Initial interpretation scores of screening mammograms and cancer detection in BreastScreen Norway

Abstract

Purpose: To explore the association between radiologists' interpretation scores, early performance measures and cumulative reading volume in mammographic screening.

Method: We analyzed 1,689,731 screening examinations (3,379,462 breasts) from BreastScreen Norway 2012-2020, all breasts scored 1-5 by two independent radiologists. Score 1 was considered negative/benign and score ≥2 positive in this scoring system. We performed descriptive analyses of recall, screen-detected cancer, positive predictive value (PPV) 1, mammographic features and histopathological characteristics by breast-based interpretation scores, and cumulative reading volume by examination-based interpretation scores.

Results: Counting breasts and not women, 3.9% (132,570/3,379,462) had a score of ≥2 by one or both radiologists. Of these, 84.8% (112,440/132,570) were given a maximum score 2. Total recall rate was 1.6% (53,735/3,379,462), 69.3% (37,220/53,735) given maximum score 2. Among the 0.3% (9733/3,379,462) diagnosed with screen-detected cancer, 34.6% (3369/9733) had maximum score 3. The percentages of recall, screen-detected cancer and PPV-1 increased by increasing the sum of scores assigned by two radiologists ($p<0.001$ for trend). Higher proportions of masses were observed among recalls and screen-detected cancers with low scores, and higher proportions of spiculated masses were observed for high scores ($p<0.001$). Proportions of invasive carcinoma, histological grade 3 and lymph node positive tumors were higher for high versus low scores ($p<0.001$). The proportion of examinations scored 1 increased by cumulative reading volume.

Conclusions:
We observed higher rates of recall and screen-detected cancer and less favorable histopathological tumor characteristics for high versus low interpretation scores. However, a considerable number of recalls and screen-detected cancers had low interpretation scores.

Highlights: (3-5 bullet points, maximum 85 characters including spaces, per bullet point)

- The majority of positive interpretations and recalls were given a low interpretation score of 2.
- A considerable proportion of screen-detected cancers were given a low interpretation score.
- Less favorable histopathological tumor characteristics were observed after a high score.
- Less experienced readers more frequently scored examinations positive.

Keywords

breast; neoplasms; mass screening; mammography; female

Abbreviations

AI – Artificial Intelligence

PPV-1 – Positive Predictive Value 1

BI-RADS – Breast Imaging – Reporting and Data System

DCIS – Ductal Carcinoma in situ

NST – No special type

HER 2 – Human Epidermal Growth Factor 2

IQR – Interquartile range

SD – Standard Deviation

Introduction

Breast cancer is a major cause of female cancer death in Norway and worldwide [1, 2], and health authorities recommend screening with mammography as secondary prevention to detect cancer in an early stage and reduce mortality from the disease [1, 3].

Screening is aimed at identifying suspicious abnormalities, leading to further assessment including supplementary imaging and needle biopsies if indicated. The radiologists' interpretation score implying likelihood of malignancy on the screening mammogram has an impact on the subsequent screening process. A low or high score may influence the decision in the setting of arbitration/consensus, as well as prioritization and procedures of the assessment. In turn, these factors may have an impact on program sensitivity and specificity.

Further, we are facing a possible paradigm shift in the interpretation of screening mammograms associated with implementation of artificial intelligence (AI). Use of AI in the screen-reading procedure has shown promising results in both retrospective and prospective studies, whether used as decision support for the radiologists, triaging or as an independent reader [4-8]. To better understand and utilize AI, we need to know how the radiologists score screening mammograms and what mammographic features are associated with the various scores.

Studies regarding interpretation scores and their association with early performance measures as recall, cancer detection, positive predictive value (PPV), mammographic and histopathological characteristics, are sparse. Further, it has been demonstrated that radiologists' reading volume influences their screening performance [9, 10], but do high volume readers score screening examinations differently than low volume readers? We took advantage of the database in BreastScreen Norway to investigate these issues. The aim of this observational cohort study was to explore the association between combinations of radiologists' interpretation scores in independent double reading and early performance measures, as well as the associations between interpretation scores and cumulative reading volume.

Materials and methods

This study had a legal basis in accordance with Articles 6 (1) (e) and 9 (2) (j) of the GDPR. The data was disclosed with a legal basis in the Cancer Registry Regulations section 3-1 and the Personal Health Filing System Act section 19 a to 19 h. [11].

BreastScreen Norway

BreastScreen Norway is an organized screening program for breast cancer inviting all women aged 50-69 to biennial two-view mammography screening. All examinations are independently interpreted by two breast radiologists, scoring each breast on a scale 1 to 5. All examinations given a score ≥2 to one or both breasts by one or both radiologists are discussed in a consensus meeting with at least two radiologists to decide whether to recall for further assessment. Each breast is considered separately at interpretation and consensus. The scoring system differs from the American College of Radiology's Breast Imaging – Reporting and Data System (BI-RADS) [12]. Score 1 indicates normal/benign findings (equals BI-RADS categories 1 and 2) and no consensus meeting or recall is needed. A score 2 indicates a probably benign finding but entails discussion in a consensus meeting to decide whether to recall. If the woman is recalled and the finding persists after further imaging at recall assessment, a needle biopsy is needed. Short-term follow-up is generally not used in our screening program. Score 3 indicates intermediate suspicion (50/50 benign/malignant finding) and should in general lead to a recall and concordant needle biopsy. Score 4 (probably malignant) and score 5 (malignant) mostly resemble BI-RADS categories 4C and 5, should always lead to a recall, and a representative needle biopsy or excision is required. According to the screening program's Quality Manual, all readers should interpret at least 4000 screening examinations every year to maintain screen-reading competence [13].

Study population

Women screened during the period 1.1.2012-31.12.2020 were included in the study population. Women participating in the screening program have their right to refuse permanent storage of data from their normal screening examinations [14]. Data from these women (1.4% of the invited) were not included in the study population.

The study population included 675,793 women, accounting for 1,901,337 screening examinations (Figure 1). We excluded screening examinations performed as part of scientific studies [15, 16] (n=180,557), after diagnosis of breast cancer (n=23,164), with recall due to technical reasons (n=1192) or self-reported symptoms (n=4354), without independent double reading (n=2165), or registered with a recall despite negative interpretation scores by both readers for both breasts (n=174). This left 1,689,731 screening examinations of 3,379,462 breasts from 649,655 women for the study sample. As digital breast tomosynthesis has only been performed as part of scientific studies in BreastScreen Norway, all mammograms from the included screening examinations were standard 2D full-film digital mammograms.

Definition of measures

All interpretations were considered independent regardless of reader and breast. Except examination-based analyses of reading volume, all analyses were breast-based with the combination of interpretation scores (score 1-5) given by the two readers for each breast as the unit of analysis. A breast-based approach was chosen as radiologists score each breast separately at screen-reading, giving the most robust results regarding analyses of concordance, mammographic findings, and histopathological characteristics. Maximum score was defined as the highest score given by the two radiologists for each breast, e.g., maximum score was 5 if radiologist A scored 5 and radiologist B scored 2 (Figure 2). Score 1 by both readers was defined as concordant negative, score 1 by one reader and ≥2 by the other as discordant, and score ≥2 by both as concordant positive interpretation (Figure 2).

We defined recall as further assessment after positive screening interpretation and consensus, and screen-detected cancer as invasive cancer and ductal carcinoma in situ (DCIS) diagnosed after recall assessment. Interval cancer was breast cancer diagnosed within 24 months after a negative screening examination, or 6-24 months after a false positive screening result. Positive predictive value 1 (PPV-1) was the percentage of screen-detected cancer diagnosed among recalled women. Mammographic features were described as mass, spiculated mass, distortion, asymmetry, mass with calcifications or calcifications alone, in accordance with the classification in the screening database of BreastScreen Norway [14]. We classified histopathological type as DCIS, invasive carcinoma of no special type (NST), invasive lobular carcinoma and other invasive carcinomas. For invasive cancer, we analyzed median tumor diameter, histological grade (1-3), lymph node status (positive/negative) and molecular subtypes based on immunohistochemistry (Luminal A like, Luminal B like human epidermal growth factor (HER) 2 negative, Luminal B like HER2 positive, HER2 positive and triple negative) [17].

Cumulative reading volume was defined as the total number of readings per radiologist, stratified in groups (<5000, 5000-9999, 10,000-19,999; 20,000-29,999; 30,000-39,999; 40,000-49,999; 50,000-99,999; ≥100,000). Thus, the same radiologist might contribute to more than one group; the first 4,999 readings were included in the first group, the next 5,000 readings in the next group etc.

Statistical analyses

We performed descriptive analyses of recall, screen-detected and interval cancer, PPV-1, mammographic features and histopathological characteristics by breast-based interpretation scores, and cumulative reading volume by examination-based interpretation scores (highest interpretation score per radiologist per examination). We presented categorical data as numbers and percentages, PPV-1 as percentage and tumor diameter (mm) as median with an interquartile range (IQR). Total and annual interpretation volumes were presented as means with standard deviation (SD) and medians with IQR. We tested for statistical significance using bivariate tests with a significance level of 0.05. All analyses were performed using Stata version 17.0 for Windows (StataCorp, TX, USA).

Results

Early performance measures

Counting breasts and not women, 3.9% (132,570/3,379,462) had a discordant or concordant positive interpretation, 1.6% (53,735/3,379,462) were recalled for further assessment, 0.3% (9733/3,379,462) were diagnosed with screen-detected cancer, and 0.09% (3200/3,379,462) with interval cancer (Table 1).

Both radiologists scored 96.1% (3,246,892/3,379,462) of all breasts concordant negative (Table 1). A maximum score 2 was given to 3.3% (112,440/3,379,462) of all breasts, constituting 84.8% (112,440/132,570) of breasts with discordant or concordant positive score. Only 0.05% (1625/3,379,462) of all and 1.2% (1625/132,570) of discordant/concordant positives had a maximum score 5. Among recalled, 69.3% (37,220/53,735) of the breasts had a maximum score 2 and 3.0% (1618/53,735) a maximum score 5.

For screen-detected cancer, 24.1% (2341/9733) of the breasts had a maximum score 2, 34.6% (3369/9733) maximum score 3, 23.6% (2296/9733) maximum score 4, and 16.1% (1531/9537) maximum score 5 (Table 1). Two percent (196/9733) of the screen-detected cancers were scored concordant negative, diagnosed after recall due to positive interpretation of the opposite breast. Among these, 54.1% (106/196) had bilateral screen-detected cancer. The vast majority (86.0%, 2753/3200) of interval cancers were diagnosed after concordant negative interpretation at screening prior to diagnosis.

For discordant scores, recall ranged from 23.7% (21,632 /91,190, score 1+2) to 94.4% (84/89, score 1+5). For concordant positive scores, recall ranged from 73.4% (15,588/21,250, score 2+2) to 100.0% (580/580, score 5+5) (Table 2).
The rate of screen-detected cancer among discordant scores ranged from 1.5% (1328/91,190, score 1+2) to 66.3% (59/89, score 1+5). For concordant positive scores, the rate ranged from 4.8% (1013/21,250, score 2+2) to 98.6% (572/580, score 5+5). PPV-1 ranged from 6.1% (score 1+2) to 70.2% (score 1+5) for discordant, and from 6.5% (score 2+2) to 98.6% (score 5+5) for concordant positive scores (Table 2). The percentages of recall, screen-detected cancer and PPV-1 increased by increasing the sum of scores assigned by 2 radiologists ($p<0.001$ for trend).

Mammographic features

Mass was the most frequent mammographic feature among recalls, 44.0% of breasts (17,136 /38,925) (Table 3). Of these, 83.1% (14,240/17,136) had a maximum score of 2, constituting 55.1% (14,240/25,852) of all recalls with maximum score 2 (Figure 3A). Asymmetry was the second most frequent feature among recalls, 22.7% (8845/38,925), of which 70.9% (6274/8845) had maximum score 2, constituting 24.3% (6274/25,852) of all recalls with maximum score 2. The highest proportions of recalls due to spiculated masses were observed in breasts with a maximum score of 4 (41.6%, 1089/2619) or 5 (55.1%, 816/1480).

The most frequent mammographic feature for screen-detected cancers was spiculated mass, 38.3% (3383/8822) (Table 3 and Figure 3B). Of these, 30.3% (1025/3383) had maximum score 4 and 23.8%, (804/3383) maximum score 5. The second most frequent feature was calcifications alone, 23.9% (2106/8822). Of these, 27.8% (585/2106) had a maximum score 2 and 45.9% (967/2106) maximum score 3.

Mammographic features stratified by all score combinations for recalls and screen-detected cancer are shown in Appendix A.1.

Histopathological characteristics

The proportion of DCIS was highest for maximum score 2 (23.4%, 547/2341) and 3 (21.6%, 729/3369); the proportion of invasive carcinoma NST was highest for maximum score 4 (75.5%, 1733/2296) and 5 (86.0%, 1316/1531) (Table 4). The highest proportion of invasive lobular carcinoma (19.4%, 38/196) was observed after concordant negative interpretation, recalled due to a positive score of the contralateral breast. Of these, 57.9% (22/38) had bilateral screen-detected cancer.

The proportion of histological grade 3 invasive tumors ranged from 16.6% (293/1768, maximum score 2) to 29.0% (418/1440, maximum score 5) (Table 4). The proportion of lymph node positive disease ranged from 16.3% (287/1763, maximum score 2) to 30.2% (430/1424, maximum score 5). Luminal A like molecular subtype ranged from 56.7% (813/1435, maximum score 5) to 62.1% (1087/1750, maximum score 2). Histopathological characteristics stratified by all score combinations are shown in Appendix A.2.

Reading volume and interpretation scores

During the study period 2012-2020, 174 radiologists performed screen reading. Mean reading volume was 46,487 (SD: 49,399) and median volume 29,196 (IQR: 6304-74,959). Mean and median annual reading volumes were 4423 (SD: 3569) and 3991 (IQR: 1563-6237), respectively.

The proportion of examinations scored 1 increased by cumulative reading volume, and the proportion of examinations scored ≥2 decreased accordingly (Table 5). Readers with a cumulative volume of <5000 examinations scored 6.0% of the examinations 2, this proportion declined gradually with increasing volume. The proportion of screen-detected cancers scored 5 ranged from 8.7% for cumulative reading volume <5000 examinations to 17.4% for ≥ 100,000 (Table 5).

Discussion

In this retrospective, breast-based observational cohort study, 96.1 % of all 3,379,462 breasts in 1,689,731 women were interpreted concordant negative. Further, 84.8% of the breasts with discordant or concordant positive interpretation were given a maximum score 2. The lowest recall rate, 23.7%, was observed after a discordant score of 1+2, while the highest recall rates, ≥89.6%, were found after a maximum score of 4 or 5. Histopathological characteristics were less favorable in screen-detected cancers with a discordant or concordant positive high score. The number of examinations scored ≥2 decreased by increasing reading volume.

Score 2 or higher in BreastScreen Norway entails discussion in a consensus meeting to decide whether to recall. The majority (66.9%) of cases with maximum score 2 were not recalled for further assessment. Further, the low PPV-1 (≤ 6.5%) illustrated a low potential for malignancy in score 2, supported by the high frequencies of masses and asymmetries, features commonly associated with benign lesions [12, 18]. Still, the absolute number of cancers was higher for maximum score 2 and 3 compared to higher scores. Score 4 or 5 is considered probably malignant or malignant, as reflected by high rates of recall and screen-detected cancer, and high PPV-1. Recall, cancer detection and PPV-1 were also high in discordant interpretations (1+4 and 1+5), indicating that discordance was caused by one reader missing the suspicious findings/cancer rather than the other incorrectly giving a high score.
The majority of interval cancers, 86%, were scored negative by both readers at screening prior to diagnosis. An earlier study from BreastScreen Norway demonstrated that about 38% of interval cancers with positive interpretation at screening prior to diagnosis were recalled [19], and another study demonstrated that 43% of interval cancers recalled at screening prior to diagnosis were recalled due to findings at the later cancer site [20].

Less favorable histopathological characteristics for screen-detected cancers with high versus low scores may relate to a more confident interpretation with larger tumor diameter, and to large tumors being more frequently associated with high histological grade and lymph node positive disease [21]. The larger proportion of DCIS with lower scores is compatible with the lower scores for calcifications alone, as the main mammographic manifestation of DCIS is calcifications [22]. However, as demonstrated in other studies, invasive cancers presenting as calcifications, in particular casting calcifications and masses with calcifications, are associated with poorer survival compared with spiculated masses [23, 24]. Thus, a low interpretation score is not necessarily indicative of low-risk findings. We observed the largest proportion of invasive lobular carcinoma in cancers with concordant negative interpretation. Lobular carcinomas represent a diagnostic challenge due to a diffuse growth pattern, and are, in line with our results, more frequently mammographically occult cancers [25, 26].

The proportion of positive scores declined by the radiologists' cumulative reading volume, mainly due to higher proportion of examinations scored 2 by low volume readers. An examination scored 2 can be dismissed at consensus without discussion, which might leave this score rather "safe" without any obligations. Further, the "filtration effect" of a consensus meeting may induce a lower threshold for positive scores among inexperienced readers. The proportion of screen-detected cancers scored 5 was doubled for cumulative volume ≥100,000 versus <5000. This may illustrate that radiologists considered a score of 5 quite definite ("malignant"), associated with more confidence and experience.

The large proportion of examinations with score 2 remains a challenge, as the screening program aims to ensure high specificity by recalling the "correct" women with cancer. Organized training, real-data learning sets and testing with feedback to the readers about their performance might be valuable tools to accomplish high specificity [27, 28], and is also welcomed by the screen-readers [29]. Further, artificial intelligence (AI) has proven to perform on par with radiologists in screen reading and is expected to become part of the screening procedure in the future [5-8]. Integration of AI in the screening workflow may include triaging, decision support, or even replacement of radiologist(s). Gaining experience on AI's potential to lower rates of recall and false positives and thus increase specificity while maintaining or increasing sensitivity is crucial for the future integration. Moreover, alignment of AI algorithms with the radiologists' interpretation scores and radiologists' interaction with AI results are important for screening logistics.

Strengths of this study were the large study sample and completeness of data. Further, as the analyses were breast-based, we ensured that registered scores were associated with the correct findings. The study included screening examinations from 16 different breast centers, interpretations by 174 radiologists over 10 years, resulting in a heterogeneous study population. More recent evidence on the association between mammographic features and breast cancer survival might affect the choice of interpretation scores [30, 31]. However, the heterogeneity and long study period may also be considered to ensure robustness of the results. Missing data may be considered a limitation. The relatively large number of recalls with no available data on mammographic features, particularly among breasts with low scores, is mainly explained by normal findings after recall assessment, as mammographic features are registered in the database after supplementary imaging. Thus, this is considered as a logic consequence of a negative finding. The number of examinations with missing data is lower among screen-detected cancers but may still have impact on the results. However, we have no information suggesting that data are missing in a non-random way.

To conclude, we observed differences in early performance measures for low versus high interpretation scores, and association between radiologists' interpretation scores and cumulative reading volume. Examinations with a low, but positive initial interpretation score represent a challenge in a population-based screening program with independent double reading and consensus. These examinations infrequently harbor cancers, but due to a large number, they still constitute a considerable proportion of screen-detected cancers. Training sets for the radiologists and artificial intelligence may be effective measures for identification of cancers in these examinations, leading to improved sensitivity and specificity of mammographic screening.

Data sharing statements.

Research data used in the analyses can be made available on request to https://helsedata.no/, given legal basis in Articles 6 and 9 of the GDPR and that the processing is in accordance with Article 5 of the GDPR.

Disclaimer.

Data from the Cancer Registry of Norway (CRN) has been used in this publication. The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by CRN is intended nor should be inferred.

References

[1] IARC, Global cancer observatory. https://gco.iarc.fr/. (Accessed 15/09/2023.

[2] Cancer Registry of Norway. Cancer in Norway 2022 - Cancer incidence, mortality, survival and prevalence in Norway, Cancer Registry of Norway, Oslo, 2023.

[3] ECIBC, Recommendations from the European Breast Cancer Guidelines https://healthcare-quality.jrc.ec.europa.eu/ecibc/european-breast-cancer-guidelines. (Accessed 15/09/2023.

[4] M. Larsen, C.F. Aglen, C.I. Lee, S.R. Hoff, H. Lund-Hanssen, K. Lang, J.F. Nygard, G. Ursin, S. Hofvind, Artificial Intelligence Evaluation of 122 969 Mammography Examinations from a Population-based Screening Program, Radiology 303(3) (2022) 502-511.

[5] A. Rodriguez-Ruiz, K. Lang, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T.H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, M.G. Wallis, I. Andersson, S. Zackrisson, R.M. Mann, I. Sechopoulos, Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists, J Natl Cancer Inst 111(9) (2019) 916-922.

[6] A.D. Lauritzen, A. Rodriguez-Ruiz, M.C. von Euler-Chelpin, E. Lynge, I. Vejborg, M. Nielsen, N. Karssemeijer, M. Lillholm, An Artificial Intelligence-based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload, Radiology 304(1) (2022) 41-49.

[7] K. Lang, V. Josefsson, A.M. Larsson, S. Larsson, C. Hogberg, H. Sartor, S. Hofvind, I. Andersson, A. Rosso, Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study, Lancet Oncol 24(8) (2023) 936-944.

[8] K. Dembrower, A. Crippa, E. Colon, M. Eklund, F. Strand, C.A.D.T.C. ScreenTrust, Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study, Lancet Digit Health (2023).

[9] S.R. Hoff, T.A. Myklebust, C.I. Lee, S. Hofvind, Influence of Mammography Volume on Radiologists' Performance: Results from BreastScreen Norway, Radiology 292(2) (2019) 289-296.

[10] M.A. Rawashdeh, W.B. Lee, R.M. Bourne, E.A. Ryan, M.W. Pietrzyk, W.M. Reed, R.C. Heard, D.A. Black, P.C. Brennan, Markers of good performance in mammography depend on number of annual readings, Radiology 269(1) (2013) 61-7.

[11] Lovdata, Regulations on the collection and processing of personal health data in the Cancer Registry of Norway (Cancer Registry Regulations), 2001. https://lovdata.no/dokument/SF/forskrift/2001-12-21-1477.

[12] Sickles E, D’Orsi CJ, Bassett LW, et al., ACR BI-RADS® Mammography. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, American College of Radiology, Reston, VA, , 2013.

[13] Kvalitetsmanual i Mammografiprogrammet, 2019. https://www.kreftregisteret.no/Generelt/Rapporter/Mammografiprogrammet/Kvalitet/.

[14] E.W. Bjørnson, A.S. Holen, S. Sagstad, M. Larsen, J. Thy, G. Mangerud, A.K. Ertzaas, S. Hofvind, BreastScreen Norway: 25 years of organized screening, 2022. https://www.kreftregisteret.no/Generelt/Rapporter/Mammografiprogrammet/25-arsrapport-mammografiprogrammet/. (Accessed 01/12/2022.

[15] S. Hofvind, T. Hovda, A.S. Holen, C.I. Lee, J. Albertsen, H. Bjorndal, S.H.B. Brandal, R. Gullien, J. Lomo, D. Park, L. Romundstad, P. Suhrke, E. Vigeland, P. Skaane, Digital Breast Tomosynthesis and Synthetic 2D Mammography versus Digital Mammography: Evaluation in a Population-based Screening Program, Radiology 287(3) (2018) 787-794.

[16] S. Hofvind, A.S. Holen, H.S. Aase, N. Houssami, S. Sebuodegard, T.A. Moger, I.S. Haldorsen, L.A. Akslen, Two-view digital breast tomosynthesis versus digital mammography in a population-based breast cancer screening programme (To-Be): a randomised, controlled trial, Lancet Oncol 20(6) (2019) 795-805.

[17] A. Goldhirsch, E.P. Winer, A.S. Coates, R.D. Gelber, M. Piccart-Gebhart, B. Thurlimann, H.J. Senn, m. Panel, Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013, Ann Oncol 24(9) (2013) 2206-23.

[18] A. Wadhwa, J.R. Sullivan, M.B. Gonyo, Missed Breast Cancer: What Can We Learn?, Curr Probl Diagn Radiol 45(6) (2016) 402-419.

[19] T. Hovda, S. Sagstad, M. Larsen, Y. Chen, S. Hofvind, Screening outcome for interpretation by the first and second reader in a population-based mammographic screening program with independent double reading, Acta Radiol (2023) 2841851231176272.

[20] S. Hofvind, S. Sagstad, S. Sebuodegard, Y. Chen, M. Roman, C.I. Lee, Interval Breast Cancer Rates and Histopathologic Tumor Characteristics after False-Positive Findings at Mammography in a Population-based Screening Program, Radiology 287(1) (2018) 58-67.

[21] E.A. Rakha, J.S. Reis-Filho, F. Baehner, D.J. Dabbs, T. Decker, V. Eusebi, S.B. Fox, S. Ichihara, J. Jacquemier, S.R. Lakhani, J. Palacios, A.L. Richardson, S.J. Schnitt, F.C. Schmitt, P.H. Tan, G.M. Tse, S. Badve, I.O. Ellis, Breast cancer prognostic classification in the molecular era: the role of histological grade, Breast Cancer Res 12(4) (2010) 207.

[22] A.K. Casasent, M.M. Almekinders, C. Mulder, P. Bhattacharjee, D. Collyar, A.M. Thompson, J. Jonkers, E.H. Lips, J. van Rheenen, E.S. Hwang, S. Nik-Zainal, N.E. Navin, J. Wesseling, P.C. Grand Challenge, Learning to distinguish progressive and non-progressive ductal carcinoma in situ, Nat Rev Cancer 22(12) (2022) 663-678.

[23] N. Moshina, H.A. Backmann, P. Skaane, S. Hofvind, Mammographic features and risk of breast cancer death among women with invasive screen-detected cancer in BreastScreen Norway 1996-2020, Eur Radiol (2023).

[24] L. Tabar, H.H. Tony Chen, M.F. Amy Yen, T. Tot, T.H. Tung, L.S. Chen, Y.H. Chiu, S.W. Duffy, R.A. Smith, Mammographic tumor features can predict long-term outcomes reliably in women with 1-14-mm invasive breast carcinoma, Cancer 101(8) (2004) 1745-59.

[25] D. Cocco, A. ElSherif, M.D. Wright, M.S. Dempster, M.L. Kruse, H. Li, S.A. Valente, Invasive Lobular Breast Cancer: Data to Support Surgical Decision Making, Ann Surg Oncol 28(10) (2021) 5723-5729.

[26] L. Domingo, D. Salas, R. Zubizarreta, M. Bare, G. Sarriugarte, T. Barata, J. Ibanez, J. Blanch, M. Puig-Vives, A. Fernandez, X. Castells, M. Sala, I.S. Group, Tumor phenotype and breast density in distinct categories of interval cancer: results of population-based mammography screening in Spain, Breast Cancer Res 16(1) (2014) R3.

[27] T.D. Geertse, E. Paap, D. van der Waal, L.E.M. Duijm, R.M. Pijnappel, M.J.M. Broeders, Utility of Supplemental Training to Improve Radiologist Performance in Breast Cancer Screening: A Literature Review, J Am Coll Radiol 16(11) (2019) 1528-1546.

[28] P.D.Y. Trieu, K. Tapia, H. Frazer, W. Lee, P. Brennan, Improvement of Cancer Detection on Mammograms via BREAST Test Sets, Acad Radiol 26(12) (2019) e341-e347.

[29] E. Michalopoulou, P. Clauser, F.J. Gilbert, R.M. Pijnappel, R.M. Mann, P.A.T. Baltzer, Y. Chen, E.M. Fallenberg, A survey by the European Society of Breast Imaging on radiologists' preferences regarding quality assurance measures of image interpretation in screening and diagnostic mammography, Eur Radiol (2023).

[30] H.S. Tsau, A.M. Yen, J.C. Fann, W.Y. Wu, C.P. Yu, S.L. Chen, S.Y. Chiu, L. Tabar, W.H. Kuo, H.H. Chen, K.J. Chang, Mammographic tumour appearance and triple-negative breast cancer associated with long-term prognosis of breast cancer death: a Swedish Cohort Study, Cancer Epidemiol 39(2) (2015) 200-8.

[31] Y. Li, J. Cao, Y. Zhou, F. Mao, S. Shen, Q. Sun, Mammographic casting-type calcification is an independent prognostic factor in invasive breast cancer, Sci Rep 9(1) (2019) 10544.

Figure legends

Figure 1. Study sample.
* The Oslo-Vestfold-Vestre Viken study. † The Tomosynthesis study in Bergen.

Figure 2. Maximum score and combinations of interpretation scores by the two radiologists per breast in BreastScreen Norway, a population-based mammographic screening program with independent double reading.

Figure 3. Distribution of mammographic features by maximum interpretation score leading to recall (A) or screen-detected cancer (B).

Tables

Table 1. Distribution (n and %) of all interpretations, positive interpretations, recall, screen-detected, and interval cancer for breasts stratified by maximum interpretation score 1-5 by one or both radiologists.

Maximum score	All interpretations n (%)	Positive interpretations† n (%)	Recall n (%)	Screen-detected cancer n (%)	Interval cancer n (%)
1	3,246,892 (96.1)	N/A	N/A	196 (2.0)*	2753 (86.0)
2	112,440 (3.3)	112,440 (84.8)	37,220 (69.3)	2341 (24.1)	365 (11.4)
3	15,463 (0.5)	15,463 (11.7)	11,931 (22.2)	3369 (34.6)	63 (2.0)
4	3042 (0.09)	3042 (2.3)	2966 (5.5)	2296 (23.6)	13 (0.4)
5	1625 (0.05)	1625 (1.2)	1618 (3.0)	1531 (16.1)	6 (0.2)
Total	3,379,462 (100.0)	132,570 (100.0)	53,735 (100.0)	9733 (100.0)	3200 (100.0)

† Discordant or concordant positive score

* Recalled due to positive score of the contralateral breast

Table 2. Breast-based rates of recall, screen-detected cancer and positive predictive value (PPV-1) for all combinations of interpretation scores by both readers.

	Recall †	Screen-detected cancer †	PPV-1†
Discordant scores			
Score 1+2	23.7% (21,632 /91,190)	1.5% (1328/91,190)	6.1% (1328/21,632)
Score 1+3	61.1% (4877/7986)	11.5% (917/7986)	18.8% (917/4877)
Score 1+4	89.6% (440/491)	44.8% (220/491)	50.0% (220/440)
Score 1+5	94.4% (84/89)	66.3% (59/89)	70.2% (59/84)
Concordant scores			
Score 2+2	73.4% (15,588/21,250)	4.8% (1013/21,250)	6.5% (1013/15,558)
Score 2+3	93.1% (5067/5441)	24.5% (1335/5441)	26.3% (1335/5067)
Score 2+4	97.8% (478/489)	59.9% (293/489)	61.3% (293/478)
Score 2+5	100.0% (59/59)	83.1% (49/59)	83.1% (49/59)
Score 3+3	97.6% (1987/2036)	54.9% (1117/2036)	56.2% (1117/1987)
Score 3+4	99.0% (1201/1213)	82.4% (1000/1213)	83.3% (1000/1201)
Score 3+5	99.5% (215/216)	88.0% (190/216)	88.4% (190/215)
Score 4+4	99.8% (847/849)	92.2% (783/849)	92.4% (783/847)
Score 4+5	99.9% (680/681)	97.1% (661/681)	97.2% (661/680)
Score 5+5	100.0% (580/580)	98.6% (572/580)	98.6% (571/580)

† p<0.001 for trend for increasing sum of scores

Table 3. Distribution of mammographic features (n and %) for discordant or concordant positive interpretations leading to recall or screen-detected cancer by maximum interpretation score.

	Maximum interpretation score					
	Score 2	Score 3	Score 4	Score 5	Total	p-value †
All recalls						
Mass	14,240 (83.1)	2487 (14.5)	317 (1.8)	92 (0.5)	17,136 (100.0)	<0.001
Spiculated mass	1194 (27.4)	1261 (28.9)	1089 (25.0)	816 (18.7)	4360 (100.0)	
Distortion	610 (56.0)	327 (30.0)	96 (8.8)	57 (5.2)	1090 (100.0)	
Asymmetry	6274 (70.9)	2044 (23.1)	364 (4.1)	163 (1.8)	8845 (100.0)	
Mass with calcifications	413 (34.7)	353 (29.7)	223 (18.8)	200 (16.8)	1189 (100.0)	
Calcifications alone	3121 (49.5)	2502 (39.7)	530 (8.4)	152 (5.2)	6305 (100.0)	
Data not available	11,368	2957	347	138	14,810	
Total	25,852 (100.0)	8974 (100.0)	2619 (100.0)	1480 (100.0)	38,925 (100.0)	
Screen-detected cancers						
Mass	423 (37.9)	402 (36.0)	207 (18.5)	84 (7.5)	1116 (100.0)	<0.001
Spiculated mass	578 (17.1)	976 (28.9)	1025 (30.3)	804 (23.8)	3383 (100.0)	
Distortion	123 (33.2)	122 (32.9)	71 (19.1)	55 (14.8)	371 (100.0)	
Asymmetry	325 (28.5)	428 (37.6)	233 (20.5)	153 (13.4)	1139 (100.0)	
Mass with calcifications	103 (14.6)	210 (29.7)	202 (28.6)	192 (27.2)	707 (100.0)	
Calcifications alone	585 (27.8)	967 (45.9)	412 (19.6)	142 (6.7)	2106 (100.0)	
Data not available	204	264	146	101	715	
Total	2137 (100.0)	3105 (100.0)	2150 (100.0)	1430 (100.0)	8822 (100.0)	

† p-value for the distribution of mammographic features

Table 4. Histopathological characteristics for screen-detected cancers by maximum interpretation score. Median tumor diameter in millimeter with an interquartile range (IQR), otherwise data are numbers with percentages in parentheses.

	Maximum interpretation score					
	Score 1*	Score 2	Score 3	Score 4	Score 5	
	n (%)	n (%)	n (%)	n (%)	n (%)	p-value †
All cancers	*n=196*	*n=2341*	*n=3369*	*n=2296*	*n=1531*	
Histological type						<0.001
Ductal carcinoma in situ	37 (18.9)	547 (23.4)	729 (21.6)	286 (12.5)	67 (4.4)	
Invasive cancer NST	112 (57.1)	1472 (62.9)	2190 (65.0)	1733 (75.5)	1316 (86.0)	
Invasive lobular carcinoma	38 (19.4)	192 (8.2)	301 (8.9)	211 (9.2)	126 (8.2)	
Other invasive carinoma	9 (4.6)	130 (5.6)	149 (4.4)	66 (2.9)	22 (1.4)	
Invasive cancer	*n=159*	*n=1794*	*n=2640*	*n=2010*	*n=1464*	
Median tumor diameter	12 (8-18)	10.5 (7-15)	11 (8-17)	14 (10-19)	17 (12-25)	
Data not available	8	26	40	38	55	
Histological grade						<0.001
Histological grade 1	55 (36.7)	631 (35.7)	786 (30.0)	492 (24.6)	274 (19.0)	
Histological grade 2	78 (52.0)	844 (47.4)	1270 (48.5)	999 (50.0)	747 (51.9)	
Histological grade 3	17 (11.3)	293 (16.6)	560 (21.4)	506 (25.3)	418 (29.0)	
Data not available	9	26	24	13	25	
Lymph node positive	23 (15.3)	287 (16.3)	455 (17.5)	427 (21.5)	430 (30.2)	<0.001
Data not available	9	31	44	23	40	
Molecular subtype						<0.001
Luminal A like	92 (62.2)	1087 (62.1)	1605 (62.4)	1128 (57.6)	813 (56.7)	
Luminal B like Her2-	24 (16.2)	258 (14.7)	351 (13.6)	259 (13.2)	162 (11.3)	
Luminal B like Her2+	24 (16.2)	279 (15.9)	422 (16.4)	373 (19.1)	326 (22.7)	
Her2+	2 (1.4)	45 (2.6)	74 (2.9)	71 (3.6)	54 (3.8)	
Triple negative	6 (4.1)	81 (4.6)	121 (4.7)	126 (6.4)	80 (5.6)	
Data not available	11	44	67	53	29	

* Recalled due to positive score of the contralateral breast

† p-value for the distribution of the variable for score 2-5

Table 5. Interpretation score (highest score per examination per radiologist) by cumulative reading volume for all interpretations and for screen-detected cancers. Interpretations performed 2012-2021 by 174 radiologists.

	Interpretation score				
Cumulative reading volume	Score 1 n (%)	Score 2 n (%)	Score 3 n (%)	Score 4 n (%)	Score 5 n (%)
All interpretations					
<5000	253,880 (93.0)	16,248 (6.0)	2350 (0.9)	357 (0.1)	136 (0.05)
5000-9999	230,452 (94.0)	12,755 (5.2)	1512 (0.6)	319 (0.1)	125 (0.05)
10,000-19,999	408,814 (94.8)	18,993 (4.4)	2519 (0.6)	579 (0.1)	226 (0.05)
20,000-29,999	301,356 (95.1)	12,672 (4.0)	2010 (0.6)	492 (0.2)	190 (0.06)
30,000-39,999	221,119 (95.6)	8300 (3.6)	1356 (0.6)	336 (0.1)	164 (0.07)
40,000-49,999	237,950 (96.1)	7999 (3.2)	1257 (0.5)	305 (0.1)	146 (0.06)
50,000-99,999	1,013,402 (95.8)	37,613 (3.6)	4782 (0.5)	1336 (0.1)	633 (0.06)
≥100,000	552,576 (95.8)	19,880 (3.4)	2928 (0.5)	821 (0.1)	574 (0.1)
Screen-detected cancer					
<5000	111 (7.7)	428 (29.6)	518 (35.8)	265 (18.3)	126 (8.7)
5000-9999	107 (8.4)	442 (34.5)	385 (30.0)	231 (18.0)	117 (9.1)
10,000-19,999	145 (6.5)	685 (30.8)	723 (32.5)	458 (20.6)	215 (9.7)
20,000-29,999	110 (6.3)	488 (27.9)	571 (32.7)	399 (22.9)	178 (10.2)
30,000-39,999	81 (6.4)	345 (27.1)	420 (33.0)	273 (21.5)	153 (12.0)
40,000-49,999	83 (7.1)	286 (24.3)	405 (34.5)	266 (22.6)	135 (11.5)
50,000-99,999	325 (6.2)	1462 (27.9)	1674 (31.9)	1167 (22.3)	616 (11.7)
≥100,000	250 (7.8)	798 (25.0)	924 (29.0)	661 (20.7)	553 (17.4)