Screening outcome for interpretation by the first and second reader in a population-based mammographic screening program with independent double reading

Tone Hovda[1], Silje Sagstad[2], Marthe Larsen[2], Yan Chen[3], Solveig Hofvind[2,4]

[1] Department of Radiology, Vestre Viken Hospital Trust, Drammen, Norway

[2] Section for Breast Cancer Screening, Cancer Registry of Norway, Oslo, Norway

[3] Translational Medical Sciences, School of Medicine, University of Nottingham, Nottingham, UK

[4] Department of Health and Care Sciences, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

Abstract

Background: Double reading of screening mammograms is associated with higher rate of screen-detected cancer than single reading, but different strategies exist regarding reader pairing and blinding. Knowledge about these aspects is important when considering strategies for use of artificial intelligence in mammographic screening.

Purpose: To investigate screening outcome, histopathological tumor characteristics and mammographic features stratified by the first and the second reader in a screening program for breast cancer.

Material and methods: The study sample consisted of data from 3,499,048 screening examinations from 834,691 women performed 1996-2018 in BreastScreen Norway. All examinations were interpreted independently by two radiologists, 272 in total. We analyzed interpretation score, recall and cancer detection, as well as histopathological tumor characteristics and mammographic features of the cancers, stratified by the first and second reader.

Results: For Reader 1, rates of positive interpretations was 4.8%, recall 2.3% and cancer detection 0.5%. The corresponding percentages for Reader 2 were 4.9%, 2.5% and 0.5% ($p<0.05$ compared with Reader 1). No statistical difference was observed for histopathological tumor characteristics or mammographic features when stratified by Reader 1 and 2. Recall and cancer detection were statistically higher and histopathological tumor characteristics less favorable for cases detected after concordant positive compared with discordant interpretations.

Conclusion: Despite reaching statistical significance, mainly due to the large study sample, we consider the differences in interpretation scores, recall and cancer detection between the first and second reader to be clinically negligible. For practical and clinical purposes, double reading in BreastScreen Norway is independent.

Keywords: breast; screening; adults; mammography; neoplasms-primary

Introduction

Breast cancer is the leading cause of cancer death in females worldwide, and health authorities recommend mammographic screening as secondary prevention to reduce mortality of the disease (1-3). Different strategies apply to screen-reading of the mammograms across programs; one (single reading) or two readers (double reading) and double reading with different approaches. Double reading is associated with higher cancer detection than single reading (4,5) and is suggested by the European commission (3). In double reading, the second reader may be blinded (independent reading) or non-blinded to the first reader's interpretation. Examinations with a positive interpretation by one or both readers may be discussed in an arbitration/consensus meeting to decide whether to recall. Arbitration/consensus has been shown to reduce recall and false positive rates with little impact on cancer detection and is thus considered cost-effective (4,6,7). Some screen-detected cancers in a program with double reading are interpreted positively by only one reader (7-9). The distribution between the first and second reader may be influenced by whether the second reader is blinded to the first reader's interpretation (4,10).

Introduction of artificial intelligence (AI) has demonstrated promising results in screen-reading of mammograms (11-14). However, the optimal strategy for screen-reading in combination with AI is yet to be determined and might vary across programs related to their inherent performance measures and reading strategies (15). Thus, knowledge is needed about the current relation between interpretation scores given by the two readers in a screening program and the screening outcome.

BreastScreen Norway, a population-based breast cancer screening program using independent double reading and consensus, offers women aged 50-69 biennial two-view mammographic screening. The readers are randomly paired, and no guidelines exist regarding who is first and second reader in the pair. However, local differences may apply, also according to availability of radiologists and their level of experience, which may in turn influence the association between screening outcome and the first and second reader. We therefore aimed to investigate screening outcome, histopathological tumor characteristics and mammographic features for screening examinations in BreastScreen Norway, stratified by the first and the second reader.

We hypothesized that no differences in the selected performance measures appeared between the readers, as the double reading should be independent according to national guidelines.

Material and methods

The Data Protection Officer for the Cancer Registry of Norway approved this retrospective cohort study. The Cancer Registry Regulations waived the requirement to obtain written informed consent (16). Women participating in the screening program have their right to request that data from their normal screening examinations be erased from the screening database after quality assurance. Data from these women (1.4% of the invited) were not included in the study population (17).

BreastScreen Norway

Two trained breast radiologists read all screening mammograms independently and assign each breast a score 1-5 (1: normal /benign, 2: probably benign, 3: intermediate suspicious, 4: probably malignant and 5: highly suspicious of malignancy). All exams scored ≥2 for one or both breasts by either radiologist are discussed in a consensus meeting to decide whether to recall the woman for further assessment. The radiologists report the findings and results of the assessment to the Cancer Registry where the data are stored in a dedicated database (17). At start-up of the program in 1996, all women were screened using screen-film mammography. From 2000, full-field digital mammography was gradually implemented, and from 2012, all screening examinations were digital.

Study population

Our study population included women who participated in BreastScreen Norway 1.1.1996-31.12.2018. To include interval cancer, the study period was 1.1.1996-31.12.2020. The population included 851,430 women, contributing with 3,753,977 screening examinations (Figure 1). We excluded examinations that were part of conducted prospective studies: the Oslo I and II studies (18,19), the Oslo Tomosynthesis Screening Trial (OTST) (20), The Oslo-Vestfold-Vestre Viken (OVVV) study (21) and the Tomosynthesis in Bergen (To-Be) trial (22), in total 214,963 examinations. Further, we excluded examinations without independent double reading, examinations performed in women reporting symptoms at screening, examinations with registered technical image errors, examinations resulting in screen-detected breast cancer without registration of laterality, and examinations from women recalled despite negative scores, in total 39,966 examinations. This left 3,499,048 examinations from 834,691

women eligible for analyses, making up the study sample (Figure 1), and the study was examination-based rather than breast-based.

Definition of measures

We defined Reader 1 as the radiologist performing the first reading of the screening mammograms, and Reader 2 as the radiologist performing the second reading. A positive interpretation was defined as an examination with an initial interpretation score ≥2 of one or both breasts; examinations with positive interpretation by one reader were defined as discordant and examinations with positive interpretation by both readers were defined as concordant positive. Recall was defined as women with discordant or concordant positive interpretation, recalled for further assessment after a consensus meeting. Screen-detected cancer was defined as ductal carcinoma in situ (DCIS) and invasive breast cancer, diagnosed within 6 months after a positive screening interpretation and recall. Interval cancer was defined as cancer (DCIS and invasive breast cancer) diagnosed ≤24 months after a negative screening examination, or >6 months after a false positive recall assessment.

Positive predictive value 1 (PPV-1) was defined as the percentage of screen-detected breast cancer cases detected among women recalled for further assessment. Positive predictive value 3 (PPV-3) was defined as the percentage of screen-detected breast cancer cases among the recalled women who underwent a needle biopsy at recall assessment. Internationally, PPV-2 usually refers to the percentage of screen-detected breast cancer cases among women recommended a biopsy. In BreastScreen Norway, almost 100% of recommended biopsies are actually performed, thus, PPV-2 and PPV-3 is almost equal, and PPV-2 is not reported in this study.

Histopathological tumor characteristics variables included tumor type (DCIS, invasive cancer of no special type, invasive lobular carcinoma or other invasive carcinomas), and for invasive cancer tumor diameter (mm), histological grade (1-3), lymph node involvement (positive/negative), as well as estrogen and progesterone receptor status (positive/negative). Mammographic features were classified as mass, spiculated mass, distortion, asymmetry, density with calcifications and calcifications alone. This classification is in accordance with the reporting to the screening database in BreastScreen Norway, and is a modification of the Breast Imaging Reporting and Data System classification of mammographic features (17).

Statistical analyses

Total and annual reading volumes were presented as medians with interquartile range (IQR). We performed descriptive analyses of screening outcome, histopathological characteristics and mammographic features stratified by readers' interpretation: Positive interpretations by Reader 1 only (discordant); positive interpretations by Reader 2 only (discordant); positive interpretations by both readers (concordant positive), all positive interpretations by Reader 1 (discordant/concordant positive), and all positive interpretations by Reader 2 (discordant/concordant positive). The analyses were examination-based. We tested for statistical significance using bivariate tests. All analyses were performed using Stata version 17.

Results

Reader characteristics

In total, 272 radiologists were involved in screen-reading in the study period. Median reading volume was 32,785 (IQR: 53,428) for Reader 1 and 29,044 (IQR: 48,377) for Reader 2. Median annual reading volume was 4059 (IQR: 5373) for Reader 1 and 3644 (IQR: 4905) for Reader 2 (Figure 2).

Screening outcome for Reader 1 and Reader 2

Of the 3,499,048 examinations, 7.6% (n=266,613) had a positive interpretation by one or both readers and were discussed in the consensus meeting whether to recall; 4.8% (n=169,092) were scored ≥2 by Reader 1 and 4.9% (n=171,302) were scored ≥2 by Reader 2 (Table 1). After the consensus meeting, the total recall rate in the study population was 3.2% (n=110,576). For examinations interpreted by Reader 1, 81,920 were recalled, representing 2.3% of all examinations and 48.4% of all with interpretation score ≥2 by Reader 1. The corresponding percentages for Reader 2 were 2.5% and 50.5 % (n=86,337). The rate of screen-detected cancer was 0.5% (n=19,007) in the total sample, for Reader 1 (n=16,445) and for Reader 2 (n=17,120). Positive predictive value (PPV) 1 was 20.1% for Reader 1 and 19.8% for Reader 2, and PPV-3 46.8% and 46.4% for Reader 1 and 2, respectively.

Concordant positive and discordant examinations

Among all screening examinations, 2.1% (n=73,781) were scored concordant positive, 2.7% (n=95,311) were scored ≥2 by Reader 1 only and 2.8% (n=97,521) by Reader 2 only (Table 1). After consensus, 21.8% of concordant positive and 72.6% of discordant cases were

dismissed (Figure 2), resulting in recall of 1.6% of all concordant positive cases, 0.7% of discordant scored ≥2 by Reader 1 only, and 0.8% of discordant scored ≥2 by Reader 2 only (Table 1).

Among those recalled after concordant positive interpretation, 25.2% were diagnosed with screen-detected cancer. The corresponding percentage for discordant interpretations was 8.4% (Figure 2). The cancer detection rate was 0.4% (n=14,558) after concordant positive interpretation, 0.05% (n=1887) after positive interpretation by Reader 1 only, and 0.07% (n=2562) after positive interpretation by Reader 2 only (Table 1). Thus, 76.6% of the screen-detected cancers had a concordant positive interpretation whereas 23.4% had a discordant interpretation (Figure 2).

Interval cancer

Eighty-two percent (n=4934) of the 6014 interval cancers were diagnosed among women negatively screened by both radiologists, whereas 11.1% (n=669) were diagnosed among women dismissed at consensus after discordant or concordant interpretation (Figure 2). The interval cancer rate among women who underwent recall assessment was 0.4 % and did not differ statistically by groups (Table 1).

Histopathological characteristics

Histopathological tumor characteristics did not differ statistically between cancers detected by Reader 1 and Reader 2, or between discordant cancers scored ≥2 by Reader 1 only and Reader 2 only (Table 2).

The proportion of DCIS was higher among discordant (Reader 1: 24.4%; Reader 2: 23.1%) compared with concordant positive cancers (16.7%) (Table 2). Invasive carcinoma NST was observed in 71.7% of concordant positive cancers compared with 62.4% (Reader 1) and 62.6% (Reader 2) of discordant cancers. A statistically higher proportion of histological grade 3 invasive tumors was observed in concordant cancers (21.5%) compared with discordant (Reader 1: 13.4%; Reader 2: 11.9%). Concordant positive cancers had a statistically higher proportion of lymph node positive tumors and a lower proportion of estrogen/progesterone receptor positive tumors than discordant (Table 2).

Mammographic features

No statistically significant differences in mammographic features were observed between cancers detected by Reader 1 and Reader 2 (Table 3). The proportion of screen-detected cancers classified as spiculated masses was higher (31.8%) for concordant positive compared with discordant cancers (Reader 1: 25.2%; Reader 2: 28.7%). The proportion of asymmetries (13.9%) and calcifications alone (21.8%) was lower for concordant positive compared with discordant cancers (Table 3).

Discussion

In this retrospective study from a screening program using independent double reading, the differences in interpretation, recall and cancer detection between Reader 1 and Reader 2 reached statistical significance at a 0.05 significance level. The statistical difference is probably due to a large sample size. It is difficult to set a threshold at which the observed differences should be considered clinically impactful in our study. However, as the observed differences were only minor, we consider that for practical and clinical purposes, our results implied the screen-reading in BreastScreen Norway to be largely independent. We observed no differences in histopathological characteristics or mammographic features between cancers detected by Reader 1 and Reader 2, but proportions of histological high-grade, hormone receptor negative and lymph node positive tumors were statistically higher in concordant positive compared with discordant cancers. A statistically higher proportion of masses and lower proportion of asymmetries and calcifications were observed in concordant cancers.

Reader concordance in double reading is influenced by whether the readers are blinded to each other's interpretation scores (4,7,9,10,23). Having access to information about the first reader's interpretation may influence the second reader's reading behavior; the second reader may, consciously or not, copy the first reader's interpretation, both negative and positive. Furthermore, having access to information about the first reader's interpretation may cause the second reader to be more attentive to subtle findings. It has been demonstrated that findings only detected by the second reader in non-blinded reading more frequently were asymmetries, and also had more favorable tumor characteristics (4,10,23). In our study, the second reader was blinded to the interpretation result, but not necessarily blinded to the identity of the first reader. This knowledge may have influenced the second reader's behavior; perhaps simply being the second reader as such sharpened the attention in screen-reading.

Our findings are in line with other studies demonstrating more favorable tumor characteristics in discordant than concordant positive cancers (9). The proportion of discordant screen-detected cancers, 23.4%, is in accordance with other studies and supports the value of double rather than single reading, (8,24). Readers may have different strengths and weaknesses in screen-reading that may complement each other and increase screening sensitivity. Our findings also demonstrate that discordant cancers possibly have a more subtle or less suspicious appearance than concordant positives, as more asymmetries and calcifications and less spiculated masses were observed among discordant cancers.

Pairing of screen-readers is often a result of convenience and randomness rather than a planned strategy of composing optimal reader pairs. Major variations have been observed in screening performance between pairs. Brennan et al (25) demonstrated better performance of double reading compared to single reading, both when readers were paired in the best possible ways as well as randomly paired. In contrast, the performance was not improved for the "worst" pairs compared with single reading. It is not feasible to let only "the best" pairs read screening mammograms, but still, being conscious about the pairing may improve screening performance. A standardized personalized training scheme may help monitoring and improving screen-readers' performance by providing self-assessment and identification of mild and severe underperforming outliers (26).

Double reading is obviously more resource intensive than single reading. However, consensus/arbitration may reduce the time consumption by reducing recall and false positive rates and increase program sensitivity (4,7). In contrast to many other mammographic screening programs only arbitrating discordant cases, the consensus meeting in BreastScreen Norway also includes discussion of concordant positive cases. This is resource intensive but may potentially lead to further reduction of recall rates and provide an additional learning/educational aspect. An alternative to consensus/arbitration in double reading is a recall procedure based on an either positive approach, in which all cases with a positive interpretation by one OR both readers, are recalled. In our study, this would have resulted in a recall rate of 7.6 %. We cannot exactly calculate the possible gain of such increase in recall for screen-detected cancer or reduction in interval cancer. However, an increased risk for interval cancer or screen-detected cancer in the consecutive screening round for women with a positive interpretation at screening prior to diagnosis has been demonstrated (24,27).

We will probably witness a future implementation of artificial intelligence (AI) in screen-reading of mammograms. AI may be implemented in different ways; as supplemental reader,

as second reader, as the only reader, or as decision support to the reading radiologists at screen-reading and/or consensus (11). Whether and how/when the radiologists have access to the AI score may influence the interpretation procedure and the screening results, possibly in the same way as in non-blinded double reading. When retrospectively comparing different scenarios for combination of one or two readers and AI with double reading, knowledge of independence between readers is essential to correctly interpret the results (15). If the interpretation by the second reader is influenced by or dependent on the first, sensitivity/specificity of interpretation by AI plus one radiologist will not be directly comparable to that of two radiologists. Further, when retrospectively analyzing and comparing screening results for AI versus the radiologists, knowledge of readers' independence is important when deciding whether to take the order of readers into account. Information on readers' independence in a program is also crucial when setting thresholds for triaging examinations with a low possibility of malignancy to be read by either one reader plus AI or solely by AI.

Our study has strengths and limitations. One major strength is the large sample size of almost 3.5 million examinations. Additionally, the high completeness of the screening database of BreastScreen Norway ensures that the study sample is highly representative for the Norwegian screening population during the 25 years' history of the screening program (17). Sparse knowledge of the pairing strategies of screen-readers at the different breast centers in Norway is a limitation. The screening volume varies substantially across centers and over the study period, as do also the available number of radiologists at each breast center. Some breast centers are small with only two or three radiologists employed, thus relatively constant reader pairs, whereas others are larger, with many affiliated radiologists resulting in a great variety of screening pairs. Possible intra-reader differences regarding whether a radiologist was first or second reader, were not analyzed in this study. Analyzing and comparing performance measures per reader might have been useful in that respect. However, we considered such analyses too extensive for the scope of this study. The examination based approach might be considered a weakness of the study and breast based analyses could have been more specific. Nevertheless, as the main objective was to compare the outcome of the first and second reader's interpretation, we consider an examination based approach sufficient.

To conclude, our results demonstrated that no clinically meaningful differences in screening outcomes were observed for Reader 1 versus Reader 2 indicating that double reading in

BreastScreen Norway is by and large independent. Whether this is the optimal strategy may be debated, also in the context of different reading strategies for implementation of AI in mammographic screening in the future.

Funding acknowledgements

The authors received no financial support for the research, authorship, and/or publication of this article.

References

1. IARC. Global cancer observatory, https://gco.iarc.fr/ (accessed 07/05/2022).
2. Perry N, Broeders MJ, de Wolf C, et al. European guidelines for quality assurance in breast cancer screening and diagnosis. Brussels, Belgium: 2006.
3. ECIBC. Recommendations from the European Breast Cancer Guidelines https://healthcare-quality.jrc.ec.europa.eu/ecibc/european-breast-cancer-guidelines (accessed 06/07/2022).
4. Taylor-Phillips S, Stinton C. Double reading in breast cancer screening: considerations for policy-making. Br J Radiol 2020;93:20190610.
5. Taylor-Phillips S, Jenkinson D, Stinton C, et al. Double Reading in Breast Cancer Screening: Cohort Evaluation in the CO-OPS Trial. Radiology 2018;287:749-757.
6. Caumo F, Brunelli S, Tosi E, et al. On the role of arbitration of discordant double readings of screening mammography: experience from two Italian programmes. Radiol Med 2011;116:84-91.

7. Klompenhouwer EG, Voogd AC, den Heeten GJ, et al. Discrepant screening mammography assessments at blinded and non-blinded double reading: impact of arbitration by a third reader on screening outcome. Eur Radiol 2015;25:2821-2829.
8. Hofvind S, Geller BM, Rosenberg RD, et al. Screening-detected breast cancers: discordant independent double reading in a population-based screening program. Radiology 2009;253:652-660.
9. Coolen AMP, Lameijer JRC, Voogd AC, et al. Characteristics of screen-detected cancers following concordant or discordant recalls at blinded double reading in biennial digital screening mammography. Eur Radiol 2019;29:337-344.
10. Coolen AMP, Voogd AC, Strobbe LJ, et al. Impact of the second reader on screening outcome at blinded double reading of digital screening mammograms. Br J Cancer 2018;119:503-507.
11. Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. Semin Cancer Biol 2021;72:214-225.
12. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol 2019;29:4825-4832.
13. Lauritzen AD, Rodriguez-Ruiz A, von Euler-Chelpin MC, et al. An Artificial Intelligence-based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload. Radiology 2022;304:41-49.
14. Dembrower K, Wahlin E, Liu Y, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health 2020;2:e468-e474.

15. Larsen M, Aglen CF, Hoff SR, et al. Possible strategies for use of artificial intelligence in screen-reading of mammograms, based on retrospective data from 122,969 screening examinations. Eur Radiol 2022;32:8238-8246.
16. Ministry of Health and Care Services, Forskrift om innsamling og behandling av helseopplysninger i Kreftregisteret (The Cancer Registry Regulation), https://lovdata.no/dokument/SF/forskrift/2001-12-21-1477 (2001, accessed 01/12/2022).
17. Bjørnson EW, Holen AS, Sagstad S, et al. BreastScreen Norway: 25 years of organized screening, https://www.kreftregisteret.no/Generelt/Rapporter/Mammografiprogrammet/25-arsrapport-mammografiprogrammet/ (2022, accessed 01/12/2022).
18. Skaane P, Skjennald A, Young K, et al. Follow-up and final results of the Oslo I Study comparing screen-film mammography and full-field digital mammography with soft-copy reading. Acta Radiol 2005;46:679-689.
19. Skaane P, Hofvind S, Skjennald A. Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: follow-up and final results of Oslo II study. Radiology 2007;244:708-717.
20. Skaane P, Bandos AI, Niklason LT, et al. Digital Mammography versus Digital Mammography Plus Tomosynthesis in Breast Cancer Screening: The Oslo Tomosynthesis Screening Trial. Radiology 2019;291:23-30.
21. Hofvind S, Hovda T, Holen AS, et al. Digital Breast Tomosynthesis and Synthetic 2D Mammography versus Digital Mammography: Evaluation in a Population-based Screening Program. Radiology 2018;287:787-794.

22. Hofvind S, Holen AS, Aase HS, et al. Two-view digital breast tomosynthesis versus digital mammography in a population-based breast cancer screening programme (To-Be): a randomised, controlled trial. Lancet Oncol 2019;20:795-805.

23. Posso M, Carles M, Rue M, et al. Cost-Effectiveness of Double Reading versus Single Reading of Mammograms in a Breast Cancer Screening Programme. PLoS One 2016;11:e0159806.

24. Martiniussen MA, Sagstad S, Larsen M, et al. Screen-detected and interval breast cancer after concordant and discordant interpretations in a population based screening program using independent double reading. Eur Radiol 2022.

25. Brennan PC, Ganesan A, Eckstein MP, et al. Benefits of Independent Double Reading in Digital Mammography: A Theoretical Evaluation of All Possible Pairing Methodologies. Acad Radiol 2019;26:717-723.

26. Gale A, Chen Y. A review of the PERFORMS scheme in breast screening. Br J Radiol 2020;93:20190908.

27. Jenkins J, Murphy AE, Edmondson-Jones M, et al. Film reading in the East Midlands Breast Screening Programme -- are we missing opportunities for earlier diagnosis? Clin Radiol 2014;69:385-390.

Tables

Table 1. Screening outcome. Stratification by readers' interpretation

	Positive interpretation (interpretation score ≥2)					
	Reader 1 only (discordant)	**Reader 2 only (discordant)**	**Both readers (concordant)**	**Reader 1 (all)**	**Reader 2 (all)**	**In total**
All interpretations	*n=3,499,048*	*n=3,499,048*	*n=3,499,048*	*n=3,499,048*	*n=3,499,048*	*n=3,499,048*
Interpretation score ≥2	95,311 (2.7)	97,521 (2.8)†	73,781 (2.1)	169,092 (4.8)	171,302 (4.9)‡	266,613 (7.6)
Recalls	24,239 (0.7)	28,656 (0.8)†	57,681 (1.6)	81,920 (2.3)	86,337 (2.5)‡	110,576 (3.2)
Biopsies	6522 (0.2)	8248 (0.2)†	28,628 (0.8)	35,150 (1.0)	36,876 (1.1)‡	43,398 (1.2)
Screen-detected cancer	1887 (0.05)	2562 (0.07)†	14,558 (0.4)	16,445 (0.5)	17,120 (0.5)‡	19,007 (0.5)
Ductal carcinoma in situ	462 (0.01)	593 (0.02)†	2428 (0.07)	2890 (0.08)	3021 (0.09)	3483 (0.1)
Invasive breast cancer	1425 (0.04)	1969 (0.06)†	12,130 (0.4)	13,555 (0.4)	14,099 (0.4)‡	15,524 (0.4)
Interpretation score ≥2	*n=95,311*	*n=97,521*	*n=73,781*	*n=169,092*	*n=171,302*	*n=266,613*
Recalls	24,239 (25.4)	28,656 (29.4)†	57,681 (78.2)	81,920 (48.4)	86,337 (50.4)‡	110,576 (41.5)
Biopsies	6522 (6.8)	8248 (8.5)†	28,628 (38.8)	35,150 (20.8)	36,876 (21.5)‡	43,398 (16.3)
Recalled	*n=24,239*	*n=28,656*	*n=57,681*	*n=81,920*	*n=86,337*	*n=110,576*
Positive predictive value 1	1887 (7.8)	2562 (8.9)†	14,558 (25.2)	16,445 (20.1)	17,120 (19.8)	19,007 (17.2)
Interval cancer	91 (0.4)	102 (0.4)	218 (0.4)	309 (0.4)	320 (0.4)	411 (0.4)
Biopsied	*n=6522*	*n=8248*	*n=28,628*	*n=35,150*	*n=36,876*	*n=43,398*
Positive predictive value 3	1887 (28.9)	2562 (31.1)‡	14,558 (50.9)	16,445 (46.8)	17,120 (46.4)	19,007 (43.8)

Data presented as numbers with percentages in parentheses

†$p<0.05$ compared with Reader 1 only (discordant)

‡$p<0.05$ compared with Reader 1 (all)

Table 2. Histopathological characteristics of screen-detected cancers (DCIS and invasive). Stratification by readers' interpretation.

	Histopathological characteristics of screen-detected cancers					
	Reader 1 only (discordant)	**Reader 2 only (discordant)**	**Both readers (concordant)**	**Reader 1 (all)**	**Reader 2 (all)**	**In total**
All tumors	*n=1887*	*n=2562*	*n=14,558*	*n=16,445*	*n=17,120*	*n=19,007*
Ductal carcinoma in situ	462 (24.4)	593 (23.1)	2428 (16.7)†	2890 (17.6)	3021 (17.6)	3483 (18.3)
Invasive carcinoma NST	1177 (62.4)	1604 (62.6)	10,442 (71.7)†	11,619 (70.7)	12,046 (70.4)	13,223 (69.6)
Invasive lobular carcinoma	175 (9.3)	252 (9.8)	1106 (7.6)†	1281 (7.8)	1358 (7.9)	1533 (8.1)
Other invasive carcinomas	73 (9.2)	113 (4.4)	582 (4.0)	655 (4.0)	695 (4.1)	768 (4.0)
Invasive tumors only	*n=1425*	*n=1969*	*n=12,130*	*n=13,555*	*n=14,099*	*n=15,524*
Median tumor diameter	11 mm	11 mm	13 mm†	13 mm	13 mm	12 mm
Data not available	19	39	288	307	327	346
Histopathologic grade 1	553 (39.5)	778 (40.4)	3583 (30.1)†	4136 (31.1)	4361 (31.5)	4914 (32.2)
Histopathologic grade 2	659 (47.1)	918 (47.7)	5765 (48.4)†	6424 (48.3)	6683 (48.3)	7342 (48.2)
Histopathologic grade 3	187 (13.4)	230 (11.9)	2562 (21.5)†	2749 (20.7)	2792 (20.2)	2979 (19.6)
Data not available	26	43	220	246	263	289
Lymph node positive	250 (18.0)	336 (17.4)	2800 (23.7)†	3050 (23.1)	3136 (22.8)	3386 (22.4)
Data not available	36	36	299	335	335	388
Estrogen receptor positive	1277 (92.7)	1745 (92.1)	10,345 (88.7)†	11,622 (89.1)	12,090 (89.2)	13,367 (89.5)
Data not available	48	75	468	516	543	591
Progesterone receptor positive	1007 (74.0)	1395 (74.2)	8136 (70.3)†	9143 (70.7)	9531 (70.8)	10,538 (71.1)
Data not available	65	89	552	617	641	706

Unless otherwise specified, data are numbers with percentages in parentheses
†p<0.05 compared with Reader 1 only (discordant) and Reader 2 only (discordant)
NST: No special type

Table 3. Mammographic features of screen-detected cancers (DCIS and invasive). Stratification by readers' interpretation.

	Mammographic features of screen-detected cancers					
	Reader 1 only (discordant)	**Reader 2 only (discordant)**	**Both readers (concordant)**	**Reader 1 (all)**	**Reader 2 (all)**	**In total**
	n=1887	*n=2562*	*n=14558*	*n=16,445*	*n=17,120*	*n=19007*
Mass	360 (20.6)	436 (18.3)	3032 (21.9) †	3392 (21.8)	3468 (21.4)	3828 (21.3)
Spiculated mass	439 (25.2) †	684 (28.7)	4387 (31.8) ‡	4826 (31.0)	5071 (31.3)	5510 (30.7)
Distortion	35 (2.0)	59 (2.5)	177 (1.3) ‡	212 (1.4)	236 (1.5)	271 (1.5)
Asymmetry	331 (19.0)	430 18.1)	1922 (13.9) ‡	2253 (14.5)	2352 (14.5)	2683 (15.0)
Density with calcifications	107 (6.1)	145 (6.1)	1293 (9.4) ‡	1400 (9.0)	1438 (8.9)	1545 (8.6)
Calcifications alone	473 (27.1)	627 (26.3)	3006 (21.8) ‡	3479 (22.4)	3633 (22.4)	4106 (22.9)
Data not available	142	181	741	883	922	1064

Data presented as numbers with percentages in parentheses
†p<0.05 compared with Reader 1 only (discordant)
‡p<0.05 compared with Reader 1 (all)

Figure legends:

Figure 1: Study population, exclusions and study sample

Figure 2: Reader volume for Reader 1 and Reader 2, interpretation results (concordant negative, discordant or concordant positive), results from consensus and recall, and interval cancer rates for 3,499,048 screening examinations in BreastScreen Norway 1996-2018.

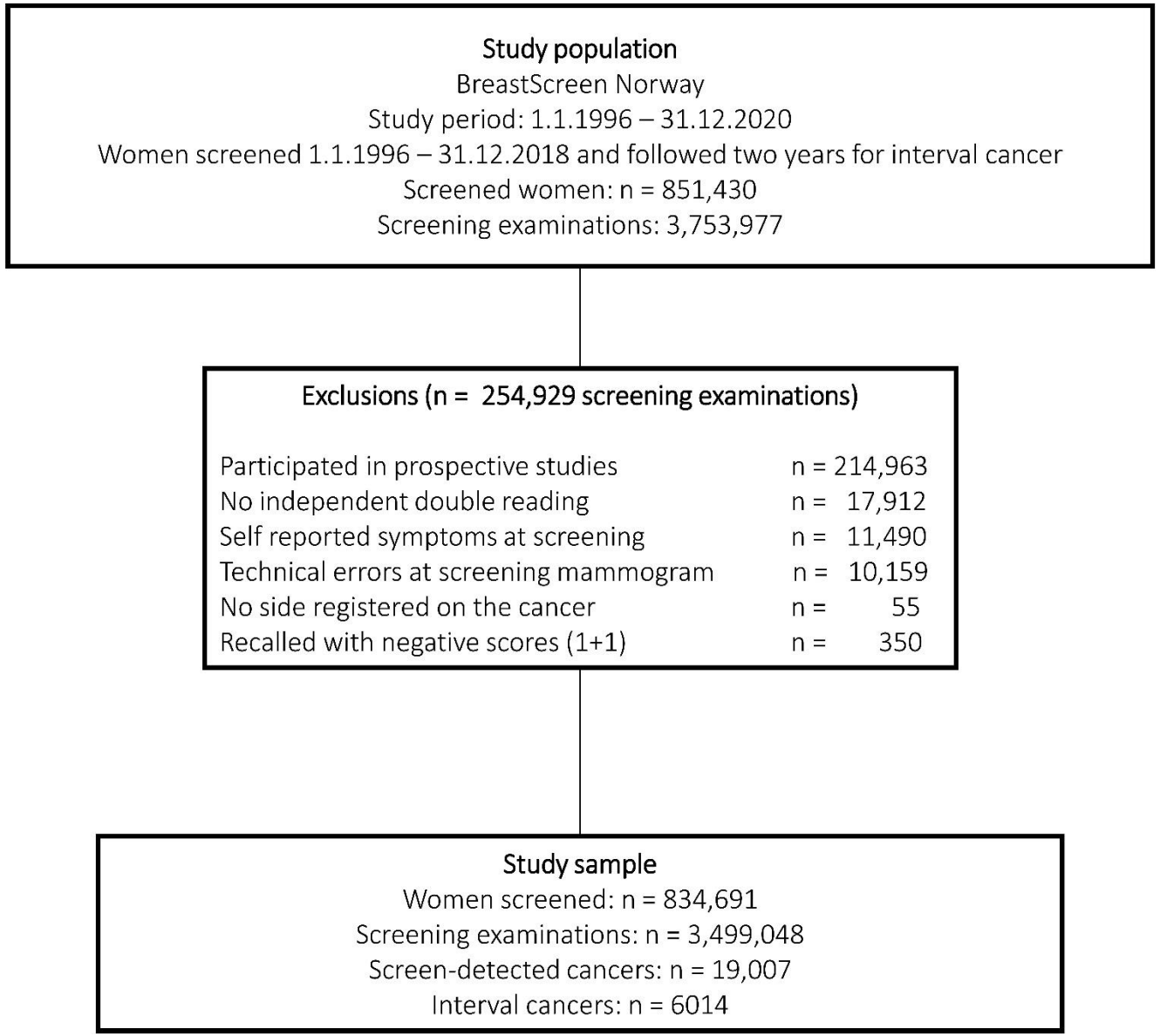
Study population
BreastScreen Norway
Study period: 1.1.1996 – 31.12.2020
Women screened 1.1.1996 – 31.12.2018 and followed two years for interval cancer
Screened women: n = 851,430
Screening examinations: 3,753,977
Exclusions (n = 254,929 screening examinations)
Participated in prospective studies n = 214,963
No independent double reading n = 17,912
Self reported symptoms at screening n = 11,490
Technical errors at screening mammogram n = 10,159
No side registered on the cancer n = 55
Recalled with negative scores (1+1) n = 350
Study sample
Women screened: n = 834,691
Screening examinations: n = 3,499,048
Screen-detected cancers: n = 19,007
Interval cancers: n = 6014

Figure 1

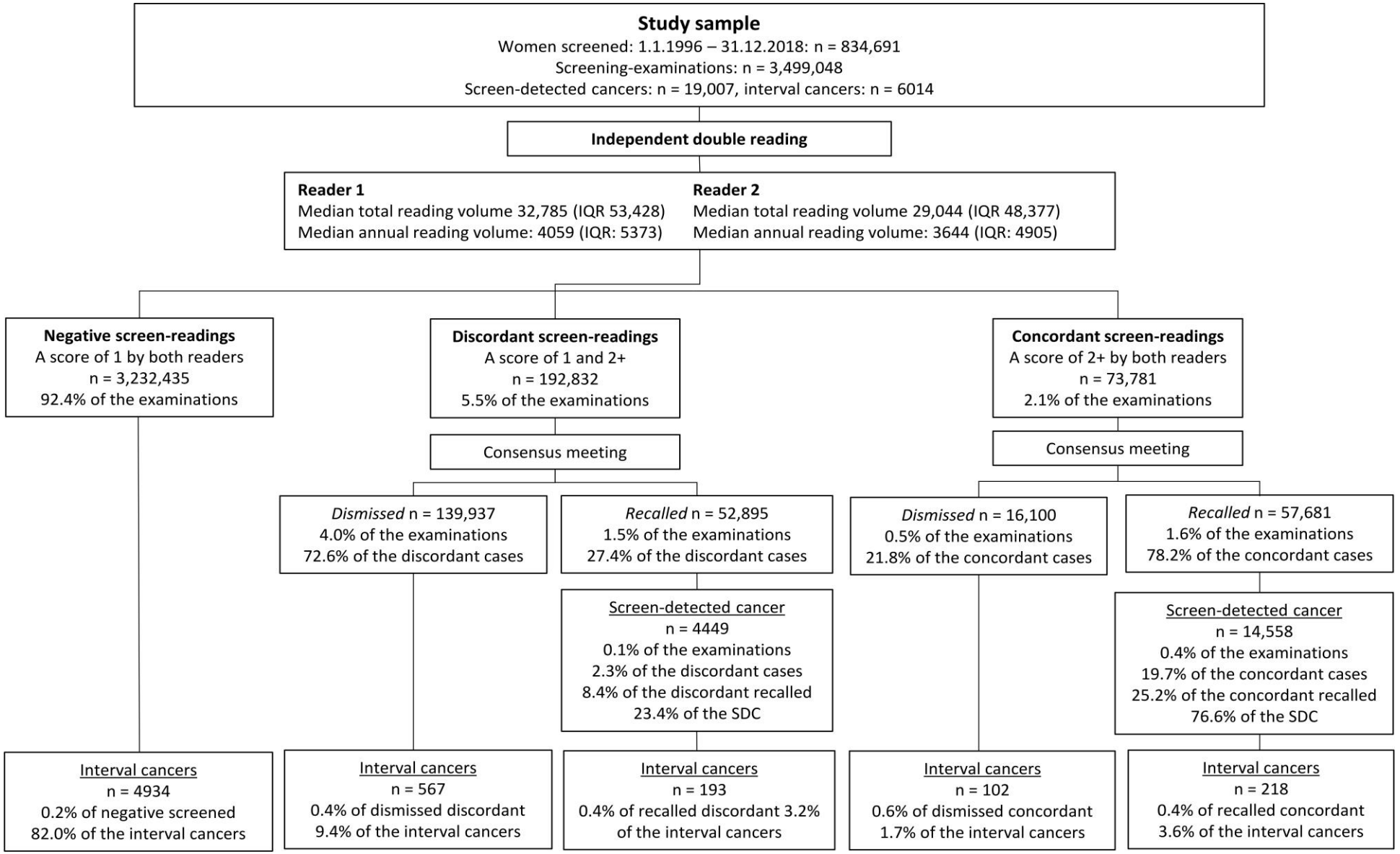
Study sample
Women screened: 1.1.1996 – 31.12.2018: n = 834,691
Screening-examinations: n = 3,499,048
Screen-detected cancers: n = 19,007, interval cancers: n = 6014
Independent double reading
Reader 1
Median total reading volume 32,785 (IQR 53,428)
Median annual reading volume: 4059 (IQR: 5373)
Reader 2
Median total reading volume 29,044 (IQR 48,377)
Median annual reading volume: 3644 (IQR: 4905)
Negative screen-readings
A score of 1 by both readers
n = 3,232,435
92.4% of the examinations
Discordant screen-readings
A score of 1 and 2+
n = 192,832
5.5% of the examinations
Concordant screen-readings
A score of 2+ by both readers
n = 73,781
2.1% of the examinations
Consensus meeting
Consensus meeting
Dismissed n = 139,937
4.0% of the examinations
72.6% of the discordant cases
Recalled n = 52,895
1.5% of the examinations
27.4% of the discordant cases
Dismissed n = 16,100
0.5% of the examinations
21.8% of the concordant cases
Recalled n = 57,681
1.6% of the examinations
78.2% of the concordant cases
Screen-detected cancer
n = 4449
0.1% of the examinations
2.3% of the discordant cases
8.4% of the discordant recalled
23.4% of the SDC
Screen-detected cancer
n = 14,558
0.4% of the examinations
19.7% of the concordant cases
25.2% of the concordant recalled
76.6% of the SDC
Interval cancers
n = 4934
0.2% of negative screened
82.0% of the interval cancers
Interval cancers
n = 567
0.4% of dismissed discordant
9.4% of the interval cancers
Interval cancers
n = 193
0.4% of recalled discordant 3.2%
of the interval cancers
Interval cancers
n = 102
0.6% of dismissed concordant
1.7% of the interval cancers
Interval cancers
n = 218
0.4% of recalled concordant
3.6% of the interval cancers

Figure 2