staining by immunohistochemistry (IHC). The overall number of males was too low to take into account in a multivariate model, however, males did not cluster in specific laboratories. To compare differences in the case mix-adjusted ORs of the individual laboratories, we calculated the positive OR difference (i.e. the difference of a laboratory-specific OR to the reference OR of 1.00) both pre- and post-feedback and compared the differences of the individual laboratories by a Wilcoxon signed rank test for both multivariable logistic regression analyses (grade I vs. grades II-III and grades I-II versus grade III).

Lastly, the effect of feedback on pathologist' level was tested by comparing the mean ODS before and after feedback, between the laboratories who received feedback both on pathologist' and laboratory level and laboratories who received feedback on laboratory level only. In addition, type of change between these groups was compared by means of a χ^2-test.

All statistical analyses were performed by using IBM SPSS Statistics version 25.0.0.2. Values of below 0.05 were considered statistically significant.

Results

Patient-, tumour- and laboratory characteristics
A total of 17,102 IBC synoptic resection specimen reports from 16,734 patients were included in our data analysis. For some patients, more than one pathology report was included as this concerns either a bilateral tumour or an ipsilateral tumour that was

reported >6 months after the first diagnosis (*Supplementary 1*). Of the included reports, 8,767 were reported before and 8,335 were reported after feedback reports were sent to the laboratories by PALGA.

All patients originated from 33/42 Dutch pathology laboratories, as four laboratories did not implement synoptic reporting between March 1 2017 and March 1 2019 and five laboratories graded less than 50 IBC lesions within the synoptic PALGA protocol per period (pre- and/or post-feedback). The number of synoptic IBC reports per laboratory ranged from 64 to 613 (median 239) in the year before the feedback reports, while the number of synoptic pathology reports per laboratory in the year after the feedback reports ranged from 52 to 637 (median 207). Characteristics of all included IBC resection specimen reports are listed in *Table 1*.

The overall mean age (SD) at diagnosis was 63.2 (11.9) years and patients were primarily female (99.2%). Breast conserving surgery was performed in approximately two-thirds of patients (68.2%). The majority of tumours were of ductal (not otherwise specified) subtype (78.2%), with a positive ER/PR status (89.8%), whereas only a small minority of tumours had a positive HER2-receptor status (7.9%). Most characteristics, including age, sex, tumour size, type of surgery, and histologic subtype, we similarly distributed pre- and post-feedback. A minimal but significant increase of hormone-receptor positive tumours was observed after the feedbacks reports, while a significant decrease was observed for HER2-receptor positive tumours (p=0.010).

Overall national proportions for IBC grades I, II and III were respectively 30.5%, 49.5% and 20.0% before the feedback reports, while IBC grades I, II and III were reported in respectively 32.0%, 49.2% and 18.8% after the feedback reports (p=0.048).

Inter-laboratory differences in histologic grading
After feedback, the total range between laboratories decreased for all grades; 3.8% for grade I (from 17.5-45.5% to 17.3-41.5%), 6.4% for grade II (from 34.3-64.5% to 35.0-58.8%) and 6.6% for grade III (from 10.9-37.1% to 9.9-29.5%) (*Figure 1*).

The mean overall ODS remained similar after feedback (13.8 vs. 13.7%), which is also reflected by the similar ODS of individual laboratories (Wilcoxon signed rank test, p=0.955). The maximum ODS, however, decreased from 34.1% to 29.4% (*Figure 2*).

Overall, 11 (33.3%), 13 (39.4%) and 16 (48.5%) of laboratories showed no shift (≤2%) after feedback for grades I, II and III (*Table 2*).

7

TABLE 1. Characteristics of the 17,102 included invasive breast cancer (IBC) resection specimen reports from the PALGA database between March 1 2017 and March 1 2019.

	Total (n=17,102)	PRE (8,767)	POST (8,335)
Histologic grade			
Grade I	5,337 (31.2%)	2,672 (30.5%)	2,665 (32.0%)
Grade II	8,445 (49.4%)	4,344 (49.5%)	4,101 (49.2%)
Grade III	3,320 (19.4%)	1,751 (20.0%)	1,569 (18.8%)
Age (y)*	63.2 (11.9)	63.0 (11.9)	63.4 (11.9)
Sex, n (%)			
Female	16,971 (99.2%)	8,962 (99.1%)	8,279 (99.3%)
Male	131 (0.8%)	75 (0.9%)	56 (0.3%)
Tumor size (cm)*	1.8 (1.3)	1.8 (1.3)	1.7 (1.3)
Type of surgery, n (%)			
Mastectomy	5,433 (31.8%)	2,844 (32.4%)	2,589 (31.1%)
Breast conserving	11,669 (68.2%)	5,923 (67.6%)	5,746 (68.9%)
Histologic subtype, n (%)			
Ductal	13,373 (78.2%)	6,886 (78.5%)	6,487 (77.8%)
Lobular	2,353 (13.8%)	1,172 (13.4%)	1,181 (14.2%)
Other	1,376 (8.0%)	709 (8.1%)	667 (8.0%)
ER/PR-receptor status, n (%)			
Negative	1,740 (10.2%)	939 (10.7%)	801 (9.6%)
Positive	15,362 (89.8%)	7,828 (89.3%)	7,534 (90.4%)
HER2-receptor status, n (%)			
Negative	15,753 (92.1%)	8,030 (91.6%)	7,723 (92.7%)
Positive	1,349 (7.9%)	737 (8.4%)	612 (7.3%)

* Mean (SD)
PRE = cases before, and POST = cases after sending feedback reports to individual laboratories.

Among laboratories that shifted >2% after feedback, the number of laboratories who became more deviant was similar to the number of laboratories that became less deviant after feedback (30.3% vs. 36.4% for grade I, 30.3% vs. 30.3% for grade II and 24.2% vs. 27.3% for grade III respectively) (*Table 2*).

For the multivariate logistic regression analysis of grade III versus grades I-II, laboratory 30 had the lowest mean deviation from the national proportion before and after feedback for grade III (0.2%) and was chosen as reference laboratory. Before feedback, adjusted ORs ranged from 0.37 (95% CI: 0.21-0.67) to 2.15 (95% CI: 1.26-3.67). After feedback, the range of adjusted ORs decreased from 0.37 (95% CI: 0.20-0.68) to 1.76 (95% CI: 1.01-3.07) (*Figure 3a*). Consequently, the absolute overall OR range decreased by 21.9% from 1.78 to 1.39. Positive OR differences of the individual laboratories did not significantly differ (Wilcoxon signed rank test, p=0.702).

For the multivariate logistic regression analysis of grades II-III versus grade I, laboratory 32 had the lowest mean deviation from the national proportion before and after feedback for grade I (0.9%) and was chosen as reference laboratory. Before feedback, adjusted ORs ranged from 0.48 (95% CI: 0.29-0.77) to 2.00 (95% CI: 1.10-3.65), resulting in the absolute overall OR range of 1.52. After feedback, the range of adjusted ORs slightly increased (10.5%), i.e. from 0.42 (95% CI: 0.26-0.67) to 2.10 (95%CI: 1.24-3.58) with corresponding absolute overall OR range of 1.68 (*Figure 3b*). Positive OR differences of the individual laboratories did not significantly differ (Wilcoxon signed rank test, p=0.640).

7

Feedback on pathologist' level
Ten of the included laboratories received feedback both on laboratory- and on pathologist'-level (*Figure 1*, *Figure 2*; striped bars). Although the mean pre-feedback ODS of these laboratories was lower (10.7%) than the mean pre-feedback ODS of laboratories who only received feedback on laboratory level (15.1%), both groups did not show noteworthy changes after feedback (10.7% and 15.0%, respectively). Furthermore, type of change in laboratories after feedback did not significantly differ for both groups (*Table 2*). Yet, a significantly higher proportion of laboratories who received feedback on pathologist' level showed no shift after feedback for grade I, whereas a similar pattern was observed for grade III (p=0.103), whereas this was not observed for grade II.

A: Before feedback reports: grade I
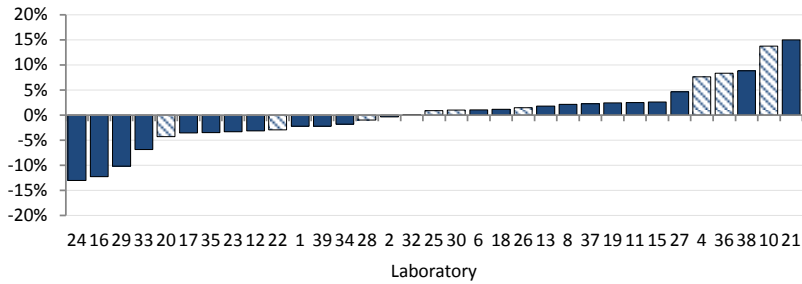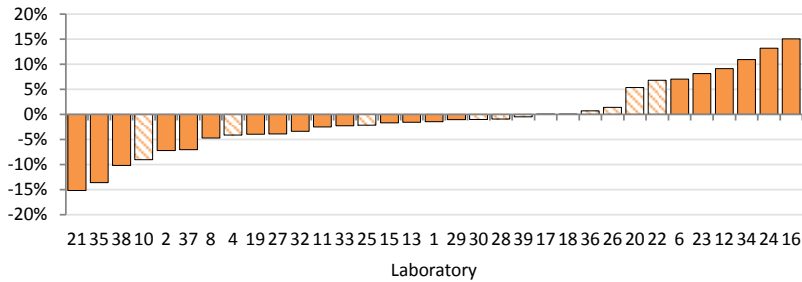
B: Before feedback reports: grade II

C: Before feedback reports: grade III
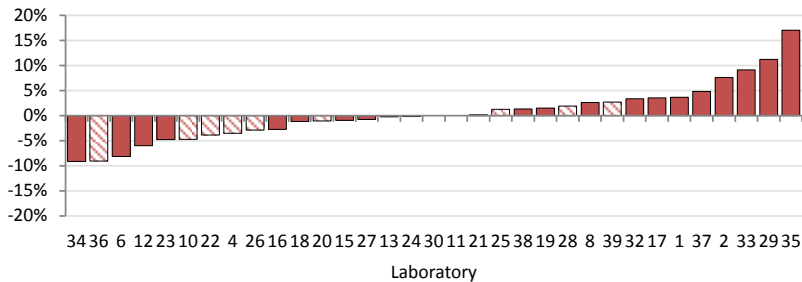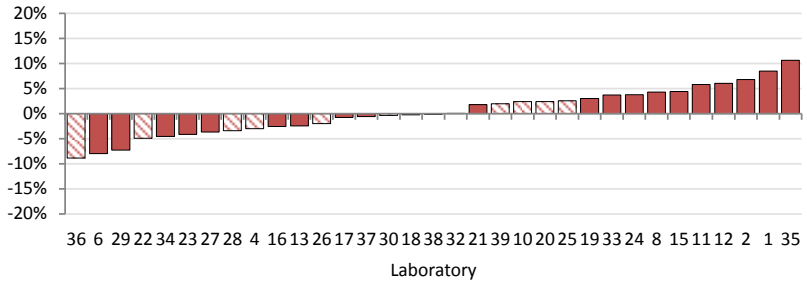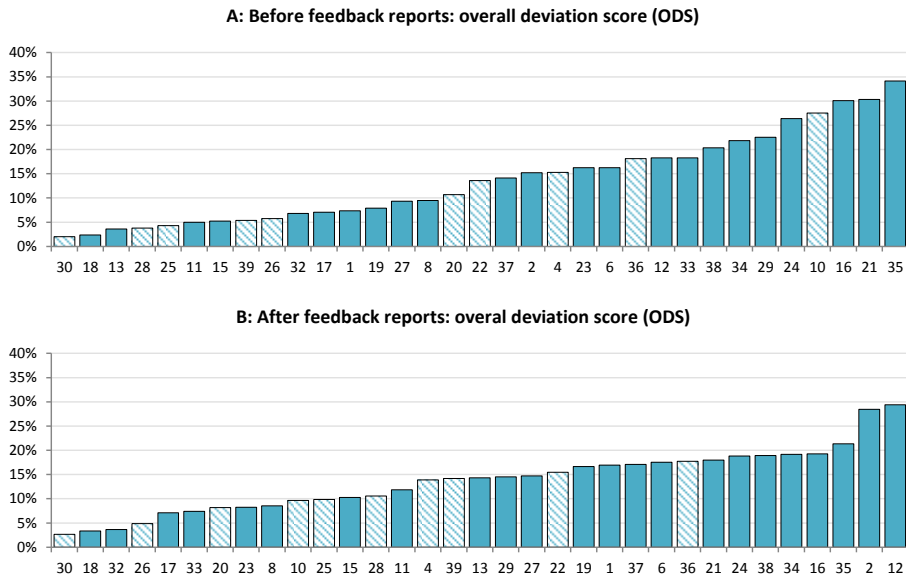

Figure 1. Inter-laboratory (n=33) variation in histologic grading of invasive breast cancer before (A-C) (n=8,767) and after feedback reports (D-F) (n=8,335). Percentages per laboratory show the absolute deviation from the national proportion per histologic grade for grade I (A+D), grade II (B+E) and grade III (C+F). Laboratory numbers for all sub-figures (A-F) correspond. All laboratories are ranged from lower (negative values) to higher proportions (positive values). Striped bars indicate laboratories who received feedback on pathologist' level (n=10).

D: After feedback reports: grade I

E: After feedback reports: grade II

F: After feedback reports: grade III

Figure 1. Continued.

A: Before feedback reports: overall deviation score (ODS)
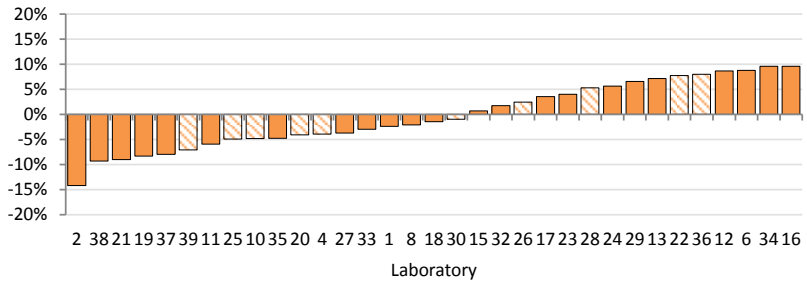
B: After feedback reports: overal deviation score (ODS)

Figure 2. Inter-laboratory (n=33) variation in histologic grading of invasive breast cancer before (A) (n=8,767) and after feedback reports (B) (n=8,335). Each bar represents the overall deviation score (ODS) per laboratory. Laboratory numbers for A & B correspond. All laboratories are ranged from lower to higher ODS. Striped bars indicate laboratories who received feedback on pathologist' level (n=10).

TABLE 2. Type of absolute change in the 33 included laboratories after feedback per histologic grade, with differentiation between laboratories who received feedback on pathologist level (n=10) and laboratories who only received feedback on laboratory-level (n=23).

Type of change	Number of laboratories (n=33)	Laboratories with feedback on pathologist' level (n=10)	Laboratories without feedback on pathologist' level (n=23)	p-value
Grade I				
No shift (≤2%)	11 (33.3%)	6 (60.0%)	5 (21.7%)	**0.032***
Shift (>2%)				
Less deviant	10 (30.3%)	3 (30.0%)	7 (30.4%)	0.190**
More deviant	12 (36.4%)	1 (10.0%)	11 (47.8%)	
Grade II				
No shift (≤2%)	13 (39.4%)	4 (40.0%)	9 (39.1%)	0.963*
Shift (>2%)				
Less deviant	10 (30.3%)	3 (30.0%)	7 (30.4%)	1.000**
More deviant	10 (30.3%)	3 (30.0%)	7 (30.4%)	
Grade III				
No shift (≤2%)	16 (48.5%)	7 (70.0%)	9 (39.1%)	0.103*
Shift (>2%)				
Less deviant	8 (24.2%)	1 (10.0%)	7 (30.4%)	0.600**
More deviant	9 (27.3%)	2 (20.0%)	7 (30.4%)	

* p-value for shift vs. no shift between laboratories with and without feedback on pathologist' level

** p-value for type of change when laboratories shifted >2% after feedback between laboratories with and without feedback on pathologist' level

7
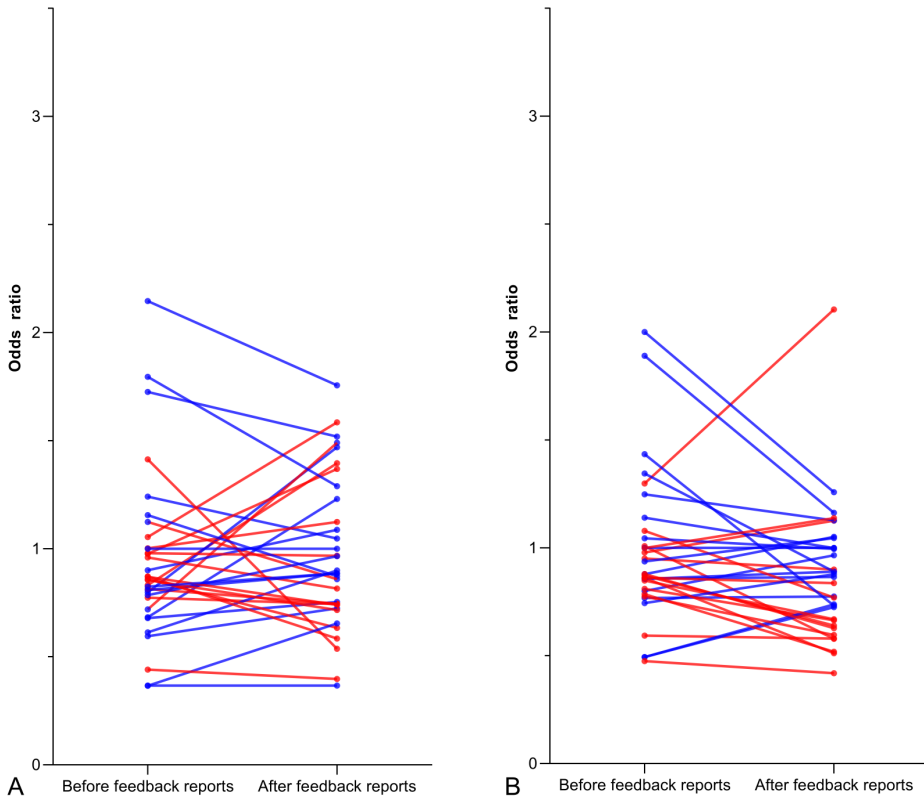
Figure 3. Case-mix adjusted odds ratios per laboratory before and after the feedback reports were calculated by multivariate logistic regression analyses (A: grade III versus grade I-II, B: grade II-III versus grade I). ORs are adjusted for age, tumor size, type of surgery, histologic subtype, hormone receptor status and HER2 receptor status. Each laboratory is represented by two dots (before and after feedback) connected by one line. The colour of the line and dots shows whether the OR after feedback shifted towards an OR of 1.00, and thus became less deviant (blue) or the OR after feedback shifted away from an OR of 1.00, and thus became more deviant (red) from the reference laboratory.

Discussion

Using nationwide data from structured (synoptic) pathology reports, we studied case-mix adjusted, laboratory-specific feedback reports as an intervention to decrease inter-laboratory variation in histologic grading of IBC. This study shows an encouraging decrease in nationwide grading variation after sending feedback reports to individual laboratories, reflected by a decrease in absolute range of grade-specific proportions after feedback for

all grades (I-III), the decrease of the maximum ODS, and the range of laboratory-specific ORs showing a notable decrease of 21.9% for grade III versus grades I-II. Overall, this shows that the most deviant laboratories became less deviant, while the overall mean ODS and positive OR-differences of individual laboratories did not significantly differ.

The primary aim of the laboratory-specific feedback reports was to create awareness among pathologists, by highlighting that grading variation in current clinical practice is substantial and improvement is warranted. It is important to stress that the aim of the feedback reports was not to just simply make 'higher' grading pathologists grade their tumours lower and vice versa. The awareness that the feedback-reports created enabled pathologists to discuss how they grade with other pathologists. Furthermore, they could perhaps conclude that they interpret the guideline differently or less strictly than other pathologists. In addition to inciting a dialogue between pathologists, we also hope that our previous paper [6] opens the dialogue between pathologists and oncologists. As we have previously shown, grade determines whether patients will get chemotherapy in approximately 30% of breast cancer patients [6], thus, awareness of grading variation is also very important to oncologists. Moreover, one could also think of peer-consultation in these cases, where grade determines whether a specific therapy is indicated.

Data included in this study were from synoptic pathology reports only, as currently over 80% of invasive breast cancer resection specimens are reported this way [32]. Moreover, besides an increased overall completeness of pathology reports [33, 34], it has recently been shown that synoptic reporting also improves patient care [34]. Besides advantages in patient care, easy data extraction from synoptic pathology reports also enables the assembly of nationwide laboratory-specific feedback reports on any chosen biomarker (histologic grade or hormone- and/or HER2-receptor status [35]). We believe that these feedback reports are an important first step towards improvement of breast cancer care by creating insight and awareness in variation of biomarker assessment, which is supported by the results of this study.

Thirty-eight of the current 42 pathology laboratories in the Netherlands implemented synoptic reporting between March 1 2017 and March 1 2019. Five of these 38 laboratories were nevertheless excluded from further data-analyses as they synoptically graded less than 50 IBC in either the pre- or post-feedback period. Two laboratories likely started using the protocol somewhere in the pre-feedback period (<50 reports) since their synoptic IBC report number increased considerably (>230) during the post-feedback period. Two other laboratories had low synoptic IBC report numbers in general (30-60 per period) and the fifth laboratory stopped reporting synoptically (425 pre-feedback, 0 post-feedback) for unknown reasons.

7

Although it seems that some laboratories (or pathologists) grade only few IBC cases annually, it is important to emphasize that these pathologists may still report IBC resection specimens narratively (i.e. outside the synoptic PALGA protocol), thus they may grade more IBC cases in clinical practice than our data may suggest. In addition, we previously showed that both laboratories who grade few and many IBC within the synoptic protocol may report significantly deviant proportions per grade [6].

However, if IBC-resection specimen numbers from our study are the true numbers per laboratory, one could argue the desirability of laboratories assessing less than 50 IBC resection specimens annually, which comes down to less than one IBC resection specimen per week. With the current situation of substantial nationwide inter-laboratory grading variation, and the potential clinical consequences in mind [6], one could argue that grading may be only be undertaken by trained, or maybe even only by expert breast pathologists. This should be the subject of future research.

Overall, grades I, II, and III IBC, were observed in 31.2%, 49.4% and 19.4% of all IBC resection specimens. This is in line with our previous study (2013-2016) and other studies [36-41], although percentages vary between the different studies. The distribution of grades did vary significantly between the pre- and post-feedback period (p=0.048); a slight increase of grade I IBC (30.5% vs. 32.0%) and a slight decrease of grade III IBC (20.0% vs. 18.8%) was apparent, whereas grade II IBC remained relatively stable (49.5% vs. 49.2%). This shift in distribution may be initiated by the feedback reports, however, it may also reflect a true change or random variation in the breast cancer population. Either way, this made it more difficult to show significant deviations towards the mean.

In addition, we found relatively low numbers of HER2-receptor positive IBC reports (7.9%) as compared to generally adopted numbers of 15-20% [42]. Furthermore, we found relatively high numbers of almost 90% ER-positive tumours, while approximately 15% of breast cancers are usually reported as triple negative [43]. Both findings are likely due to the fact that we excluded patients who received neoadjuvant treatment, which is the preferred initial approach in HER2-positive and triple-negative breast cancer patients [44].

We analysed the data in an absolute and a relative manner, comparing laboratories both to the national proportion and a reference laboratory. Overall, the mean ODS and changes within individual laboratories were non-significant (ODS, positive-OR differences). However, all analyses did show a decrease of the extremes (absolute range per grade, maximum ODS, and the 21.9% decrease in ORs for grade III versus grades I-II) after feedback. In addition, the slight increase (10.5%) of the absolute range of OR before and after feedback for grades II-III versus I seems to be mainly caused by a single extremely deviating laboratory (#12),

as this laboratory became significantly more deviant in the same direction after feedback. An impressive decrease in absolute OR range of 45.0% can be observed for the remainder of the laboratories. Overall, this shows that the most deviant laboratories became less deviant, while the majority of the other laboratories remained stable. We therefore believe that these results show an encouraging decrease in breast cancer grading variation after feedback.

Interestingly, we found no differences between laboratories who only received feedback on laboratory level and laboratories who additionally received feedback on pathologist' level. This may be due to the multistep, semi-objective way of the grading according to the modified Bloom and Richardson guideline [30, 31], with scores on the three components of grading, resulting in a total score and subsequent grade. This makes it more objective and thus more robust to direct influences on overall grade. In this light, it may be interesting to reflect on the three different subcategories (tubular differentiation, nuclear polymorphism, mitotic count) at pathologist' level. In addition, although the mean ODS did not change in both groups, the mean ODS of laboratories who also received feedback on pathologist' level was notably lower than the mean ODS of laboratories who received feedback on laboratory level only. Hence at the starting point, laboratories with feedback on the pathologist' level were already less deviant from the overall mean, which may have influenced their urgency to adjust their grading practices. This may be reflected by the significantly higher proportion of laboratories showing no shift for grade I among laboratories that received feedback on pathologist' level as compared to laboratories who received feedback on laboratory-level only. Lastly, the relatively low number of laboratories who received feedback on pathologist' level (n=10), makes it difficult to draw definite conclusions. However, literature does suggest that feedback is more effective when individual rather than general data are provided [26-29].

As with all interventions implemented in an uncontrolled environment, such as daily clinical practice, the observed decrease in grading variation can probably not solely be attributed to the feedback reports. However, within the timeframe of our previous study [6] and this study (December 31 2016 to March 1 2019), no guideline changes and no major other interventions or events on a nationwide scale, took place.

Although feedback reports were sent by March 1 2018, it is very likely that they were discussed in the laboratories somewhere in the week(s) thereafter, for example in a regular staff meeting. Hence, this means that the actual post-feedback period may have started somewhat later than March 1 2018, which could have clouded the effect of the feedback reports. Hence, the actual effect of the feedback reports may be even greater.

7

Despite the encouraging decrease in nationwide grading variation, we also showed that grading variation remains substantial. Besides continuous monitoring and benchmarking [26] of histologic grading (and other crucial biomarkers like ER, PR and HER2), which is already being considered by PALGA and the Dutch Society of Pathology, future research might focus on developing an e-learning module to train pathologists and residents in histologic grading of IBC in a standardized way to further decrease grading variation. This is underlined by Elston and Ellis, who state that grading of IBC should only be undertaken by specifically trained pathologists [45].

Conclusion

An encouraging decrease in nationwide Dutch grading variation of invasive breast cancers was observed after feedback. As feedback reports were sent to the laboratories for the first time, this was not (yet) a closed quality loop. Therefore, although our results are encouraging, the full potential of these feedback reports is still unknown. As overall grading variation remains substantial, it seems worthwhile to monitor this by continuing with feedback reports. Closing the quality loop and further training of pathologists, for example by e-learning, may help to further decrease grading variation and improve clinical decision making and thereby the outcome of our patients.

Acknowledgements

References

1. The Netherlands Comprehensive Cancer Organisation (IKNL), Oncoline: Breast Cancer Guideline, 2017. Available from: https://www.oncoline.nl/borstkanker. Accessed October 2 2019.

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(19):3153-8.

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 research : BCR 2010;12(4):207.

4. Curigliano G, Burstein HJ, Winer EP, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. Ann Oncol. 2017;28(8):1700-12.

5. Metzger-Filho O, Ferreira AR, Jeselsohn R, et al. Mixed Invasive Ductal and Lobular Carcinoma of the Breast: Prognosis and the Importance of Histologic Grade. Oncologist 2019;24(7):e441-e49.

6. van Dooijeweert C, van Diest PJ, Willems SM, et al. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: A nationwide study of 33,043 patients in the Netherlands. Int J Cancer. 2020;146)3):769-780.

7. Smith BD, Bellon JR, Blitzblau R, et al. Radiation therapy for the whole breast: Executive summary of an American Society for Radiation Oncology (ASTRO) evidence-based guideline. Pract Radiat Oncol. 2018;8(3):145-52.

8. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med 2016;375(8):717-29.

9. Krop I, Ismaila N, Andre F, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. J Clin Oncol. 2017;35(24):2838-47.

10. 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.

11. Boiesen P, Bendahl PO, Anagnostaki L, et al. Histologic grading in breast cancer reproducibility between seven pathologic departments. South Sweden Breast Cancer Group. Acta Oncol. 2000;39(1):41-5.

12. Frierson HF Jr., Wolber RA, Berean KW, et al. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. Am J Clin Pathol. 1995;103(2):195-8.

13. Italian Network for Quality Assurance of Tumour Biomarkers (INQAT) Group. Quality control for histological grading in breast cancer: an Italian experience. Pathologica. 2005;97(1):1-6.

14. Meyer JS, Alvarez C, Milikowski C, et al. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. Mod Pathol. 2005;18(8):1067-78.

15. van Dam PA, Verkinderen L, Hauspy J, et al. Benchmarking and audit of breast units improves quality of care. Facts Views Vis Obgyn. 2013;5(1):26-32.

16. van Bommel AC, Spronk PE, Vrancken Peeters MT, et al. Clinical auditing as an instrument for quality improvement in breast cancer care in the Netherlands: The

7

national NABON Breast Cancer Audit. J Surg Oncol. 2017;115(3):243-49.

17. Srigley J, Lankshear S, Brierley J, et al. Closing the quality loop: facilitating improvement in oncology practice through timely access to clinical performance indicators. J Oncol Pract. 2013;9(5):e255-61.

18. Sacerdote C, Bordon R, Pitarella S, et al. Compliance with clinical practice guidelines for breast cancer treatment: a population-based study of quality-of-care indicators in Italy. BMC Health Serv Res. 2013;13:28.

19. Laronga C, Gray JE, Siegel EM, et al. Florida Initiative for Quality Cancer Care: improvements in breast cancer quality indicators during a 3-year interval. J Am Coll Surg .2014;219(4):638-45.e1.

20. Veerbeek L, van der Geest L, Wouters M, et al. Enhancing the quality of care for patients with breast cancer: seven years of experience with a Dutch auditing system. Eur J Surg Oncol. 2011;37(8):714-8.

21. Casparie M, Tiebosch ATMG, Burger G, et al. Pathology Databanking and Biobanking in The Netherlands, a Central Role for PALGA, the Nationwide Histopathology and Cytopathology Data Network and Archive. Cell Oncol. 2007;29(1):19-24.

22. Adams AL, Eltoum I, Krontiras H, et al. The effect of neoadjuvant chemotherapy on histologic grade, hormone receptor status, and HER2/neu status in breast carcinoma. Breast J. 2008;14(2):141-6.

23. Yin HF, Wang YH, Qin XQ, et al. [Effect of neoadjuvant chemotherapy on histologic grade and expression of biological markers in breast cancer]. Zhonghua zhong liu za zhi [Chinese journal of oncology] 2009;31(11):858-62.

24. Zheng S, Zhang BL, Xiao T, et al. [Comparison of histopathologic changes and expression of biomarkers in breast carcinoma before and after neoadjuvant chemotherapy]. Zhonghua bing li xue za zhi = Chinese journal of pathology 2011;40(7):465-70.

25. Deckers IAG (PALGA Foundation), [Internet], Feedback variation in grading of invasive breast cancer and DCIS (in Dutch), 2018. Available from: https://www.palga.nl/assets/uploads/Spiegelinfo/PALGA%20rapport%20-%20Terugkoppeling%20variatie%20in%20gradering%20mammacarcinoom_algemeen.pdf. Accessed 1 October 2019.

26. Brehaut JC, Colquhoun HL, Eva KW, et al. Practice Feedback Interventions: 15 Suggestions for Optimizing Effectiveness. Ann Int Med. 2016;164(6):435-41.

27. Kluger AN, DeNisi A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin. 1996;119(2):254-84.

28. Ilgen DR, Fisher CD, Taylor MS. Consequences of individual feedback on behavior in organizations. Journal of Applied Psychology 1979;64(4):349-71.

29. Archer-Kath J, Johnson DW, Johnson RT. Individual versus Group Feedback in Cooperative Groups. The Journal of Social Psychology. 1994;134(5):681-94.

30. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer 1957;11(3):359-77.

31. 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(5):403-10.

32. PALGA Foundation, Annual report 2018 (in Dutch), Available from: https://www.palga.nl/assets/uploads/Jaarverslag/2018%20Palga%20web.pdf. Accessed 1 October 2019.

33. Sluijter CE, van Lonkhuijzen LR, van Slooten HJ, et al. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows

Arch. 2016;468(6):639-49.

34. Sluijter CE, van Workum F, Wiggers T, et al. Improvement of Care in Patients With Colorectal Cancer: Influence of the Introduction of Standardized Structured Reporting for Pathology. JCO Clin Cancer Inform. 2019;3:1-12.

35. van Dooijeweert C, Deckers IAG, Baas IO, et al. Hormone- and HER2-receptor assessment in 33,046 breast cancer patients: a nationwide comparison of positivity rates between pathology laboratories in the Netherlands. Breast Cancer Res Treat. 2019;175(2):487-97.

36. Blamey RW, Hornmark-Stenstam B, Ball G, et al. ONCOPOOL- a European database for 16,944 cases of breast cancer. Eur J Cancer. 2010;46(1):56-71.

37. Kiaer HW, Laenkholm AV, Nielsen BB, et al. Classical pathological variables recorded in the Danish Breast Cancer Cooperative Group's register 1978-2006. Acta Oncol. 2008;47(4):778-83.

38. Moller H, Henson K, Luchtenborg M, et al. Short-term breast cancer survival in relation to ethnicity, stage, grade and receptor status: national cohort study in England. Br J Cancer. 2016;115(11):1408-15.

39. Puig-Vives M, Sanchez MJ, Sanchez-Cantalejo J, et al. Distribution and prognosis of molecular breast cancer subtypes defined by immunohistochemical biomarkers in a Spanish population-based study. Gynecol Oncol. 2013;130(3):609-14.

40. Seneviratne S, Lawrenson R, Scott N, et al. Breast cancer biology and ethnic disparities in breast cancer mortality in new zealand: a cohort study. PloS One. 2015;10(4):e0123523.

41. Sun J, Chen C, Wei W, et al. Associations and indications of Ki67 expression with clinicopathological parameters and molecular subtypes in invasive breast cancer: A population-based study. Oncol Lett. 2015;10(3):1741-48.

42. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med. 2005;353(16):1659-72.

43. Denkert C, Liedtke C, Tutt A, et al. Molecular alterations in triple-negative breast cancer- the road to new treatment strategies. Lancet . 2017;389(10087):2430-42.

44. Burstein HJ, Curigliano G, Loibl S, et al. Estimating the Benefits of Therapy for Early Stage Breast Cancer The St Gallen International Consensus Guidelines for the Primary Therapy of Early Breast

45. Cancer 2019. Ann Oncol. 2019;30(10):1541-1557

46. 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. 2002;41(3a):154-61.

```
┌─────────────────────────────────────┐
│   Synoptic pathology reports of      │
│  25,420 invasive breast tumors (IBC) │
│       from PALGA database            │
│   (March 1 2017-March 1 2019)        │
│       (22,476 patients)              │
└─────────────────────────────────────┘
                 │
                 │──1,497 reports──▶  ╭─────────────────────╮
                 │                    │ Small laboratories  │
                 ▼                    │  (< 50 IBC per year) │
┌─────────────────────────────────────┐╰─────────────────────╯
│   23,941 IBC                         │
│   (21,175 patients)                  │
└─────────────────────────────────────┘
                 │
                 │──4,052 reports──▶  ╭─────────────────────╮
                 │                    │ Neoadjuvant treatment│
                 ▼                    ╰─────────────────────╯
```
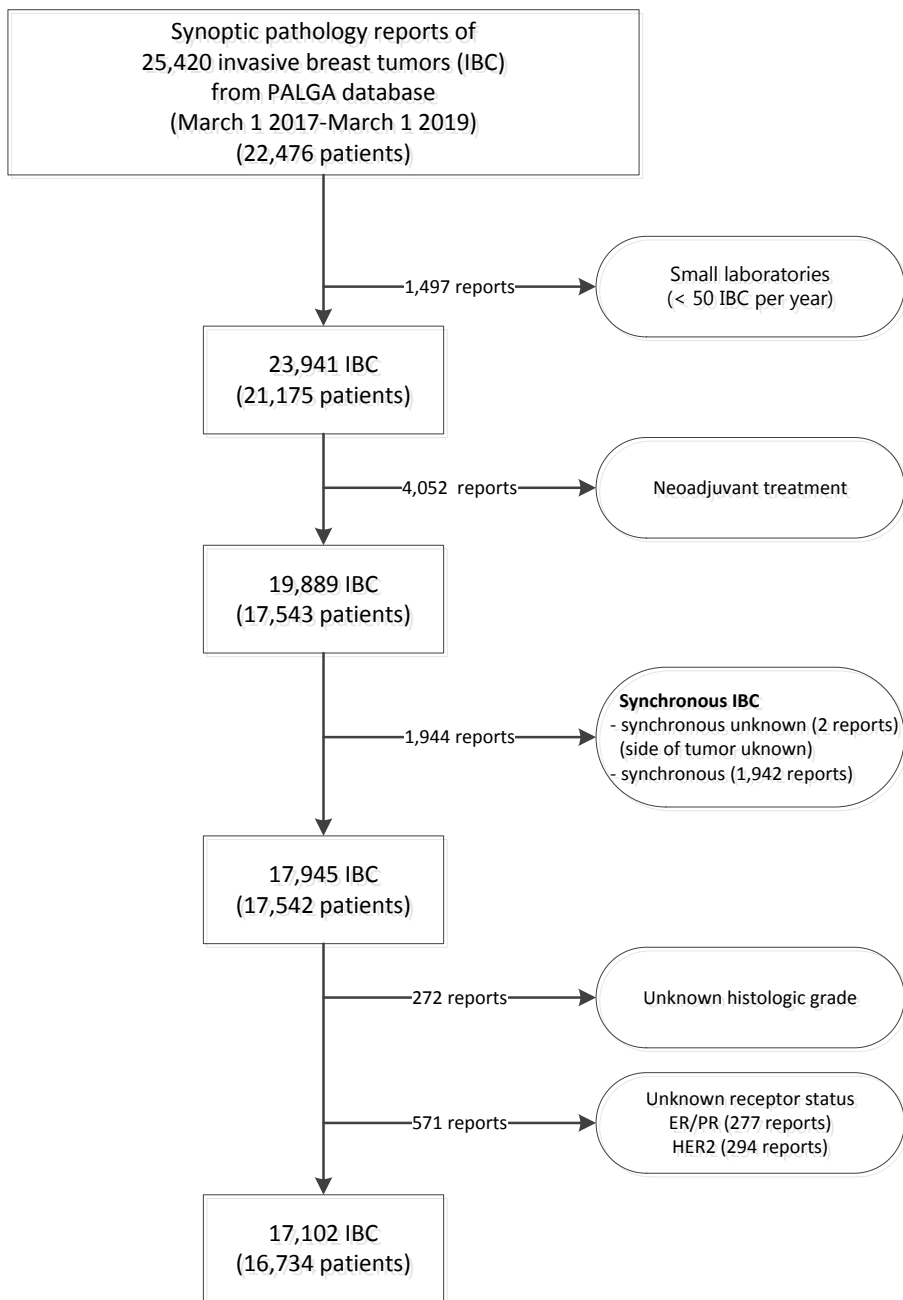
8

The effect of an e-learning on grading variation of (pre)malignant breast lesions

Carmen van Dooijeweert
Ivette A. G. Deckers
Emma J. de Ruiter
Natalie D. ter Hoeve
Celien P. H. Vreuls
Elsken van der Wall
Paul J. van Diest

Abstract

Histologic grade is a biomarker that is widely used to guide treatment of invasive breast cancer (IBC) and ductal carcinoma *in situ* of the breast (DCIS). Yet, currently, substantial grading variation between laboratories and pathologists exists in daily pathology practice. This study was conducted to evaluate whether an e-learning may be a feasible tool to decrease grading variation of (pre)malignant breast lesions.

An e-learning module, representing the key-concepts of grading (pre)malignant breast lesions through gold standard digital images, was designed. Pathologists and residents could take part in either or both the separate modules on DCIS and IBC. Variation in grading of a digital set of lesions before and after the e-learning was compared in a fully-crossed study-design. Multiple outcome measures were assessed: inter-rater reliability (IRR) by Light's kappa, the number of images graded unanimously, the number of images with both extreme scores (i.e. grade I and grade III), and the average number of discrepancies from expert-consensus. Participants were included as they completed both the pre- and post e-learning set (DCIS-module: n=36, IBC-module: n=21).

For DCIS, all outcome measures improved after e-learning, with the IRR improving from fair (kappa: 0.532) to good (kappa: 0.657). For IBC, all outcome measures for the subcategories tubular differentiation and mitosis improved, with >90% of participants agreeing on almost 90% of the images after the e-learning. In contrast, the IRR for the subcategory of nuclear pleomorphism remained fair (kappa: 0.523 vs. kappa: 0.571).

This study shows that an e-learning module, in which pathologists and residents are trained in histologic grading of DCIS and IBC, is a feasible and promising tool to decrease grading variation of (pre)malignant breast lesions. This is highly relevant given the important role of histologic grading in clinical decision making of (pre)malignant breast lesions.

Introduction

Histologic grade is an important prognostic biomarker, which is widely used to guide breast cancer treatment [1-3]. In the Netherlands, grade currently indicates the need for adjuvant chemotherapy in nearly a third of breast cancer patients [1, 4]. For ductal carcinoma *in situ* of the breast (DCIS), grade influences radiotherapy decisions [1, 5] and indicates the need for a sentinel lymph node procedure [1]. Moreover, histologic grade may become the single biomarker that decides whether DCIS patients should or should not be treated, as this is now being investigated by multiple clinical trials [6-9].

Subsequently, accurate, and reproducible grading is of major clinical importance. However, significant inter- and intra-laboratory variation in histologic grading of invasive breast cancer (IBC) and DCIS exists in daily pathology practice [4, 10]. Providing laboratories with feedback reports, in which laboratory-specific case-mix adjusted proportions per grade were benchmarked against other laboratories, resulted in a promising, yet small, decrease in grading variation (van Dooijeweert et al, *submitted*). As substantial differences in grade-specific proportions were observed between pathologists within individual laboratories [4, 10], we hypothesized that variation in grading may be primarily explained by differences in grading practices of individual pathologists.

Therefore, we believe that training of pathologists in the assessment of histologic grade could attribute to better synchronization, and thereby a decrease in grading variation of both IBC and DCIS. This is further supported by Elston and Ellis, who emphasize that grading should only be performed by trained pathologists [11]. Yet, in the Netherlands, pathologists or pathology residents are currently not specifically trained in histologic grading of (pre) malignant breast lesions.

Training of pathologists by e-learning seems feasible, as e-learnings have been shown to decrease grading variation of dysplasia in colorectal adenomas [12] and to improve consistency in the histopathological diagnosis of sessile serrated colorectal lesions [13]. In addition, an e-learning module is easily accessible online for medical professionals throughout the country, or even worldwide, without the need for a live tutor or planned course days [14]. Thus, it enables pathologists to train when and where they want.

The aim of this study was to evaluate whether an e-learning module may be an effective tool to decrease grading variation of (pre)malignant breast lesions (i.e. DCIS and IBC) by studying variation in grading of a digital set of lesions by pathologists before and after the e-learning.

8

Materials and Methods

E-learning design
The e-learning module was designed by the co-authors and agreement on its content was reached by three expert breast pathologists (PvD, CV, RG) and one expert breast pathologist' assistant (NtH). The e-learning presents the key-concepts of histologic grading of (pre)malignant breast lesions, including background information, discussion of the specific grading classifications for DCIS and IBC, and an extensive review of the these criteria with example images.

The e-learning module consists of two separate modules, one on grading of IBC and one on grading of DCIS. Pathologists and residents could participate in either one or both modules. Separate analyses were conducted for the DCIS- and IBC-module.

All images in the e-learning module were derived from breast cancer cases from daily pathology practice in our institute based on consensus . All patient–related information was removed from all images in the e-learning module to comply with the General Data Protection Regulation (GDPR).

Classifications of histologic grading
For grading of DCIS, we used the classification of Holland [15] in our e-learning (Supplementary 1), as this is the guideline recommended by the Dutch breast cancer guideline, and, as we know from our previous survey among pathologists [10], this seems to be the most widely used classification in daily pathology practice.

The modified Bloom and Richardson guideline (Elston-Ellis modification of Scarff-Bloom-Richardson grading system, also known as the Nottingham grading system) [16, 17] (Supplementary 2) was used in this e-learning module for IBC, as it is the most widely used grading system for IBC and it is globally incorporated in breast cancer guidelines [1, 3, 18, 19]. This classification combines the assessment of cell morphology (nuclear pleomorphism), the proliferation (mitotic count), and differentiation (tubule formation), resulting in a total score and subsequent grade [17].

Recruitment of participants
Pathologists and pathology residents were invited to participate in this e-learning through the news-bulletin of the Dutch Society for Pathology (NVvP).

Study design

Baseline grading variation

Before they could start the e-learning module, participants were obliged to assign a grade to 52 snapshots of DCIS lesions, as they would in daily practice, to determine the baseline variation in grading between participants for DCIS. As agreed upon by consensus of the expert-panel (PvD, CV, RG, NtH,) 16, 19, and 17 of these snapshots were DCIS grade I, II, and III, respectively. Several examples of these snapshots can be found in Supplementary 3. All architectural patterns (solid, cribriform, papillary and micropapillary were present in these tests, and some cases included comedonecrosis.

For IBC, participants were obliged to score snapshots per subcategory of the modified Bloom and Richardson guideline [16, 17], i.e. tubular differentiation (19 snapshots; 7 category 1, 5 category 2, 7 category 3), nuclear pleomorphism (19 snapshots; 6 category 1, 6 category 2, 7 category 3), and mitosis (30 mitoses, 20 non-mitoses). With regard to mitoses, participants were asked to state whether they would count 50 figures as a mitosis or not. This was done because we hypothesized that recognizing mitosis is the bottleneck, rather than counting itself. Furthermore, this way, snapshots were sufficient, while whole slides are used for this in daily clinical practice (and pathologists select the field where they count). Several examples of these snapshots can be found in Supplementary 4. These snapshots were primarily derived from invasive ductal carcinoma cases and tubular, cribriform carcinomas for tubule formation, and some invasive lobular carcinoma cases. Since snapshots for tubule formation, nuclear pleomorphism and mitoses were taken from different cases, only improvement for individual subcategories could be assessed.

After scoring the pre e-learning set of DCIS lesions and/or all images of the IBC subcategories, participants would start the e-learning itself.

Post e-learning grading variation

After completion of the e-learning, participants were asked to re-grade the same DCIS lesions, and/or IBC lesions to determine the post e-learning variation in grading between participants. All lesions were presented in a different order than prior to the e-learning. Only after this second round of grading, participants received feedback on their answers.

Outcome measures and statistical analyses

Variation in grading was measured according to several outcome measures both pre- and post e-learning. We included scores of participants who completed grading of both the pre- and post-e-learning set of lesions, for DCIS and/or IBC. Hence, a fully-crossed design was used [20].

8

As primary outcome measure, the inter-rater reliability (IRR) was calculated by Light's kappa [21], which is a type of kappa-statistic suitable for a fully-crossed study-design [20]. This overall kappa represents the arithmetic mean of kappa scores for all coder pairs [20, 21]. The 95% confidence interval for kappa was calculated using bootstrapping from 1000 replications. Interpretation of the kappa-statistics was performed according to the proposed interpretation of Cicchetti [22] (K <0.40: poor, K 0.40-0.59: fair, K 0.60-0.74: good, K 0.75-1.00: excellent). Secondary outcome measures were the number of images scored unanimously (i.e. agreement by 90% and 100% of scoring participants) and the number of single lesions scored as both grade I and III (for DCIS) or both category one and three (for IBC: tubular differentiation, nuclear pleomorphism). Lastly, the average number of discrepancies with grade (DCIS) or sub-score (IBC) from the reference score by the expert panel (PvD, CV, RG, NtH) was determined, overall, and for subgroups of participants (expert breast pathologists, general pathologists, and residents).

All statistical analyses were performed using IBM SPSS Statistics version 25 and R [23].

Table 1. Characteristics of e-learning participants

Characteristics	DCIS (n=36)	IBC (n=21)
Function, n (%)		
Expert breast pathologist	10 (27.8%)	7 (33.3%)
General pathologist	20 (55.6%)	9 (42.9%)
Resident	6 (16.7%)	5 (23.8%)
Years of experience, mean (SD)		
Expert breast pathologist	13.2 (7.3)	13.1 (8.3)
General pathologist	9.5 (6.2)	8.6 (7.5)
Resident	2.7 (1.2)	2.8 (1.3)

Results

Participants

Thirty-six (DCIS) participants were included in the data-analyses for DCIS, while 21 participants were included in the data-analyses for IBC, as they completed both the pre- and post-e-learning grading sets for the separate modules. The majority of participants identified themselves as general pathologist, followed by expert breast pathologists, and residents (Table 1). Expert breast pathologists executed their function for the largest numbers of years, with a mean of >13 years (Table 1).

DCIS

All outcome measures for DCIS improved after the e-learning (Table 2). Overall agreement (IRR) improved from fair (kappa: 0.532) to good (kappa: 0.657) [22], and the amount of lesions graded unanimously by >90% (33/36) and 100% (36/36) of participants increased after e-learning, with best observed agreement for DCIS grade III. Interestingly, before the e-learning, almost 30% of DCIS images (n=15) were graded as both grade I and grade III by the different participants, which decreased to less than 10% (n=5) after the e-learning (Table 2). Lastly, the average number of discrepancies with the reference score by the expert panel decreased for all subgroups of participants.

Table 2. Variation in grading of ductal carcinoma *in situ* of the breast before and after e-learning (36 participants)

Outcome measure (total questions = 52)	PRE e-learning	POST e-learning
IRR (Light's kappa (95% confidence interval))	**0.532 (0.457-0.591)**	**0.657 (0.582-0.730)**
Cases graded 100% unanimous (36/36)	**4 (7.7%)**	**9 (17.3%)**
Grade I (n=16)	-	-
Grade II (n=19)	-	2 (10.5%)
Grade III (n=17)	4 (23.5%)	7 (41.2%)
Cases graded >90% unanimous (33/36)	**13 (25.0%)**	**24 (46.2%)**
Grade I (n=16)	-	6 (37.5%)
Grade II (n=19)	3 (15.8%)	6 (31.6%)
Grade III (n=17)	10 (58.8%)	12 (70.6%)
Cases graded both grade I and III	**15 (28.8%)**	**5 (9.6%)**
Grade I (n=16)	4 (25.0%)	-
Grade II (n=19)	11 (57.9%)	5 (26.3%)
Grade III (n=17)	-	-
Average number of discrepancies	**12.8 (24.6%)**	**8.1 (15.6%)**
Expert breast pathologists	10.4 (20.0%)	7.6 (14.6%)
General pathologists	12.7 (24.4%)	7.6 (14.6%)
Residents	17.0 (32.7%)	10.8 (20.8%)

8

Invasive breast cancer

All outcome measures that contribute to grading, except for two outcome measures for nuclear pleomorphism, improved after e-learning (Table 3).

For tubular differentiation, overall agreement (IRR) improved from good (kappa: 0.653) to excellent (kappa: 0.846) [22], while the number of images graded unanimously by >90% (19/21) and 100% (21/21) participants increased as well (from 63.2% to 89.5%, and from 0.0% to 47.4%, respectively). The number of images scored both as category 1 and 3 was notably high before e-learning (78.9%), yet a substantial decrease was observed after e-learning (10.5%). Finally, the average number of discrepancies decreased for all subgroups of participants.

For nuclear pleomorphism, overall agreement (IRR) remained only fair (kappa: 0.523 vs. kappa: 0.571) after the e-learning module. The number of images graded 100% unanimous slightly increased (from 15.8% to 26.3%), while the number of images graded unanimously by >90% of participants remained stable (31.6% pre- and post-e-learning). Overall, for both outcome measures, best agreement was observed for nuclear pleomorphism category 3. After e-learning, an increase was observed for the number of images scored as both category 1 and 3 (0.0% to 15.8%), all of which were deemed category 2 by consensus of the expert-panel. Hence, for these lesions, the same lesion may be treated differently in daily clinical practice, based upon which pathologist graded it, as this is a difference of two points on the overall score, and most probably would lead to a different overall grade (3-5 = grade I, 6-7 = grade II, 8-9 = grade III). Lastly, the average number of discrepancies did show improvement after e-learning (27.1% to 18.8%).

For mitoses, overall agreement (IRR) showed improvement from good (kappa: 0.699) to excellent (kappa: 0.826). After the e-learning, an increase was observed for the number of lesions scored unanimously by >90% and 100% of participants (from 36.0% to 56.0%, and from 74.0% to 86.0%, respectively). The overall number of discrepancies decreased from 10.5% before e-learning to 6.6% after e-learning.

A few example images with low- and high- concordance after e-learning for both DCIS and IBC can be found in Supplementary 5.

Table 3. Variation in scoring of the 3 constituents of invasive breast cancer grading before and after e-learning (21 participants)

Tubular differentiation (total cases = 19)	PRE e-learning	POST e-learning
IRR (Light's kappa (95% confidence interval))	**0.653 (0.529-0.739)**	**0.846 (0.745-0.918)**
Cases graded 100% unanimous (21/21)	**-**	**9 (47.4%)**
Category 1 (n=7)	-	4 (21.1%)
Category 2 (n=5)	-	2 (10.5%)
Category 3 (n=7)	-	3 (15.8%)
Cases graded >90% unanimous (19/21)	**12 (63.2%)**	**17 (89.5%)**
Category 1 (n=7)	5 (71.4%)	6 (85.7%)
Category 2 (n=5)	2 (40.0%)	4 (80.0%)
Category 3 (n=7)	5 (71.4%)	7 (100.0%)
Cases scored both category 1 and 3	**15 (78.9%)**	**2 (10.5%)**
Category 1 (n=7)	5 (26.3%)	1 (14.3%)
Category 2 (n=5)	4 (24.1%)	-
Category 3 (n=7)	6 (31.6%)	1 (14.3%)
Average number of discrepancies	**3.0 (15.5%)**	**1.0 (5.3%)**
Expert breast pathologists	4.6 (24.1%)	1.3 (6.8%)
General pathologists	1.8 (9.4%)	0.8 (4.1%)
Residents	2.8 (14.7%)	1.0 (5.3%)
Nuclear pleomorphism (total cases = 19)	PRE e-learning	POST e-learning
IRR (Light's kappa (95% confidence interval))	**0.523 (0.372-0.6314)**	**0.571 (0.443-0.672)**
Cases graded 100% unanimous (21/21)	**3 (15.8%)**	**5 (26.3%)**
Category 1 (n=6)	-	-
Category 2 (n=6)	-	-
Category 3 (n=7)	3 (42.9%)	5 (71.4%)
Cases graded >90% unanimous (19/21)	**6 (31.6%)**	**6 (31.6%)**
Category 1 (n=6)	1 (16.7%)	-
Category 2 (n=6)	1 (16.7%)	1 (16.7%)
Category 3 (n=7)	4 (57.1%)	5 (71.4%)
Cases scored both category 1 and 3	**-**	**3 (15.8%)**
Category 1 (n=6)	-	-
Category 2 (n=6)	-	3 (50.0%)
Category 3 (n=7)	-	-
Average number of discrepancies	**5.1 (27.1%)**	**3.6 (18.8%)**
Expert breast pathologists	6.3 (33.1%)	4.0 (21.1%)
General pathologists	4.4 (23.2%)	2.8 (14.7%)
Residents	4.0 (21.1%)	3.8 (20.0%)

8

Table 3. Continued

Mitoses (total lesions = 50)	PRE e-learning	POST e-learning
IRR (Light's kappa (95% confidence interval))	**0.699 (0.586-0.785)**	**0.826 (0.746-0.891)**
Cases graded 100% unanimous (21/21)	**18 (36.0%)**	**28 (56.0%)**
Mitosis (n=30)	10 (33.3%)	18 (60.0%)
Non-mitosis (n=20)	8 (40.0%)	10 (50.0%)
Cases graded >90% unanimous (19/21)	**37 (74.0%)**	**43 (86.0%)**
Mitosis (n=30)	21 (70.0%)	25 (83.3%)
Non-mitosis (n=20)	16 (80.0%)	18 (90.0%)
Average number of discrepancies	**5.2 (10.5%)**	**3.3 (6.6%)**
Expert breast pathologists	4.6 (9.1%)	3.9 (7.7%)
General pathologists	4.7 (9.3%)	2.1 (4.2%)
Residents	7.2 (14.4%)	4.2 (8.2%)

Discussion

This study shows that an e-learning, in which pathologists and residents are trained in histologic grading of DCIS and IBC, is a feasible and promising tool to decrease grading variation, as 18/20 outcome measures improved after e-learning. This is highly relevant considering the substantial grading variation of (pre)malignant breast lesions in current daily clinical practice [4, 10] bearing in mind that grading plays a decisive role in clinical decision-making such as the indication for chemotherapy in IBC, and for a sentinel lymph node procedure, and partial breast irradiation in DCIS. Increased consensus with regard to grading will diminish variation in treatment and thereby most likely also patient outcome [4].

To evaluate the effect of our e-learning, we used multiple outcome measures (Table 2, Table 3). The IRR was chosen as outcome measure as this provides a way of quantifying the degree of agreement [20] between the e-learning participants on histologic grading. It is important to acknowledge that kappa-values may be influenced by the choice of the specific kappa statistic, and there is no clear guideline which specific kappa statistic to use for a fully-crossed design with multiple (≥2) coders and an ordinal outcome measure (grade I-III, category 1-3) [20, 24-26]. Here, we chose Light's kappa [20], which uses the arithmetic mean of the kappa's of all possible coder pairs.

To provide further insights, we assessed the number of lesions scored unanimously (both by all, and by >90% of participants), and the number of lesions which were scored both grade I and III or category 1 and 3, as this outlines where difficulties may and may not lay (specific grades/subcategories).

For DCIS, almost 30% of the pre-e-learning images were scored as both grade I and grade III (of which the majority was deemed grade II by our expert panel). Although most discrepancies were only one grade apart, this may point out a challenge for current clinical trials, who focus on the safety of active-surveillance in grade I and grade II DCIS patients [6-9], especially since central pathology review is not carried out in all trials [27]. This means that, in current daily pathology practice, grade III DCIS lesions may erroneously be deemed grade II and randomized to active surveillance. This underlines that central pathology review is essential for these trials, and raises the question whether histologic grade, in its current state, should be the single identifying biomarker for low-risk DCIS, as is also pointed out by Cserni [28].

For IBC, most difficulties were observed for scoring nuclear pleomorphism, which is in line with previous studies [4, 29-32], and may be due to the fact that scoring of this subcategory is least quantitative, leaving most room for variation in interpretation. It should also be mentioned that we used snapshots of lesions (of very similar magnification), which makes interpreting nuclear pleomorphism somewhat more difficult, especially since comparison to other cells (for example epithelium) by zooming out was not possible. In contrast, for tubular differentiation and mitoses, >90% of participants agreed on almost 90% of the questions after the e-learning, indicating that for these two constituents of grade the learning effect is much greater.

A limitation of this study may be that participants graded the same digital set of lesions after finishing the e-learning. Participants could go through the e-learning (and both tests) at their own pace, even in multiple sessions, so it is unknown what the 'wash-out time' was between the pre- and post-test. Therefore, it cannot be ruled out that, although the lesions were presented in a different order, participants 'recognized' some pictures. However, we would like to emphasize that participants only received extensive feedback after the final test. In addition, grading the same set of lesions, was deemed the best way to observe the effect of the e-learning on grading variation as this enables comparing 'baseline' variation before e-learning with grading variation after e-learning. Grading a different set of lesions after the e-learning would not enable us to distinguish between an effect of the e-learning or simply because grading of the post-e-learning set may have be 'easier' of 'more difficult'.

Furthermore, as complete consensus was not always obtained after e-learning, the images chosen based on 'true grade/category' consensus of our expert panel may still not be the perfect examples. Nonetheless, we would like to emphasize that the average number of discrepancies from this 'true grade/category' decreased after e-learning for both DCIS and all subcategories of IBC. Therefore, if this type of training by e-learning is to be enrolled on

8

a larger scale, one could think of a larger consensus panel of expert breast pathologists, especially since, for tubular differentiation and nuclear pleomorphism, most discrepancies from our expert panel were observed among this subgroup of participants.

A final limitation may be our decision to use mitosis identification in snapshots, rather than the selection of high power fields and actual mitosis counting. This was done because of practical reasons, however, we would like to emphasize that this indeed leaves out selection of high power fields, which may be a challenge, especially in a heterogeneous tumor, leaving room for grading variation within this subcategory.

Supported by the results of this study, we believe that training of pathologists in the assessment of histologic grade could further attribute to better synchronization and thereby a decrease in grading variation of both IBC and DCIS. An e-learning module as described in this study, in combination with monitoring variation in histologic grading of IBC and DCIS in daily clinical practice by laboratory- and pathologist specific feedback, may be considered by the Dutch Society of Pathology as one of the pathways for those pathologists that want to qualify as breast pathologist. Likewise, pathologists who participate in the Dutch national colorectal cancer screening program are already obliged to participate in an e-learning. In addition, the use of peer-learning programs for training and quality was identified as one of the strategies to improve patient access to high-quality oncologic pathology by Nass et al [33]. Lastly, the e-learning may well be implemented in training programs of pathology residents.

While both laboratory-specific feedback and training of pathologists by e-learning may be promising tools to decrease grading variation, we believe that it is also important to emphasize that histologic grade is not a fact, nor the truth, but merely a model or tool that consists of a statistically computed set of cut-off values in a spectrum of histopathologic features to classify expected tumor behavior. Yet, these cut-offs are often used as hard criterion by clinicians in clinical decision making. For example, according to the Dutch breast cancer guideline, for a large group of patients (≥35 years, N0-, HER2-, 1.1-2cm) (1), ≥grade II indicates that they are eligible for chemotherapy, whereas grade I does not. Yet, the majority of patients (>75%) has a score on the switch point of grades, that is scores 5 or 6, and scores 7 or 8 [4]. Thus, for these patients, the difference of only one point on any of the subcategories of the Bloom- and Richardson classification may already alter their overall histologic grade, and thus may have different therapeutic implications. Therefore, we believe that awareness, and understanding of the difficulties of histologic grading, among clinicians is crucial as well to improve clinical decision making for patients. Furthermore, in the current era of shared-decision-making, these difficulties should also be discussed with patients, where relevant.

In conclusion, this study shows that an e-learning module, in which pathologists and residents are trained in histologic grading of DCIS and IBC, is a feasible and promising tool to decrease grading variation of (pre)malignant breast lesions. This is highly relevant given the important role of histologic grading in clinical decision making of DCIS and invasive breast cancer whereby an influence on outcome cannot be ruled out.

Acknowledgements

8

References

1. The Netherlands Comprehensive Cancer Organization (IKNL). Breast Cancer Guideline 2017 [updated 2012-07-02]. Available from: https://www.oncoline.nl/borstkanker. Accessed 17 January 2019.

2. Rakha EA, El-Sayed ME, Lee AH, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol. 2008;26:3153-8.

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. van Dooijeweert C, van Diest PJ, Willems SM, et al. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: A nationwide study of 33,043 patients in the Netherlands. Int J Cancer. 2020;146)3):769-780.

5. Smith BD, Bellon JR, Blitzblau R, et al. Radiation therapy for the whole breast: Executive summary of an American Society for Radiation Oncology (ASTRO) evidence-based guideline. Pract Radiat Oncol. 2018;8:145-52.

6. 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;51:1497-510.

7. Francis A, Thomas J, Fallowfield L, et al. Addressing overtreatment of screen detected DCIS; the LORIS trial. Eur J Cancer. 2015;51:2296-303.

8. Groen EJ, Elshof LE, Visser LL, et al. Finding the balance between over- and under-treatment of ductal carcinoma in situ (DCIS). Breast. 2017;31:274-83.

9. Youngwirth LM, Boughey JC, Hwang ES. Surgery versus monitoring and endocrine therapy for low-risk DCIS: The COMET Trial. Bull Am Coll Surg. 2017;102:62-3.

10. van Dooijeweert C, van Diest PJ, Willems SM, Kuijpers C, Overbeek LIH, Deckers IAG. Significant inter- and intra-laboratory variation in grading of ductal carcinoma in situ of the breast: a nationwide study of 4901 patients in the Netherlands. Breast Cancer Res Treat. 2019;174:479-88.

11. 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. 2002;41:154-61.

12. Madani A, Kuijpers C, Sluijter CE, et al. Decrease of variation in the grading of dysplasia in colorectal adenomas with a national e-learning module. Histopathology. 2019;74(6):925-932.

13. IJspeert JE, Madani A, Overbeek LI, Dekker E, Nagtegaal ID. Implementation of an e-learning module improves consistency in the histopathological diagnosis of sessile serrated lesions within a nationwide population screening programme. Histopathology. 2017;70:929-37.

14. Ruiz JG, Mintzer MJ, Leipzig RM. The impact of E-learning in medical education. Acad Med. 2006;81:207-12.

15. Holland R, Peterse JL, Millis RR, et al. Ductal carcinoma in situ: a proposal for a new classification. Semin Diagn Pathol. 1994;11:167-80.

16. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer. 1957;11:359-77.

17. 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;19:403-10.

18. Harris L, Fritsche H, Mennel R, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol. 2007;25:5287-312.

19. Singletary SE, Allred C, Ashley P, et al. Revision of the American Joint Committee on Cancer staging system for breast cancer. J Clin Oncol. 2002;20:3628-36.

20. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol. 2012;8:23-34.

21. Light RJ. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin. 1971;76:365-77.

22. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment. 1994;6:284-90.

23. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

24. Warrens MJ. Inequalities between multi-rater kappas. Advances in Data Analysis and Classification. 2010;4:271-86.

25. Berry KJ, Johnston JE, Mielke PW, Jr. Weighted kappa for multiple raters. Percept Mot Skills. 2008;107:837-48.

26. Conger AJ. Integration and generalization of kappas for multiple raters. Psychological Bulletin. 1980;88:322-8.

27. van Bockstal MR, Agahozo MC, Koppert LB, van Deurzen CHM. A retrospective alternative for active surveillance trials for ductal carcinoma in situ of the breast. Int J Cancer. 2020;146(5):1189-1197.

28. Cserni G, Sejben A. Grading Ductal Carcinoma In Situ (DCIS) of the Breast - What's Wrong with It? Pathol Oncol Res 2019. doi: 10.1007/s12253-019-00760-8. [Epub ahead of print]

29. Boiesen P, Bendahl PO, Anagnostaki L, et al. Histologic grading in breast cancer-- reproducibility between seven pathologic departments. South Sweden Breast Cancer Group. Acta Oncol. 2000;39:41-5.

30. Frierson HF Jr., Wolber RA, Berean KW, et al. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. Am J Clin Pathol. 1995;103:195-8.

31. Italian Network for Quality Assurance of Tumour Biomarkers (INQAT). Quality control for histological grading in breast cancer: an Italian experience. Pathologica. 2005;97:1-6.

32. Meyer JS, Alvarez C, Milikowski C, et al. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. Mod Pathol. 2005;18:1067-78.

33. Nass SJ, Cohen MB, Nayar R, et al. Improving Cancer Diagnosis and Care: Patient Access to High-Quality Oncologic Pathology. Oncologist. 2019;24:1287-90.

8

Supplementary 1. DCIS grading classification derived from Holland et al (15), and the Dutch breast cancer guideline (1).

Defining characteristics	DCIS grade I	DCIS grade II	DCIS grade III
Primary			
Nuclei	Monomorphic	Pleomorphic+	Pleomorphic ++
Chromatin	Fine	Fine-vesicular	Vesicular-coarse
Nucleoli	Insignificant	Evident	Prominent
Mitoses	Rare	Occasionally	Often
Secondary			
Orientation (polarization)	Marked	Some	Little-none
Frequently associated characteristics			
Central necrosis	Absent-minimal	Variable	Often (prominent)
Apoptosis	Absent	Focally (if present)	Usually present
Growth pattern	Micropapillary, cribriform, rarely solid	All patterns	Solid, cribriform, micropapillary
Calcifications	Psammoma-like (rarely amorph)	Amorph	Amorph

Supplementary 2. Invasive breast cancer grading classification derived from the modified Bloom and Richardson guideline (16, 17) and the Dutch breast cancer guideline (1).

	Description
Tubular formation	
Score 1	>75% of the tumor
Score 2	10-75% of the tumor
Score 3	<10% of the tumor
Nuclear pleomorphism	
Score 1	Small, regular uniform cells, little different than normal epithelium
Score 2	Moderate increase in size and variability, small nucleoli, vesicular
Score 3	Marked variation, large nucleoli, vesicular
Mitotic count (per 2mm^2)	
Score 1	0-7 mitoses
Score 2	8-12 mitoses
Score 3	≥13 mitoses
Overall grade	
Grade 1	3-5
Grade 2	6-7
Grade 3	8-9

Supplementary 3. Some example images of an e-learning module for grading of ductal carcinoma *in situ* (DCIS)
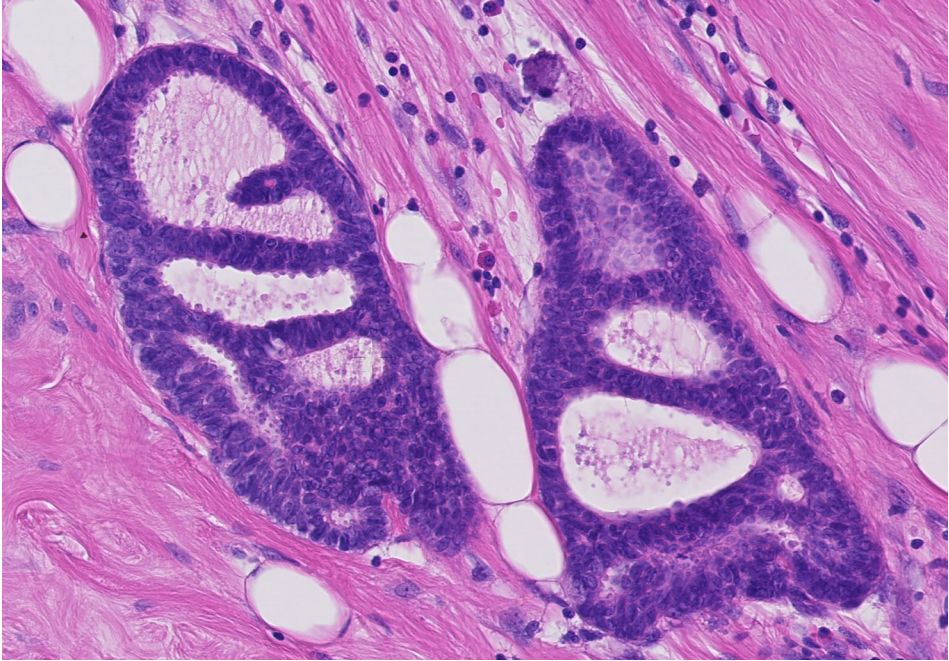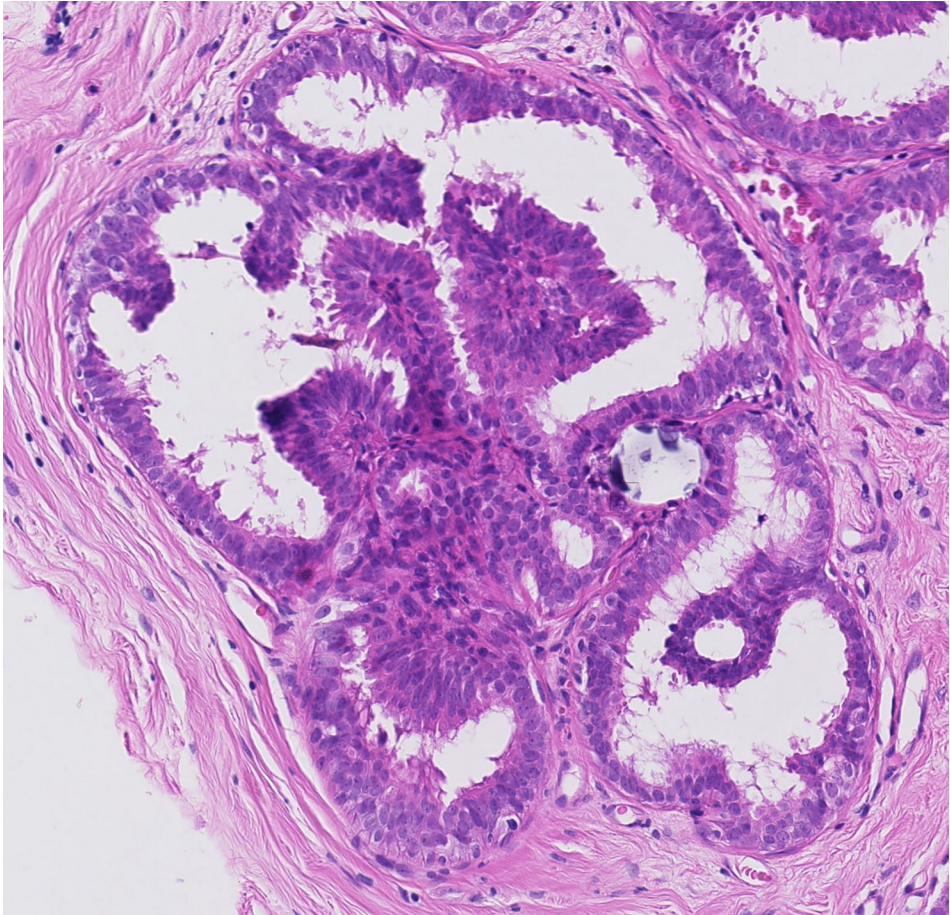
Grade I.

Grade II.
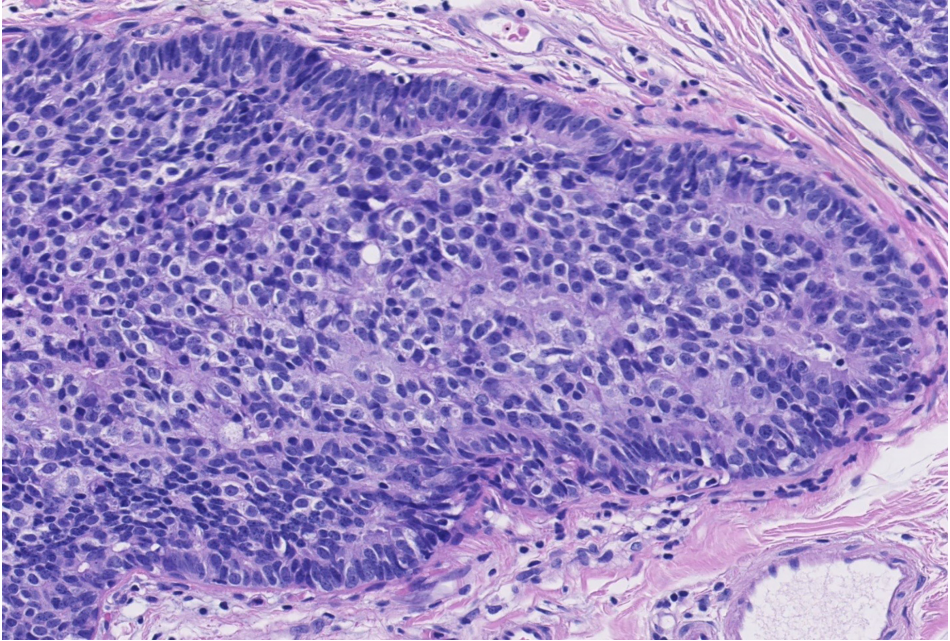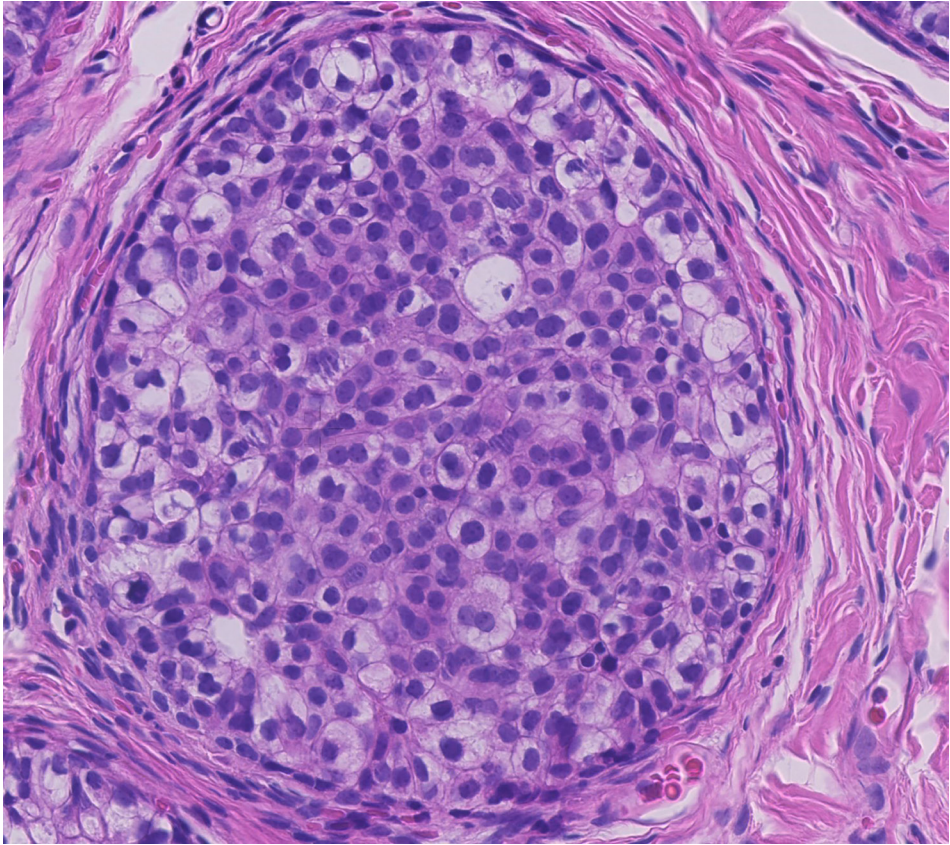
Grade III.
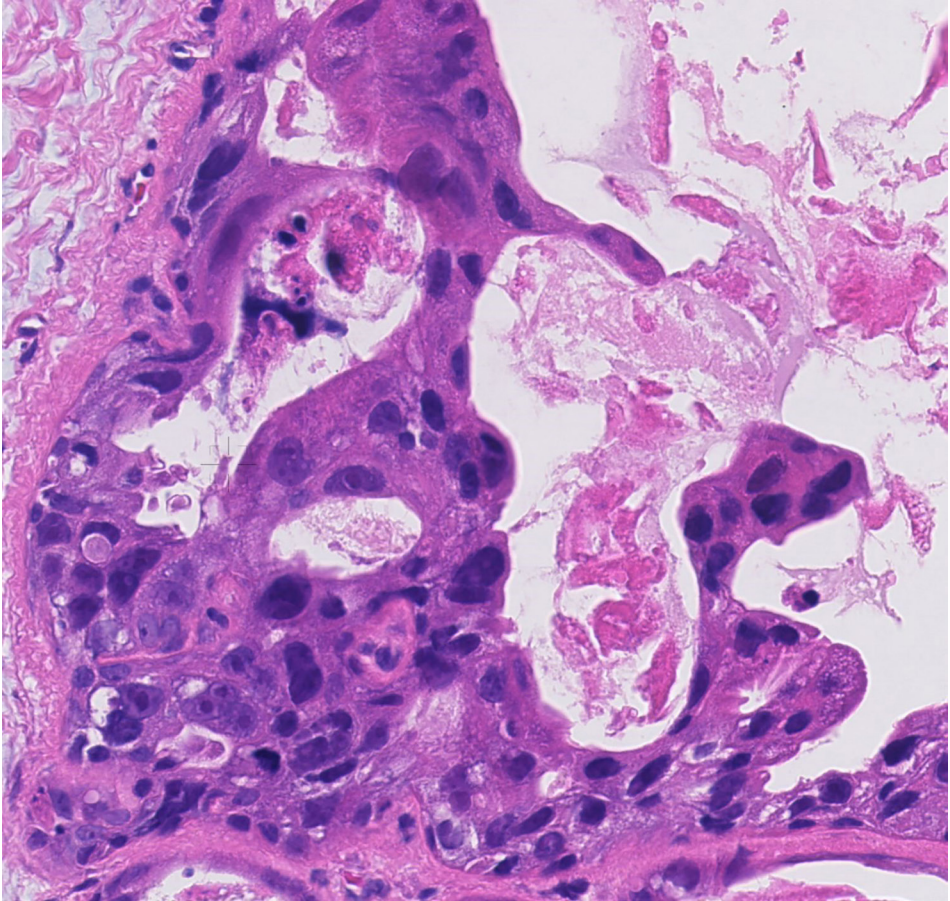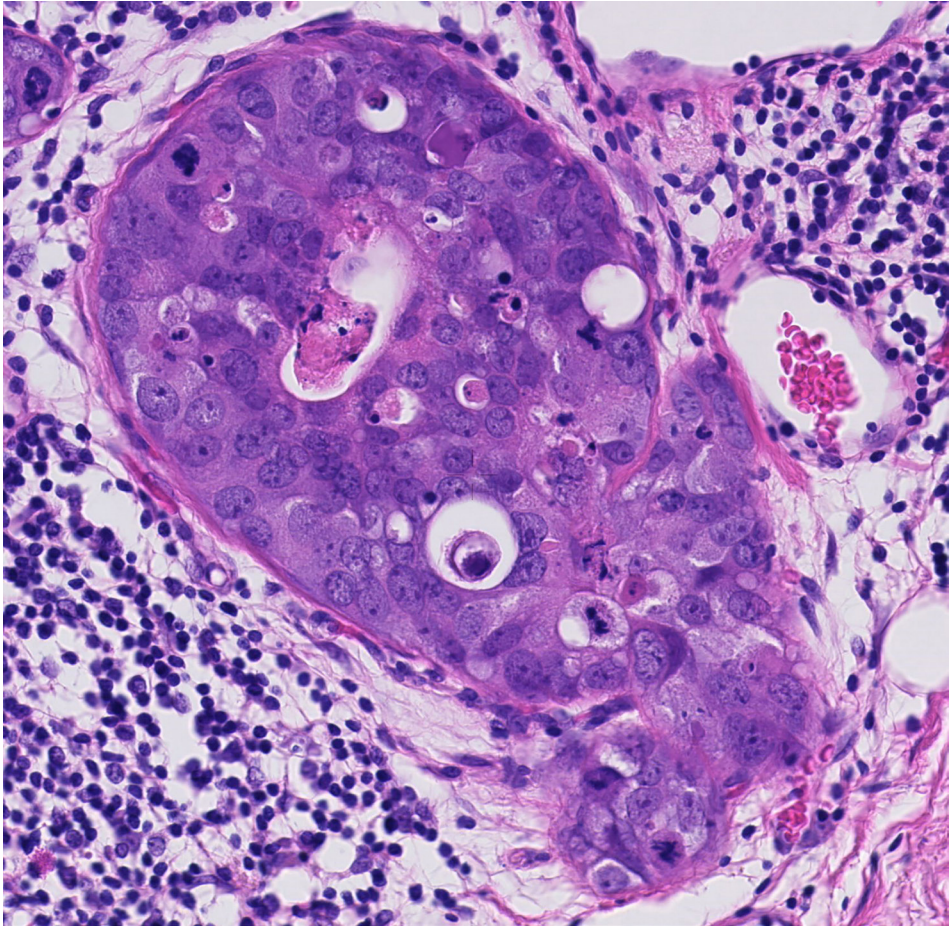
Supplementary 4. Some example images of an e-learning module for Nottingham grading of infiltrating breast cancer

<u>Tubular differentiation</u>, score: 1 point
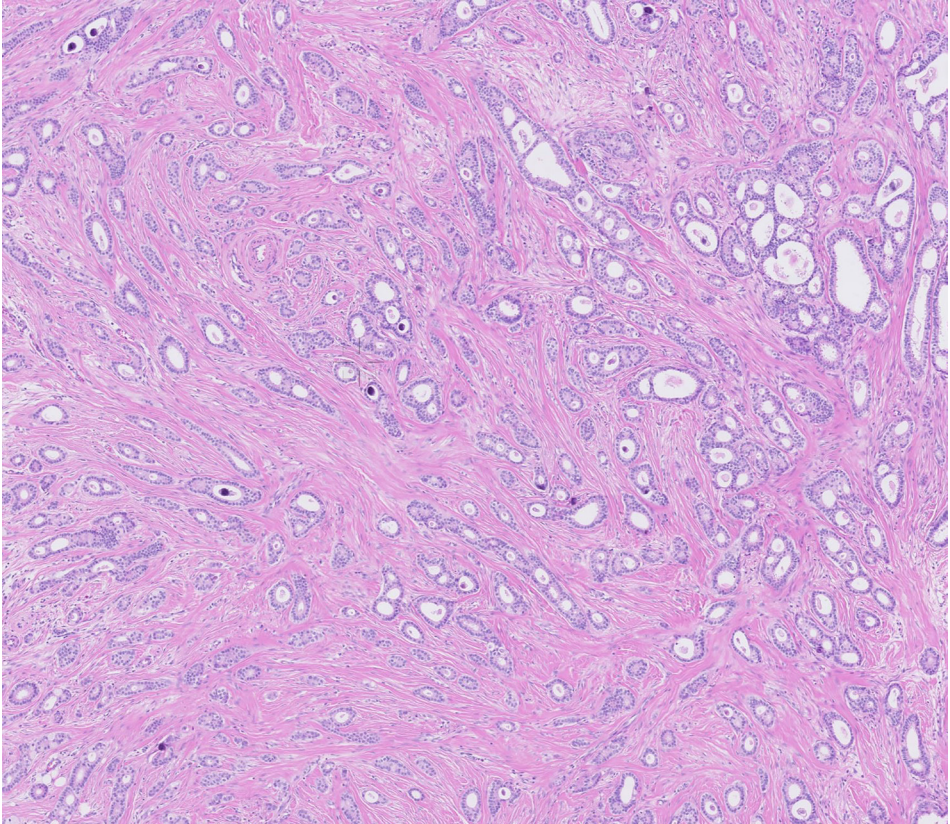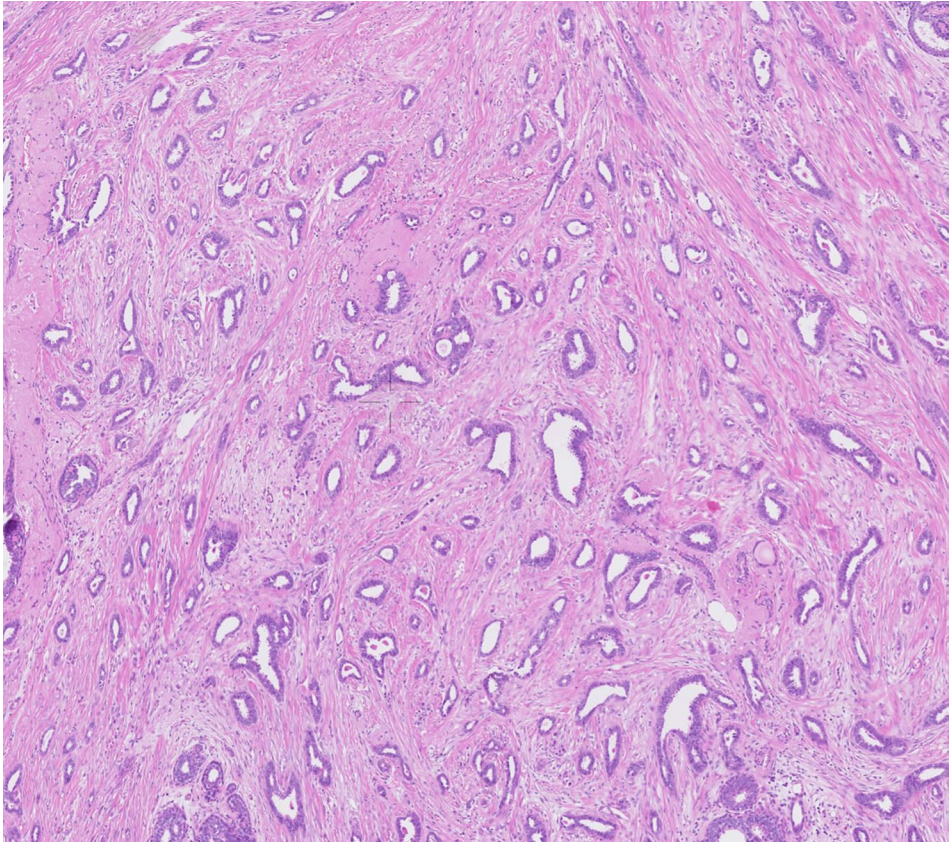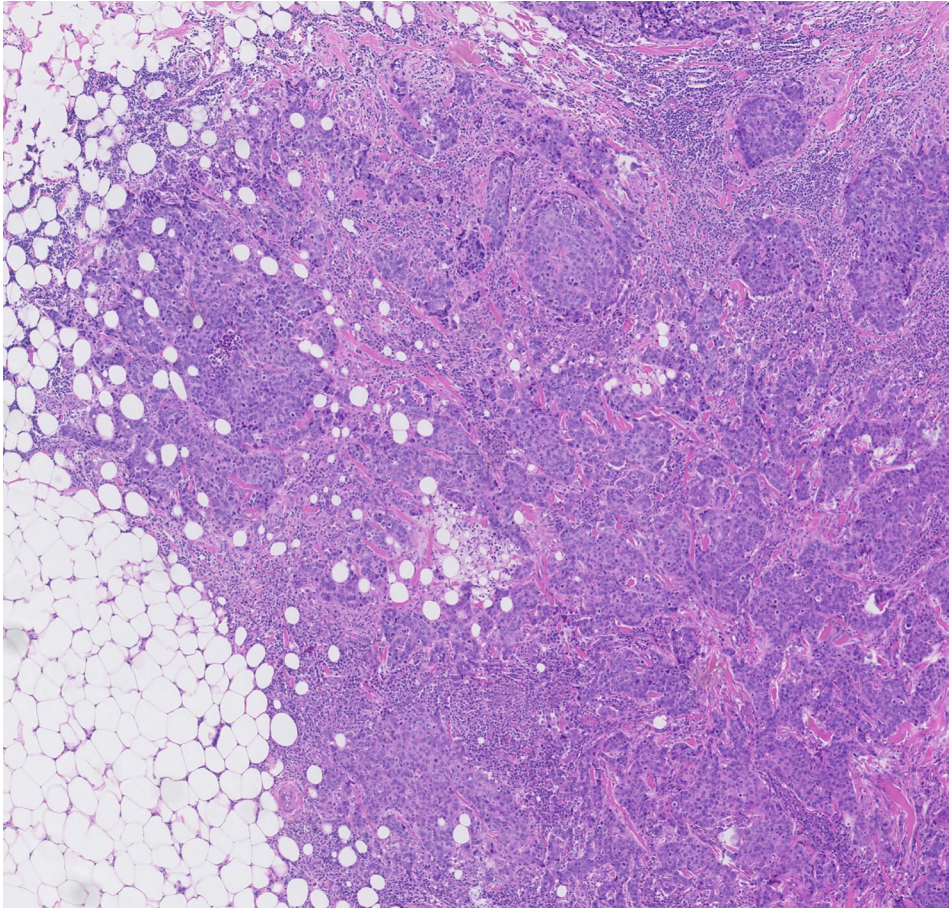
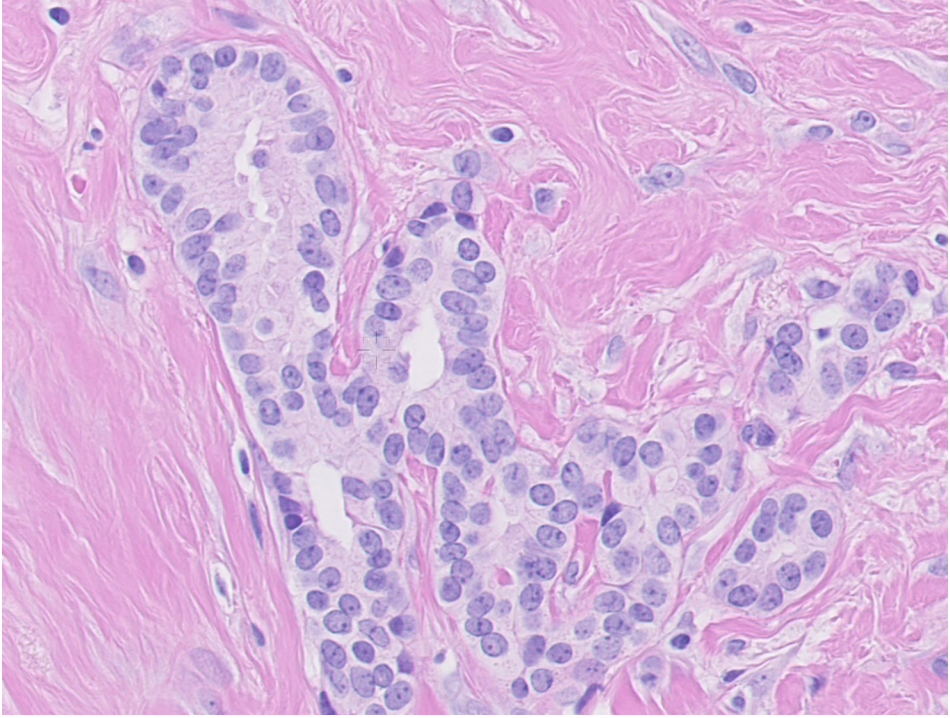**Tubular differentiation**, score: 2 points
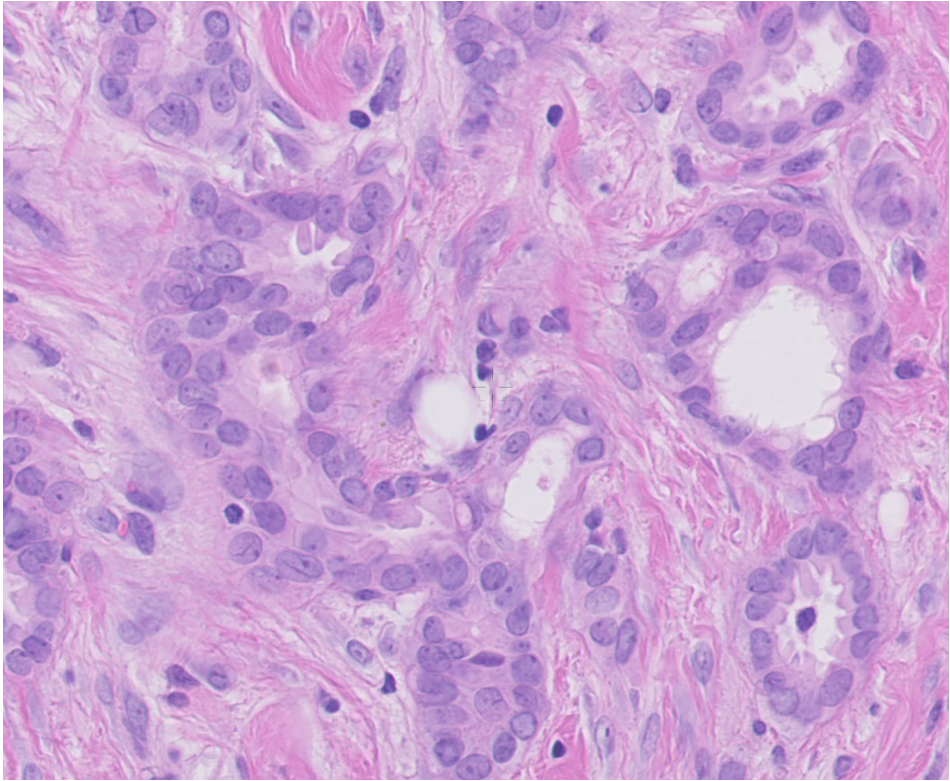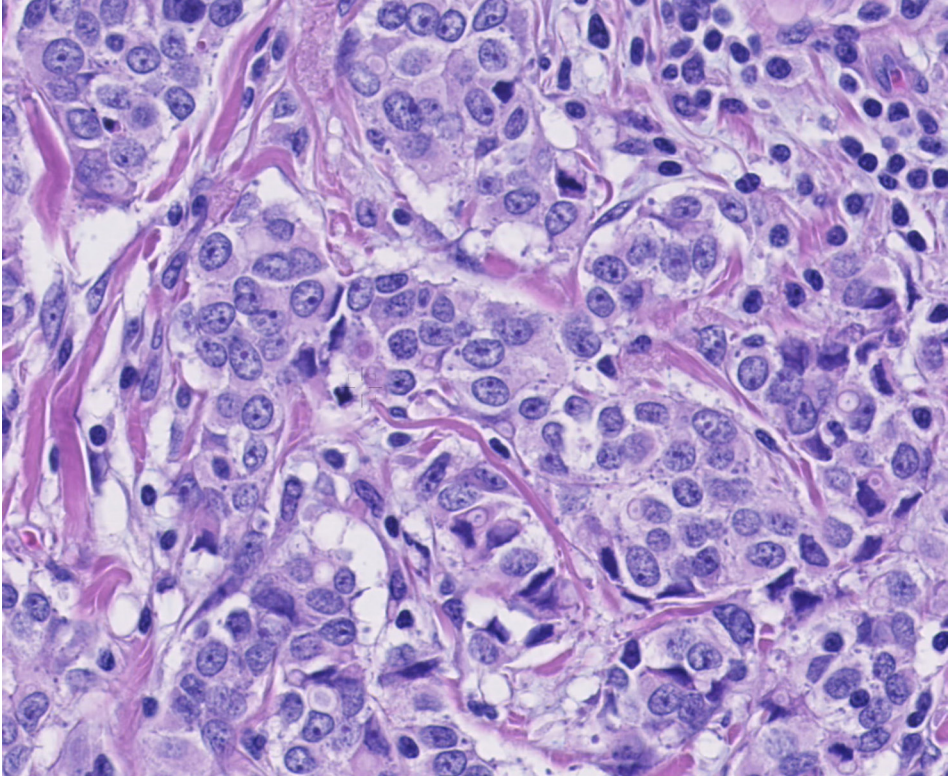

8
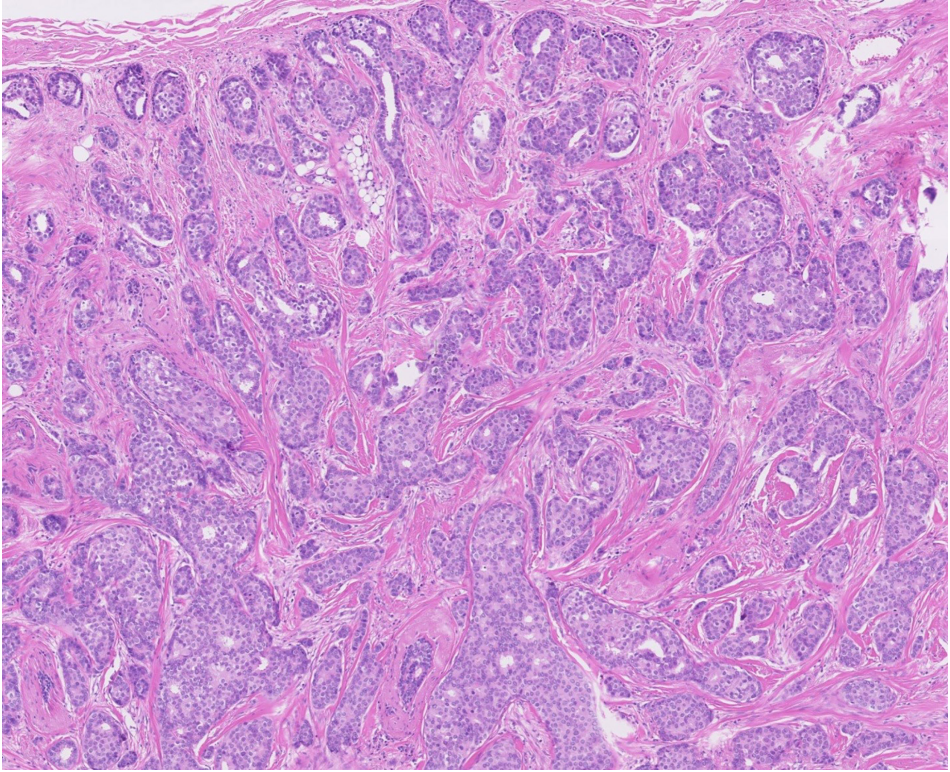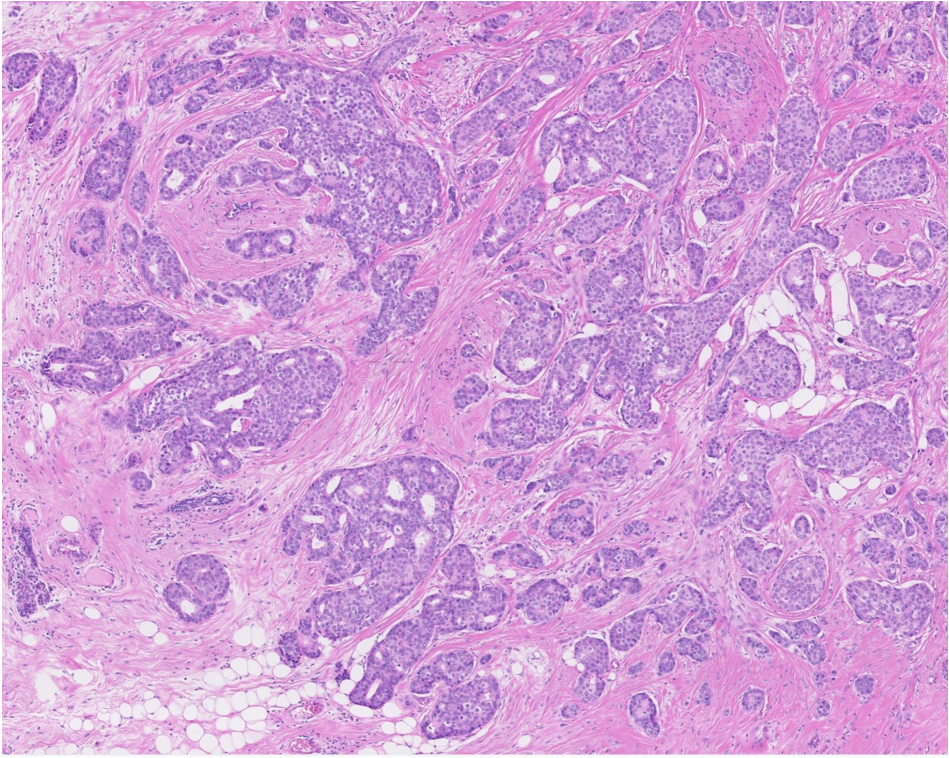
Tubular differentiation, score: 3 points
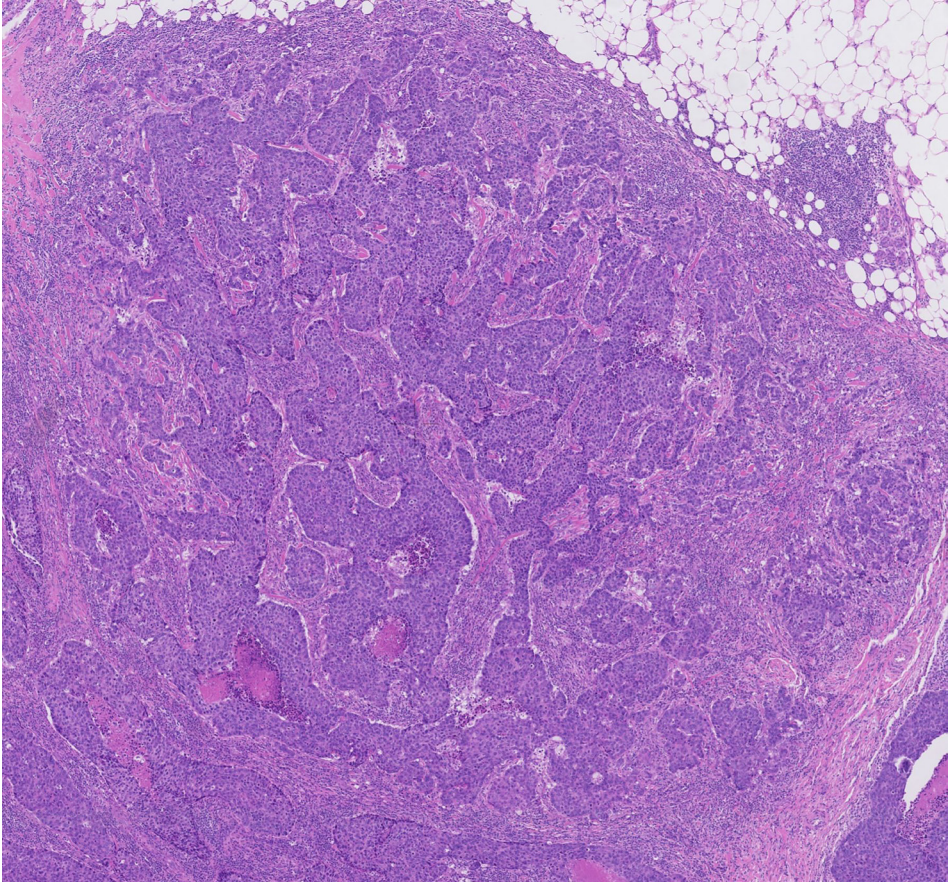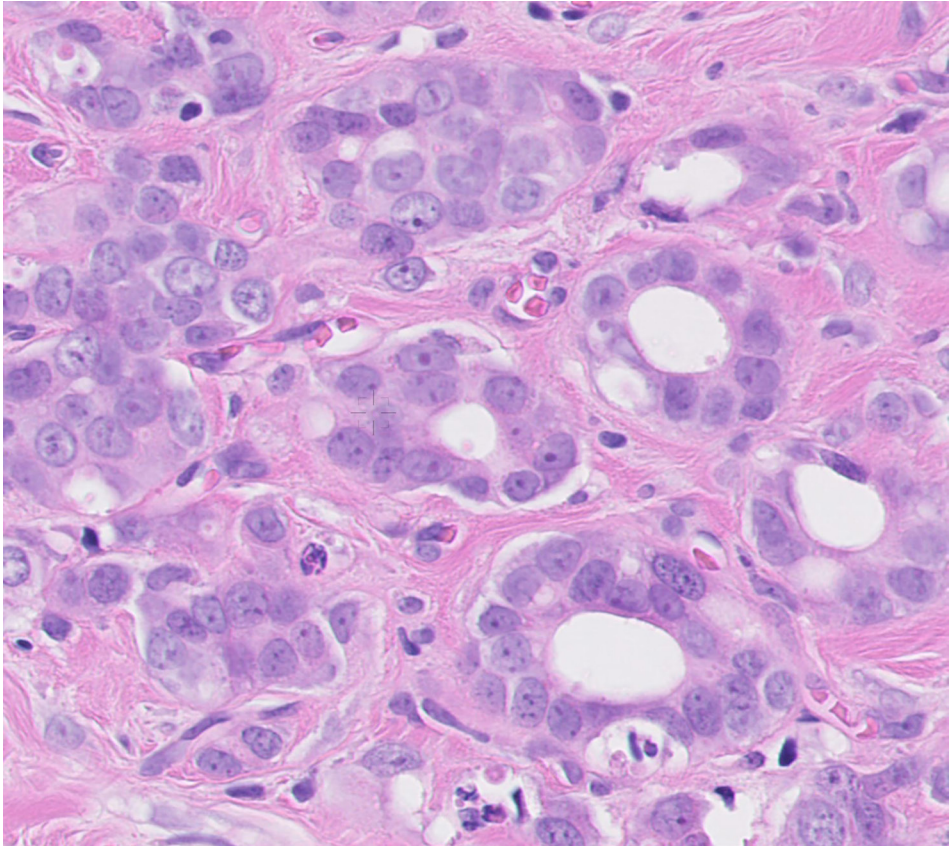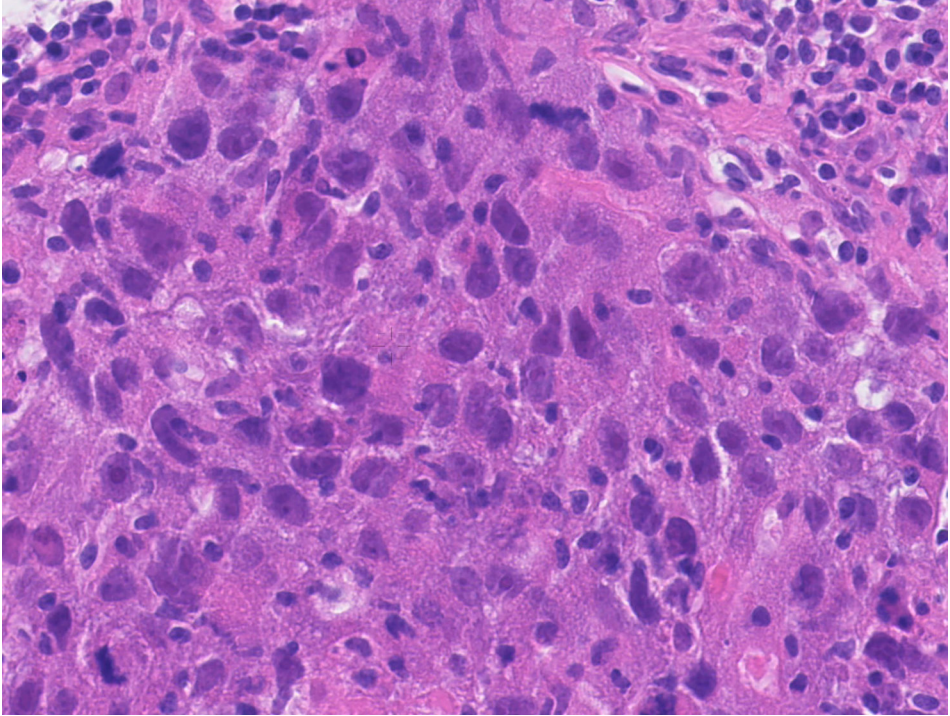
Nuclear pleomorphism, score: 1 point
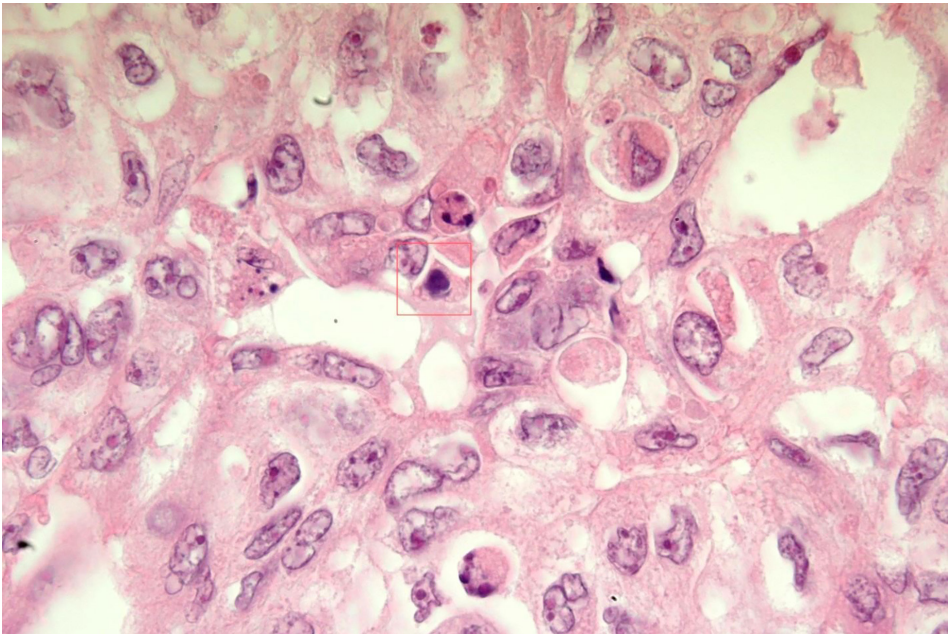
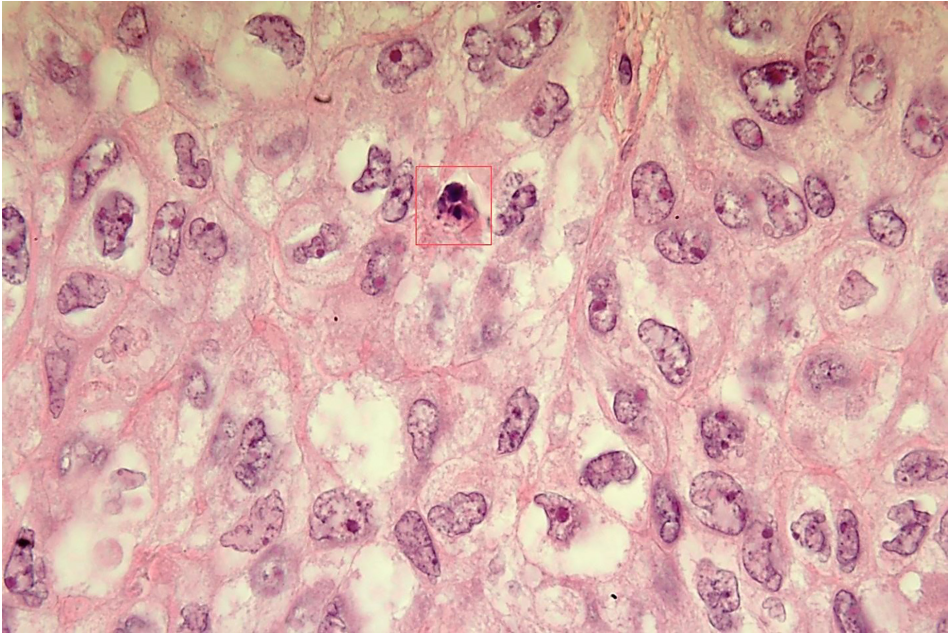**Nuclear pleomorphism**, score: 2 points

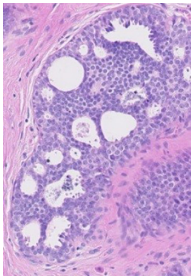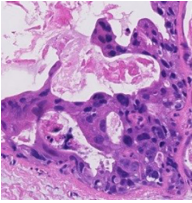**<u>Nuclear pleomorphism</u>**, score: 3 points
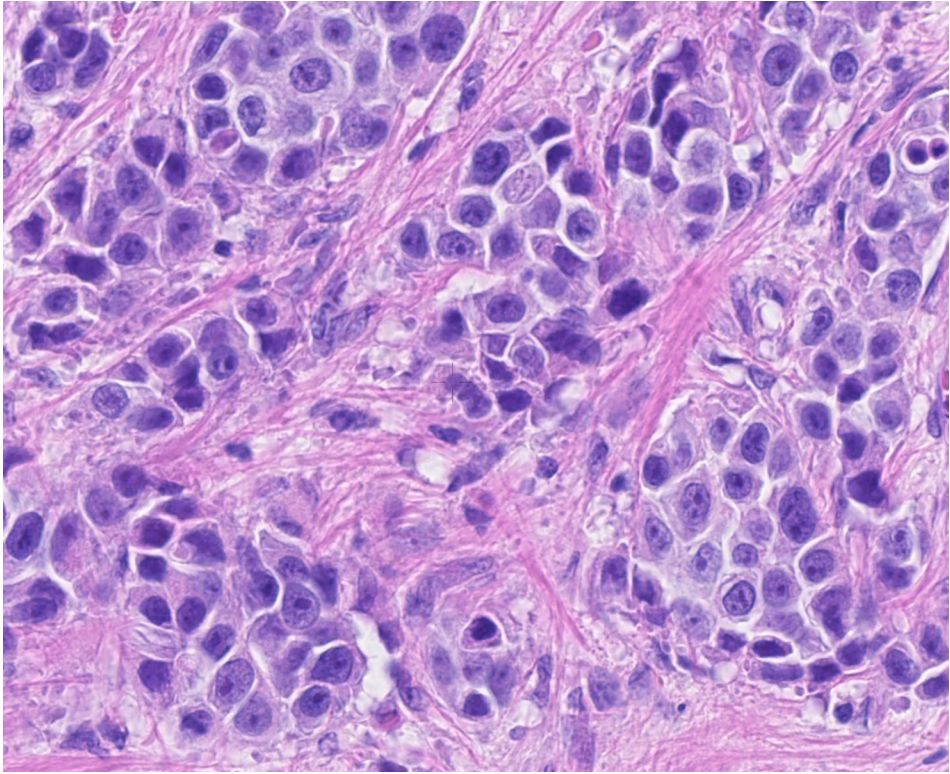
Mitosis (within red square)
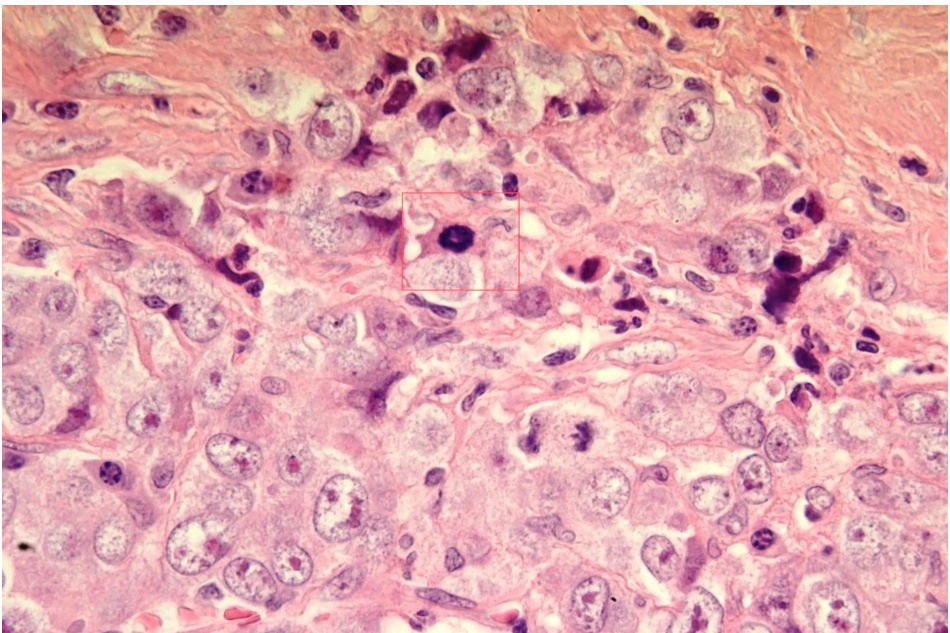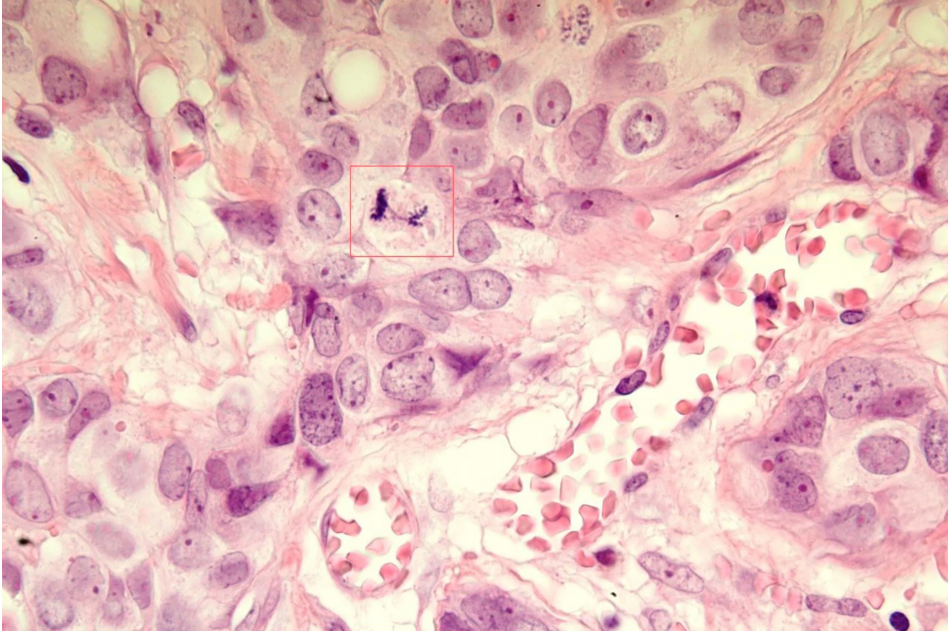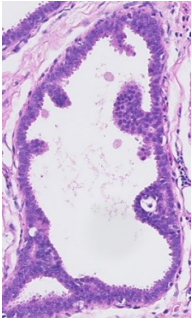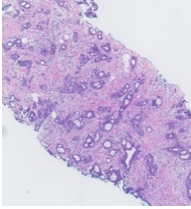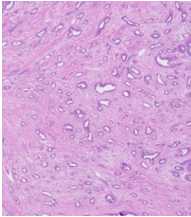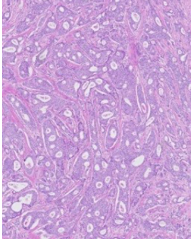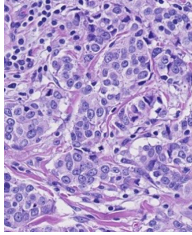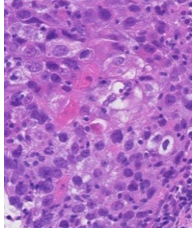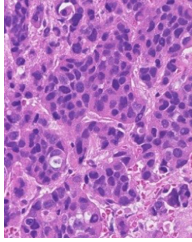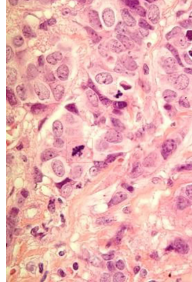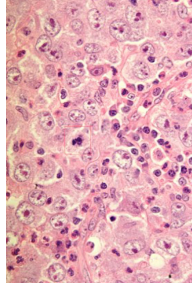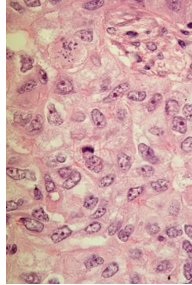
No mitosis (within red square)

Supplementary 5: Concordance before and after e-learning to grade DCIS and infiltrating cancers of the breast: a few examples

	Low concordance both pre- and post e-learning	High concordance both pre- and post e-learning	Low concordance pre-e-learning, high concordance post- e-learning
DCIS			
IBC: tubular differentiation			
IBC: nuclear pleomorphism			
IBC: mitoses			

8

9

Summarizing discussion
and future perspectives

Summarizing discussion

Treatment of individual breast cancer patients is primarily guided by biomarkers, including histologic grade and receptor-status for ER, PR, and HER2, which stresses the importance of high-quality biomarker assessment by pathologists. While it is well known that reproducibility of biomarker assessment is at best moderate in study settings [1-16], it is important to assess reproducibility in a real-world setting as this creates insight in daily practice, where patient populations may differ from study participants. Moreover, pathologists may feel best addressed by their own data. In this thesis, variation in biomarker assessment in daily pathology practice was analyzed on a nationwide level, thereby focusing on ER-, PR- and HER2-receptor assessment and histologic grading of both DCIS and IBC to create insight and awareness, which is a first condition for improvement. Subsequently, two initiatives were launched to minimize grading variation in daily pathology practice.

Part 1 - Biomarker assessment in daily pathology practice
In **chapter 2** we studied the inter-laboratory variation in case-mix adjusted receptor positivity rates for ER, PR, and HER2, in a nationwide cohort of 33,046 breast cancer patients, using real-life data from synoptic pathology reports between 2013 and 2016. Absolute differences between laboratories were reassuringly limited. While all Dutch pathology laboratories participate in mandatory external audits, like SKML, NordiQC and/or UK-Neqas, it has been argued that these programs only render a temporary and incomplete assessment of testing performance [17], as tissue fixation and processing steps are not covered.

As we observed that variation between individual pathologists within laboratories was minimal, we hypothesized that factors other than pathologists' interpretation of the fixed and stained tissue slides, for example tissue fixation techniques or the use of different antibodies, may explain the observed, though rather limited, differences between laboratories. We therefore believe that, in addition to the external audits, continuous monitoring and benchmarking of positivity rates may help to maintain the current high standard of receptor assessment in the Netherlands. This could for example be performed within the existing nationwide multidisciplinary breast cancer audit (NBCA) [18], which currently only reports on one pathology indicator, i.e. whether the synoptic PALGA protocol is used [19].

In **chapter 3** we evaluated variation in histologic grading of DCIS in daily clinical practice within 4,901 patients with synoptically reported pure DCIS lesions (i.e. without an invasive component) between 2013 and 2016. Substantial variation was observed between laboratories (ranging up to 40%) and between pathologists within individual laboratories

(ranging up to 70%). The latter implies that even within laboratories, histologic grading is not performed in a standardized manner, hence, the observed inter-laboratory variation may predominantly be the result of different grading practices of individual pathologists. This is further supported by the results of our questionnaire, which showed that numerous reference classifications (as well as 'intuition') were mentioned as guideline for DCIS grading by pathologists. This calls for (inter)national consensus on the grading system and criteria to improve reproducibility in view of the therapeutic consequences.

The results of this study emphasize that histologic grading of DCIS is currently not meeting acceptable clinical standards. Awareness is an important first step towards improvement. This is highly relevant since the future management of DCIS may solely depend on histologic grade, as this is currently being investigated by multiple clinical trials [20-24]. Our study also stresses the importance of central pathology review within these trials to justify their conclusions, which, however, is not always carried out [25].

In **chapter 4** we assessed the inter- and intra-laboratory variation in histologic grading of IBC within 33,043 patients with a synoptically reported IBC resection specimen between 2013 and 2016. Again, we found substantial variation, both between laboratories (ranging up to 30%) and between pathologists within individual laboratories (range up to 50%). Of the three components that comprise grade (tubular formation, nuclear pleomorphism, mitotic count), most variation between laboratories was observed in scoring nuclear pleomorphism, which is in line with previous studies [13-16]. This may be explained by the fact that scoring of this subcategory is least quantitative, as compared to the other component that comprise grade (e.g. scoring the percentage of tubular formation and mitosis counting).

Interestingly, more than three-quarters of patients showed an overall grading score on a tipping point of grades (i.e. scores 5 or 6, and scores 7 or 8), which indicates that variation of only one point (on any of the three subcategories) may already alter the overall grade, and may therefore influence patient management. This is supported by our finding that grade determines whether adjuvant chemotherapy (aCT) is indicated for about one in every three IBC patients. It is therefore presumable that variation in grading influences treatment decisions, and subsequently may influence outcome, including exposure to unnecessary toxicity of individual breast cancer patients. Since we also found significant differences in grading between pathologists in five out of the eight analyzed laboratories (62.5%) shows that even within laboratories histologic grading may not be performed in a similar manner. Although, in contrast to DCIS grading, a single guideline (modified Bloom- and Richardson classification) is used for IBC grading. In view of its clinical relevance, we strongly believe that every effort should be taken to minimize the observed variation. One

9

option could be to develop an individual training program for pathologists, for example by e-learning, followed by a regular evaluation whether this results in a better harmonization of histologic grading of IBC.

In **chapter 5** we studied the increasing relevance of histologic grading in tailoring adjuvant systemic therapy in light of the previous and current Dutch breast cancer guidelines. By linking two nationwide databases (PALGA and the Dutch Cancer Registry) we were able to analyze real-world pathology and treatment data from 30,843 breast cancer patients. Next to a high degree of guideline non-adherence, in which histologic grade played a key role, these data illustrate the increasing relevance of histologic grade in tailoring adjuvant systemic therapy. We showed that histologic grade is decisive for the aCT-indication in a third of IBC patients, as compared to a quarter of IBC patients in the previous guideline. In addition, the role of histologic grading has even extended to de-escalation of aET, as grade is now decisive in about a third of IBC patients as well. Lastly, whether patients are eligible for the 70-gene-signature (70-GS) is (partially) guided by histologic grade.

Of note, although the indication for aCT according to the previous guideline depended on histologic grade in approximately 25% of patients, we were unable to show that patients who had their tumor graded in 'higher grading laboratories' subsequently received chemotherapy more often. This is most likely due to the high degree of guideline non-adherence in patients with a guideline aCT indication (40%). If guideline adherence would be 100%, grading more breast cancers as high grade would, for those 25% of patients where grade is the decisive clinical factor, result in more frequent administration of aCT. This again stresses the need for optimizing grading by pathologists as much as possible, to minimize the risk of worse patient outcome due to under-treatment, and unnecessary costs and suffering from side effects due to overtreatment.

Part 2 - Initiatives to decrease variation in biomarker assessment
In **chapter 6** we studied the effect of laboratory-specific feedback reports as a first intervention to decrease variation in grading of DCIS. Within these feedback reports, case-mix adjusted laboratory-specific proportions per grade were benchmarked against other laboratories and the nationwide average. As feedback reports were sent on March 1 2018, we compared grading variation between the year before (March 1, 2017-March 1, 2018) and the year after feedback (March 1, 2018-March 1, 2019), using real-life nationwide data from 2,934 patients with synoptic resection specimen pathology reports. In addition, some laboratories, who provided specific pathologist data themselves, also received feedback on pathologist-level. A promising decrease in grading variation was observed for DCIS grades II and III, while this was not observed for DCIS grade I.

We hypothesize that the lack of consensus on a grading classification [26, 27] may be the explanation for these mixed results as uniform grading starts with the use of single grading classification. It seems that histologic grade is currently not a clinically optimal biomarker for DCIS, let alone that it should be the single factor that decides whether patients should be treated or not, as overall grading variation remains substantial for all grades. Improvement and standardization is adamant in light of current ongoing trials [20-24], especially since central pathology review is not carried out in all trials [25].

We believe that feedback reports may be a useful tool to monitor grading variation in daily pathology practice, thereby creating awareness. This may open the much needed discussion on a single nationwide DCIS grading classification. As several studies have highlighted the complexity of DCIS grading, even among expert breast pathologists [28,29], adequate training of (expert breast) pathologists, according to a single grading classification, may help to establish uniform grading in clinical practice over time.

In **chapter 7** we analyzed the effect of laboratory-specific feedback reports on IBC grading variation within a cohort of 16,734 breast cancer patients. We observed an encouraging decrease in grading variation in the year after feedback. It is important to emphasize that the primary aim of the feedback reports was to create awareness on grading variation among pathologists by showing them their own data, and not to simply make 'higher' grading pathologists grade their tumors lower (or vice versa). Yet, these feedback reports enabled pathologists to discuss how they grade with other pathologists. This could lead to a mutual exchange of thoughts on how to interpret the guideline, likely resulting in more consensus. Therefore we strongly believe that feedback reports on a regular basis are an important first step towards improvement of breast cancer grading, which is supported by the results of this study. Nevertheless, overall grading variation remains substantial, even after feedback, which shows that other interventions, such as additional training of pathologists, are urgently required.

In **chapter 8** we investigated the effect of an e-learning module as a second intervention to decrease variation in histologic grading of both DCIS and IBC. We showed that our e-learning module, in which pathologists and pathology residents were trained in histologic grading of DCIS and/or IBC, is a feasible and promising tool for this purpose. This is highly relevant considering the substantial grading variation in daily pathology practice as described above, bearing in mind the important role of grade in clinical decision-making of DCIS and IBC. An increased consensus on grading of both DCIS and IBC among pathologists will subsequently diminish variation in treatment, and thereby most likely also patient outcome. Implementing an e-learning as presented in this study, in combination with regular feedback reports to monitor variation in biomarker assessment in daily clinical

9

practice, should be considered by the Dutch Society of Pathology (NVvP) as a way to optimize breast cancer biomarker assessment for those that want to qualify as breast pathologist, and for training residents. This is in line with the guideline for pathologists who diagnose Dutch colorectal cancer screening cases who are already obliged to participate in an e-learning to obtain their certification.

Conclusions of this thesis

We showed that nationwide variation for ER-, PR-, and HER2-receptor assessment in daily clinical practice is limited. In contrast, our data on nationwide variation in histologic grading of both DCIS and IBC in routine care cannot be classified any other than substantial. Minimizing grading variation is highly relevant given the crucial role of grading in clinical decision-making. Therefore, two promising initiatives, laboratory-specific feedback reports and an e-learning module were launched. Although these initiatives showed encouraging results, overall grading variation remained substantial. Despite the indisputable need to further diminish variation in grading, we believe that it is also important to emphasize that histologic grade is not a fact, nor the truth, but merely a model or tool that consists of a statistically computed set of cut-off values in a spectrum of histopathologic features to classify expected tumor behavior. Yet, these cut-offs are often translated into hard criteria by clinicians in therapeutic decision-making. We therefore believe that awareness among clinicians, and understanding of the difficulties of histologic grading, is crucial to improve clinical decision making for patients as well. Furthermore, in the current era of shared-decision-making, where relevant, these difficulties should also be discussed with patients.

Future perspectives

Regular feedback reports and large-scale use of the e-learning module

We believe that it is important to continue both improvement initiatives, not in the least because improvement takes time. We analyzed the effect of the first round of feedback reports, hence, this was not yet a closed quality loop, while laboratories were also not used to this type of feedback. Routine (for example annual) feedback reports will close the quality loop, and pathology laboratories will get familiar with these reports, while grading variation on a nationwide level is monitored in the meantime. In addition, unrolling our e-learning, for example among pathologists who participate in the national breast cancer screening program, preferably as an obligatory tool to obtain certification, and among pathology residents, may result in better standardization on a larger scale. These items would preferably be an integral part of the quality system of pathology labs dealing with breast specimens, and get externally audited by e.g. the Dutch Society of Pathology audit teams or DICA (www.dica.nl).

Additional avenues to ensure high quality oncological care

Next to the above mentioned ongoing initiatives, additional avenues should be explored to ensure high quality oncological care [30]. It may for example be interesting to explore whether and how artificial intelligence could support pathologists in biomarker assessment [31-33]. However, it is important to note that this requires well annotated, consensus-based, training datasets to begin with. The grading sets that were prepared for the e-learning could be quite useful here.

Another initiative may be systematic double reading by experts or peer consultation in specific cases, for example when grade determines whether adjuvant systemic therapy is indicated. In this light, the Dutch Pathology Image Exchange (PIE) [34] project may provide a useful platform for fast review. However, as we have seen that grade is the decisive factor in a substantial amount of patients, one could argue the practicability of peer consultation in all of these cases. Reimbursement should be in place and fair, which should be doable since costs of double reading will be far less than gene-expression profiling (GEP).

Lastly, although we have clearly observed the added value of GEP in aCT de-escalation, it is important to emphasize that their use, in specific subgroups, is primarily dependent on histopathologic factors. Also, their worldwide availability is limited because of their costs. Furthermore, multiple studies have described that Oncotype DX scores may be accurately predicted by PR-receptor status and histologic grade in specific subgroups, omitting the need for expensive GEPs [35-41].

9

Specialized breast pathologists?

Another question that may be raised by our research is whether it is desirable that some laboratories assess only few IBC or DCIS cases per year. Considering the complexity and clinical importance of histologic grading, one could argue that this should only be performed by expert breast pathologists [42]. However, we would like to emphasize that our data are not suitable for this type of analysis, since we used synoptic pathology reports only, thereby excluding approximately 20% of Dutch breast cancer cases, as these were reported narratively [43]. Hence, the absolute number of cases per laboratory cannot be drawn from our data. However, this is an interesting discussion that should be conducted based on the proper data.

Outcome analyses: was aCT guideline non-adherence justified?

Whether de-escalation of aCT in 40% of aCT-eligible women in our cohort did not negatively affect patient outcome remains to be elucidated after longer follow-up. This analysis will provide additional important clinical data and will be conducted by our group in the future.

References

1. Cuadros M, Cano C, López FJ, et al. HER2 status in breast cancer: experience of a Spanish National Reference Centre. Clin Transl Oncol. 2011;13(5)335-340.

2. Denkert C, Huober J, Loibl S, et al. HER2 and ESR1 mRNA expression levels and response to neoadjuvant trastuzumab plus chemotherapy in patients with primary breast cancer. Breast Cancer Res. 2013;15(1):R11.

3. McCullough AE, Dell'orto P, Reinholz MM, et al. Central pathology laboratory review of HER2 and ER in early breast cancer: an ALTTO trial [BIG 2-06/NCCTG N063D (Alliance)] ring study. Breast Cancer Res Treat. 2014;143(3):485-92.

4. Orlando L, Viale G, Bria E, et al. Discordance in pathology report after central pathology review: Implications for breast cancer adjuvant treatment. The Breast. 2016;30:151-5.

5. Paik S, Bryant J, Tan-Chiu E, et al. Real-world performance of HER2 testing -National Surgical Adjuvant Breast and Bowel Project experience. J Natl Cancer Inst. 2002;94(11):852-4.

6. Perez EA, Suman VJ, Davidson NE, et al. HER2 testing by local, central, and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial. J Clin Oncol. 2006;24(19):3032-8.

7. Roche PC, Suman VJ, Jenkins RB, et al. Concordance between local and central laboratory HER2 testing in the breast intergroup trial N9831. J Natl Cancer Inst. 2002;94(11):855-7.

8. Rosa M, Khazai L. Comparison of HER2 testing among laboratories: Our experience with review cases retested at Moffitt Cancer Center in a two-year period. The Breast Journal. 2017;24(2):139-47.

9. Viale G, Regan MM, Maiorano E, et al. Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1-98. J Clin Oncol. 2007;25(25):3846-52.

10. Douglas-Jones AG, Morgan JM, Appleton MA, et al. Consistency in the observation of features used to classify duct carcinoma in situ (DCIS) of the breast. J Clin Pathol. 2000;53(8):596-602.

11. Elston CW, Sloane JP, Amendoeira I, et al. Causes of inconsistency in diagnosing and classifying intraductal proliferations of the breast. European Commission Working Group on Breast Screening Pathology. Eur J Cancer. 2000;36(14):1769-72.

12. 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(10):705-11.

13. Boiesen P, Bendahl PO, Anagnostaki L, et al. Histologic grading in breast cancer--reproducibility between seven pathologic departments. South Sweden Breast Cancer Group. Acta Oncol. 39;(1):41-45.

14. Frierson HF Jr., Wolber RA, Berean KW, et al. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. Am J Clin Pathol. 103(2):195-198.

15. Italian Network for Quality Assurance of Tumour Biomarkers (INQAT) Group. Quality control for histological grading in breast cancer: an Italian experience. Pathologica. 2005;97(1):1-6.

16. Meyer JS, Alvarez C, Milikowski C, et al. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation

9

index: reproducibility of grade and advantages of proliferation index. Mod Pathol. 2005;18(8):1067-78.

17. Choritz H, Busche G, Kreipe H. Quality assessment of HER2 testing by monitoring of positivity rates. Virchows Arch. 2011;459(3):283-9.

18. van Bommel AC, Spronk PE, Vrancken Peeters MT, et al. Clinical auditing as an instrument for quality improvement in breast cancer care in the Netherlands: The national NABON Breast Cancer Audit. J Surg Oncol. 2017;115(3):243-9.

19. Dutch Institute for Clinical Auditing (DICA)/ Netherlands Comprehensive Cancer Organisation/ NABON Breast Cancer Audit (NBCA) (2018) Factsheet Indicators NABON Breast Cancer Audit (NBCA) 2018, version 2018.3.

20. 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(3):e026797.

21. 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(12):1497-510.

22. Francis A, Thomas J, Fallowfield L, et al. Addressing overtreatment of screen detected DCIS; the LORIS trial. Eur J Cancer. 2015;51(16):2296-303.

23. Groen EJ, Elshof LE, Visser LL, et al. Finding the balance between over- and under-treatment of ductal carcinoma in situ (DCIS). Breast. 2017;31:274-83.

24. Lippey J, Spillane A, Saunders C. Not all ductal carcinoma in situ is created equal: can we avoid surgery for low-risk ductal carcinoma in situ? ANZ J Surg. 2016;86(11):859-60.

25. Van Bockstal MR, Agahozo MC, Koppert LB, van Deurzen CHM. A retrospective alternative for active surveillance trials for ductal carcinoma in situ of the breast. Int J Cancer. 2020;146(5):1189-1197.

26. van Dooijeweert C, van Diest PJ, Willems SM, Kuijpers C, Overbeek LIH, Deckers IAG. Significant inter- and intra-laboratory variation in grading of ductal carcinoma in situ of the breast: a nationwide study of 4901 patients in the Netherlands. Breast Cancer Res Treat. 2019;174(2):479-88.

27. 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(4):285-92.

28. Harrison BT, Hwang ES, Partridge AH, Thompson AM, Schnitt SJ. Variability in diagnostic threshold for comedo necrosis among breast pathologists: implications for patient eligibility for active surveillance trials of ductal carcinoma in situ. Mod Pathol. 2019;32(9):1257-1262.

29. Dano H, Altinay S, Arnould L, et al. Interobserver variability in upfront dichotomous histopathological assessment of ductal carcinoma in situ of the breast: the DCISion study. Mod Pathol. 2019;33(3):354-366.

30. Nass SJ, Cohen MB, Nayar R, et al. Improving Cancer Diagnosis and Care: Patient Access to High-Quality Oncologic Pathology. Oncologist. 2019;24(10):1287-90.

31. BE Bejnordi , Veta M, van Diest PJ, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017;318(22):2199-210.

32. Veta M, Heng YJ, Stathonikos N, et al. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. Med Image Anal. 2019;54:111-21.

33. Veta M, van Diest PJ, Jiwa M, Al-Janabi S, Pluim JP. Mitosis Counting in Breast Cancer: Object-Level Interobserver Agreement and

Comparison to an Automatic Method. PloS One. 2016;11(8):e0161286.

34. van Diest PJ, Huisman A, van Ekris J, et al. Pathology Image Exchange: The Dutch Digital Pathology Platform for Exchange of Whole-Slide Images for Efficient Teleconsultation, Televerision, and Virtual Expert Panels. JCO Clin Canc Inform. 2019;3:1-7.

35. Wu SG, Zhang WW, Wang J, et al. Progesterone receptor status and tumor grade predict the 21-gene recurrence score of invasive lobular breast cancer. Biomark Med. 2019;13(12):1005-12.

36. Allison KH, Kandalaft PL, Sitlani CM, Dintzis SM, Gown AM. Routine pathologic parameters can predict Oncotype DXTM recurrence scores in subsets of ER positive patients: who does not always need testing? Breast Cancer Res Treat. 2012;131(2):413-24.

37. Huang JL, Kizy S, Marmor S, et al. Tumor grade and progesterone receptor status predict 21-gene recurrence score in early stage invasive breast carcinoma. Breast Cancer Res Treat. 2018;172(3):671-7.

38. Yoo SH, Kim T-Y, Kim M, et al. Development of a Nomogram to Predict the Recurrence Score of 21-Gene Prediction Assay in Hormone Receptor–Positive Early

Breast Cancer. Clinical Breast Cancer. 2020;20(2):98-107.

39. Orucevic A, Bell JL, King M, McNabb AP, Heidel RE. Nomogram update based on TAILORx clinical trial results - Oncotype DX breast cancer recurrence score can be predicted using clinicopathologic data. Breast. 2019;46:116-25.

40. Davis BA, Aminawung JA, Abu-Khalaf MM, et al. Racial and Ethnic Disparities in Oncotype DX Test Receipt in a Statewide Population-Based Study. J Natl Compr Canc Netw. 2017;15(3):346-54.

41. Zhang L, Hsieh MC, Petkov V, Yu Q, Chiu YW, Wu XC. Trend and survival benefit of Oncotype DX use among female hormone receptor-positive breast cancer patients in 17 SEER registries, 2004-2015. Breast Cancer Res Treat. 2020;180(2):491-501.

42. 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. 2002;41(3a):154-61.

43. Foundation PALGA. Annual report 2018. Available from: https://www.palga.nl/assets/uploads/Jaarverslag/2018%20Palga%20web.pdf. Accessed October 10, 2019.

9

APPENDICES

List of abbreviations
Summary in Dutch / Nederlandse samenvatting
Acknowledgements / dankwoord
Curriculum Vitae
List of Publications

List of abbreviations

aCT	Adjuvant chemotherapy
aET	Adjuvant endocrine therapy
AOR	Adjusted odds ratio
CI	Confidence interval
CISH	Chromogenic in situ hybridization
COMET	Comparison of Operative versus Monitoring and Endocrine Therapy trial
DCIS	Ductal carcinoma *in situ*
ER	Estrogen receptor
FISH	Fluorescence in situ hybridization
GDPR	General Data Protection Regulation
GEP	Gene-Expression-Profiling
HER2	Human Epidermal growth factor Receptor 2
IBC	Invasive breast cancer
IHC	Immunohistochemistry
IKNL	Integraal Kankercentrum Nederland/ The Netherlands Comprehensive Cancer Organisation
IRR	Inter-rater reliability
ISH	In situ hybridization
Ki67	Antigen Ki-67
LARRIKIN	Low And InteRmediate RIsK ductal carcinoma IN situ study
LORD	Low Risk DCIS trial
LORIS	LOw RISk DCIS trial
MINDACT	Microarray In Node-negative (or 1-3 positive lymph nodes) Disease may Avoid ChemoTherapy
NVvP	Nederlandse Vereniging voor Pathologie/ the Dutch Society of Pathology
ODS	Overall deviation score
OR	Odds ratio
PALGA	Pathologisch-Anatomisch Landelijk Geautomatiseerd Archief/ The nationwide network and registry of histo- and cytopathology in the Netherlands
PIE	Pathology Image Exchange
PR	Progesterone receptor
SD	Standard deviation
SISH	Silver in situ hybridization
SKMS	Stichting Kwaliteitsgelden Medisch Specialisten/ The Quality Foundation of the Dutch Association of Medical Specialists
TNBC	Triple negative breast cancer
21-GS	21-gene recurrence score assay
70-GS	70-gene signature

Nederlandse samenvatting

Borstkanker in Nederland

Met 2,1 miljoen nieuwe gevallen per jaar is borstkanker wereldwijd de meest voorkomende vorm van kanker bij vrouwen. In Nederland betekent dit dat ongeveer 1 op de 7 vrouwen gedurende haar leven borstkanker ontwikkelt, wat neer komt op 15.000 vrouwen met een borstkankerdiagnose per jaar. Daarnaast wordt bij meer dan 2.000 vrouwen per jaar een voorstadium van borstkanker geconstateerd, het zogenaamde ductaal carcinoom *in situ* ofwel DCIS (*in situ:* 'op zijn plek', de tumor groeit niet door in het omliggende weefsel). Het is echter onduidelijk hoeveel van deze DCIS zich, indien onbehandeld, zouden ontwikkelen tot invasieve borstkanker (invasief: de tumor groeit in het omliggende weefsel).

De overleving van (invasieve) borstkanker is in verhouding tot andere vormen van kanker relatief goed. Zo is 88% van de patiënten met borstkanker na vijf jaar nog in leven, en geldt dit na tien jaar voor 79% van de patiënten. Desalniettemin overlijden ruim 3.000 vrouwen in Nederland per jaar aan borstkanker.

Alle vrouwen tussen de 50 jaar en 75 jaar worden in Nederland uitgenodigd voor deelname aan het bevolkingsonderzoek, wat bestaat uit een tweejaarlijkse mammografie. Het doel hiervan is het detecteren van borstkanker in een vroeg stadium, opdat dit leidt tot een betere prognose, tot een minder invasieve behandeling en een betere overleving. Als gevolg van deze screening is het aantal ontdekte gevallen van borstkanker, alsmede het aantal ontdekte vrouwen met DCIS, aanzienlijk gestegen. Op dit moment wordt ongeveer de helft van de invasieve borstkankers ontdekt via het bevolkingsonderzoek en dit geldt zelfs voor 70% van de patiënten met DCIS.

Biomarkers in borstkanker

De prognose van borstkankerpatiënten wordt bepaald door patiënt-gerelateerde factoren, zoals leeftijd en algehele gezondheidstoestand, en tumor-gerelateerde factoren oftewel biomarkers (meetbare indicatoren), zoals het histologisch subtype (het 'uiterlijk' van de tumor onder de microscoop), de grootte van de tumor, de graad (hoe agressief de tumor is), de hormoon- en HER2-receptor status (m.a.w. gebruikt de tumor deze receptoren en stofjes om te groeien), en het aantal lymfeklieren waarin zich tumorweefsel bevindt. Deze biomarkers worden door de patholoog bepaald, op basis van het weefselonderzoek van de tumor.

Er worden ruwweg drie borstkanker-subtypen onderscheiden op basis van de receptor status (oestrogeen (ER) en progesteron (PR) en de HER2-receptor) van de tumor. Het gaat dan om hormoon-gedreven borstkanker (ER+ of PR +), HER2-gedreven borstkanker (HER2+) en triple negatieve borstkanker (ER-, PR-, HER2-).

De behandeling van borstkanker bestaat uit vele verschillende modaliteiten, waaronder het wegsnijden van de tumor (chirurgie), het bestralen van de tumor (radiotherapie), het toedienen van celdodende medicijnen (chemotherapie), anti-hormonale medicijnen (voor hormoongevoelige borstkanker) en anti-HER2 medicijnen (doelgerichte therapie voor HER2+ borstkanker). Welke behandeling geïndiceerd is voor welke patiënt hangt, naast de algemene gezondheidsstatus, af van de hiervoor genoemde biomarkers. Dit onderstreept de cruciale rol van een hoogwaardige en uniforme beoordeling van deze biomarkers door pathologen.

In dit proefschrift hebben we ons specifiek gericht op de beoordeling van de receptor-status van de tumor (ER, PR en HER2) en het vaststellen van de histologische graad van de tumor door pathologen.

Receptoren

De bepaling van de receptor-status speelt een cruciale rol in de behandeling van borstkanker. Naast hun prognostische rol, vormen deze receptoren namelijk aangrijpingspunten voor hormoon- en anti-HER2 therapie. In de dagelijkse praktijk wordt de receptorstatus bepaald door middel van een zogenaamde immuun-histochemische analyse van het tumor weefsel (voor ER, PR en HER2) en/of *in situ* hybridisatie (ISH) (voor HER2). Het accuraat uitvoeren van deze analysen en het interpreteren van de resultaten daarvan op een betrouwbare en reproduceerbare manier, is cruciaal voor de patiënt. Vals-negatieve resultaten kunnen leiden tot het onthouden van een effectieve behandeling, terwijl vals-positieve resultaten kunnen leiden tot overbehandeling met ineffectieve (en kostbare) therapie, waarbij patiënten onnodig worden blootgesteld aan ongewenste bijwerkingen (zowel direct als op de lange termijn).

Gradering

De histologische graad is een biomarker die sterk geassocieerd is met overleving (hoe lager de graad, hoe beter de overleving). De graad van de tumor speelt daarom een belangrijke rol in de behandeling van invasieve borstkanker, bijvoorbeeld bij het selecteren van patiënten voor wie aanvullende chemotherapie en/of hormoontherapie nodig is. De graad wordt vastgesteld door pathologen aan de hand van de gemodificeerde Bloom en Richardson classificatie, waarbij zij, kijkend naar het tumorweefsel onder de microscoop (of digitaal), een score van 1-3 geven voor drie factoren. Die factoren bestaan uit 1) de mate van buisvorming (in hoeverre lijkt de tumor nog op normaal borstweefsel, wat normaal gesproken klierbuizen vormt), 2) het uiterlijk van de celkernen (kernpolymorfie) en 3) het aantal waarneembare kerndelingen (mitosen). De optelsom van de scores op de subcategorieën levert een uiteindelijke score op, die resulteert in de graad van de tumor (score 3-5 = graad 1, score 6-7 = graad 2, score 8-9 = graad 3).

Voor DCIS speelt graad op dit moment een kleinere klinische rol, maar hier gaat mogelijk verandering in komen. Op dit moment worden alle DCIS patiënten geopereerd. In het geval van een borst-besparende behandeling volgt dan bestraling en soms volgt ook nog hormoontherapie. Omdat over het algemeen wordt aangenomen dat een groot deel van de DCIS uiteindelijk niet zal uitgroeien tot een invasief carcinoom, worden veel patiënten met DCIS nu waarschijnlijk overbehandeld. Daarom wordt in vier klinische studies onderzocht of in patiënten met een laaggradig DCIS, de behandeling veilig achterwege kan worden gelaten. Daarbij wordt er vanuit gegaan dat áls een laaggradig DCIS uitgroeit tot een invasieve tumor, dit een laaggradige invasieve tumor (een minder agressieve tumor) zal zijn die goed te behandelen is. Een DCIS die geclassificeerd wordt als zijnde hooggradig blijft daarentegen behandeld worden, zoals eerder beschreven. Daarmee wordt graad in de toekomst wellicht de belangrijkste biomarker voor patiënten met DCIS.

Reproduceerbaarheid van de beoordeling van biomarkers

Accurate en reproduceerbare beoordeling van biomarkers is cruciaal voor het selecteren van de juiste behandeling voor de juiste patiënt. We weten echter uit eerdere studies, waarbij hetzelfde tumorweefsel door meerdere pathologen werd beoordeeld, dat er substantiële verschillen bestaan voor de receptorbepalingen (ER, PR en HER2) en dat de reproduceerbaarheid voor graad op zijn hoogst matig kan worden genoemd. Dit roept de vraag op of dit ook geldt voor de beoordeling van biomarkers in de dagelijkse klinische praktijk.

Het Pathologisch-Anatomisch Landelijk Geautomatiseerd Archief (PALGA)

PALGA speelt een belangrijke rol in de dagelijkse praktijk binnen de Nederlandse pathologie. Naast het beheren van alle pathologie verslagen in Nederland (sinds 1991), waarmee een enorme databron voor onderzoek wordt gevormd, faciliteert PALGA ook protocolmodules. Pathologen hoeven daardoor hun verslagen niet meer te dicteren, of zelf uit te schrijven, maar kunnen aan de hand van een zogenaamd synoptisch protocol, door middel van vinkjes te zetten, een gestandaardiseerd en compleet pathologieverslag genereren. Deze meer complete rapporten hebben bij darmkankerpatiënten al aangetoond te resulteren in verbeterde patiëntenzorg en zelfs in een verbeterde overleving. Daarnaast kan er eenvoudig onderzoek worden gedaan met grote aantallen patiënten, omdat alle variabelen worden opgeslagen op een gestandaardiseerde manier. Ruwweg 80% van de Nederlandse borstkankers wordt op dit moment aan de hand van een synoptisch protocol verslagen.

DEEL 1

De beoordeling van biomarkers in de dagelijkse praktijk

In het eerste deel van dit proefschrift hebben wij de variatie in biomarker-beoordelingen (voor zowel DCIS als het invasieve borstkanker) tussen pathologielaboratoria, en tussen pathologen binnen eenzelfde laboratorium, geëvalueerd door gebruik te maken van landelijke pathologie data uit de dagelijkse praktijk. Daarbij hebben wij gebruik gemaakt van de protocollaire pathologieverslagen.

In **hoofdstuk 2** hebben we de percentages ER, PR en HER2-positieve tumoren tussen Nederlandse pathologielaboratoria vergeleken, waarbij we hebben gecorrigeerd voor patiënt- en tumorkarakteristieken (case-mix). Hierdoor werd het mogelijk voor pathologielaboratoria om hun eigen (case-mix gecorrigeerde) percentages vergelijken met andere laboratoria. In een landelijk cohort van 33.046 patiënten met een borstkankeroperatie tussen 2013 en 2016, zagen we dat de absolute verschillen in de percentages ER, PR en HER2-positieve tumoren per laboratorium beperkt waren. Naast de bestaande externe audits (die niet alle stappen in het beoordelingsproces meenemen), kan continue monitoring en benchmarking, zoals in onze studie, mogelijk bijdragen aan het in stand houden van de huidige hoge standaard van receptorbepalingen in Nederland.

In **hoofdstuk 3** hebben we de landelijke variatie in gradering van DCIS geëvalueerd. In een cohort van 4.901 patiënten met een operatie vanwege DCIS tussen 2013 en 2016, zagen we aanzienlijke variatie tussen de pathologielaboratoria (range tot 40%) en tussen pathologen binnen individuele laboratoria (range tot 70%). Dat laatste impliceert dat zelfs binnen de laboratoria, dus tussen de verschillende pathologen binnen één laboratorium, de beoordeling van de graad niet op eenzelfde manier wordt uitgevoerd. Dat blijkt ook uit onze enquête onder pathologen, waarbij talrijke classificaties (evenals intuïtie) werden genoemd als gebruikte richtlijn voor DCIS gradering. Dit vraagt om (inter)nationale consensus over een classificatie om de reproduceerbaarheid te verhogen. Die verbetering is zeer relevant omdat de beslissing om DCIS patiënten wel of niet te behandelen in de toekomst mogelijk uitsluitend afhankelijk is van de graad, zoals eerder benoemd.

In **hoofdstuk 4** analyseerden we de landelijk variatie in gradering van invasieve borstkanker. Hiervoor gebruikten we de protocollaire pathologieverslagen van 33.043 patiënten met een operatie vanwege invasieve borstkanker tussen 2013 en 2016. Ook hier zagen we dat er aanzienlijke variatie bestaat tussen de laboratoria (range tot 30%) en tussen de pathologen binnen individuele laboratoria (range tot 50%). In lijn met eerdere studies observeerden wij de meeste variatie in het scoren van kernpolymorfie (een van de drie subcategorieën). Dit wordt mogelijk verklaard door het feit dat deze beoordeling het minst kwantitatief is, in vergelijking met het tellen van de mitosen en het beoordelen van het percentage

A

buisvorming. Daarnaast bleek dat meer dan driekwart van de patiënten een score 5 of 6, dan wel 7 of 8 had. Dit geeft aan dat zij bij slechts één punt verschil op de totaalscore al een andere graad, en daaraan gekoppeld een andere behandelindicatie zouden kunnen hebben. Het is daarom zeer waarschijnlijk dat variatie in gradering, zoals waargenomen in onze studie, leidt tot variatie in behandeling, en daardoor mogelijk de uitkomst van borstkankerpatiënten beïnvloedt. Het feit dat we ook significante verschillen vonden tussen pathologen binnen laboratoria (in vijf van de acht geanalyseerde labs), toont aan dat zelfs binnen laboratoria, gradering mogelijk niet op een vergelijkbare manier wordt uitgevoerd, hoewel, in tegenstelling tot de gradering van DCIS, één richtlijn (gemodificeerde Bloom- en Richardson-classificatie) wordt gebruikt. Gezien de klinische relevantie zijn wij ervan overtuigd dat alles in het werk moet worden gesteld om de waargenomen variatie te verminderen.

In **hoofdstuk 5** hebben we gekeken hoe belangrijk graad is in de dagelijkse praktijk. Met andere woorden, hoe vaak hangt een behandelbeslissing af van de graad die door de patholoog aan een tumor is gegeven. Daarbij hebben we gekeken naar de rol van gradering binnen de oude Nederlandse borstkanker richtlijn (geldig gedurende de studieperiode van 2013 tot en met 2016) en de huidige Nederlandse borstkanker richtlijn (geldig sinds 2019). Door het koppelen van twee landelijke databases (PALGA en het Nederlands Kankerregister), konden we de pathologie- en behandelgegevens van 30.843 patiënten met een borstkanker operatie tussen 2013 en 2016 analyseren. Deze data lieten zien dat graad een zeer belangrijke rol speelt bij behandelbeslissingen rondom aanvullende systeemtherapie. Zo zagen we dat graad in de oude richtlijn voor een kwart van de patiënten bepaalde of zij wel of geen chemotherapie zouden moeten krijgen, terwijl dit in de huidige richtlijn zelfs het geval is voor een derde van de patiënten. Voor een groot deel van de borstkanker bepaalt graad dus of zij wel of geen chemotherapie moeten krijgen. Daarbij komt dat graad, in tegenstelling tot de voorgaande richtlijn, in de huidige richtlijn ook een rol speelt in de besluitvorming rondom hormoontherapie. Voor bijna 30% van de patiënten bepaalt momenteel alleen de graad of aanvullende hormoontherapie geïndiceerd is. Daarnaast zagen wij in **hoofdstuk 5** dat de richtlijn vaak niet werd nageleefd. Dat wil zeggen, van alle patiënten met een richtlijn-indicatie voor aanvullende chemotherapie (ongeveer de helft van alle patiënten), ontving 40% in de praktijk geen chemotherapie. Dat kan deels worden verklaard door anticipatie op de nieuwe richtlijn, waarbij er over het algemeen minder vaak chemotherapie wordt gegeven (de-escalatie). Anderzijds speelt de Mammaprint (70-GS), een genexpressie test, daarbij een belangrijke rol (met name sinds 2015-2016). Deze test kan bij een specifieke subgroep van patiënten, bij wie op basis van de reguliere 'clinico-pathologische' variabelen wordt voorspeld dat zij een hoog risico hebben op terugkeer van de borstkanker, aantonen of chemotherapie daadwerkelijk geïndiceerd is. In het geval van een lage Mammaprint score kan aanvullende chemotherapie, ondanks een hoog 'klinisch'

risico, namelijk veilig achterwege worden gelaten. Of patiënten in aanmerking komen voor een Mammaprint wordt echter primair ook bepaald door de graad. Daarnaast zagen wij dat graad ook bij het 'niet-naleven' van de richtlijn een belangrijke rol heeft gespeeld. Patiënten met een tumor van een hogere graad kregen significant vaker aanvullende chemotherapie.

Hoewel de indicatie voor aanvullende chemotherapie volgens de destijds geldende richtlijn bij ongeveer 25% van de patiënten afhing van de graad, konden wij niet aantonen dat patiënten bij wie de tumor werd beoordeeld in laboratoria die aanzienlijk hoger gradeerden dan andere laboratoria, vervolgens ook vaker chemotherapie kregen. Dit komt waarschijnlijk door de hoge mate van niet-naleving van de richtlijn bij patiënten met een richtlijn-indicatie voor aanvullende chemotherapie (40%). Als de richtlijn in 100% van de gevallen zou zijn nageleefd, zou het beoordelen van meer tumoren als hooggradig, voor die 25% van de patiënten waarbij graad de doorslaggevende factor is, resulteren in een frequentere toediening van aanvullende chemotherapie. Dit benadrukt nogmaals de noodzaak om de beoordeling door pathologen zo veel mogelijk te optimaliseren, om het risico op een slechtere uitkomst van de patiënt als gevolg van onderbehandeling en onnodige kosten en bijwerkingen als gevolg van overbehandeling te minimaliseren.

DEEL 2
Initiatieven om de variatie in biomarker-beoordelingen te verminderen
De data uit de **hoofdstukken 2, 3 en 4** zijn teruggekoppeld aan de pathologielaboratoria in de vorm van zogenaamde spiegelrapporten. In deze spiegelrapporten werden de laboratorium-specifieke proporties (gecorrigeerd voor case-mix) afgezet tegen andere geanonimiseerde laboratoria en tegen het landelijk gemiddelde. Enkele laboratoria, die hiervoor zelf gegevens hebben aangeleverd, hebben daarnaast ook spiegelrapporten op patholoogniveau ontvangen.

In **hoofdstuk 6** hebben we gekeken naar het effect van de spiegelrapporten op de variatie in de gradering van DCIS. We hebben daarbij een vergelijking gemaakt tussen de variatie in DCIS gradering in het jaar vóór en het jaar na het versturen van de spiegelrapporten. Daarbij hebben we gebruik gemaakt van protocoldata van 2.934 DCIS patiënten. We constateerden een hoopgevende afname in variatie in gradering voor DCIS graad II en DCIS graad III, welke niet kon worden aangetoond voor DCIS graad I. Over het algemeen bleef de variatie in gradering tussen de laboratoria aanzienlijk. Wij vermoeden dat het gebrek aan consensus over de classificatie voor DCIS gradering hierbij een belangrijke rol speelt. Uniforme gradering begint met het gebruik van één classificatie door alle pathologen. De spiegelrapporten kunnen een nuttig hulpmiddel zijn om de variatie in gradering in de praktijk ten minste te monitoren om zo bewustzijn te creëren en de discussie over de te gebruiken classificatie te openen.

A

In **hoofdstuk 7** hebben we een vergelijkbare analyse gedaan naar het effect van de spiegelrapporten op de variatie in gradering van invasieve borstkanker. Wij zagen, in ons cohort van 16.734 patiënten, een bemoedigende afname in variatie in gradering (voor zowel graad I, II, als III). Spiegelrapporten op een routinematige basis (bijvoorbeeld jaarlijks) vormen dus een belangrijke eerste stap naar de vermindering van variatie in gradering van borstkanker. Desalniettemin blijft de algehele variatie in gradering, ook na spiegelrapporten aanzienlijk. Met het oog op de klinische consequenties, lijken andere interventies, zoals bijvoorbeeld aanvullende training van pathologen, vereist.

In **hoofdstuk 8** hebben we het effect van een e-learning module op de variatie in histologische gradering van zowel DCIS als invasieve borstkanker geanalyseerd. We hebben daarbij laten zien dat onze e-learning module, waarin pathologen en pathologen in opleiding getraind worden in het graderen van DCIS en/of invasieve borstkanker, een haalbaar en veelbelovend instrument is voor het verminderen van variatie in gradering. Dit is zeer relevant gezien de aanzienlijke variatie in gradering in de dagelijkse pathologiepraktijk zoals beschreven in **hoofdstukken 3 en 4**, rekening houdend met de belangrijke rol van graad in de klinische besluitvorming zoals o.a. beschreven in **hoofdstuk 5**.

Conclusie

In dit proefschrift hebben wij laten zien dat de landelijke variatie voor ER-, PR-, en HER2-receptoren in de dagelijkse praktijk beperkt is. De landelijke variatie in gradering van zowel DCIS als invasieve borstkanker daarentegen kan niet anders dan als substantieel worden bestempeld. Het beperken van deze variatie is belangrijk gezien de cruciale rol van gradering in de therapeutische besluitvorming. Hiervoor werden twee initiatieven gelanceerd, de laboratorium-specifieke spiegelrapporten en een e-learning module. Hoewel deze initiatieven bemoedigende resultaten opleverden, bleef de algehele variatie in gradering aanzienlijk.

Ondanks de onbetwistbare noodzaak om variatie in gradering verder te verminderen, zijn we van mening dat het ook belangrijk is om te benadrukken dat de graad van een tumor geen feit of de waarheid is, maar slechts een model of hulpmiddel, bestaande uit statistisch berekende afkapwaarden in een spectrum van histopathologische kenmerken om het verwachte tumorgedrag te classificeren. Toch worden deze cut-offs door de behandelend artsen vaak vertaald in harde criteria in de therapeutische besluitvorming (graad I geen chemotherapie, graad II wel chemotherapie). We zijn daarom van mening dat bewustzijn en begrip van de moeilijkheden van gradering onder clinici (oncologen, radiotherapeuten) cruciaal is om ook de klinische besluitvorming voor patiënten te verbeteren. Bovendien moeten deze moeilijkheden in het huidige tijdperk van gedeelde besluitvorming ('shared-decision making'), waar relevant, ook met patiënten worden besproken.

TOEKOMSTPERSPECTIEVEN

Regelmatige spiegelrapporten en grootschalig gebruik van de e-learning module

Het voortzetten van beide verbeterinitiatieven is belangrijk, niet in de laatste plaats omdat verbetering tijd kost. Wij analyseerden het effect van de eerste ronde van de spiegelrapporten. Dit was dus nog geen gesloten kwaliteitsloop, en pathologielaboratoria waren bovendien ook niet gewend aan dit soort feedback. Routinematige (bijvoorbeeld jaarlijkse) spiegelrapporten sluiten de kwaliteitsloop en geven pathologielaboratoria de kans om bekend te raken met deze rapporten, terwijl de variatie in gradering op landelijk niveau tegelijkertijd wordt gemonitord. Daarnaast kan het uitrollen van onze e-learning, bijvoorbeeld onder pathologen die deelnemen aan het landelijke bevolkingsonderzoek borstkanker, leiden tot betere standaardisatie op grotere schaal.

Andere initiatieven om hoogwaardige oncologische zorg te garanderen

Naast de twee lopende initiatieven, moeten aanvullende wegen worden verkend om oncologische zorg van hoge kwaliteit te garanderen. Het zou interessant kunnen zijn om te onderzoeken of en hoe kunstmatige intelligentie pathologen zou kunnen ondersteunen bij de beoordeling van biomarkers. Het is echter belangrijk op te merken dat dit, om te beginnen, goed geannoteerde, op consensus gebaseerde trainingsdatasets vereist. De digitale beoordelingssets die voor de e-learning zijn gebruikt, kunnen hier van waarde zijn.

Een ander initiatief is het systematisch dubbel lezen door experts of intercollegiaal overleg ('peer-consultation') in specifieke gevallen, bijvoorbeeld wanneer de graad bepaalt of adjuvante systemische therapie is geïndiceerd. Het Nederlandse Pathology Image Exchange (PIE) project kan hierbij een nuttig platform bieden voor snelle beoordeling. Aangezien we echter hebben aangetoond dat de graad doorslaggevend is bij een substantieel aantal patiënten, kan de uitvoerbaarheid van systematisch dubbel lezen worden betwist. Het toekennen van een passende vergoeding (voor het dubbel-lezen) lijkt in ieder geval haalbaar, omdat de kosten van dubbel lezen veel lager zullen zijn dan bijvoorbeeld het inzetten van een gen-expressie test (een Mammaprint kost ongeveer €2.500).

Gen-expressie testen, zoals de Mammaprint, hebben een duidelijk toegevoegde waarde in de klinische besluitvorming. Het is echter belangrijk om hierbij te realiseren dat het inzetten van deze test alleen is geïndiceerd voor een specifieke groep borstkankerpatiënten. Daarnaast wordt de inzet van de Mammaprint binnen die groep in de eerste plaats ook bepaald door tumorkenmerken, zoals de graad.

A

Gespecialiseerde borst-pathologen?

Een andere discussie die door ons onderzoek kan worden aangewakkerd, is of het wenselijk is dat sommige laboratoria slechts enkele gevallen van DCIS of invasieve borstkanker per jaar beoordelen. Gezien de complexiteit en het klinische belang van gradering, zou men kunnen stellen dat dit alleen mag worden uitgevoerd door deskundige borstpathologen. Wij willen echter benadrukken dat onze gegevens niet geschikt zijn voor dit type analyse, omdat wij alleen protocollaire pathologierapporten hebben gebruikt, waardoor ongeveer 20% van de Nederlandse gevallen van borstkanker zijn geëxcludeerd. Daarom kan het absolute aantal gevallen per laboratorium niet worden afgeleid uit onze gegevens. Het is echter een interessante discussie die op basis van de juiste gegevens zou moeten worden gevoerd.

Acknowledgements / dankwoord

Lieve familie, vrienden en collega's,

Aan het begin van dit traject, ruim drie jaar geleden, was het niet mijn bedoeling om te gaan promoveren, maar begon ik 'gewoon' aan een mooi, relevant en toch vooral ook tijdelijk project. Daar dachten sommigen (gelukkig!) anders over, en daar ben ik nu ontzettend blij om. Ik heb in de afgelopen jaren de kans gekregen om de wereld van de wetenschap te ontdekken. Een wereld die veel leuker bleek dan ik dacht. Zodanig leuk dat het wetenschappelijk avontuur niet gaat eindigen na vandaag. Terugkijkend kan ik niet anders zeggen dan dat ik heb genoten van mijn 'verlengde studententijd'. Een tijd waarin ik veel vrijheid en mooie kansen heb gekregen. Daarvoor ben ik een hoop mensen dank verschuldigd. Dat gaat vast niet allemaal lukken in dit dankwoord, maar ik ga toch een poging wagen.

Geachte **prof. dr. van Diest**, beste **Paul**, met 'food for thought' vertrok ik naar de Alpen en na een tweede Skype-gesprek over een hele matige wifi-verbinding durfde jij het aan om mij te laten starten op het project 'mammagradering'. Jij zag het promotietraject al voor je toen ik dat zelf nog lang niet zag. Allereerst veel dank voor dat vertrouwen en die visie! Je neemt jouw taak als promotor zeer serieus en geeft daar op jouw eigen, unieke, enthousiaste en motiverende wijze invulling aan. Je staat altijd klaar vóór en vierkant áchter jouw promovendi (en dat zijn er inmiddels bijna 100!). Daarnaast was er vaak tijd voor een goed gesprek. Ik houd mij overigens nog altijd aanbevolen voor een (ski)congres in de Rockies!

Geachte **prof. dr. van der Wall**, beste **Elsken** en beste 'Promotor van het jaar 2019'! Je bent voor mij een voorbeeld als mens, als dokter én als wetenschapper. Met altijd aandacht voor de 'persoon achter de promovendus' en voor alle andere belangrijke zaken in het leven. Ik heb dat ontzettend gewaardeerd. Jij zet jouw promovendi voorop en als het even kan met het grootste plezier, en vol trots, in de spotlight. Na onze wekelijkse meetings ging ik altijd met frisse moed weer door. Jouw warme betrokkenheid, steun en vertrouwen zijn voor mij van grote waarde geweest.

Geachte **dr. Deckers**, beste **Ivette**, zonder jou was ik waarschijnlijk in zeven statistische sloten tegelijk gelopen. Jouw kritische blik en begeleiding, en vooral ook jouw eindeloze geduld, hebben mij ontzettend veel geleerd en ver gebracht. Ik ben trots op de projecten die wij samen hebben neergezet en ik vind het heel gaaf dat jij de spiegelinformatie op grote schaal uitrolt bij PALGA. Veel succes daarbij en ik hoop dat dit niet het einde is van onze samenwerking!

Geachte **dr. Baas**, beste **Inge**, toen ik in 2016 bij jou in het Meander aanklopte voor een wetenschapsstage had ik nooit kunnen bedenken dat dit daarvan het gevolg zou zijn. Gelukkig stimuleerde jij mij om eens met Paul te gaan praten en er gewoon voor te gaan. Gedurende mijn promotietraject keerde je terug naar de academie en raakte je opnieuw betrokken bij mijn onderzoek. Met jou als copromotor is het cirkeltje rond! Veel dank voor jouw begeleiding, vertrouwen en support (tot in Nice aan toe!). Ik hoop dat we nog lang blijven samenwerken en dat we nog eens samen op de piste staan.

Geachte leden van de leescommissie: **prof. dr. Borel Rinkes, prof. dr. van Gils, prof. dr. Linn, dr. Sonke en prof. dr. van de Vijver**. Hartelijk dank voor het zitting nemen in mijn leescommissie en het beoordelen van mijn proefschrift.

Geachte **opponenten**, hartelijk dank voor het bijwonen van mijn verdediging en het voeren van de oppositie.

Lieve **Willy**, wat ook er ook geregeld moest worden, ik kon altijd bij jou terecht. Van het regelen van afspraken in de overvolle agenda van Paul tot het boeken van vliegtickets en hotels, maar het leukste was toch wel het organiseren van het P75 feest! Veel dank voor alles!

Lieve **PRL'ers,** jullie hebben mijn PhD-tijd zóveel leuker gemaakt. Na mijn eerste jaar op de 4^e verdieping zag ik op tegen mijn verhuizing naar het 'kippenhok' op de 3^e. Dat was een inschattingsfout, want met *noise cancelling headphones* was het daar prima uit te houden. Bovendien steeg de gezelligheid op logaritmische schaal. De dagelijkse Pitstop-koffie-pauzes, de vele borrels, lunchruns, escaperooms, de voetbalpoules, en als kers op de taart ons gezamenlijke congres in Nice en natuurlijk de PRL-skitrip! Ik ga jullie ontzettend missen en kijk uit naar de post-PRL borrels, feestjes en ski-trip(s)!

Emma en Bregje, veel dank dat jullie naast mij staan als paranimf. **Emma**, onze bridgende PRL-veteraan. Ik ken niemand met meer enthousiasme voor R en zoveel bijkomend geduld om dat aan anderen uit te leggen. Wat zijn de Light's kappa's mooi geworden! Maar bovenal en vooral heel veel dank voor alle gezelligheid. **Bregje,** begonnen als PRL'er en inmiddels patholoog in spé. Onze werkplekken werden gescheiden door een bierkrat, maar dat heeft ons niet weerhouden van veel sparren over inter-laboratorium variatie en alle andere boeiende dingen in het leven met een kop thee. Dank voor alles! **Quirine**, mijn San Antonio-maatje. Dank voor alle gezelligheid ook op Nederlandse én Oostenrijkse bodem. En je mag uiteraard altijd mee op de post-PRL ski-trip, maar nog wel even oefenen met nagelen! **Cathy**, dank ook voor jouw bijdrage aan de San Antonio avonturen. En sorry dat ik uiteindelijk toch jouw onderzoek niet ben gaan doen! **Natalie**, waar zou de afdeling zijn

A

zonder jou? Ik heb veel respect voor alles wat jij doet en hoop ooit voor jou in deze zaal in het publiek te mogen zitten! **Folkert**, Folkie, pater familias van de PRL. Altijd in voor een praatje, goede fietsadviezen en natuurlijk strava-kudos. Ik ga je grote glimlach missen! **Lilian**, onze autofanaat. Jij had alles behalve een makkelijke tijd, maar wat is het goed om te zien hoe sterk jij je staande weet te houden en wat was je goed in Nice! **Aernoud**, aanstichter van de lunchruns, voetbalpoules en om een uur of vier "is het al tijd voor bier?" Veel succes met de afronding van jouw proefschrift. **Wenzel,** onbetwist de PRL'er met het hoogste werktempo, maar nooit afwezig bij een borrel! **Selena**, mijn maatje tijdens mijn 1e PhD-jaar. Jij maakte daarna een dappere keuze en koos een ander pad. Ik vind het mooi om te zien hoe jij jouw droom volgt. Veel dank ook voor alle personal coaching tijdens onze hardlooprondjes. **Sebastiaan**, onze sportfanaat. De uitnodiging om mee te gaan op de ski-trip blijft gewoon staan! **Gwen**, de enige echte chef pistekwaliteit. Veel succes met jouw laatste loodjes. **Susana**, ongekend organisatietalent en sluipschutter van het lasergamen. Ook jij veel succes met het afronden van jouw proefschrift. **Liling** and **Shuang**, thank you for holding out with all of us at the PRL and for hosting that Chinese dinner. **Betzabel, Willem en Floris**, PRL'ers van de 'cave' op de 2e, hou vol daar! En tenslotte natuurlijk de nieuwste PRL'er **Rachel**, jij heel veel succes met het project 'prostaatgradering'!

Lieve **Willemijn, Mimount en Birgit**, wat begon met een keer koffie drinken omdat we allemaal promotieonderzoek deden, is geëindigd in een hechte groep. Jullie nuchterheid en 'trijf'-radar zijn echt ongeëvenaard. Promoveren is soms ook ventileren en daarvoor had ik mij geen fijner clubje mensen kunnen wensen. Veel dank ook voor het proofreaden! Lieve **Willemijn**, je compenseert je gebrek aan richtingsgevoel en coördinatie gelukkig met een ongekend talent voor spelling. Over twee weken staan we hier (eindelijk) voor jou! Lieve **Mimount**, yes we can! Ik bewonder jouw kijk op het leven en de manier waarop jij het afgelopen jaar hebt doorstaan. Ik hoop dat we deze koffie-bitch sessies nog lang zullen voortzetten!

Dear **Tyger Tigers**, safe to say we came a long way. All doctors now! Baie dankie for all the good times and see you in Cape Town for the 2024-edition!

Liebe **Max und Waltraud**, Danke für ein warmes Zuhause weg von Zuhause seit 20 Jahren. Es ist immer wieder eine Große Freude.

Lieve **ski-familie**. **Simone**, de wegkapitein. Niet bij te houden op ski's, snowboard of de racefiets. Dank voor alle ski-, après-ski-, en wielren-avonturen! **Elise**, eigenlijk onze enige echte ski-pro. Succes met jouw mannenhuishouden! **Carola**, mijn 'Flachau-moeder'. Dank voor alle gezellige skitochten, goede gesprekken en vele etentjes. En dank dat je mij naar de SHOS hebt gehaald. **Juffies en meesters uit Nieuwegein,** in het bijzonder **Thijs, Jip, Erwin,**

Debbie, Vivian, Hans en Gerda, dank voor alle gezelligheid op en naast de borstelbaan en niet te vergeten in Breckenridge, Ischgl en Myrkdalen! Lieve **Daniël en Marjolein**, dank dat ik altijd welkom ben bij de SSCU-Family, op de trainingen in Hintertux en natuurlijk in Soest.

Lieve **Jennifer, Emma, Olaf, Wouter, Pepijn, Lars, Hylke, Timion, Karin en Birgit,** de raddraaiers- en vensterbank-clan van het OZL is toch maar goed terecht gekomen. Fijn dat we nog altijd samen komen!

Lieve **Karin**, vriendinnetjes sinds de brugklas en gelukkig is die band er nog steeds. Inmiddels geen huisgenoten meer, maar wat is het fijn om altijd bij jou terecht te kunnen. Dank voor wie jij bent!

Lieve **Eline**, mijn grote 'kleine' zusje, die voor een semester naar Glasgow vertrok. Je bent uitgegroeid tot iemand die sterk in haar schoenen staat en nu zelfs een tweede master gaat doen. Welke richting jij uiteindelijk ook kiest, ik weet zeker dat jij op je plekje terecht komt. En wat heb ik genoten van onze roadtrip door de Schotse Hooglanden! Lieve **Birgit**, lieve twinnie. Nog een zus die sterk in haar schoenen staat en waar ik heel trots op ben. Na bijna 30 jaar gaan de wegen (eindelijk) een beetje scheiden, maar beste maatjes blijven we altijd. You jump, I jump! En lieve **zusjes** wat was met zijn drietjes rondcrossen door Nieuw Zeeland onvergetelijk.

Lieve **Pap** en **Mam**, allerliefste 'bazen van het bospad'. Jullie zijn mijn grote fundament. Zonder jullie onvoorwaardelijke steun had ik hier niet gestaan. Hard werken en klaarstaan voor een ander, maar vooral ook nú léven en daarvan genieten. Dat is wat jullie ons altijd hebben geleerd. En daarbij staan jullie echt altijd klaar voor ons. Ik vind het mooi om te zien hoe jullie zelf ook genieten van het leven en ik hoop nog héél lang van jullie beiden te mogen genieten. En wat ben ik blij dat we hier vandaag met zijn allen staan!

A

Curriculum Vitae

Carmen van Dooijeweert was born on May 9, 1991 in Zeist, the Netherlands, to Emmy van Ommen and Rick van Dooijeweert, followed twenty-eight minutes later by her twin sister, Birgit, and five years later by Eline.

After obtaining her Gymnasium diploma at the Openbaar Zeister Lyceum in 2009, she shortly studied Political Science in Nijmegen, before spending the winter of 2009-2010 in Flachau, Austria, as a ski-instructor at Skischule Fun & Pro (Rupert Pichler). In 2010 she started medical school at Utrecht University, which included a rotation at Tygerberg Hospital in Cape Town, South Africa, during her 5th year of studies.

After graduating from medical school in 2016, with a senior rotation at the Department of Internal Medicine of the UMC Utrecht and a scientific internship at the Department of Medical Oncology of the Meander Medical Center (supervised by dr. Inge Baas), Carmen was introduced to prof. dr. Paul van Diest. After another ski-season in Austria, Carmen started her PhD-track in April 2017, which, in addition to Paul, was supervised by prof. dr. Elsken van der Wall, dr. Ivette Deckers, and Inge herself.

Research from this thesis was presented at several (inter)national conferences, among which the San Antonio Breast Cancer Symposium (Texas, USA) (2x), and the European Congress of Pathology (Nice, France), where Carmen was awarded best oral presentation of the breast session, resulting in a plenary presentation during the closing ceremony.

As of August 2020 Carmen has started working as a resident not-in-training at the department of Internal Medicine of the Meander Medical Center in Amersfoort.

A

List of Publications

Thesis

van Dooijeweert C, Deckers IAG, Baas IO, van der Wall E, van Diest PJ. Hormone- and HER2-receptor assessment in 33,046 breast cancer patients: a nationwide comparison of positivity rates between pathology laboratories in the Netherlands.
Breast Cancer Res Treat. 2019; 175(2): 487-497.

van Dooijeweert C, van Diest PJ, Willems SM, Kuijpers CCHJ, Overbeek LIH, Deckers IAG. Significant inter- and intra-laboratory variation in grading of ductal carcinoma in situ of the breast: a nationwide study of 4901 patients in the Netherlands.
Breast Cancer Res Treat. 2019; 174(2): 479-488.

van Dooijeweert C, van Diest PJ, Willems SM, Kuijpers CCHJ, van der Wall E, Overbeek LIH, Deckers IAG. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: A nationwide study of 33,043 patients in the Netherlands.
Int J Cancer. 2020; 146(3): 769-780.

van Dooijeweert C, van Diest PJ, Baas IO, van der Wall E, Deckers IAG. Grading variation in 2,934 patients with ductal carcinoma in situ of the breast: the effect of laboratory- and pathologist-specific feedback reports.
Diagn Pathol. 2020;15(52): doi: 10.1186/s13000-020-00970-8 (Epub ahead of print)

van Dooijeweert C, van Diest PJ, Baas IO, van der Wall E, Deckers IAG. 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: doi: 10.1136/jclinpath-2019-206362 (Epub ahead of print)

van Dooijeweert C, Deckers IAG, de Ruiter EJ, ter Hoeve ND, Vreuls CPH, van der Wall E, van Diest PJ. The effect of an e-learning on grading variation of (pre)malignant breast lesions.
Mod Pathol. 2020: doi: 10.1038/s41379-020-0556-6 (Epub ahead of print)

van Dooijeweert C, Baas IO, Deckers IAG, Siesling S, van Diest PJ, van der Wall E. The increasing relevance of histologic grading in tailoring adjuvant systemic therapy in 30,843 breast cancer patients.
Manuscript submitted

Other

van Dooijeweert C, van der Wall E, Baas IO. Chemotherapy-induced febrile neutropenia: primary G-CSF prophylaxis indicated during docetaxel cycles.
Neth J Med. 2019; 77(9):310-316.

Suelmann BBM, **van Dooijeweert C**, van der Wall E, Zweemer R, Linn S, van Diest PJ. The aggressive histopathologic profile of pregnancy associated breast cancer (PABC); an analysis of the nationwide Dutch Pathology Registry.
Manuscript submitted

Koomen BM, Voorham QJM, Epskamp-Kuijpers CCHJ, **van Dooijeweert C**, van Lindert A, Deckers IAG, Willems SM. Considerable interlaboratory variation in PD-L1 positivity in a nationwide cohort of non-small cell lung cancer patients.
Manuscript submitted

A