Optimizing the Pairs of Radiologists That Double Read Screening Mammograms

Jessie J.J. Gommers, MSc; Craig K. Abbey, PhD; Fredrik Strand, MD, PhD; Sian Taylor-Phillips, PhD; David J. Jenkinson, PhD; Marthe Larsen, MSc; Solveig Hofvind, PhD; Ioannis Sechopoulos, PhD; Mireille J.M. Broeders, PhD

From the department of Medical Imaging (J.J.J.G., I.S.) and Department for Health Evidence (M.J.M.B.), Radboud University Medical Center, Geert Grooteplein 10, 6525 GA, Post 766, Nijmegen, The Netherlands; Department of psychological and brain sciences, University of California, Santa Barbara, United States (C.K.A.); Department of Oncology-Pathology, Karolinska Institute, Stockholm, Sweden (F.S.); and Breast Radiology, Karolinska University Hospital, Stockholm, Sweden (F.S.); Warwick Medical School, University of Warwick, Coventry, United Kingdom (S.T.P, D.J.J.); Section for Breast Cancer Screening, Cancer Registry of Norway, Oslo, Norway (M.L., S.H.); Department of Health and Care Sciences, UiT The Arctic University of Norway, Tromsø, Norway (S.H.); Dutch expert center for screening (LRCB), Nijmegen, The Netherlands (I.S., M.J.M.B.); Technical Medicine Center, University of Twente, Enschede, The Netherlands (I.S).

Address correspondence to M.J.M.B.
Mireille.Broeders@radboudumc.nl

Funding
aiREAD – Accurate and Intelligent Reading for EArlier breast cancer Detection (project number 17912). This research is supported by KWF Kankerbestrijding and NWO Domain AES, as part of their joint strategic research program Technology for Oncology II. The collaboration project is co-funded by the PPP Allowance made available by Health-Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.

Abbreviations
AI = artificial intelligence, AIR = abnormal interpretation rate, CDR = cancer detection rate, HH = reader characterized by a high CDR and high AIR, HL = reader characterized by a high CDR and low AIR, LH = reader characterized by a low CDR and high AIR, LL = reader characterized by a low CDR and low AIR

Summary
Performance characteristics of mammography readers influenced the performance of pairs, but specific pairing strategies did not result in significantly different overall performance compared with that resulting from random pairing strategies.

Key Results

■ Retrospective data (3,592,414 examinations) from three population-based breast cancer screening programs (Sweden, England, and Norway) showed variation in cancer detection and abnormal interpretation rates among radiologists.

■ Performance of specific pairs of radiologists was influenced by what types of individual readers were involved.

■ Specific radiologist pairing strategies were not significantly different from the random radiologist pairing strategies; data in which all radiologists read all examinations are needed to explore if there is an optimal pairing strategy that maximizes performance.

Abstract

Background:

Despite variation in performance characteristics among radiologists, the pairing of radiologists for the double reading of screening mammograms is performed randomly. It is unknown how to optimize pairing to improve screening performance.

Purpose:

To investigate whether radiologist performance characteristics can be used to determine the optimal set of pairs of radiologists to double read screening mammograms for improved accuracy.

Materials and Methods:

This retrospective study was performed with reading outcomes from breast cancer screening programs in Sweden (2008–2015), England (2012–2014), and Norway (2004–2018). Cancer detection rates (CDRs) and abnormal interpretation rates (AIRs) were calculated, with AIR defined as either reader flagging an examination as abnormal. Individual readers were divided into performance categories based on their high and low CDR and AIR. The performance of individuals determined the classification of pairs. Random pair performance, for which any type of pair was equally represented, was compared with the performance of specific pairing strategies, which consisted of pairs of readers who were either opposite or similar in AIR and/or CDR.

Results:

Based on a minimum number of examinations per reader and per pair, the final study sample consisted of 3,592,414 examinations (Sweden, n = 965,263; England, n = 837,048; Norway, n = 1,790,103). The overall AIRs and CDRs for all specific pairing strategies (Sweden AIR range, 45.5–56.9 per 1000 examinations and CDR range, 3.1–3.6 per 1000; England AIR range, 68.2–70.5 per 1000 and CDR range, 8.9–9.4 per 1000; Norway AIR range, 81.6–88.1 per 1000 and CDR range, 6.1–6.8 per 1000) were not significantly different from the random pairing strategy (Sweden AIR, 54.1 per 1000 examinations and CDR, 3.3 per 1000; England AIR, 69.3 per 1000 and CDR, 9.1 per 1000; Norway AIR, 84.1 per 1000 and CDR, 6.3 per 1000).

Conclusion:

Pairing a set of readers based on different pairing strategies did not show a significant difference in screening performance when compared with random pairing.

Introduction

Population-based breast cancer screening programs with mammography have proven to be effective in reducing breast cancer–specific mortality (1, 2). Nevertheless, breast cancer is the most commonly diagnosed cancer and a leading cause of cancer death among women worldwide (3). Radiologists miss 3%–40% of mammographically visible cancers (4-6). At the same time, false-positive screening results lead to unnecessary work-up, participant anxiety, and a reduction in cost-effectiveness (7).

A potential avenue for reducing the rate of errors in a screening program may be to optimize the double reading of screening mammograms. Double reading facilitates the interpretation of mammograms by two individuals with different cognitive, perceptual, and decision-making expertise. Previous studies have shown that double reading increases the cancer detection rate (CDR) when compared with single reading (8-10). As a result, the European Commission Initiative on Breast Cancer recommends double reading, and breast cancer screening programs in Australia, Europe, and New Zealand have implemented double reading (11). For breast cancer screening programs where mammograms are currently single read, it is possible that artificial intelligence (AI) may be added as a second reader in the future (12-15).

The pairing of radiologists who double read screening mammograms is currently assigned randomly, out of convenience, or to balance the workload. However, screening performance among radiologists varies (16). A previous multi-reader multi-case study with an enriched case set in a laboratory setting demonstrated that it was possible to improve the accuracy of mammography interpretation with double reading when the set of paired radiologists was optimized (17). Optimal pairing may thus be feasible, but no data were available on what factors determined the optimized set of pairs. To our knowledge, only one previous study investigated what prospective criteria could be used to pair radiologists optimally (18). In that study, Gandomkar et al. (18) suggested pairing radiologists with different cognitive eye-tracking metrics to optimize the pairings. However, eye-tracking data are not yet routinely available in screening programs.

Therefore, the purpose of this study was to determine if pairing optimization can be achieved based on the individual readers' performance characteristics that are routinely available in the screening program. The optimal set of pairs is defined as the one that results in the best overall screening performance, characterized by a high CDR and a low abnormal interpretation rate (AIR), for the entire case set. The pair composed of the two most accurate radiologists of the whole program may yield the best overall outcome if they read all examinations, but this is not realistic since the caseload needs to be divided evenly. Therefore, the aim of this study was to investigate whether radiologist performance characteristics can be used to determine the optimal set of pairs within a group of radiologists to double read screening mammograms.

Materials and Methods

This retrospective study was performed with deidentified retrospectively collected screening reading outcomes. Three data sets were used: the Swedish Cohort of Screen-Age Women, or CSAW, data set (19), the Changing case Order to Optimize patterns of Performance in Screening, or CO-OPS, data set from England (20, 21), and registry data from BreastScreen Norway to be able to compare different screening practices. All analyses were carried out separately on each data set, with no pooling performed at any point. The CSAW and CO-OPS data sets were used in previously published articles (15, 19-26), but these studies did not investigate different pairing strategies. The use of the CSAW data set for the purpose of research has been approved by the regional ethical review board, which waived the requirement for written informed consent. For the CO-OPS data set, ethical approval for the original trial was obtained from the Coventry and Warwickshire National Health Service Research Ethics Committee, and informed consent was obtained from each director of breast screening. The Norwegian data set was disclosed with legal bases in the Cancer Registry of Norway Regulations of December 21, 2001, no. 47, and no approval by an ethical board or informed consent was needed, as the project was considered a research quality assurance project.

Screening procedures

The screening programs of the three countries differ in several aspects (Fig 1).

The CSAW data set consists of women from Stockholm County who attended mammography screening between 2008 and 2015. Details of the data set have been described elsewhere (19, 22). Women 40–74 years of age were invited for two-view digital screening mammography every 18 or 24 months. Women older than 49 years were invited every 24 months, whereas younger women were invited every 18 months. The examinations were assessed with independent double reading. At most centers, the second radiologist was blinded to the assessment performed by the first radiologist. Screening examinations with concordant negative assessments were considered normal. Examinations with one or two abnormal assessments were flagged for consensus, where the final recall decision was made.

The CO-OPS data set includes women (predominantly aged 47–73 years) invited to two-view digital mammography screening between 2012 and 2014 at 46 breast screening centers throughout England. Women were invited every 3 years, and the mammograms were assessed by two expert readers (radiologists, advanced radiography practitioners, or breast clinicians), who independently decided whether the woman should be recalled. There are local variations in practice regarding blinding, but at most centers, the second reader was not blinded to the decision of the first. Screening examinations with concordant negative assessments were considered normal, and examinations with concordant positive readings were recalled. Discrepant readings were resolved by a third reader or group arbitration. At some centers with high recall rates, arbitration was also applied when both readers indicated recall. The original trial and protocol are published elsewhere (20, 21).

The data set from BreastScreen Norway includes data from women who attended digital mammography screening between 2004 and 2018. BreastScreen Norway offers women aged 50–69 years biennial two-view digital mammography. The screening examinations were independently assessed by two radiologists who were blinded to each other's assessment (27). Both radiologists assigned a score from 1 to 5 indicating the suspiciousness of mammographic findings (1, negative for malignancy; 5, high suspicion of malignancy). If both readers assigned a score of 1, the screening examination was considered normal. If either or both radiologists assigned a score of 2 or higher, a consensus meeting determined whether the woman should be recalled. A score of 2 or higher was also the threshold used in this study as a positive assessment.

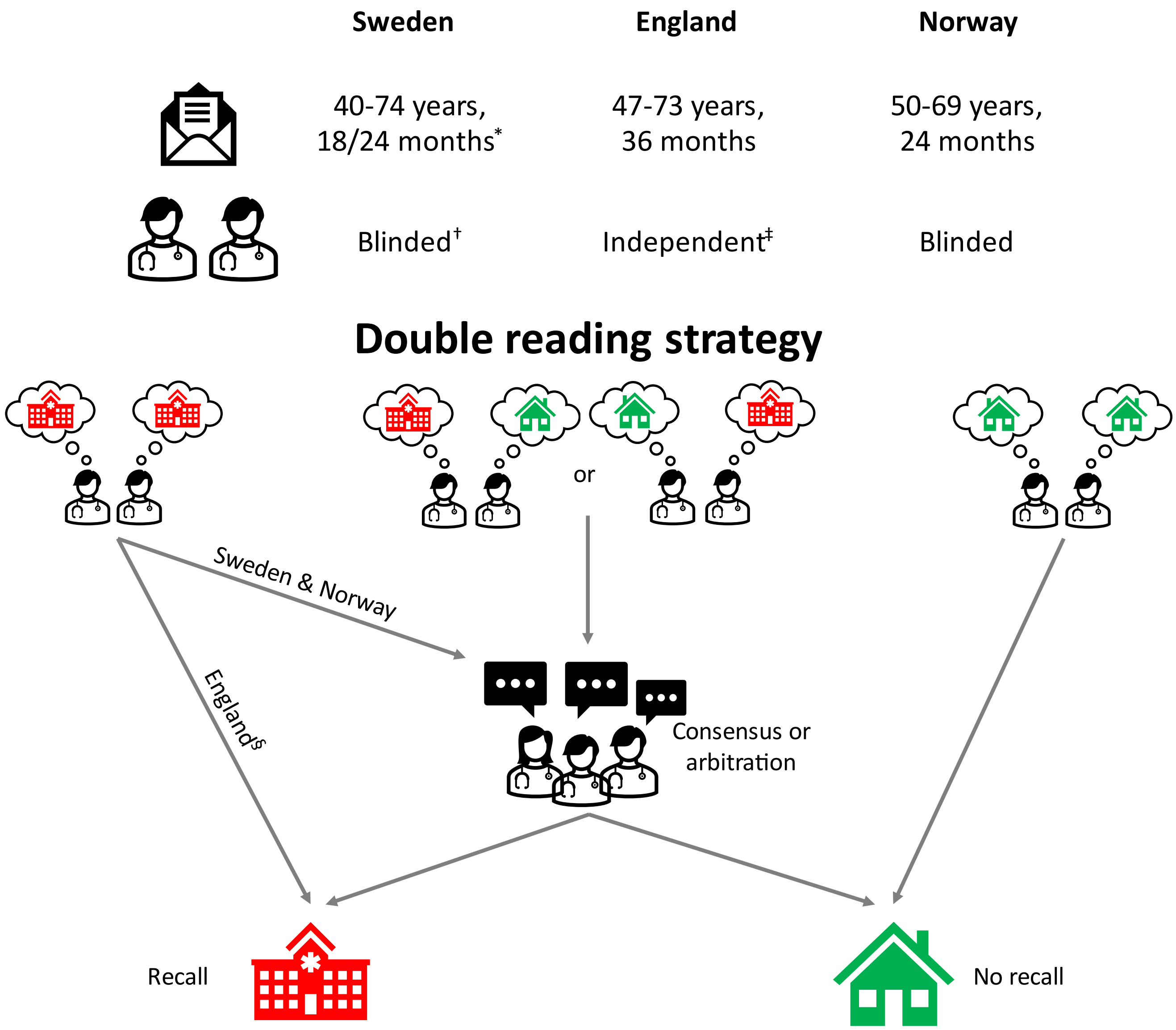

Figure 1: Schematic shows screening policy and double-reading strategy for the different data sets. * = Women older than 49 years were invited every 24 months, whereas younger women were invited every 18 months. † = There are local variations in practice regarding blinding, but at most centers, the second reader was blinded to the decision of the first. ‡ = There are local variations in practice regarding blinding, but at most centers, the second reader was not blinded to the decision of the first. § = At most centers, only discrepant readings were resolved with a third reader or group arbitration, but at some centers with high recall rates, arbitration was also applied when both readers indicated recall.

Study population

All data sets consisted of more than 1 million screening examinations (Table 1). Screening performance, based on the final recall decision, varied among the data sets. The differences were at least partly to be expected due to differences in screening policies. The recall rate and CDR of the Swedish data set were the lowest, and the recall rate and CDR of the English data set were the highest.

Table 1: Study Population Characteristics and Screening Performance for the Different Data Sets

Characteristic	Swedish Data Set	English Data Set	Norwegian Data Set
Study population characteristics			
No. of women screened	416,861	1,194,147	694,740
Age at screening (y)*	53 (46-62)†	59 (53-65)‡	59 (54-64)
Screening performance			
No. of screening examinations	1,180,828	1,194,147	2,230,225
Recalls	23.0§	41.5	34.2
Cancer detection rate	3.4§‖	8.8#	5.9#
Interval cancers	1.8†§**	1.9††	1.8††

Note.—Unless otherwise specified, data are numbers per 1000 examinations.
* Data are medians, with IQRs in parentheses.
† Age was unknown for seven screening examinations.
‡ Age was unknown for six screening examinations .
§ Final screening assessment was unknown for 106 screening examinations.
‖ Recall and breast cancer detection within 12 months after screening.
Breast cancer detected before the next screening examination as a result of recall at screening.
** Women who did not have a screening-detected cancer but had a breast cancer diagnosed within 18-24 months after screening.
†† Women who did not have a screening-detected cancer but had a breast cancer diagnosed before the next screening round. For the English data set, interval cancers were incomplete because not all data for interval cancers was known at the time of data extraction.

Breast cancer was defined as needle biopsy or surgery samples that tested positive for ductal carcinoma in situ or invasive cancer and was diagnosed either after further assessment in screening or clinically before the next screening examination (ie, interval cancer). For the English and Norwegian data sets, any breast cancer detected before the next screening examination was used for the analyses. For the Swedish data set, breast cancers detected within 18 and 24 months after screening for women aged 49 years or younger and older than 49 years, respectively, were used, as information on the exact date of the next screening examination was not available.

Analyses were performed on each data set separately. Reliable performance measures in a screening program with a low event rate can only be established in sufficiently large data sets (28). Data from readers with fewer than 500 interpretations in total, examinations with unknown reader data, examinations with recall because of clinical symptoms or inadequate images, and pairs with a relatively low volume in the data set (fewer than the mean number of interpretations per pair) were excluded. The mean number of interpretations per pair was used to have one consistent selection criterion for all three data sets.

Statistical analysis

The CDR (true-positive findings per 1000 examinations) and AIR ([true-positive findings + false-positive findings] per 1000 examinations) were calculated for the individual readers and pairs, and the performance of different pairing strategies was hypothesized (Fig 2). For the purposes of this study, AIR refers to the decision of the individual or the pair of readers, while recall rate refers to the final decision according to the program policy, including consensus or arbitration, if necessary.

Individual readers

The CDR and AIR of the individual readers were calculated. The readers were divided into four performance categories, using the individual weighted mean CDR and AIR as cutoffs (top left of Fig 2). These cutoffs were calculated for each of the three country data sets separately, and the weights were the number of examinations for each reader. Readers were characterized by high CDR and low AIR (HL), high CDR and AIR (HH), low CDR and AIR (LL), or low CDR and high AIR (LH).

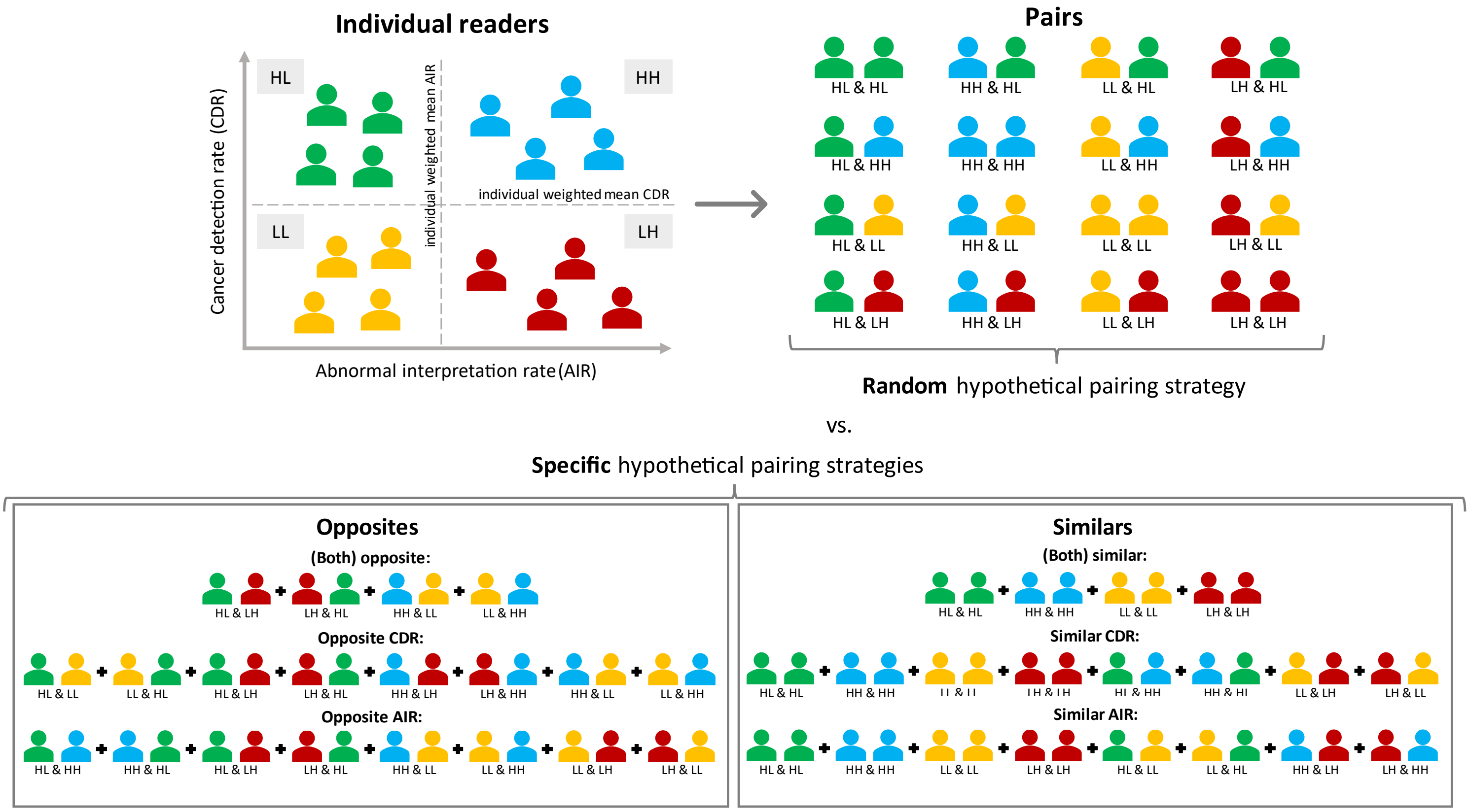

Figure 2: Diagram shows explanation for the evaluation of the screening performance of the individual readers and pairs, as well as an explanation of the hypothetical pairing strategies used to investigate the potential to improve outcomes with an appropriate double-reading strategy. HH = reader characterized by a high CDR and high AIR, HL = reader characterized by a high CDR and low AIR, LH = reader characterized by a low CDR and high AIR, LL = reader characterized by a low CDR and low AIR.

Pairs

CDR and AIR were also calculated for the pairs of readers. For paired reader assessments, the pairing rule shown in Figure 3 was used, where a positive assessment was defined by either or both readers flagging an examination as abnormal (ie, all concordant positive assessments or discrepant assessments were defined as a positive paired assessment). The use of this pairing rule was meant to optimize the performance in terms of cancer cases being, at least, sent to consensus or arbitration after double reading. Consensus discussion or arbitration itself was not considered, as the goal of the study was optimizing the original paired reading only. Pairs were classified based on the performance of the involved individuals. Based on the four different types of individual readers, 16 different types of pairs exist (top right of Fig 2). For each of these 16 types of pairs, the average CDR and AIR were calculated by taking the unweighted mean CDR and AIR of the unique pairs involved in that specific type (eg, the average AIR of the HL+HL pair type was calculated by averaging over the pairing rule–defined AIRs of the pairs that were classified as HL+HL).

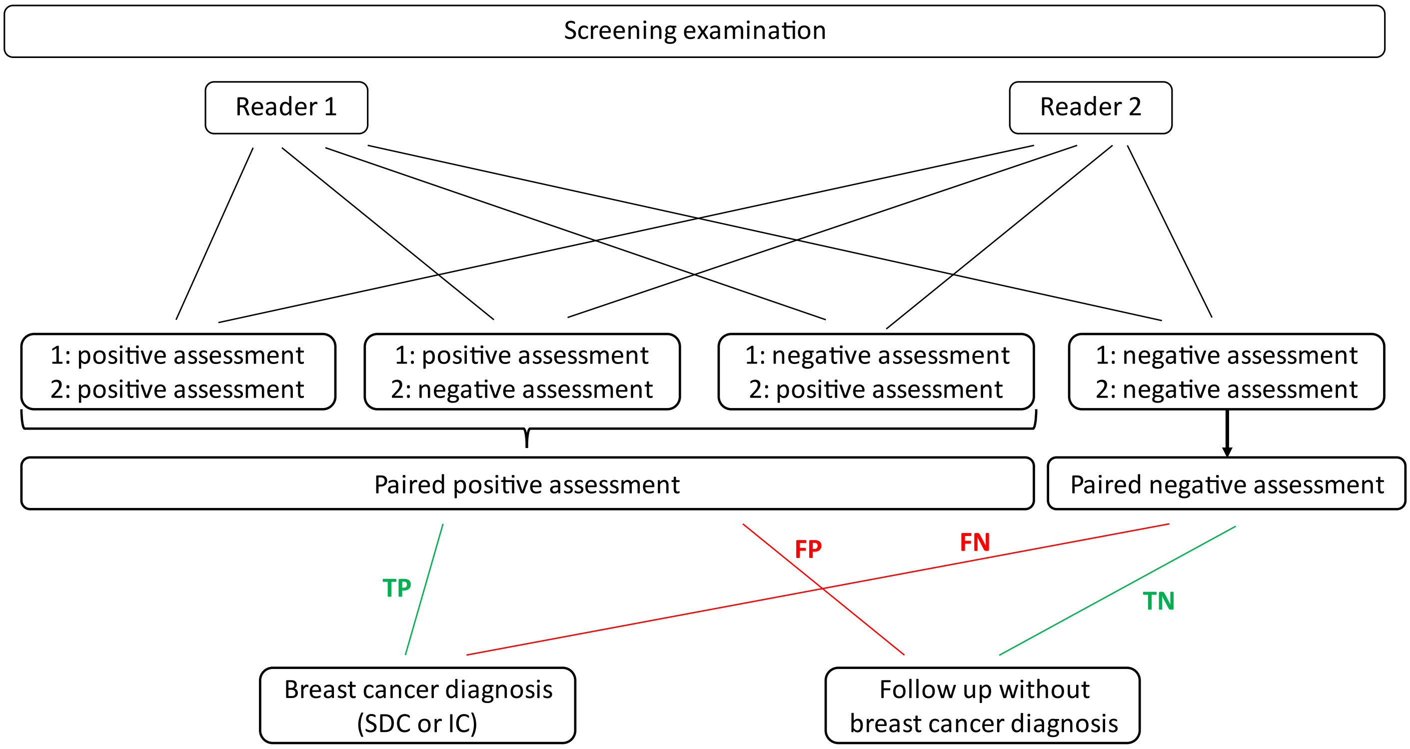

Figure 3: Flowchart shows the possible screening reading outcomes for the pairs in this study. Discrepant paired readings were classified as positive assessments. FN = false-negative, FP = false-positive, IC = interval cancer, SDC = screening-detected cancer, TN = true-negative, TP = true-positive.

Hypothetical pairing strategies

Random hypothetical pair performance was compared with the performance of specific hypothetical pairing strategies. For the analyses, the main interest was whether pairing readers with (a) opposite or (b) similar performance characteristics was beneficial. Based on the two performance measures (AIR and CDR), six specific pairing strategies were tested against the random pairing strategy (bottom of Fig 2):

- opposite AIR and CDR
- opposite CDR
- opposite AIR
- similar AIR and CDR
- similar CDR
- similar AIR

Random pair performance was estimated by averaging over the CDR and AIR of the 16 types of pairs, assuming each of the 16 types was equally represented (top right of Fig 2). Specific pair performance was estimated by averaging over the CDR and AIR of the types of pairs that belong to the specific pairing strategy (bottom of Fig 2). For example, a pairing strategy with opposite CDR readers in a pair consists of eight types of pairs (HL&LL, LL&HL, HL&LH, LH&HL, HH&LH, LH&HH, HH&LL, and LL&HH), which all involve one reader characterized by high CDR and one reader characterized by low CDR. The grouped screening performance measures of the six specific pairing strategies were compared with the performance measures of the random pairing strategy.

Bootstrap resampling of the screening examinations with corresponding readers (n = 1000) was used to obtain 95% CIs. For each bootstrap sample, all analysis steps were performed as described earlier, including the assessment of individual and paired performance, the classification of readers and pairs, and the evaluation of the hypothetical pairing strategies. Bonferroni-corrected P < .008 (.05/6) was regarded as indicating a statistically significant difference, and all analyses were performed in RStudio, version 4.1.0 (PBC).

Results

The final study sample of the Swedish, English, and Norwegian data sets included 965,263, 837,048, and 1,790,103 screening examinations, respectively (Fig 4).

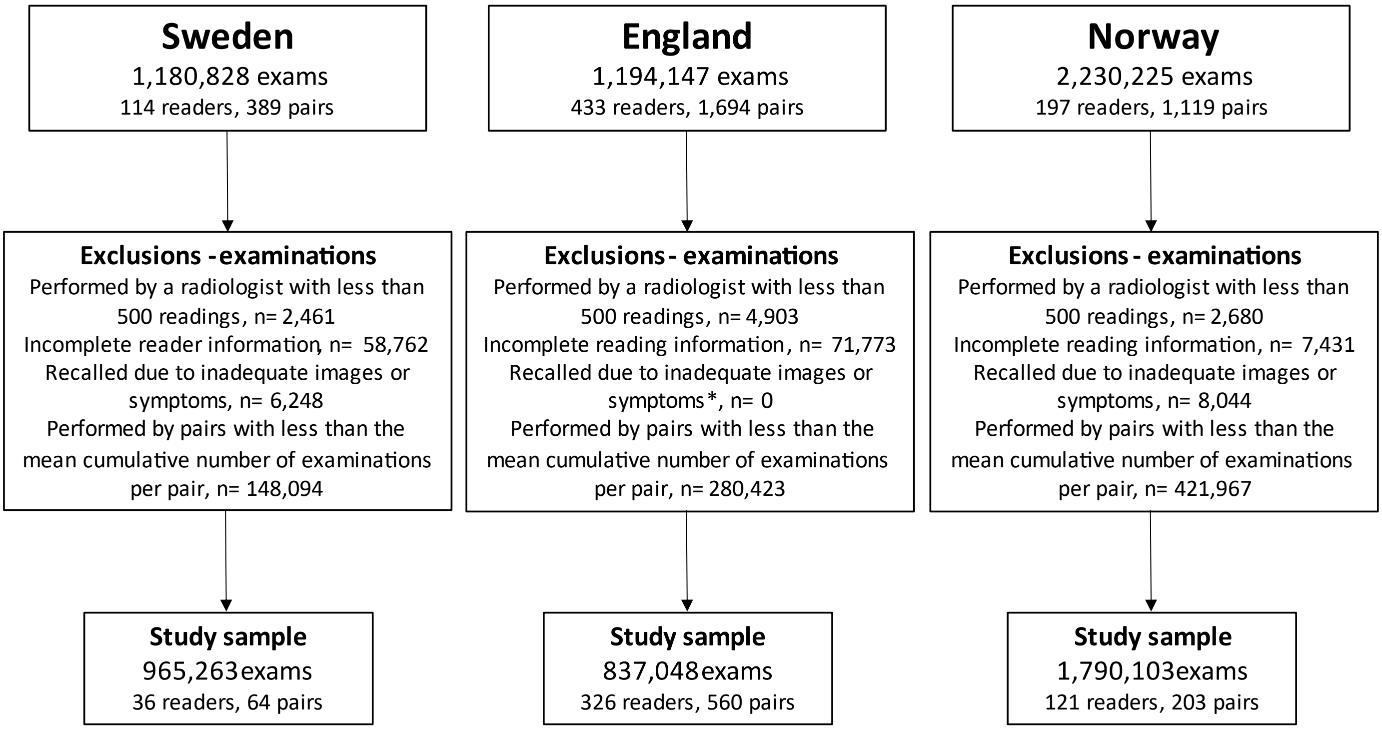

Figure 4: Flowchart of screening examinations after applying exclusion criteria. * = Examinations of women in the English data set who were recalled due to inadequate images or symptoms were already excluded before the data were received for this study.

Study sample characteristics

The study subsamples involved different numbers of readers and pairs with different screening performance (Table 2). The Swedish subsample included the youngest study sample. The readers in Sweden had the lowest individual weighted mean AIR and CDR (36.3 and 3.1 per 1000 examinations, respectively). The readers in the English subsample had the highest individual CDR, with 8.3 per 1000 examinations.

Paired AIR and CDR were higher than individual AIR and CDR for all three study subsamples. This is explained by the pairing rule for which any discrepant reading was classified as a positive reading, thereby increasing the AIR and CDR of the pairs when there is disagreement. Paired AIR and CDR were lowest for the Swedish subsample (47.8 and 3.4 per 1000 examinations, respectively), while paired CDR was highest for the English subsample (8.9 per 1000). The Norwegian subsample showed the most disagreement between the readers (5.6%), resulting in the highest paired AIR of 77.0 per 1000 examinations.

Table 2: Study Sample, Reader, and Pair Characteristics for the Study Subsamples after Selection Criteria

Characteristic	Swedish Data Set (n = 965,263)	English Data Set (n = 837,048)	Norwegian Data Set (n = 1,790,103)
Study sample characteristics			
No. of women screened	396,193	837,048	647,275
Age at screening (y)*	53 (46-62)	59 (53-65)†	59 (54-64)
Reader characteristics			
No. of readers	36	326	121
Individual weighted mean AIR	36.3	48.2	49.0
Individual weighted mean CDR	3.1‡	8.3	5.2
Cumulative reading volume*	27,346 (8,980-81,599)	4,524 (2,916-6,402)	20,259 (7,788-42,686)
Pair characteristics			
No. of pairs	64	560	203
Paired weighted mean AIR	47.8	63.7	77.0
Paired weighted mean CDR	3.4	8.9	6.0
Disagreement between two readers in a pair (%)	2.3	3.1	5.6
Cumulative reading volume*	10,120 (5,529-17,119)	1,240 (938-1,727)	5,612 (3,445-10,344)

Note.—Unless otherwise specified, data are numbers per 1000 examinations. AIR = abnormal interpretation rate, CDR = cancer detection rate.

* Data are medians, with IQRs in parentheses.

† Age was unknown for five screening examinations.

‡ Recall and breast cancer diagnosed within 18-24 months after screening, 18 months for women 49 years old or younger and 24 months for women older than 49 years.

Individual performance

Figure 5 shows the classification of the individual readers into four performance groups. The individual weighted mean AIR (Sweden, 36.3 per 1000 examinations; England, 48.2 per 1000; and Norway, 49.0 per 1000) and CDR (Sweden, 3.1 per 1000; England, 8.3 per 1000; and Norway, 5.2 per 1000) were used as cutoffs for the respective data sets (Table 2).

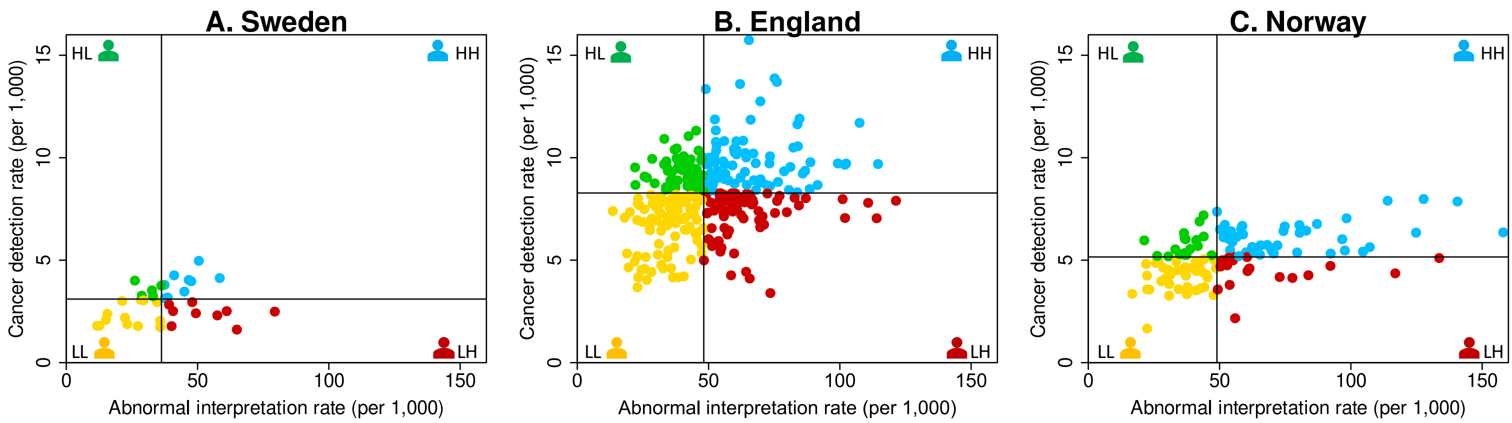

Figure 5: Quadrant graphs for individual screening performance in the (A) Swedish, (B) English, and (C) Norwegian data sets. The individual weighted mean performance was used as the cutoff. Readers were characterized by a high cancer detection rate (CDR) and low abnormal interpretation rate (AIR) (HL), high CDR and high AIR (HH), low CDR and low AIR (LL), or low CDR and high AIR (LH).

Paired performance

Figure 6 shows the resulting paired AIRs and CDRs for the 16 specific pair types. The pattern for all three study subsamples looks similar. Pairs consisting of two high-AIR readers (blue and red) resulted in high paired AIR values (>63, >87, and >103 per 1000 examinations for the Swedish, English, and Norwegian subsamples, respectively), and pairs consisting of two low-AIR readers (green and yellow) resulted in low paired AIR values (<38, <51, and <59 per 1000 for the Swedish, English, and Norwegian subsamples, respectively). The same applies to pairs consisting of two high-CDR readers (green and blue) or two low-CDR readers (yellow and red), resulting in high (>3.8, >10.5, and >6.2 per 1000 examinations for the Swedish, English, and Norwegian subsamples, respectively) or low (<3.1 , <7.6, <5.7 per 1000 for the Swedish, English, and Norwegian subsamples, respectively) paired CDRs, respectively. Pairs consisting of opposite AIR and/or CDR readers resulted in average paired AIRs and CDRs compared with the other pair types.

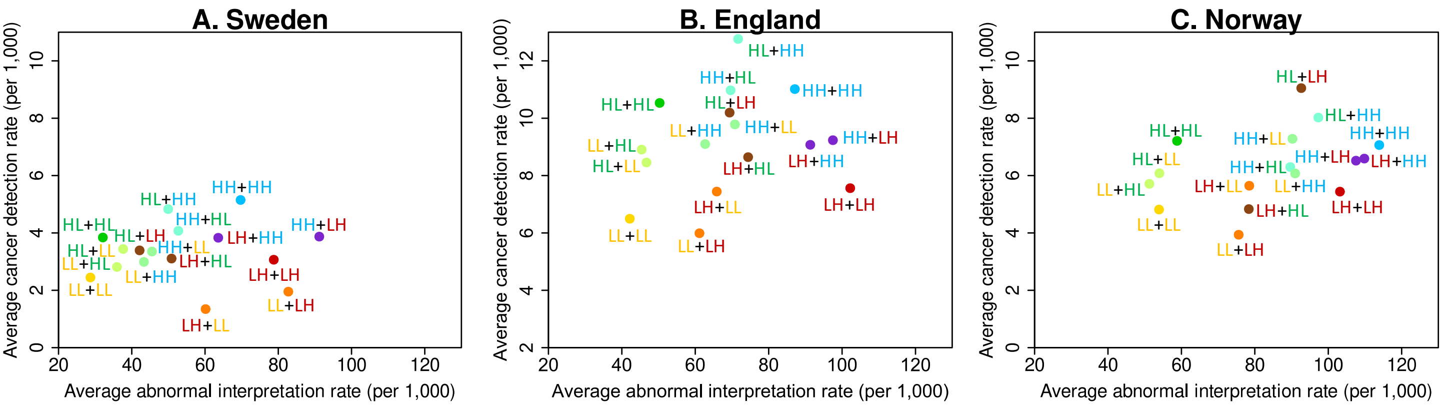

Figure 6: Scatterplots with the average cancer detection rate (CDR) and abnormal interpretation rate (AIR) of the 16 specific pairs in the (A) Swedish, (B) English, and (C) Norwegian data sets. HH = reader characterized by a high CDR and high AIR, HL = reader characterized by a high CDR and low AIR, LH = reader characterized by a low CDR and high AIR, LL = reader characterized by a low CDR and low AIR.

Hypothetical set of pairs

To find out if individual reader performance characteristics can be leveraged to identify the optimal set of pairs, random hypothetical pair performance was compared with the hypothetical group performance of six specific pairing strategies. The group AIR and CDR of the random pairing strategies were 54.1 per 1000 examinations (95% CI: 46.1, 62.1) and 3.3 per 1000 (95% CI: 2.8, 3.9) for the Swedish subsample, 69.3 per 1000 (95% CI: 63.7, 74.9) and 9.1 per 1000 (95% CI: 8.3, 9.9) for the English subsample, and 84.1 per 1000 (95% CI: 80.3, 88.0) and 6.3 per 1000 (95% CI: 5.9, 6.7) for the Norwegian subsample (Fig 7). The CIs from the grouped AIRs and CDRs of the specific pairing strategies overlapped with those from the random pairing strategy. The group AIRs and CDRs for the specific pairing strategies were thus not statistically significantly different from the random pairing strategy in all three study subsamples (Table 3).

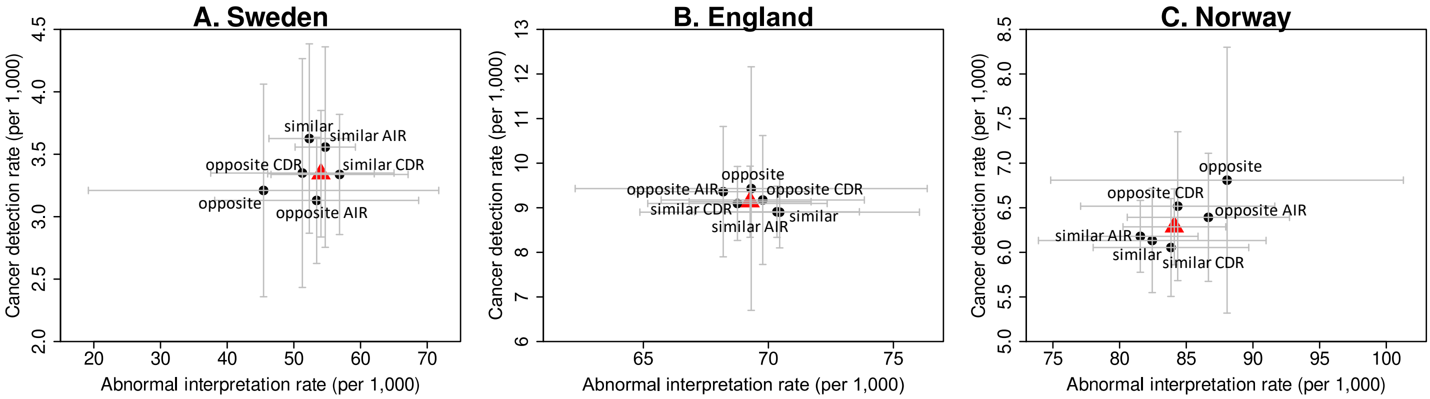

Figure 7: Group screening performance for the different pairing strategies in the (A) Swedish, (B) English, and (C) Norwegian data sets. The triangles (red) represent the average screening performance for the random hypothetical pairing strategy, and the dots represent the performance for the specific hypothetical pairing strategies (black). Error bars are Bonferroni-adjusted 95% CIs obtained by means of bootstrap resampling (n = 1000). Please note that the axes are different due to the differences in cancer detection rate (CDR) and abnormal interpretation rate (AIR) for the data sets.

Table 3: Group Screening Performance for the Different Pairing Strategies

	Swedish Data Set		English Data Set		Norwegian Data Set	
Pairing strategy	AIR	CDR	AIR	CDR	AIR	CDR
Random	54.1 (46.1, 62.1)	3.3 (2.8, 3.9)	69.3 (63.7, 74.9)	9.1 (8.3, 9.9)	84.1 (80.3, 88.0)	6.3 (5.9, 6.7)
Both opposite	45.5 (19.2, 71.7)	3.2 (2.4, 4.1)	69.3 (65.7, 72.9)	9.4 (6.7, 12.2)	88.1 (74.8, 101.3)	6.8 (5.3, 8.3)
Opposite CDR	51.3 (37.5, 65.0)	3.3 (2.4, 4.3)	69.8 (66.5, 73.1)	9.2 (7.7, 10.6)	84.4 (77.1, 91.6)	6.5 (5.7, 7.4)
Opposite AIR	53.4 (38.1, 68.7)	3.1 (2.6, 3.6)	68.2 (61.2, 75.2)	9.4 (7.9, 10.8)	86.7 (80.6, 92.7)	6.4 (5.7, 7.1)
Both similar	52.3 (46.3, 58.4)	3.6 (2.9, 4.4)	70.5 (66.4, 74.5)	8.9 (8.1, 9.7)	82.4 (73.9, 91.0)	6.1 (5.5, 6.7)
Similar CDR	56.9 (46.6, 67.1)	3.3 (2.9, 3.8)	68.8 (65.2, 72.3)	9.1 (8.3, 9.9)	83.9 (78.0, 89.7)	6.1 (5.5, 6.6)
Similar AIR	54.7 (50.2, 59.2)	3.6 (2.8, 4.4)	70.3 (67.9, 72.8)	8.9 (8.3, 9.5)	81.6 (77.2, 85.9)	6.2 (5.8, 6.6)

Note.—Data in parentheses are 95% CIs. Abnormal interpretation rate (AIR) and cancer detection rate (CDR) are given per 1000 examinations, and 95% CIs are Bonferroni-adjusted ($P < .05/6$) and were obtained by bootstrap resampling (n = 1000).

Discussion

Our retrospective study showed that performance characteristics of mammography readers influenced the performance of pairs, but specific pairing strategies did not result in significantly different overall performance compared with that resulting from random pairing strategies. The specific pairing strategies included some higher-performing pairs as well as lower-performing pairs that together balanced the overall screening performance of the pairing strategies to abnormal interpretation rates and cancer detection rates that were very similar.

In most countries, the picture archiving and communication system provides opportunities for strategic pairing, as readers can assess the screening mammograms from different locations. Although a previous study by Brennan and colleagues (17) demonstrated that some pairing schemes were better than others, their study was not designed to identify the criteria that can be used prospectively to determine the best pairing schemes. In our study, we attempted to determine if this pairing optimization can be achieved based on the readers' individual performance, but we were unable to identify a pairing rule that consistently and significantly improved the overall program performance. This could be due to the actual underlying optimal criterion not being the individual performance of the readers, the inconsistencies across screening programs introducing differences in the predicted outcomes, the classification of the readers, or the data sets having too few reads per examination to exhaustively explore the different pairing strategies.

First, the true optimal pairing criteria may be related to factors other than individual performance. For example, different readers may be better at detecting specific types of findings (eg, calcifications vs soft-tissue lesions vs architectural distortions), and therefore, pairing based on those differing abilities would yield the optimal program performance, as opposed to the criteria investigated herein. The concept of AI as a second reader is also an upcoming and promising method that has gained strong attention in the past decade, and this could be a way to incorporate double reading into single-reading breast cancer screening programs. A potential implementation of AI as a second reader could involve adjusting the AI operating point settings to the performance of the paired human reader to counter the reader's operating point and hence optimize screening performance. Future research should therefore focus on what pairing criteria may optimize screening performance for both double-reading programs as well as human single-reading programs with AI as a second reader.

Second, the different policies of the screening programs may introduce differences in performance measures among the three subsamples, which then confound the results of the optimization process investigated herein. Swedish readers had the lowest individual AIR and CDR, probably due to the younger group of women who were screened every 18 months. The English readers had the highest CDR, probably because of the screening interval of 36 months. Nevertheless, individual AIR was not the highest for the English subsample, but for the Norwegian subsample. One reason for this may be that radiologists in England are more reluctant to flag an examination as abnormal because they know that the woman will be recalled if the other radiologist also decides to recall, whereas in Norway, consensus will always decide on recall, even after two positive reader assessments. Paired performance showed that Norwegian pairs disagreed more than Swedish and English pairs. This may be because all Norwegian readers were blinded to each other's assessment, whereas for the Swedish and English subsamples, not all readers were blinded.

Furthermore, the performance measures of some of the individual readers are close to the predefined cutoff line. Therefore, although dichotomized as high or low CDR and/or AIR, those readers might actually perform very similarly to others who are just on the other side of the corresponding threshold. Therefore, if there actually were an impact on overall performance by pairing readers with specific CDR and/or AIR characteristics, these would be challenging to tease out with data sets where the actual CDR and/or AIR differences are small.

Finally, although this study consisted of large study samples, the individual examinations were read by only one pair of readers, and therefore, there might not have been enough pair realizations to identify prospective selection criteria for the optimization of pairs. Thus, it could be helpful to exhaustively evaluate all theoretically possible pairs. Research should therefore focus on simulating individual radiologist assessments, making it possible to analyze data consisting of results of all readers reading all cases, and on exploring possible optimal pairing strategies.

This study has several strengths and limitations. A major strength is that this study involved data from actual screening programs, in which reading behavior influenced care. Furthermore, this study was able to compare three screening programs. However, our study was affected by incorporation bias, because if a reader selected recall, then cancer was more likely detected because of follow-up tests. This bias was reduced by taking into account interval cancers and therefore identifying the false-negative findings in the screening data sets. While this biased overall cancer detection accuracy upwards, it was unlikely to impact the comparisons in this study, as this effect would have applied to all readers. Furthermore, the hypothetical pairing strategies rely on the assumption that each type of reader was equally represented in the pairings. This allowed us to make a fair comparison of the different pairing strategies without being influenced by what type of readers were included (eg, including a larger number of high CDR and low AIR than low CDR and high AIR readers would automatically result in better performance independent of the pairing strategy). In actual screening practice, there will be variation in the number and type of readers involved, which should be considered when interpreting our results. In addition, there is some variation in the blinding of the second reader, both across and within the different study subsamples. We did not have information on which readers were blinded and how much experience the readers had, so we could not control for these differences.

In conclusion, this study shows that the type of readers involved in a pair influences paired screening performance. Nevertheless, pairing strategies based on cancer detection rate and abnormal interpretation rate performance characteristics for the full set of pairs did not consistently show significant differences in grouped screening performance. Future studies should include data sets with screening examinations read by more than two readers and test pairing strategies with a variable number of reader types to explore the possibility of improving overall screening performance with different pairing strategies.

Acknowledgements

The CSAW data set was provided by Fredrik Strand, MD, PhD, and has been approved by the regional ethical review board for research. The CO-OPS data set, funded by an NIHR Postdoctoral Fellowship and an NIHR Career Development Fellowship (CDF-2016-09-018), was provided by Sian Taylor-Phillips, PhD, and David Jenkinson, PhD. The data set from BreastScreen Norway was disclosed with legal bases in the Cancer Registry of Norway regulations and provided by Solveig Hofvind, PhD, and Marthe Larsen, MSc.

Author contributions

Guarantors of integrity of entire study, J.J.J.G., M.J.M.B.; study concepts/study design or data acquisition or data analysis/ interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, J.J.J.G., S.H., I.S., M.J.M.B.; statistical analysis, J.J.J.G., D.J.J.; and manuscript editing, J.J.J.G., C.K.A., S.T.P., M.L., S.H., I.S., M.J.M.B.

Data sharing

Data analyzed during the study were provided by a third party. Requests for data should be directed to the provider indicated in the Acknowledgements.

Disclosures of conflicts of interest

J.J.J.G. No relevant relationships. C.K.A. Consultant at Canon Medical; advisory board member for Izotropic. F.S. Payment for lectures from Lunit; president of the Research and Education Committee of the Swedish Society of Breast Imaging. S.T.P. Funded by an NIHR Postdoctoral Fellowship and an NIHR Career Development Fellowship (CDF2016-09-018). D.J.J. Grants from the National Institute of Health and Care Research. M.L. No relevant relationships. S.H. Head of BreastScreen Norway. I.S. Grants or grants pending from Siemens Healthcare, Canon Medical, ScreenPoint Medical, Sectra Benelux, Hologic, Volpara Solutions, Lunit, and iCAD; payment for lectures including service on speakers' bureaus from Siemens Healthcare; member of the scientific advisory board of Koning; member of the Radiology editorial board. M.J.M.B. Grants or grants pending from ScreenPoint Medical, Sectra Benelux, Hologic, Volpara Solutions, Lunit, and iCAD; personal speaker fees from Hologic and Siemens Healthcare.

References

1. Broeders M, Moss S, Nyström L, Njor S, Jonsson H, Paap E, et al. The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies. J Med Screen. 2012;19 Suppl 1:14-25.
2. Myers ER, Moorman P, Gierisch JM, Havrilesky LJ, Grimm LJ, Ghate S, et al. Benefits and Harms of Breast Cancer Screening: A Systematic Review. Jama. 2015;314(15):1615-1634.
3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-249.
4. Broeders MJ, Onland-Moret NC, Rijken HJ, Hendriks JH, Verbeek AL, Holland R. Use of previous screening mammograms to identify features indicating cases that would have a possible gain in prognosis following earlier detection. Eur J Cancer. 2003;39(12):1770-1775.
5. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. Radiographics. 2003;23(4):881-895.
6. Weber RJ, van Bommel RM, Louwman MW, Nederend J, Voogd AC, Jansen FH, et al. Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. Breast Cancer Res Treat. 2016;158(3):471-483.
7. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. Ann Intern Med. 2007;146(7):502-510.
8. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. AJR Am J Roentgenol. 2003;180(5):1461-1467.
9. Coolen AMP, Voogd AC, Strobbe LJ, Louwman MWJ, Tjan-Heijnen VCG, Duijm LEM. Impact of the second reader on screening outcome at blinded double reading of digital screening mammograms. Br J Cancer. 2018;119(4):503-507.
10. Taylor P, Potts HW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer. 2008;44(6):798-807.
11. European Commission Initiatives on Breast and Colorectal Cancer. Use of double reading in mammography screening. https://healthcare-quality.jrc.ec.europa.eu/ecibc/european-breast-cancer-guidelines/organisation-of-screening-programme/double-reading-in-mammography-screening. Published 2019. Accessed September 2022.
12. Marinovich ML, Wylie E, Lotter W, Pearce A, Carter SM, Lund H, et al. Artificial intelligence (AI) to enhance breast cancer screening: protocol for population-based cohort study of cancer detection. BMJ Open. 2022;12(1):e054005.
13. Sharma N, Ng AY, James JJ, Khara G, Ambrozay E, Austin CC, et al. Retrospective large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. medRxiv. 2022:2021.02.26.21252537. Posted June 10, 2022. Accessed January 2023.
14. Larsen M, Aglen CF, Hoff SR, Lund-Hanssen H, Hofvind S. 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(12):8238-8246.

15. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncol. 2020;6(10):1581-1588.
16. Klompenhouwer EG, Duijm LE, Voogd AC, den Heeten GJ, Nederend J, Jansen FH, et al. Variations in screening outcome among pairs of screening radiologists at non-blinded double reading of screening mammograms: a population-based study. Eur Radiol. 2014;24(5):1097-1104.
17. Brennan PC, Ganesan A, Eckstein MP, Ekpo EU, Tapia K, Mello-Thoms C, et al. Benefits of Independent Double Reading in Digital Mammography: A Theoretical Evaluation of All Possible Pairing Methodologies. Acad Radiol. 2019;26(6):717-723.
18. Gandomkar Z, Tay K, Brennan PC, Kozuch E, Mello-Thoms C. Can eye-tracking metrics be used to better pair radiologists in a mammogram reading task? Med Phys. 2018;45(11):4844-4856.
19. Dembrower K, Lindholm P, Strand F. A Multi-million Mammography Image Dataset and Population-Based Screening Cohort for the Training and Evaluation of Deep Neural Networks-the Cohort of Screen-Aged Women (CSAW). Journal of digital imaging. 2020;33(2):408-413.
20. Taylor-Phillips S, Wallis MG, Parsons H, Dunn J, Stallard N, Campbell H, et al. Changing case Order to Optimise patterns of Performance in mammography Screening (CO-OPS): study protocol for a randomized controlled trial. Trials. 2014;15:17.
21. Taylor-Phillips S, Wallis MG, Jenkinson D, Adekanmbi V, Parsons H, Dunn J, et al. Effect of Using the Same vs Different Order for Second Readings of Screening Mammograms on Rates of Breast Cancer Detection: A Randomized Clinical Trial. Jama. 2016;315(18):1956-1965.
22. Salim M, Dembrower K, Eklund M, Lindholm P, Strand F. Range of Radiologist Performance in a Population-based Screening Cohort of 1 Million Digital Mammography Examinations. Radiology. 2020;297(1):33-39.
23. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, 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(9):e468-e474.
24. Dembrower K, Liu Y, Azizpour H, Eklund M, Smith K, Lindholm P, et al. Comparison of a Deep Learning Risk Score and Standard Mammographic Density Score for Breast Cancer Risk Prediction. Radiology. 2020;294(2):265-272.
25. Taylor-Phillips S, Jenkinson D, Stinton C, Wallis MG, Dunn J, Clarke A. Double Reading in Breast Cancer Screening: Cohort Evaluation in the CO-OPS Trial. Radiology. 2018;287(3):749-757.
26. Cooper JA, Jenkinson D, Stinton C, Wallis MG, Hudson S, Taylor-Phillips S. Optimising breast cancer screening reading: blinding the second reader to the first reader's decisions. Eur Radiol. 2022;32(1):602-612.
27. Wølner Bjørnson E, Sørlien Holen Å, Hofvind S, Sagstad S, Thy J, Larsen M, et al. BreastScreen Norway: 25 years of organized screening. Oslo, Norway: Cancer Registry of Norway; 2022. https://www.kreftregisteret.no/globalassets/ mammografiprogrammet/rapporter-og-publikasjoner/2022- 25-arsrapport_webversjon.pdf.
28. Burnside ES, Lin Y, Munoz del Rio A, Pickhardt PJ, Wu Y, Strigel RM, et al. Addressing the challenge of assessing physician-level screening performance: mammography as an example. PLoS One. 2014;9(2):e89418.