Improving Screening Mammography Interpretation

Real Impact Within Reach

Jessie J.J. Gommers

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Improving Screening Mammography Interpretation

Real Impact Within Reach

Jessie Joke José Gommers

The research described in this thesis was carried out at the Advanced X-ray Tomographic Imaging (AXTI) group, Radboud university medical center, Nijmegen, the Netherlands. Financial support for the research described in this thesis was provided by KWF Dutch Cancer Society and NWO Domain AES, as part of their joint strategic research program Technology for Oncology II (project number 17912). The collaboration project was co-funded by the PPP Allowance made available by Health Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.

Improving Screening Mammography Interpretation: Real Impact Within Reach
Jessie Gommers

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Proefschrift AIO | Annelies Lips
Cover: Proefschrift AIO | Guntra Laivacuma
Printing: DPN Rikken/Pumbo

ISBN: 9789465152790
DOI: 10.54195/9789465152790
Free download at: https://doi.org/10.54195/9789465152790

RADBOUD
UNIVERSITY
PRESS

Improving Screening Mammography Interpretation

Real Impact Within Reach

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 6 juli 2026
om 12.30 uur precies

door

Jessie Joke José Gommers
geboren op 27 mei 1997
te Eindhoven

Promotoren:
Prof. dr. I. Sechopoulos
Prof. dr. M.J.M. Broeders

Copromotoren:
Dr. C.K. Abbey (University of California - Santa Barbara, Verenigde Staten)
Dr. L.E.M. Duijm (CWZ)

Manuscriptcommissie:
Prof. dr. I.D. Nagtegaal
Prof. dr. C.H. van Gils (Universiteit Utrecht)
Dr. F. Kilburn-Toppin (Cambridge University Hospital, Verenigd Koninkrijk)

Improving Screening Mammography Interpretation

Real Impact Within Reach

Dissertation to obtain the degree of doctor

from Radboud University Nijmegen

on the authority of the Rector Magnificus prof. dr. J.M. Sanders,

according to the decision of the Doctorate Board

to be defended in public on

Monday, July 6, 2026

at 12:30 pm

by

Jessie Joke José Gommers

born on May 27, 1997

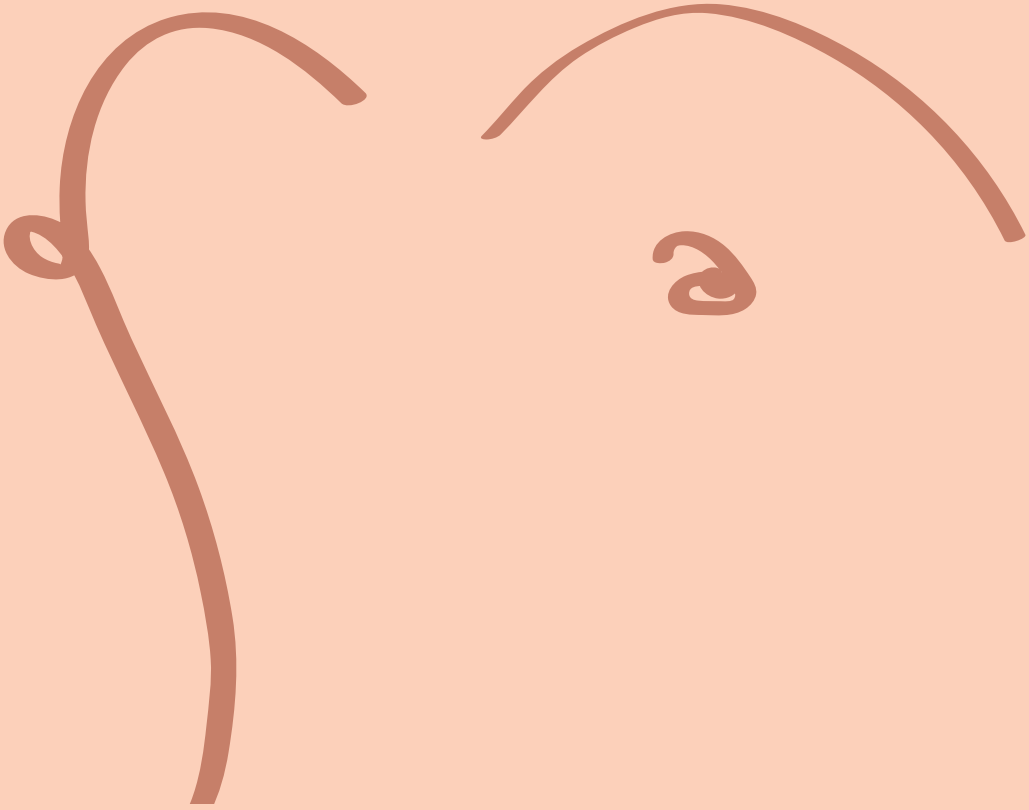
in Eindhoven

PhD supervisors:
Prof. dr. I. Sechopoulos
Prof. dr. M.J.M. Broeders

PhD co-supervisors:
Dr. C.K. Abbey (University of California, Santa Barbara, United States)
Dr. L.E.M. Duijm (CWZ)

Manuscript Committee:
Prof. dr. I.D. Nagtegaal
Prof. dr. C.H. van Gils (Utrecht University)
Dr. F. Kilburn-Toppin (Cambridge University Hospital, United Kingdom)

Table of Contents



Chapter 1

General Introduction

Breast cancer is a major public health problem, primarily because it is the most frequent cancer in women, with over 2 million new cases and more than 650,000 deaths in 2022 worldwide [1]. According to the latest numbers, 17,620 women were diagnosed with breast cancer, and 3,065 women died as a result of breast cancer in the Netherlands in 2023 [2]. It is estimated that 1 in 7 women in the Netherlands will develop breast cancer by the age of 80, making it the most prevalent cancer among women [2–4]. The survival prospects depend on the stage and characteristics of the breast tumor at diagnosis. For patients diagnosed with early-stage breast cancer, 96% are still alive after 10 years, whereas for patients diagnosed with breast cancer that has spread outside the breast, the 10-year relative survival rate is only 13% [2].

1.1 Breast cancer screening

Breast cancer screening programs aim to reduce breast cancer-related morbidity and mortality through early detection and treatment of breast cancer in asymptomatic women. The goal is to find clinically relevant cancers early, when they are still small and less likely to have spread outside the breast, thereby maximizing the chances of successful and less aggressive treatment. The first breast cancer screening programs started in the late 1980s, and since then, many developed countries have implemented nationwide programs. A systematic review has shown that breast cancer screening in Europe reduced breast cancer mortality in participants versus non-participants by between 12% and 58% [5]. In the Netherlands, more than 30 years after the introduction of breast cancer screening in 1989 and alongside several treatment improvements, breast cancer mortality has been reduced by 45.6% [6].

1.1.1 Screening with digital mammography

Digital mammography (DM) is currently the most common method for the early detection of breast cancer [7]. It is a quick, non-invasive procedure that is relatively affordable and widely available. The technique uses low-dose x rays to create two-dimensional (2D) images of the breast.

During a mammographic examination, a woman's breast is placed between two parallel plates: a support table at the bottom and a compression paddle at the top (Figure 1.1). The support table helps keep the breast steady while the compression paddle applies force to compress the breast from above. This compression reduces the breast thickness, thereby lowering the required radiation dose, decreasing

scattered radiation, and reducing tissue overlap, which improves visualization of the breast tissue [8]. It also stabilizes the breast, preventing movement during the procedure. Once the breast is properly positioned, an x-ray source emits a low dose of x rays that passes through the compressed breast, with different tissues absorbing them to varying degrees [8]. The detector on the opposite side of the breast captures the attenuated x-ray signal and converts it into electronic signals, which are sent to a computer and processed to create 2D images of the breast [8].

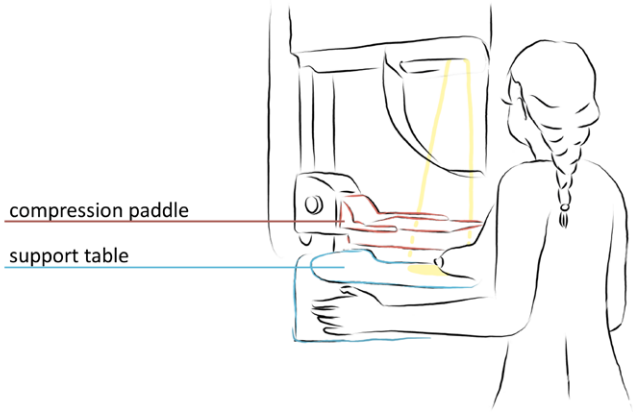


Figure 1.1: Illustration of a mammography system, depicting how a woman's breast is positioned between the support table and compression paddle.

Routine screening typically involves two views of each breast: the mediolateral oblique (MLO) view (angled from the side) and the craniocaudal (CC) view (from top to bottom) (Figure 1.2). Capturing the breast from two angles helps to reduce tissue overlap, which in a single 2D image could either hide a lesion or create the illusion of one [9]. Additional views of the breast from different angles may also be obtained if, for example, certain areas of the breast are not optimally visualized, or to determine if a suspicious-looking area consists of overlapping tissue or a true lesion.

When interpreting screening mammograms, radiologists examine both the CC and MLO views, along with any additional views obtained, to identify lesions suspicious for breast cancer. A lesion may be visible in one or more views, and normal tissue may mimic a lesion in some views but not others, highlighting the importance of using all available views for accurate interpretation. Radiologists describe their findings using standardized morphological descriptors, including 1) masses, high-density areas with different shape and margin descriptors; 2) architectural distortions, distorted patterns of breast tissue with no definite mass visible; 3) asymmetries, dense tissue patterns in one of the breasts not conforming to the definition of a

mass and not corresponding with the dense patterns in the contralateral breast; and 4) calcification clusters, tiny, bright, white deposits of calcium [10] (Figure 1.3). A thorough interpretation of mammograms is important to detect and distinguish between benign and malignant lesions. Whenever possible, radiologists compare the current mammograms to the prior ones to see if morphologic findings are new or have changed over time. Suspicious findings that remain unchanged over time are generally considered less concerning [9].

Figure 1.2: Example of standard mammographic views. The top two images show the right and left breasts in the mediolateral (MLO) view, and the bottom two images show the right and left breasts in the craniocaudal (CC) view.

Figure 1.3: Examples of mammographic abnormalities. Shown from left to right are: mass, architectural distortion, asymmetry, and calcifications.

If a suspicious finding is detected during screening, the woman is recalled for additional diagnostic workup. This workup may involve additional imaging, including DM, tomosynthesis (pseudo-3D digital mammography), and/or ultrasound, or, in some cases, contrast-enhanced mammography, and/or contrast-enhanced magnetic resonance imaging (MRI). If imaging confirms a suspicious abnormality, a biopsy is performed. During a biopsy, a small sample of the suspicious lesion is extracted. This is done to be able to send the extracted tissue to a pathology laboratory for analysis, where the results determine whether breast cancer is present.

1.1.2 Breast cancer screening programs around the world

Screening mammography can be offered as part of an organized program or through opportunistic screening [11]. In organized screening programs, a national or regional government body holds overall responsibility, delegating to the screening organization the task of periodically inviting the target population of women to a dedicated screening center or hospital for screening mammography. In opportunistic screening, women visit private practices or hospital-affiliated institutions to undergo screening mammography, either at their own request or upon referral from a physician. Most high-income countries, especially in Europe, have implemented mammographic screening through organized, population-based screening programs. Other countries, such as the United States, primarily rely on opportunistic screening [9].

Breast cancer screening policies and guidelines vary across countries, differing in the starting and stopping ages, the time interval between screening rounds, and the reading strategies used to interpret the mammograms [12]. In most European countries, screening is offered from the age of 40–50 years until the age of 69–75 years, with intervals of 1.5 to 3 years, and mammograms being double-read by two radiologists [11,13,14]. This aligns with the Guidelines of the European Commission

Initiative on Breast Cancer (ECIBC), which recommend biennial mammography screening for women aged 50–69 years and suggest biennial or triennial screening for women aged 45–49 and 70–74 years, with all mammograms double-read by two radiologists [15]. Similar to most European countries, Australia follows a double-reading approach, inviting women aged 50-74 years every two years [16]. The United States does not have an organized population-based screening program and has multiple guidelines, most of which recommend annual screening for women aged 45–54 years and biennial screening for those aged 55 years and older, with single instead of double reading, and mostly using digital breast tomosynthesis instead of DM [17].

1.1.3 Breast cancer screening in the Netherlands

In the Netherlands, women aged between 50 and 75 years are invited to participate in breast cancer screening. The aim is to invite women for DM every two years. However, due to the COVID-19 pandemic and a shortage of screening radiographers, the screening interval can be extended up to 3 years since 2020 [18]. In 2024, the average screening interval was 28 months [6]. In that same year, 1,338,520 women were eligible for participation, of whom 874,391 chose to participate, resulting in a participation rate of 65.3% [6].

Screening is performed at one of the 71 screening centers, most of which are mobile mammography units that travel around the country to ensure accessibility for all women [3]. During the screening process, trained radiographers take mammograms of both breasts in the CC and MLO views, with additional views obtained as necessary. Within one or two days after acquisition, two trained breast radiologists independently assess the mammograms and decide whether a woman needs to be recalled to a hospital for further diagnostic workup [19]. To standardize the assessment, the Breast Imaging Reporting and Data System, or BI-RADS, is used [10]. The BI-RADS was developed by the American College of Radiology to facilitate communication among radiologists. The Dutch Breast Cancer Screening Program uses five BI-RADS categories: category 0 means insufficient information for a definitive assessment; category 1 indicates a negative result with no signs of breast cancer; category 2 indicates benign findings; category 4 indicates the presence of a finding suspicious of malignancy; and category 5 indicates the presence of a finding highly suggestive of malignancy. BI-RADS category 3, which indicates a finding that typically requires follow-up, is not used in the Dutch Program. Each of the two radiologists assigns a BI-RADS category to each breast separately. Further diagnostic workup at a hospital is required for examinations with BI-RADS categories 0, 4, or 5, whereas those with categories 1 or 2 do not require recall. If the two radiologists disagree on whether the woman should be

recalled for further workup, one of them reassesses the examination. It can then be decided that additional workup is needed or that a third radiologist is needed to make the final recall decision (Figure 1.4).

Participating women receive a letter with the outcome of the assessment within approximately 10 days after the screening, indicating 1) no abnormal findings (BI-RADS 1 or 2); 2) inconclusive findings (BI-RADS 0); or 3) abnormal findings (BI-RADS 4 or 5) [20]. Women with inconclusive results or abnormal findings are recalled to a hospital for additional diagnostic workup.

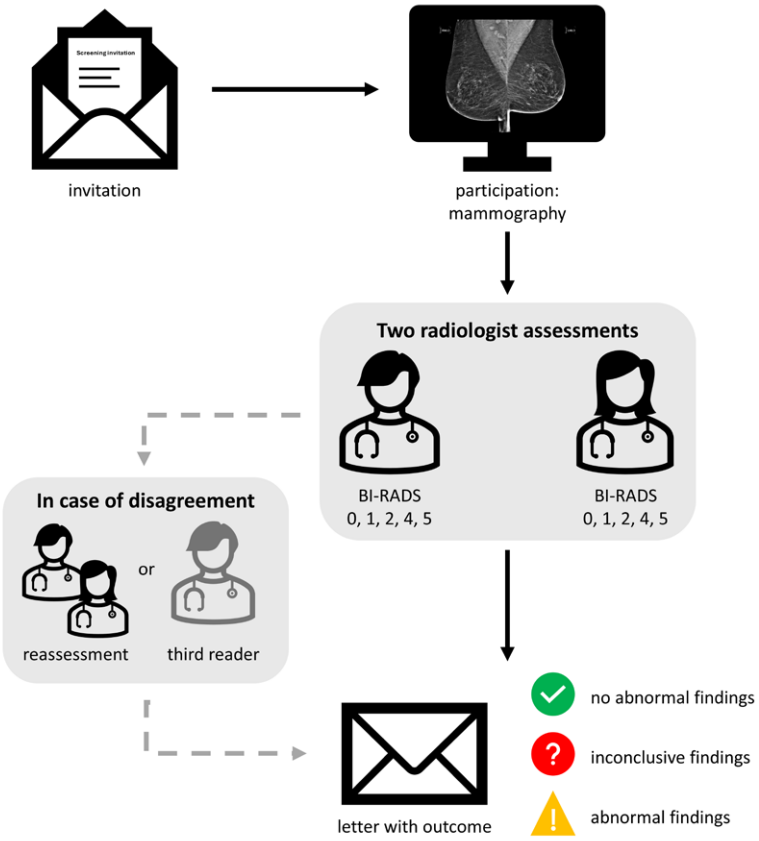


Figure 1.4: Illustration of the breast cancer screening workflow in the Dutch Screening Program.

In 2024, 20,002 (2.3%) women were recalled to a hospital for further diagnostic workup, resulting in 6,303 confirmed breast cancer diagnoses and a cancer detection of 0.72% [6]. The recall rate in the Dutch Program is among the lowest worldwide [21]. For comparison, the recall rate in the United States is approximately

10%, which can be partly attributed to differences in screening practices and concerns about the potential legal consequences of missing a cancer [9,22]. Despite finding more than 6,300 breast cancers in the Dutch Program annually, an additional 2,200 breast cancers are diagnosed within two years of a negative screening result [6,23]. This means that approximately one in four breast cancers goes undetected at screening. Additionally, despite the low recall rate, only one in three women recalled for further diagnostic workup is ultimately diagnosed with breast cancer, whereas the remaining two-thirds receive a false alarm [6]. While the Dutch Breast Cancer Screening Program demonstrates good performance, with high cancer detection and low recall rates, there is potential for further improvements.

1.2 Points for improvement in breast cancer screening

While mammography screening programs have been shown to detect cancers at an earlier stage, thereby reducing the need for aggressive treatment, increasing the possibility of breast-conserving surgery, reducing breast cancer-related mortality, and improving quality of life [24–27], there is still room for better outcomes. Opportunities for improvement include reducing underdiagnosis, overdiagnosis, and false alarms.

1.2.1 Underdiagnosis

Underdiagnosis happens when a clinically relevant breast cancer is present in the breast but is not detected during screening. This can happen for two main reasons: the cancer may not be visible on the mammogram, or radiologists may overlook or misinterpret visible signs of the cancer. Underdiagnosis becomes apparent when cancers are detected later, either at the next screening round or in between screening rounds, by which time they may have progressed, making treatment more aggressive and less likely to be curative [11].

Underdiagnosis can lead to interval cancers, which are breast cancers diagnosed between screening rounds. Compared to screen-detected cancers, interval cancers generally have less favorable prognostic characteristics and lower survival rates [28]. Studies estimate that approximately 17% to 30% of all breast cancers occurring in biennial screening participants are interval cancers [28].

Some interval cancers are not visible on prior screening mammograms. These cases may arise due to rapidly growing cancers that develop between screening rounds, cancers hidden due to tissue overlap, poor mammographic positioning, or lesions

located outside the target area of the mammogram. Since these cancers were not visible on prior mammograms, either the current screening interval was too long or the use of mammography as the sole screening modality may have been insufficient for these women.

However, not all interval cancers are new or invisible on prior mammograms. Approximately 17% to 58% of the interval cancers are retrospectively visible on prior mammograms [28–30]. Interval cancers with visible mammogram findings in hindsight are often classified as minimal signs or missed, depending on how suspicious the mammographic findings appear. Determining how many cancers fall in the missed category is challenging, since visibility is subjective. However, literature shows that approximately 14% to 35% of the interval cancers are considered missed [28–30]. Reasons for missing cancers are failure to detect an abnormality or incorrect interpretation of it [31]. Radiologist performance plays a significant role, with inter-reader differences in interpretive performance [32–34]. Additionally, the low prevalence of breast cancer within the screened population makes detection more challenging. When cancer is rare, it is easier to miss and more difficult to spot, as explained by Professor Jeremy Wolfe and Associate Professor Karla Evans with the statement: "*If you don't find it often, you often don't find it*" [35].

Given that a substantial proportion of interval cancers are retrospectively visible on prior mammograms, improving the interpretation of these images represents one of the most immediate and widely applicable strategies for reducing underdiagnosis.

1.2.2 Overdiagnosis

Whereas underdiagnosis reflects missed cancers, screening may also detect breast cancers that, without screening, would not have become clinically evident and would not have resulted in adverse consequences for the individual. This is known as overdiagnosis. Overdiagnosis can occur when cancers progress slowly or not at all and would not have caused symptoms, or when the individual dies from another cause before the cancer would have become symptomatic [36].

The magnitude of overdiagnosis in breast cancer screening is controversial, mainly because it is difficult to quantify. For any individual screen-detected cancer, it is impossible to determine whether it represents overdiagnosis. Various methods have been developed to estimate overdiagnosis, and although estimates vary widely, it is generally estimated that up to 10% of the breast cancers detected through screening are overdiagnosed [26,37]. The consequences of overdiagnosis include the psychosocial impact of being diagnosed with breast cancer and the burden of

unnecessary diagnostic workup [36]. Since current guidelines recommend treatment for all breast cancers, overdiagnosis also results in unnecessary overtreatment.

Overdiagnosis may be reduced by identifying women with low-risk cancers who do not require recall or further diagnostic workup at screening, as well as those who could be managed with active monitoring rather than immediate surgery [11]. However, accurately identifying low-risk cases in mammography screening remains challenging, highlighting the need for further research.

1.2.3 False alarms

The majority of women who are recalled for further diagnostic workup after a screening examination do not have breast cancer. In the Netherlands, about one-third of those recalled will be diagnosed with breast cancer, while the remaining two-thirds represent false-positive findings, meaning they were recalled but no breast cancer was diagnosed [6]. False-positive findings could result from benign abnormalities, overlapping tissue that mimics suspicious patterns, or inconclusive images that require further imaging to rule out breast cancer. Such findings are most common in the first round of screening, when radiologists have no prior mammograms for comparison [6,11,38].

False-positive findings are considered a harm of breast cancer screening [26]. A recall can be considered appropriate if the imaging finding was reasonably suspicious at the time, even when it later proves to be a false positive. However, whether appropriate or not, such outcomes often lead to participant anxiety, a lower quality of life, unnecessary diagnostic procedures that compromise cost-effectiveness, and a reduced likelihood of future participation in screening programs [26,39–41]. This decreased likelihood of screening participation could theoretically increase the risk of delayed breast cancer diagnosis, which is particularly concerning given that women with a false-positive mammogram are at a significantly higher risk of developing breast cancer compared to those without a prior false-positive result [39,42,43].

The recall rate is positively associated with the risk of a false-positive finding [38]. Research has shown that, compared with single reading, double reading can significantly lower recall rates without compromising cancer detection, thereby reducing the number of false-positive findings [44,45]. Nevertheless, not all screening programs have implemented double reading [14,44].

1.3 Evaluating screening performance with Signal Detection Theory

Improving breast cancer screening involves balancing the detection of clinically relevant cancers while minimizing false alarms. Signal Detection Theory, a statistically based decision framework introduced in the field of psychology, can help to understand and quantify this balance [46].

Breast cancer screening interpretation involves a radiologist extracting a probability of malignancy (PoM) from each screening examination. The PoM will depend on the presence (or not) of signal(s) in a mammogram (e.g., mass, architectural distortion, asymmetry, calcification) that may suggest, more or less strongly, the presence of breast cancer. However, this PoM does not inherently fall into a binary category of cancer or noncancer. Instead, it will range from low probability (clearly normal) to high probability (clearly malignant). Subconsciously, the radiologist must set a threshold along this continuum to distinguish between "positive" examinations, that require additional workup, and "negative" examinations, that do not require further workup. Figure 1.5 illustrates the Signal Detection Theory with the two overlapping bell-shaped curves representing the distribution of PoMs resulting from the screening examinations of two groups of women. The orange curve represents women with breast cancer, who generally have higher PoM scores, and the blue curve represents women without breast cancer, who tend to have lower PoM scores. The overlap between the curves represents the uncertainty in screening, where some women with breast cancer appear to have normal findings on mammography and some women without breast cancer appear to have abnormal findings. The vertical threshold on these curves represents the decision threshold of the radiologist. When the PoM falls above the threshold, the radiologist decides to recall, otherwise, no recall is indicated. Therefore, the location of this threshold determines the distribution of the four possible decision outcomes: true positives, false negatives, true negatives, and false positives. A woman with breast cancer with a radiologist PoM score above the threshold is correctly classified as a true positive, whereas if the score falls below the threshold, the cancer is missed, resulting in a false negative. Conversely, a woman without breast cancer whose score falls below the threshold is correctly identified as a true negative, while a score above the threshold results in a false-positive outcome. Moving the threshold along the x-axis changes the proportion of the four different outcomes. Increasing the threshold will reduce both true and false positives and increase true and false negatives, while lowering the threshold will have the opposite effect. There is not just one correct threshold; the optimal threshold depends on the consequences of one error over another.

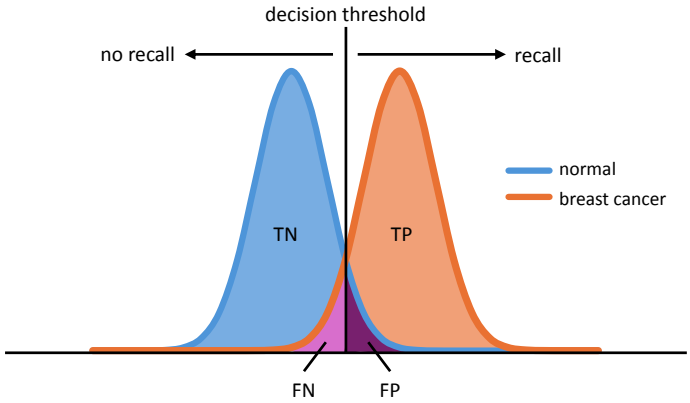


Figure 1.5: Illustration of the Signal Detection Theory. The blue bell-shaped curve represents the examinations of women without breast cancer, and the orange curve represents the examinations of women with breast cancer. The vertical line represents the decision threshold for recall. Examinations above this threshold are recalled, whereas examinations below this threshold are not recalled. The color-filled areas represent the different screening outcomes for this specific threshold. FN = false negative, FP = false positive, TN = true negative, TP = true positive.

This balance between correct and incorrect classifications can be quantified using two metrics: sensitivity and specificity. Sensitivity is the proportion of true- positive results among women with breast cancer (i.e., among the true-positive plus false-negative results). Specificity is the proportion of true negatives among the women without breast cancer (i.e., among the true-negative plus false-positive results). As the threshold is changed, sensitivity and specificity move in opposite directions. Lowering the threshold improves sensitivity but reduces specificity, while increasing the threshold improves specificity but lowers sensitivity.

The relationship between sensitivity and specificity across all possible thresholds is summarized by the receiver operating characteristic (ROC) curve, which plots sensitivity on the y-axis and 1-specificity on the x-axis (Figure 1.6). Each point on the curve represents a threshold with a corresponding sensitivity and specificity. If you could perfectly distinguish between women with and without breast cancer, the ROC curve would reach the top left corner of the figure with a sensitivity and specificity of unity. If the curve falls along the diagonal line from the bottom left to the top right, the test would be no better than chance. The area under the ROC curve (AUC) summarizes the performance across all possible thresholds. An AUC of unity indicates perfect accuracy, whereas an AUC of 0.5 indicates random guessing. In medical imaging research, ROC curves and AUC values are widely used to evaluate and compare different interpretation strategies, providing a comprehensive assessment that is independent of decision thresholds and cancer prevalence.

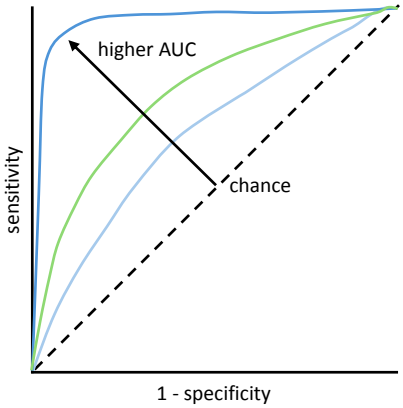


Figure 1.6: Receiver operating characteristic (ROC) curves with sensitivity plotted on the y-axis and 1-specificity plotted on the x-axis. The dashed diagonal line represents the performance expected by random chance for distinguishing between examinations with and without breast cancer. The curve that is closest to the top left corner of the graph has the highest area under the ROC curve (AUC).

1.4 Strategies to improve mammography interpretation

The effectiveness of screening largely relies on the ability of readers to accurately interpret mammograms, a process that involves searching for, perceiving, and making decisions about potential abnormalities suspicious for breast cancer. Errors can arise at any stage of this process and are influenced by various factors, including breast and lesion characteristics, the underlying cancer prevalence, technical constraints, reader characteristics, and external influences, such as fatigue and distraction [33,35]. Several strategies for improving breast cancer screening are being discussed, including the use of new imaging technologies (e.g., breast tomosynthesis, contrast-enhanced mammography, or contrast-enhanced MRI) in more personalized screening programs. However, since mammography remains the most commonly used screening method worldwide, improving the accuracy of mammography interpretation is particularly important. This thesis searches for immediate and accessible opportunities that may potentially benefit screening participants. Specifically, the aim is to investigate whether small adjustments within the existing screening infrastructure can improve the interpretation of screening mammography, highlighting opportunities for *real impact within reach*. Since evidence suggests that interpretation accuracy is affected by human limitations, potential solutions could include optimizing the individual- and double-reading process, and exploring the potential impact of artificial intelligence (AI) in improving screening outcomes.

1.4.1 Double reading

Double reading accounts for differences in human search, perception, and decision-making abilities. In breast cancer screening, this approach works on the principle that if one reader misses a cancer, there is a chance that the second reader will detect and interpret it correctly, and likewise, that if one reader has a false-positive finding, the other reader may discard it. In other words, *two heads are better than one*, especially since research has shown that mammography interpretation varies substantially among radiologists [33,34]. In double reading, each reader independently evaluates the mammograms and decides whether further diagnostic workup is needed. In case of disagreement between readers, a third reader may serve as an arbitrator, or a consensus meeting may be held to reach a joint decision. In some screening programs, for example in Norway, a consensus meeting is done if either reader recommends recall, including cases where both readers agree on recall [47]. Previous studies found that double reading improves cancer detection rates and reduces the number of interval cancers compared with single reading in screening mammography [44,45,48–50]. Additionally, when disagreements arise, an arbitrator or a consensus meeting may help make the final recall decision and thereby lower recall rates [44,45,48,49]. In the Dutch Breast Cancer Screening Program, an additional read by the radiographer is also performed. Immediately after image acquisition, the radiographer can mark suspicious findings, triggering a pop-up for the radiologist if they initially decide not to recall. The radiologist can then ignore the pop-up or change their recall decision. Building on prior work [51], ongoing studies are evaluating whether this approach of additional reading by a radiographer improves screening outcomes.

The ECIBC and, previous to that, the European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis have recommended independent double reading for breast cancer screening, with the second reader blinded to the first reader's interpretation and discordant opinions being resolved either through consensus or by involving an arbitrator [52,53]. Most screening programs have implemented double reading [14,44]. However, countries employing double reading generally pair readers based on their availability, without attempting to optimize the pairs for improved screening outcomes. Although double reading reduces the group variability by accounting for individual differences, significant variations in outcomes among pairs of radiologists still exist [32]. Therefore, it is important to organize reader pairings that maximize overall group performance while balancing the radiologists' workload. In some breast screening centers, experienced and inexperienced readers, as well as high and low recall readers, are paired. However, this is done based on intuition, and currently, evidence on pairing

criteria that have proven to improve screening outcomes is limited. Nevertheless, strategic pairing may be worth investigating, especially since telemedicine makes implementing such pairing strategies across locations feasible.

A previous retrospective study has shown that, in theory, it is possible to optimize reader pairs for better group performance [54]. Brennan and his colleagues explored all possible ways of dividing a group of 12 radiologists into unique pairs and evaluated the diagnostic accuracy when reading a set of mammograms for all pairings. Results showed variability in diagnostic performance between sets of pairings, suggesting that the method of pairing readers affects the effectiveness of double reading. By comparing the diagnostic accuracy of the best, worst, and random pairings against that of single reading, they demonstrated that the best pairings, and to some extent the random pairings, outperformed single reading, whereas the worst pairings showed no significant difference in sensitivity compared to single reading. Therefore, to maximize the benefits of double reading, reader pairs should be optimized.

However, in order to optimize pairings, prospective selection criteria are needed. One approach for pairing radiologists could be based on radiologists' visual search patterns, for example, by pairing readers with different search patterns to ensure the second reader extracts complementary information. To quantify search patterns, eye trackers are commonly used. Eye trackers record eye movements and measure fixations, which are moments when the eyes pause to extract detailed information from where they are focused on. An example of eye-tracking data, obtained while a radiologist interpreted two mammograms, is shown in Figure 1.7. Such data can provide a better understanding of how radiologists explore mammograms.

Gandomkar et al. quantified 14 different eye-tracking metrics and paired readers from the most different to the most similar eye-tracking metric [55]. They found that this method of pairing was comparable to ranking readers from best to worst based on the AUC of the pairs, suggesting that pairing readers with different eye-tracking patterns may improve diagnostic accuracy. However, eye trackers are not yet available in most screening programs, and unlike Brennan et al., this study focused on pairing individual readers, not assessing the best pairing scheme for the entire group of readers. Therefore, additional studies investigating this pairing strategy at the group level are needed.

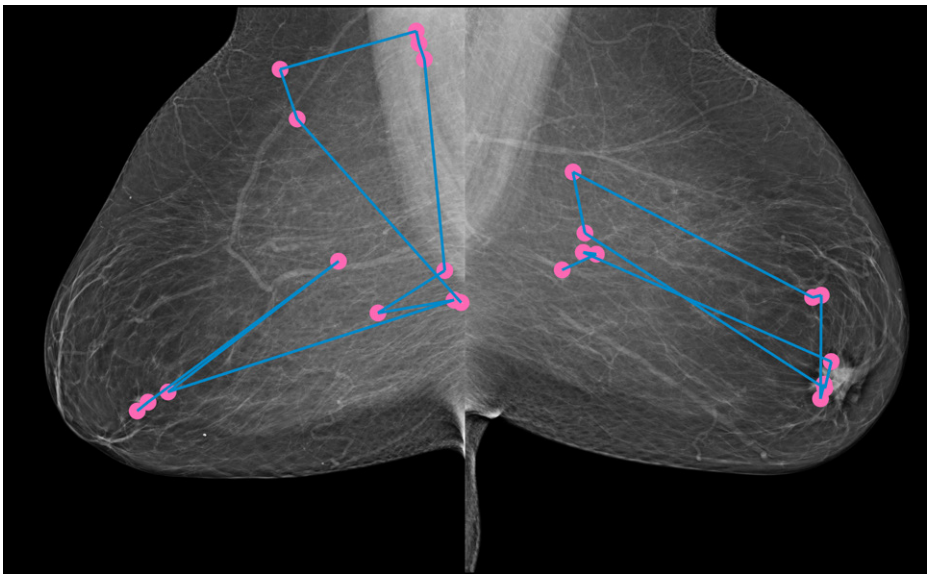


Figure 1.7: Example of eye-tracking data from a radiologist viewing the right and left breast in the mediolateral view. Fixations are shown as pink circles and are connected by blue lines, which depict the movements of the eyes between them.

Another approach to optimize double reading would be to pre-assess each reader's diagnostic accuracy and pair them based on it [56]. For quality assurance purposes, readers' performance metrics, such as recall and cancer detection rates, are commonly available in screening programs. However, studies that explore radiologist pairing based on performance metrics are lacking.

1.4.2 Ordering mammograms

In large screening programs, radiologists typically interpret screening mammograms in batches, usually within one to two days after acquisition. Studies have shown that, compared to interrupted reading, batch reading can improve screening performance by reducing recall rates, without compromising cancer detection rates [57,58]. This is in line with sequential reading effects, for which it has been shown that recall rates, false alarms, and reading times decrease with time on task [59–63], while cancer detection rates remain stable [62,63], though findings on cancer detection are mixed [60].

It is not entirely clear what causes these sequential reading effects. Possible factors include task engagement, fatigue, and an increase in the recall threshold (defined by Signal Detection Theory) as readers become more conservative after encountering many normal examinations [35,63]. Another potential explanation is that readers' visual perception adjusts as they assess multiple screening mammograms. This

phenomenon, called "visual adaptation", has received limited attention in the medical image perception literature and warrants further investigation.

Visual adaptation is a process in our sensory systems that adjusts sensitivity to certain visual stimuli after prolonged exposure [64]. This results in temporary changes, reducing sensitivity to stimuli similar to the adaptor but not to those sufficiently different. It is linked to the core processes of our visual system, where photons reflected off objects enter the eye, interact with photoreceptors in the retina, and are converted into electrical signals that are transmitted through the optic nerve to visual neurons in the brain [65]. Almost every aspect of visual perception can be influenced by prior visual experience, often without conscious awareness. For example, by fixating for a few seconds on the center dot in the red square in Figure 1.8 and then shifting your gaze to the dot in the gray field next to it, you may experience a visual aftereffect, with the gray area appearing slightly greenish [66].

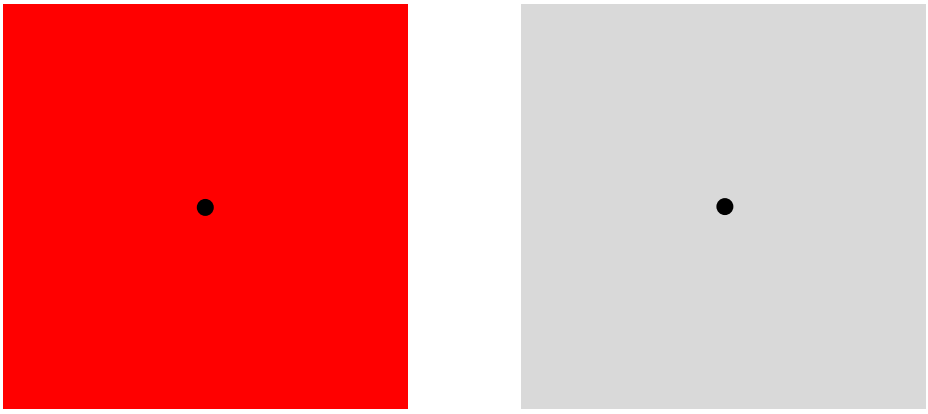


Figure 1.8: Example to showcase the effect of visual adaptation. By first focusing on the black dot within the red square for a few seconds and then shifting focus to the black dot in the gray square, you may notice a visual aftereffect in which the gray square appears slightly greenish.

These adaptation effects may be important in screening interpretation, since radiologists spend a lot of time examining mammograms, making it likely that they adapt to the images they view. Research has already shown that exposure to mammograms can induce visual adaptation effects that influence both image perception and visual search efficiency [66–69]. These adaptation effects are likely induced by the visual appearance of mammograms. One key variable that influences mammographic appearance is breast density, which refers to the proportion of fibroglandular tissue compared to fatty tissue in the breast. Dense breasts are generally more difficult to interpret on mammography as they are characterized by a high proportion of fibroglandular tissue, which can obscure lesions on

mammography. Previous studies have found that prior exposure to dense or fatty images caused an intermediate-density image, created by blending dense and fatty images with different weights, to appear more fatty or dense, respectively [66,69]. This is consistent with negative aftereffects in the perceived texture of mammograms. With prolonged viewing, the adapting images themselves were also perceived as more neutral (less dense or fatty) [66]. Moreover, prior exposure to blurred images influenced the perceived blur in images viewed next (i.e., to appear less blurry) [68]. These findings reflect perceptual renormalization, wherein the visual system predicts and discounts expected properties of an image, thereby enhancing the salience of unexpected properties. This is somewhat similar to the challenge faced by radiologists, who search mammograms with complex backgrounds for the presence of suspicious features. In line with this, a previous study reported that adaptation to dense or fatty images improved search times for detecting simulated lesions embedded within images with those background textures [67]. This demonstrates that adaptation may not only influence the appearance of images, but potentially also how accurately or quickly readers can find lesions in those images.

If adaptation plays a role in batch reading, then adjusting the batch reading process to facilitate these adaptation mechanisms may offer a way to improve performance. Adaptation aftereffects imply that the perception of an image is shaped by those viewed earlier, suggesting that the order in which images are interpreted within a batch could affect interpretation. However, no guidelines exist regarding the order in which mammograms should be read, and most reading centers follow the order in which images are acquired. From an adaptation perspective, however, ordering images according to their features may help reduce abrupt transitions between examinations, allowing radiologists to remain in a more stable adaptation state adjusted to the characteristics of the mammograms they are about to interpret. No studies have directly tested the effect of ordering mammograms to maximize visual adaptation. However, a large clinical trial in England tested how changing reading order could impact vigilance decrement [62]. In this study, the second reader reviewed the same batch of mammograms in reverse order from the first, with the hypothesis that by varying the reading order, readers experience peak vigilance at different points within the batch. However, the study found no significant improvements in double-reading screening outcomes, although they did observe a gradual decrease in false-positive recalls with no change in cancer detection with time on task [62]. Sequential effects were thus present, although the prevailing effect(s) was not visual adaptation. Studies are needed to investigate how the ordering of mammograms for interpretation, based on the concept of visual adaptation, affects radiologists' performance. A logical starting point would

be to order by breast density, which can be determined by automated algorithms. However, other texture-based features could also be explored.

1.4.3 Artificial intelligence for decision support

One of the most promising advancements in breast cancer screening is the potential of AI in interpreting mammograms. Although there are numerous potential applications for AI, this thesis primarily investigates its potential as a decision-support tool, with the discussion addressing additional future applications.

In the early 1990s, computer-aided detection (CAD) systems were developed and introduced as support tools for radiologists, designed to mark suspicious areas in mammograms and thereby improve clinical decision making [70]. However, evidence on the effectiveness of CAD is controversial. Early studies found that CAD-assisted single reading could be a potential alternative to double reading [71–74]. Nonetheless, other large-scale studies found that CAD did not improve screening outcomes in clinical practice and was associated with increased recall, false-positive, and biopsy rates, without improvements in cancer detection [75–77]. The inability of CAD to show improved screening outcomes is often related to the low specificity of CAD, with false-positive markers leading to an increase in the time needed to interpret an examination and lower radiologist trust in CAD [78,79].

Recently, AI-based CAD has shown high potential to improve screening outcomes, outperforming traditional CAD systems and showing comparable or even superior performance to radiologists [80,81]. The main advantage of AI lies in being trained on large datasets to identify self-learned patterns, instead of relying on predefined, human-interpretable descriptors, as CAD systems do. For decision support, AI provides traditional lesion markers that highlight suspicious areas, along with a score for each marked region and an overall score for the entire examination (Figure 1.9). The region scores for most AI systems lie between 1 and 100, with higher scores representing a higher probability of malignancy (PoM). The overall score for the entire examination is usually equal to the score of the region assigned the highest PoM score and is sometimes further defined into a suspiciousness category (e.g., low, intermediate, elevated).

Multiple retrospective reader studies have shown that AI-based decision support in mammography improves radiologist performance by increasing cancer detection, while maintaining or reducing false positives, without prolonging reading time [82–85]. This shows potential for AI to improve breast cancer screening outcomes. Although reader studies are important for evaluating new strategies, they have limitations. In reader studies, radiologists' decisions do not affect outcomes for women as they do

in real-world screening, and these studies are often conducted with cancer-enriched datasets that differ from typical screening populations. The promising results from reader studies, however, have recently been followed by prospective studies conducted in population-based breast cancer screening programs, comparing radiologists reading with and without AI decision support [86–88]. These large-scale prospective studies found that AI-aided reading resulted in improvements in cancer detection rates [86–88], without negatively affecting recall rates [86,87], providing growing evidence for the positive effects of AI decision support on screening outcomes.

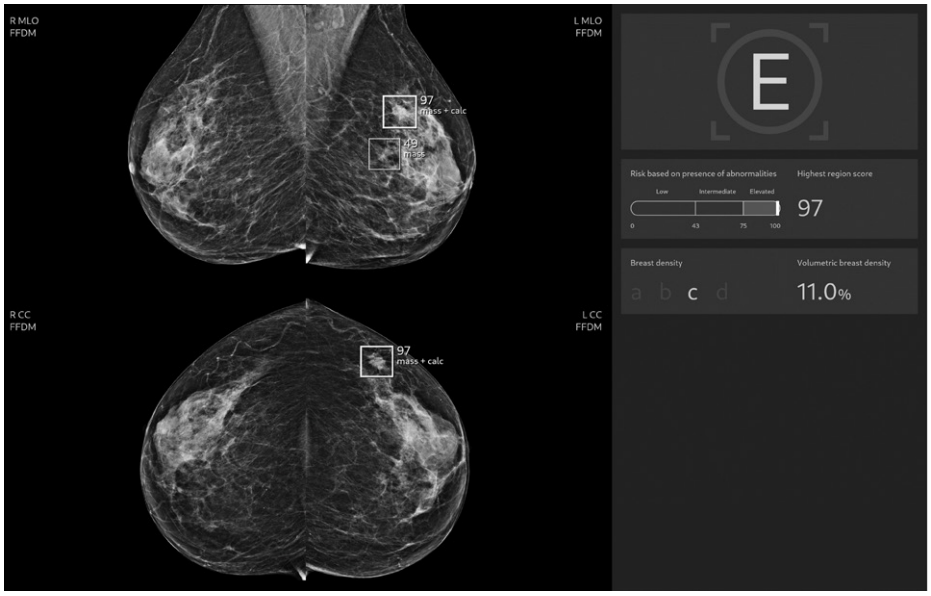


Figure 1.9: Example of a screening examination with a spiculated mass in the left breast, diagnosed as invasive carcinoma of no special type. The malignancy was marked by the artificial intelligence (AI) system with lesion markers and high probability of malignancy (PoM) scores. The overall score of the examination is 97, representing the highest region score. This examination was categorized as "E", which stands for "Elevated".

Although the impact of AI decision support on radiologists' performance has been studied, little is known about how AI suggestions influence their visual search behavior. To optimally integrate AI decision support into the radiological workflow, it is important to understand how radiologists interact with AI. Several visual search studies have investigated radiologists' search behavior in mammography to identify patterns associated with accurate diagnoses and sources of search-related errors, but studies have not yet addressed the impact of AI decision support on visual search [89]. Evidence from earlier studies using CAD shows that CAD cues guided observers' eye movements to marked locations, resulting in less exploration

of unmarked areas [90,91]. These CAD findings suggest that if AI decision support similarly influences radiologists, and its accuracy is high, it could help ensure critical findings are not overlooked. However, there is also a risk that radiologists may become overly reliant on AI cues and stop critically evaluating its output. This phenomenon, known as automation bias, can potentially lead to missed cancers and unnecessary recalls, particularly when the AI is not accurate [92–95]. Future research should explore how radiologists interact with AI to guide the development of best practices and optimize its role in decision support. As AI systems become more widely implemented and demonstrate high accuracy, understanding their effects on visual search behavior becomes increasingly important.

1.5 Thesis overview

This thesis is part of the aiREAD project (Accurate and Intelligent Reading for EArlier breast cancer Detection) and explores several strategies to improve the interpretation of screening mammography. As the title of this thesis suggests, the focus is on strategies that aim for *real impact within reach*. Potential approaches could include optimizing double reading, adjusting the order in which mammograms are interpreted, and using AI for decision support.

Chapters 2 and 3 focus on optimizing radiologist pairs for double reading. **Chapter 2** introduces an exploratory approach to evaluate whether pairing strategies based on individual reading performance can improve group outcomes. **Chapter 3** builds on this by using logistic regression to model screening assessments, allowing for a more comprehensive evaluation of different performance-based pairing strategies.

Chapter 4 investigates the effect of ordering mammograms for radiologist interpretation. This study builds on the hypothesis that visual adaptation may help radiologists improve their screening performance and includes a reader study with different ordering criteria.

Chapters 5 and 6 examine the impact of AI decision support. **Chapter 5** describes a reader study comparing radiologists' screening performance and visual search behavior when reading with and without AI support. **Chapter 6** examines how automation bias and the timing of AI decision support influence radiologists' performance.

Finally, **Chapter 7** discusses the findings of this thesis, their implications for clinical practice, and possible future directions for breast cancer screening.

Chapter 2

Optimizing the Pairs of Radiologists That Double Read Screening Mammograms

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

Published in Radiology 2023; 309(1):e222691.
DOI: 10.1148/radiol.222691

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 1,000 examinations and CDR range, 3.1–3.6 per 1,000; England AIR range, 68.2–70.5 per 1,000 and CDR range, 8.9–9.4 per 1,000; Norway AIR range, 81.6–88.1 per 1,000 and CDR range, 6.1–6.8 per 1,000) were not significantly different from the random pairing strategy (Sweden AIR, 54.1 per 1,000 examinations and CDR, 3.3 per 1,000; England AIR, 69.3 per 1,000 and CDR, 9.1 per 1,000; Norway AIR, 84.1 per 1,000 and CDR, 6.3 per 1,000).

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.

2.1 Introduction

Population-based breast cancer screening programs with mammography have proven to be effective in reducing breast cancer–specific mortality [96,97]. Nevertheless, breast cancer is the most commonly diagnosed cancer and a leading cause of cancer death among women worldwide [98]. Radiologists miss 3%–40% of mammographically visible cancers [29,99,100]. At the same time, false-positive screening results lead to unnecessary work-up, participant anxiety, and a reduction in cost-effectiveness [101].

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 [45,50,102]. 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 [103]. 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 [104–107].

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 [32]. A previous multireader multicase 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 [54]. 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 [55]. In that study, Gandomkar et al. [55] 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.

2.2 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 [108]; the Changing case Order to Optimize patterns of Performance in Screening, or CO-OPS, data set from England [62,109]; 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 [48,62,107–113], 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.

2.2.1 Screening Procedures

The screening programs of the three countries differ in several aspects (Figure 2.1).

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 [108,110]. 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.

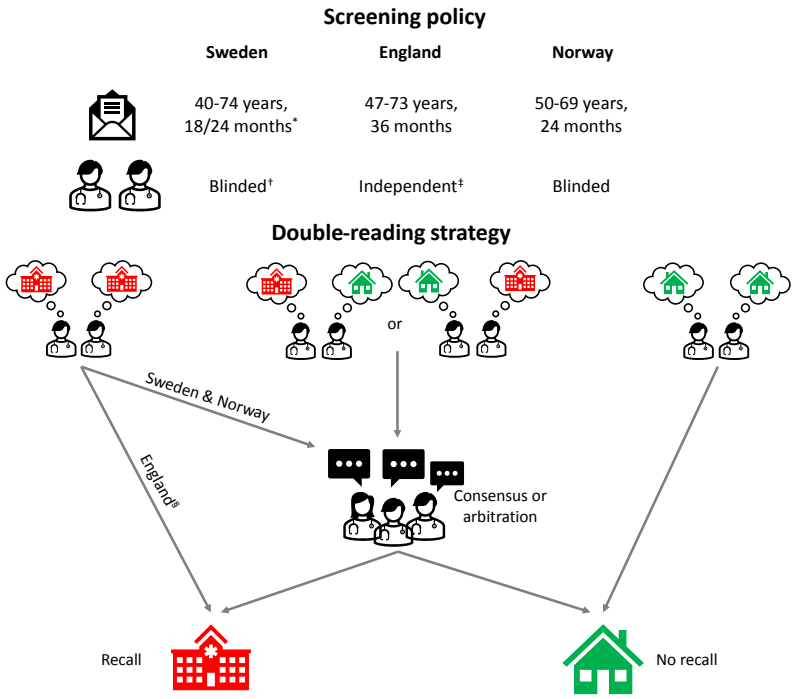


Figure 2.1: Schematic shows screening policy and double-reading strategy for the different data sets. * = Women older than 49 years were invited every 24 months, whereas younger women were invited every 18 months. † = There are local variations in practice regarding blinding, but at most centers, the 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.

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 [62,109].

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 [47]. 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.

2.2.2 Study Population

All data sets consisted of more than 1 million screening examinations (Table 2.1). 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.

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 (i.e., 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 [114]. 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.

Table 2.1: 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 (years)[*]	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 1,000 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 detected 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 screen-detected cancer but had a breast cancer diagnosed within 18–24 months after screening. [††] = Women who did not have a screen-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.

2.2.3 Statistical Analysis

The CDR (true-positive findings per 1,000 examinations) and AIR ([true-positive findings + false-positive findings] per 1,000 examinations) were calculated for the individual readers and pairs, and the performance of different pairing strategies was hypothesized (Figure 2.2). 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 Figure 2.2). 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).

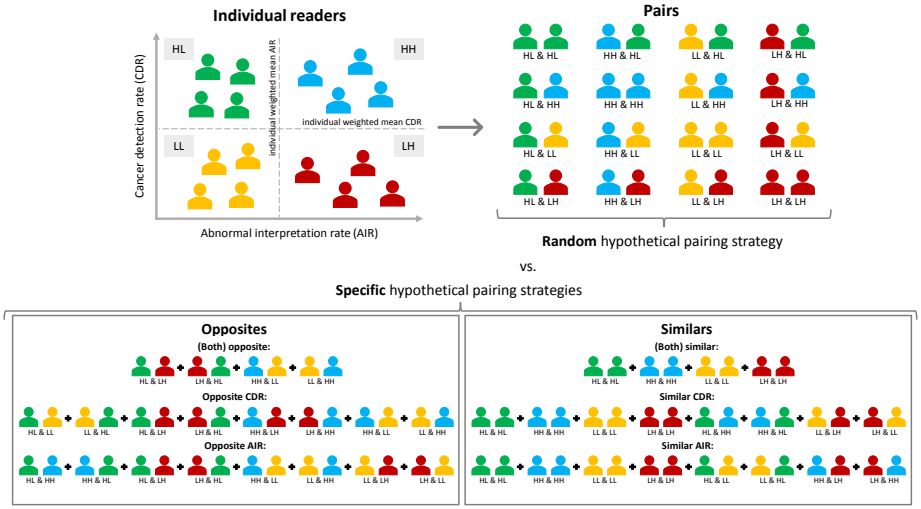


Figure 2.2: Diagram shows explanation for the evaluation of the screening performance of the individual readers and pairs, as well as an explanation of the hypothetical pairing strategies used to investigate the potential to improve outcomes with an appropriate double-reading strategy. HH = reader characterized 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 2.3 was used, where a positive assessment was defined by either or both readers flagging an examination as abnormal (i.e., 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 Figure 2.2). 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 (e.g., 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).

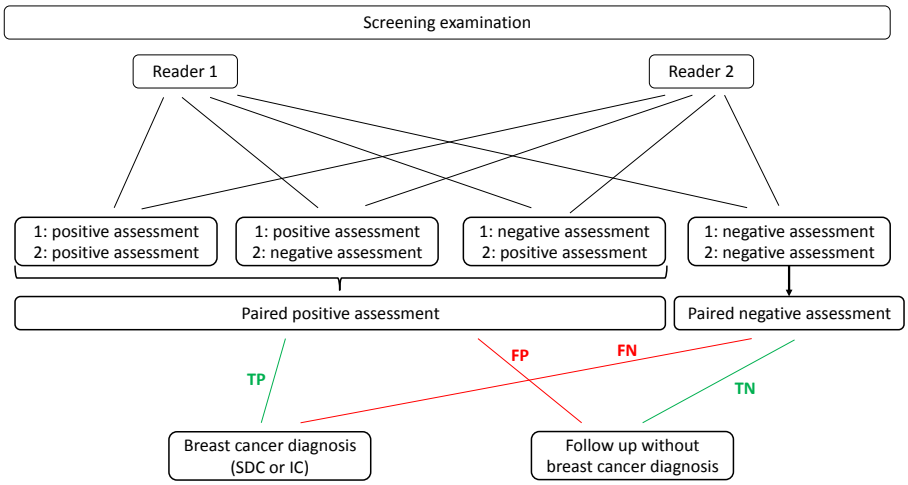


Figure 2.3: Flowchart shows the possible screening reading outcomes for the pairs in this study. Discrepant paired readings were classified as positive assessments. FN = false negative, FP = false positive, IC = interval cancer, SDC = screen-detected cancer, TN = true negative, TP = true positive.

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 Figure 2.2): opposite AIR and CDR, opposite CDR, opposite AIR, similar AIR and CDR, similar CDR, and 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 Figure 2.2). 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 Figure 2.2). 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 = 1,000) 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 statistical analyses were performed in RStudio, version 4.1.0 (PBC).

2.3 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 (Figure 2.4).

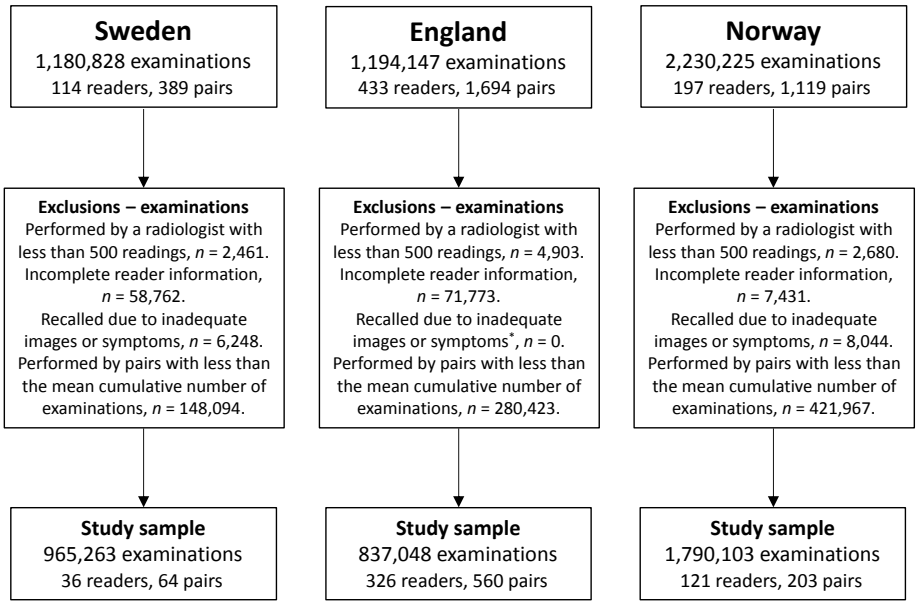


Figure 2.4: Flowchart of screening examinations after applying exclusion criteria. * Examinations of women in the English data set who were recalled due to inadequate images or symptoms were already excluded before the data were received for this study.

2.3.1 Study Sample Characteristics

The study subsamples involved different numbers of readers and pairs with different screening performance (Table 2.2). 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 1,000 examinations, respectively). The readers in the English subsample had the highest individual CDR, with 8.3 per 1,000 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 1,000 examinations, respectively), while paired CDR was highest for the English subsample (8.9 per 1,000). The Norwegian subsample showed the most disagreement between the readers (5.6%), resulting in the highest paired AIR of 77.0 per 1,000 examinations.

Table 2.2: 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 (years)*	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 1,000 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.

2.3.2 Individual Performance

Figure 2.5 shows the classification of the individual readers into four performance groups. The individual weighted mean AIR (Sweden, 36.3 per 1,000 examinations; England, 48.2 per 1,000; and Norway, 49.0 per 1,000) and CDR (Sweden, 3.1 per 1,000; England, 8.3 per 1,000; and Norway, 5.2 per 1,000) were used as cutoffs for the respective data sets (Table 2.2).

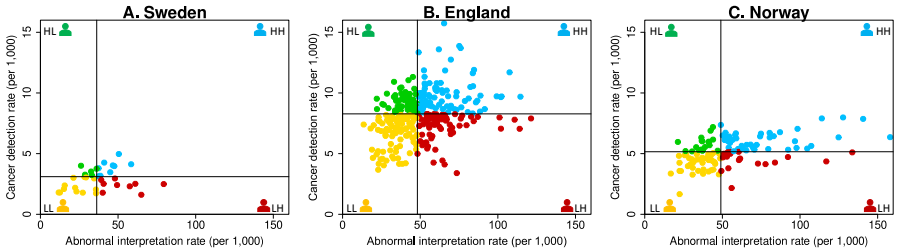


Figure 2.5: Quadrant graphs for individual screening performance in the (A) Swedish, (B) English, and (C) Norwegian data sets. The individual weighted mean performance was used as the cutoff. Readers were characterized by a high cancer detection rate (CDR) and low abnormal interpretation rate (AIR) (HL), high CDR and high AIR (HH), low CDR and low AIR (LL), or low CDR and high AIR (LH).

2.3.3 Paired Performance

Figure 2.6 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 1,000 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 1,000 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 1,000 examinations for the Swedish, English, and Norwegian subsamples, respectively) or low (< 3.1, < 7.6, and < 5.7 per 1,000 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.

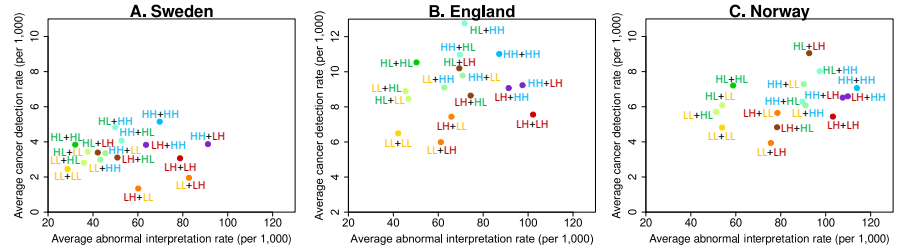


Figure 2.6: Scatterplots with the average cancer detection rate (CDR) and abnormal interpretation rate (AIR) of the 16 specific pairs in the (A) Swedish, (B) English, and (C) Norwegian data sets. HH = reader characterized by a high CDR and high AIR, HL = reader characterized by a high CDR and low AIR, LH = reader characterized by a low CDR and high AIR, LL = reader characterized by a low CDR and low AIR.

2.3.4 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 1,000 examinations (95% CI: 46.1, 62.1) and 3.3 per 1,000 (95% CI: 2.8, 3.9) for the Swedish subsample, 69.3 per 1,000 (95% CI: 63.7, 74.9) and 9.1 per 1,000 (95% CI: 8.3, 9.9) for the English subsample, and 84.1 per 1,000 (95% CI: 80.3, 88.0) and 6.3 per 1,000 (95% CI: 5.9, 6.7) for the Norwegian subsample (Figure 2.7). 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 2.3).

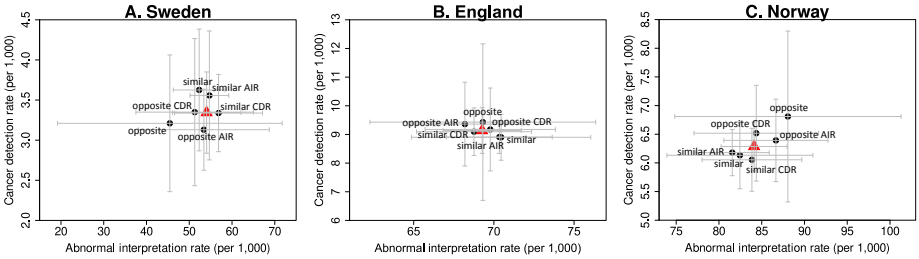


Figure 2.7: Group screening performance for the different pairing strategies in the (A) Swedish, (B) English, and (C) Norwegian data sets. The triangles (red) represent the average screening performance for the random hypothetical pairing strategy, and the dots represent the performance for the specific hypothetical pairing strategies (black). Error bars are Bonferroni-adjusted 95% CIs obtained by means of bootstrap resampling (n = 1,000). 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 2.3: 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)

Table 2.3: Continued

	Swedish Data Set		English Data Set		Norwegian Data Set	
Pairing Strategy	**AIR**	**CDR**	**AIR**	**CDR**	**AIR**	**CDR**
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 1,000 examinations, and 95% CIs are Bonferroni-adjusted ($P < .05/6$) and were obtained by bootstrap resampling ($n = 1,000$).

2.4 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 [54] 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 (e.g., 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

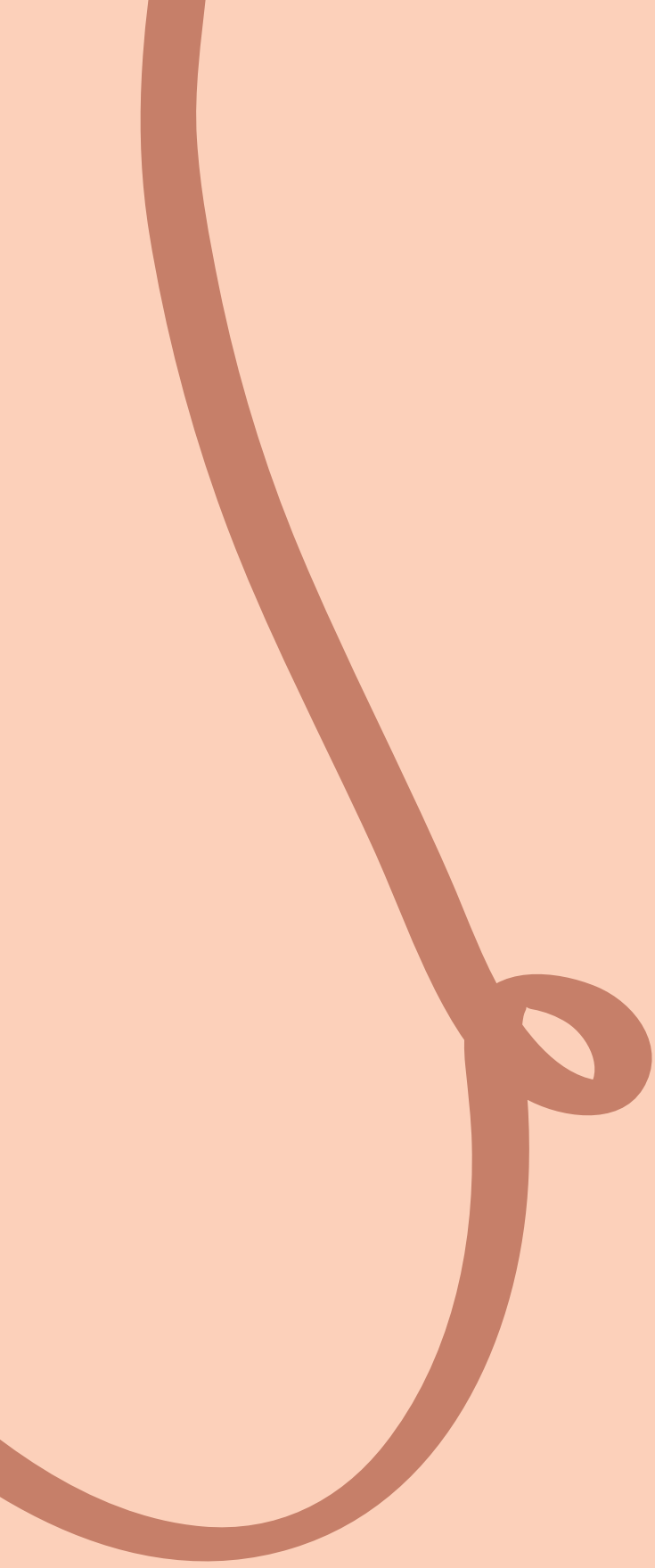
for the optimization of pairs. Thus, it could be helpful to exhaustively evaluate all theoretically possible pairs. Research should therefore focus on simulating individual radiologist assessments, making it possible to analyze data consisting of results of all readers reading all cases, and on exploring 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 (e.g., 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.



Chapter 3

Modeling Radiologists' Assessments to Explore Pairing Strategies for Optimized Double Reading of Screening Mammograms

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

Published in Medical Decision Making 2024 Oct; 44(7):828–42.
DOI: 10.1177/0272989X241264572

Abstract

Purpose

To develop a model that simulates radiologist assessments and use it to explore whether pairing readers based on their individual performance characteristics could optimize screening performance.

Materials and Methods

Logistic regression models were designed and used to model individual radiologist assessments. For model evaluation, model-predicted individual performance metrics and paired disagreement rates were compared against the observed data using Pearson correlation coefficients. The logistic regression models were subsequently used to simulate different screening programs with reader pairing based on individual true-positive rates (TPR) and/or false-positive rates (FPR). For this, retrospective results from breast cancer screening programs employing double reading in Sweden, England, and Norway were used. Outcomes of random pairing were compared against those composed of readers with similar and opposite TPRs/FPRs, with positive assessments defined by either reader flagging an examination as abnormal.

Results

The analysis data sets consisted of 936,621 (Sweden), 435,281 (England), and 1,820,053 (Norway) examinations. There was good agreement between the model-predicted and observed radiologists' TPR and FPR ($r \geq 0.969$). Model-predicted negative-case disagreement rates showed high correlations ($r \geq 0.709$), whereas positive-case disagreement rates had lower correlation levels due to sparse data ($r \geq 0.532$). Pairing radiologists with similar FPR characteristics (Sweden: 4.50% [95% confidence interval: 4.46%–4.54%], England: 5.51% [5.47%–5.56%], Norway: 8.03% [7.99%–8.07%]) resulted in significantly lower FPR than with random pairing (Sweden: 4.74% [4.70%–4.78%], England: 5.76% [5.71%–5.80%], Norway: 8.30% [8.26%–8.34%]), reducing examinations sent to consensus/arbitration while the TPR did not change significantly. Other pairing strategies resulted in equal or worse performance than random pairing.

Conclusion

Logistic regression models accurately predicted screening mammography assessments and helped explore different radiologist pairing strategies. Pairing readers with similar modeled FPR characteristics reduced the number of examinations unnecessarily sent to consensus/arbitration without significantly compromising the TPR.

3.1 Introduction

Breast cancer remains a major global health concern, and early detection plays a crucial role in improving patient outcomes. Population-based mammography screening programs have shown to be effective in reducing breast cancer–related mortality due to earlier detection of the disease [96,97,115,116]. However, the challenges posed by the complexity of mammographic images and the potential for human error underline the need for continuous efforts to improve the interpretation of screening mammography. One approach to improving mammography interpretation is the implementation of double reading. Screening programs in Europe, Australia, and New Zealand have implemented double reading, in which each screening examination is assessed by 2 independent readers, increasing the cancer detection rate (CDR) for combined assessments [48,50,117]. Usually, discordant assessments between the 2 readers are referred to an arbitrator or discussed at a consensus meeting, so that a final recall decision can be made. In some screening programs, concordant positive assessments are also referred for a final check by an arbitrator or by consensus. In countries in which screening mammograms are single read, a potential double-reading strategy is to use artificial intelligence (AI) as a second independent reader [105–107,118].

The success of double reading, however, may rely on how the radiologists are paired. Countries currently employing double reading generally pair readers randomly without considering the characteristics of the readers. However, matching radiologists with different strengths can potentially maximize the detection of breast cancer while minimizing the chances of false positives. A previously published study showed that the accuracy of mammography interpretation with double reading can indeed be improved by optimizing the set of paired radiologists [54]. However, this study did not identify what the a priori pairing strategy should be to achieve this optimization. A potential pairing optimization strategy for screening may be based on individual reader performance characteristics, which vary considerably among radiologists [32]. The present study was triggered by the findings of another previous study [119], which found that radiologist screening performance characteristics influenced the performance of radiologist pairs. However, that study was not able to detect significant variations in overall group performance resulting from different pairing strategies based on individual characteristics. This might be due to the limitation that with existing real-world screening data, each screening examination is read by only 2 readers, and therefore exhaustive exploration of different pairing strategies is not possible.

Previous studies have used mathematical models of the human observer, known as model observers, to investigate the diagnostic potential of different images and consequently the performance of radiologists [120,121]. However, to the best of our knowledge, there are no studies that used retrospective screening data to model radiologist assessments and immediately use this for generating a new data set. While several studies have explored the association between radiologist characteristics and screening performance, they did not focus on modeling individual radiologists explicitly [122,123].

Therefore, this study aims to develop an explanatory logistic regression model designed to simulate individual radiologists' assessments and use it to investigate if there is a strategy for pairing radiologists based on individual screening performance characteristics that would optimize overall screening performance. By modeling radiologists' assessments and creating interpretations involving all readers for all examinations, we are able to explore various radiologist pairing strategies against random pairing with sufficient statistical power.

3.2 Materials and Methods

3.2.1 Reader Model

Model definition.—A logistic model is derived to fit observed data that is binary in nature. Let Y^- and Y^+ represent the screening outcomes (1 = recall, 0 = return to screening) for negative (noncancer) and positive (cancer) cases, respectively. When Y^- has a value of 1 for a given case and reader, the decision is defined as a false positive, and when Y^+ has a value of 1, the decision is defined as a true positive.

The model posits a latent decision variable for each reader and case. Let the readers be indexed by $j = 1, \ldots, J$, negative cases by $k = 1, \ldots, K_i^-$, and positive cases by $k = 1, \ldots, K_j^+$. Note that positive and negative cases are considered to be completely independent ($K^+ \ll K^-$). The decision variables, q_{jk}^- for negative cases and q_{jk}^+ for positive cases are assumed to be the sum of a reader-specific effect, r_j^- or r_j^+ , and a case-specific effect, C_k^- or C_k^+,

$$q_{jk}^- = r_j^- + C_k^- \quad \text{and} \quad q_{jk}^+ = r_j^+ + C_k^+ \tag{1}$$

The reader effects, parameterized by r_j^- and r_j^+, are considered to be fixed effects (and they include any global intercept parameter). The case effects, C_k^- and C_k^+, are assumed to be random effects, modeled as realizations of a normal random process

with a mean of 0 and a standard deviation of σ_{C^-} or σ_{C^+} for the negative and positive examinations, respectively. A logistic link function converts the q_{jk} variables into abnormal-interpretation probabilities,

$$a_{jk}^- = \frac{\exp\left(q_{jk}^-\right)}{1+\exp\left(q_{jk}^-\right)} \quad \text{and} \quad a_{jk}^+ = \frac{\exp\left(q_{jk}^+\right)}{1+\exp\left(q_{jk}^+\right)}. \tag{2}$$

Note that a_{jk}^- represents the probability of a false-positive interpretation and a_{jk}^+ represents a true-positive interpretation, conditioned on a specific reader and a specific case. For a given reader, the unconditional probability of an abnormal interpretation (across cases) is the expected value of the probabilities in equation 2 over the distribution of case effects,

$$a_j^- = E\left(a_{jk}^-\right)_{C_k^- \sim N\left(0,\sigma_{C^-}\right)} \quad \text{and} \quad a_j^+ = E\left(a_{jk}^+\right)_{C_k^+ \sim N\left(0,\sigma_{C^+}\right)} \tag{3}$$

These probabilities represent the reader's false-positive rate (1 − specificity) and true-positive rate (sensitivity) over the population of cases. Let A_j^- and A_j^+ represent the number of abnormal findings for negative and positive cases, respectively; these are presumed to be related to the reader parameters through binomial distributions,

$$A_j^- \sim B\left(a_j^-, K_j^-\right) \quad \text{and} \quad A_j^+ \sim B\left(a_j^+, K_j^+\right) \tag{4}$$

Equations 1 to 4 relate the model parameters (equation 1) to observed single-reader data (equation 4).

Up to this point, we have analyzed only single-reader performance, but we can use a similar approach to extend the approach to disagreement rates between paired readers. For a pair of readers, indexed by j and j' ($j \neq j'$), let the negative cases be indexed by $k = 1, \ldots, K_{jj}^-$ and the positive cases be indexed by $k = 1, \ldots, K_{jj}^+$. We can define case-specific disagreement rates based on the probabilities in equation 2 as

$$d_{jj'k}^- = a_{jk}^-\left(1-a_{j'k}^-\right) + a_{j'k}^-\left(1-a_{jk}^-\right) \quad \text{and} \quad d_{jj'k}^+ = a_{jk}^+\left(1-a_{j'k}^+\right) + a_{j'k}^+\left(1-a_{jk}^+\right) \tag{5}$$

These case-specific probabilities can be marginalized into reader-pair–specific probabilities by taking the expectation across cases,

$$d_{jj'}^- = E\left(d_{jj'k}^-\right)_{C_k^- \sim N\left(0,\sigma_{C^-}\right)} \quad \text{and} \quad d_{jj'}^+ = E\left(d_{jj'k}^+\right)_{C_k^+ \sim N\left(0,\sigma_{C^+}\right)} \tag{6}$$

Let $D_{jj'}^{-}$ and $D_{jj'}^{+}$ represent the number of disagreements for negative and positive cases, respectively; these are presumed to be related to the reader parameters through binomial distributions,

$$D_{jj'}^{-} \sim B\left(d_{jj'}^{-}, K_{jj'}^{-}\right) \quad \text{and} \quad D_{jj'}^{+} \sim B\left(d_{jj'}^{+}, K_{jj'}^{+}\right) \tag{7}$$

Note that both the abnormal-interpretation probabilities, a_j^- and a_j^+, and the disagreement probabilities, $d_{jj'}^-$ and $d_{jj'}^+$, are related to the data through binomial distributions. But the disagreement component of the model plays a crucial role, since the model is not identifiable (i.e., multiple parameter values produce identical abnormal-interpretation probabilities) based on the abnormal interpretation component alone.

Model fitting.—Our implementation of the model computed expectations in equations 3 and 6 using Monte Carlo integration with 1,000,000 sample case effects for each integral. This resulted in 1,000,000 sample abnormal interpretation and disagreement probabilities in equations 2 and 5, and these were averaged to get high-precision estimates of reader abnormal-interpretation probabilities and paired disagreement rates. For rates that were near the estimated values, the Monte Carlo coefficient of variation on abnormal interpretation rates was less than 1%.

Estimation of the reader parameters and the case variances was performed by maximizing the log-likelihood defined by the binomial distributions in equations 4 and 7. This approach allows the computation of individual reader performance characteristics as well as paired disagreement probabilities. We describe this for positive cases, but there is a corresponding log-likelihood for negative cases. The resulting optimization function is

$$\begin{aligned}\lambda\left(r_1^+, \ldots, r_{N_R}^+, \sigma_{C^+}\right) = &\sum_{j=1} \left(K_j^+ \, ln\left(a_j^+\right) + \left(K_j^+ - A_j^+\right) ln\left(1 - a_j^+\right)\right) \\ &+ \sum_{(j,j')\in\Pi} \left(\mathrm{K}_{\mathrm{jj'}}^{+} \, ln\left(\mathrm{d}_{\mathrm{jj'}}^{+}\right) + \left(\mathrm{K}_{\mathrm{jj'}}^{+} - \mathrm{D}_{\mathrm{jj'}}^{+}\right) ln\left(1 - \mathrm{d}_{\mathrm{jj'}}^{+}\right)\right) + X\end{aligned} \tag{8}$$

where X represents constant terms unrelated to the parameters (log-factorial components) and Π represents the set of observed reader pairs in the data. An optimization algorithm implementing Powell's method [124] with standard convergence criterion ($f_{Tol} = 10^{-4}$) was used to find maximum likelihood estimates of the reader parameters (r_j^- and r_j^+) and case standard deviations (σ_{C^-} and σ_{C^+}) that maximize this function.

For graphical evaluations of model fit, estimated reader abnormal interpretation rates (a_j^- and a_j^+, equation 3) and paired disagreement rates ($d_{jj'}^-$ and $d_{jj'}^+$, equation 6) were compared to direct estimates of the quantities from the data. Direct estimates for the abnormal interpretation rates are given by

$$FPR_j = A_j^-/K_j^- \quad \text{and} \quad TPR_j = A_j^+/K_j^+ \tag{9}$$

and for the disagreement rates by

$$DR_{jj'}^- = D_{jj'}^-/K_{jj'}^- \quad \text{and} \quad DR_{jj'}^+ = D_{jj'}^+/K_{jj'}^+ \tag{10}$$

These rates were adjusted using the Agresti-Coull procedure [125], which was also used to obtain 95% confidence intervals.

3.2.2 Double-Reading Simulation

Assuming the logistic regression models adequately accounted for single-reader performance and disagreement rates between paired readers, the reader parameters (r_j^- and r_j^+) and case standard deviations (σ_{c^-} and σ_{c^+}) were used to simulate double reading and investigate the effect of different pairings of readers computationally. A random pairing strategy was compared against 2 main other pairing strategies, including 1) pairing readers with *similar* performance characteristics and 2) pairing readers with *opposite* performance characteristics. In total, 7 different pairing strategies were simulated:

1. Similar true-positive rate (TPR)
2. Similar false-positive rate (FPR)
3. Opposite TPR
4. Opposite FPR
5. Total similars = combination of TPR and FPR (TPR + FPR × slope)
6. Total opposites = combination of TPR and FPR (TPR + FPR × slope)
7. Random

For the pairing of similar and opposite readers, the modeled reader-specific TPR and FPR from equation 3 were used. For the 3 similar pairing strategies, the available reader with the lowest performance value in question (TPR, FPR, TPR + FPR × slope) was selected and paired with the reader closest in corresponding performance (i.e., had the second lowest value, Figure 3.1A). This pairing continued for the readers with the third and fourth lowest values, and so forth, until all readers were assigned. For the 3 opposite pairing strategies, readers were split into 2 groups, as above and below the median of that performance metric. The available reader with the lowest performance value in question (TPR, FPR, TPR + FPR × slope) in one group (i.e.,

below the median) was paired with the available reader with the lowest value in the other group (i.e., above the median), until all readers were assigned (Figure 3.1B). For the total similar and opposite pairing strategies, the same pairing strategy as shown in Figure 3.1A and 3.1B was used, but now the slopes of the linear regression lines of the individual TPR and FPR were used for pairing (TPR + FPR × slope). For random pairing, readers were randomly sampled without replacement, making sure that no reader was paired with themselves.

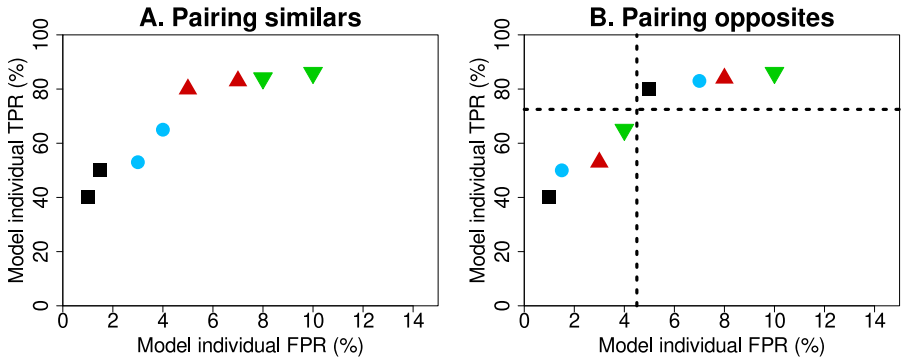


Figure 3.1: Explanation for the pairing of readers being either (A) similar or (B) opposite in their performance characteristics. The colors and symbols represent the pairs of readers, and the dashed lines are the median true-positive rate (TPR) and false-positive rate (FPR) of the readers.

For pairs, case-specific paired abnormal interpretation rates based on the probabilities in equation 2 are defined as

$$p^{-}_{jj'(k)} = a^{-}_{jk} + a^{-}_{j'k} - (a^{-}_{jk})(a^{-}_{j'k}) \quad \text{and} \quad p^{+}_{jj'(k)} = a^{+}_{jk} + a^{+}_{j'k} - (a^{+}_{jk})(a^{+}_{j'k}) \tag{11}$$

where $jj'(k)$ is defined as the pair of readers assigned a given examination. Equation 11 defines disagreement between 2 readers as an abnormal interpretation for the pair. This maintains a high sensitivity in which cancer cases are, at least, being discussed in a consensus meeting or send to a third reader for arbitration [126].

3.2.3 Data Sets

For this study, the above-described model was derived using the screening reading results from 3 different breast cancer screening programs: the Swedish CSAW (Cohort of Screen-Age Women) data set [108], the English CO-OPS (Changing case Order to Optimize patterns of Performance in Screening) data set [62,109], and a data set from BreastScreen Norway. These data sets were also used in a previously published article that investigated pairing strategies based on reader performance [119]. However,

the previous publication included analysis performed on the observed data, rather than using the data to derive a reader model to exhaustively explore all possible pairings. The study population and screening procedures of the 3 data sets have been previously described [47,62,108–110,119]. Briefly, the CSAW data set consists of women aged 40 to 74 years who participated in 2-view digital mammographic screening in the Stockholm region in Sweden between 2008 and 2015. The regional Ethical Review Board in Sweden approved the use of the CSAW data and waived the need for informed consent. The CO-OPS data set consists of women aged 47 to 73 years who attended 2-view digital mammography screening between 2012 and 2014 in England. Institutional Review Board approval for the original CO-OPS trial was obtained from Coventry and Warwickshire National Health Service Research Ethics Committee (June 27, 2012), and each breast screening director gave informed consent. The nationwide data set from the Norwegian screening program consists of women aged 50 to 69 years who underwent 2-view digital screening mammography between 2004 and 2018. The Norwegian data were obtained in accordance with the legal bases outlined in the Norwegian Cancer Registry Regulations of December 21, 2001, No. 47, and the requirement for informed consent was waived. In the 3 screening programs, mammograms are double read by 2 readers, who have access to prior mammograms if available. There are some variations in recall protocols, as depicted in Figure 3.2. Across all programs, agreements on negative assessments are deemed normal, while discrepancies are resolved either through a third reader's evaluation or a consensus meeting. Notably, in Sweden and Norway, agreed positive assessments prompt discussion in a consensus meeting to assess the necessity of further diagnostic evaluation, whereas in England, such agreements typically lead to immediate recall.

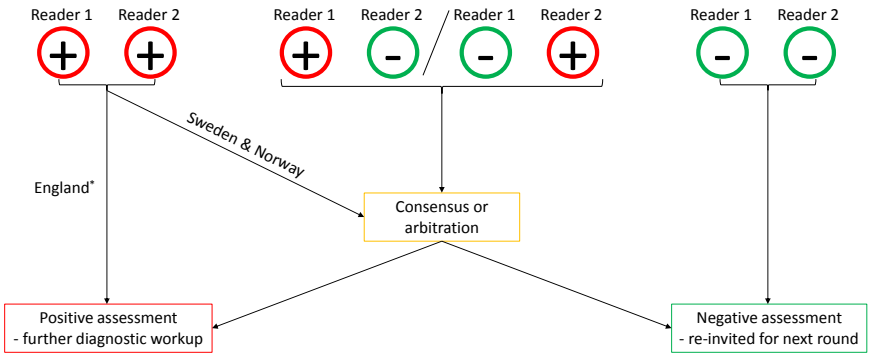


Figure 3.2: Explanation of the recall procedures in the involved data sets. The red circle (+) represents a positive assessment warranting further diagnostic workup, while the green circle (–) represents a negative assessment. * = At most centers, only discrepant readings were resolved with a third reader or consensus meeting. However, at certain centers with elevated recall rates, arbitration was also applied when both readers had a positive assessment.

Breast cancers included invasive cancers or ductal carcinoma in situ that were diagnosed with needle biopsy or surgery after diagnostic workup in screening or clinically in between screening rounds (i.e., interval cancer). Entries from symptomatic women and readings with missing reader data or inadequate images were excluded. For an appropriate fit of the logistic regression model, readers had to have at least 1 true-positive and false-positive assessment, and the number of true positives and false positives should not be equal to the total number of positive and negative examinations, respectively. To meet these criteria, pairs with a relatively low volume in the data set were also excluded.

3.2.4 Simulation to Optimize Screening Performance

Double-reading strategies.—The logistic regression models were used to explore radiologist pairing strategies based on individual screening performance, as described above. To mimic real-world breast cancer screening, each of the 7 above-described pairing strategies was simulated by sampling 365 screening days with 4,000 examinations distributed over 32 batches per day. For each screening day, a random set of 16 readers was chosen. Within the simulations, each of the batches was read by 2 of the 16 randomly selected radiologists reading that day, with the constraint that each reader read about the same number of batches. Each day, 8 unique pairs were created, ensuring a similar number of pairs per day for all pairing strategies. All 7 simulations involved the same examinations and readers; the only difference was the pairing of the readers. In the end, all 7 simulations resulted in 1,460,000 examinations ($4{,}000 \times 365$) with each having 2 individual outcome probabilities (reader 1 and reader 2) from equation 2 and a paired probability from equation 11. The main endpoint was the expected group TPR and FPR of each pairing strategy, given by

$$FPR_{total} = \sum_{k^-=1}^{K^-} p^-_{jj'(k^-)} \Big/ K^- \quad \text{and} \quad TPR_{total} = \sum_{k^+=1}^{K^+} p^+_{jj'(k^+)} \Big/ K^+ \tag{12}$$

Bootstrap resampling ($n = 1{,}000$) was used to obtain 95% confidence intervals for the group performance. The 95% confidence intervals were Bonferroni corrected for 6 comparisons and used to compare the 3 similar and 3 opposite pairing strategies against the random pairing strategy. TPR and FPR were plotted, and nonoverlapping confidence intervals were regarded as statistically significant.

In addition, differences in the number of true-positive (TP) and false-positive (FP) examinations between the different pairing strategies and the random pairing strategy were illustrated. For this, individual outcome probabilities (equation 2) were sampled as absolute values of 0 (indicating no suggested recall) or 1

(indicating suggested recall) according to the Bernoulli distribution, resulting in 3 potential paired assessments—concordant positive, concordant negative, and discordant assessments—facilitating the evaluation of differences in the number of TP and FP examinations.

Individual reading.—To compare the performance of the double-reading pairing strategies to that of individual readers, individual reader simulations were also performed. The simulations involved the same examinations and readers as for the paired simulations, but this time each examination was interpreted by only 1 of the 16 random readers reading that day, again with the constraint that each reader interpreted the same number of batches. Individual abnormal interpretation probabilities were obtained from equation 2 and used to calculate grouped TPR and FPR of examinations read by 1 reader only

$$FPR_{single} = \sum_{k^-=1}^{K^-} a^-_{j(k^-)} \Big/ K^- \quad \text{and} \quad TPR_{single} = \sum_{k^+=1}^{K^+} a^+_{j(k^+)} \Big/ K^+ \tag{13}$$

where $j(k)$ indicates the reader assigned to examination k. The 95% confidence intervals, obtained from bootstrap resampling ($n = 1{,}000$), were Bonferroni corrected for 7 comparisons contrasting the TPR and FPR of the 7 double-reading strategies against the TPR and FPR of the individual reading.

All statistical analyses were performed in each of the 3 data sets separately. Model fitting was performed in the Interactive Data Language version 8.2.3 (IDL, L3Harris Geospatial, Broomfield, CO, USA), and the resulting model parameters were subsequently used to simulate pairing strategies in R studio version 4.1.0 (RStudio, PBC, Boston, MA, USA). The funding source ensured the authors' independence in designing the study, interpreting the data, and writing the article.

3.3 Results

Figure 3.3 summarizes how the final study samples were assembled. For an appropriate fit of our logistic regression model, an exclusion criterion of at least 17 positive reads per pair was needed to make sure that the number of true and false positives was not equal to the total number of positive and negative examinations, respectively. The final study samples consisted of $n = 936{,}621$ (Sweden), $n = 435{,}281$ (England), and $n = 1{,}820{,}053$ (Norway) screening examinations.

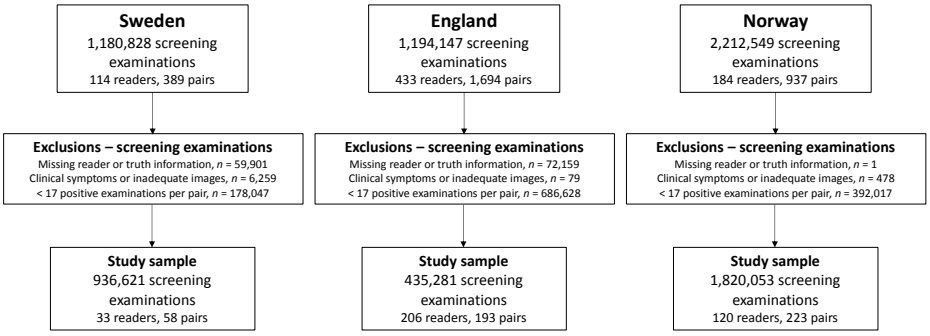


Figure 3.3: Flowchart of screening examinations after applying exclusion criteria.

3.3.1 Study Sample Characteristics

The population and reader characteristics varied among the study samples (Table 3.1). The Swedish study sample included the youngest population of women with a median age of 53 years at screening. Paired TPR and FPR in all study samples were higher than individual TPR and FPR as a result of our pairing rule, where disagreement between 2 readers was defined as a positive assessment. The English study sample showed the highest average individual TPR (78.2%) and paired TPR (83.0%), whereas the Norwegian study sample had the highest individual and paired FPRs (5.1% and 8.0%, respectively). Norway's study sample also showed most disagreement between readers in a pair (5.6%).

Table 3.1: Population, reader, and pair characteristics for the study samples after selection criteria.

	Sweden	England	Norway
	n = 936,621 examinations	*n* = 435,281 examinations	*n* = 1,820,053 examinations
Population characteristics			
Women, *n*	390,680	435,281	653,161
Median age at screening, years (IQR)	53 (46–62)	59 (53–65)[*]	59 (54–64)
Screening programs (double reading)			
Screening interval, months	18–24[†]	36	24
Recalled examinations, per 1,000	17.8[‡]	41.1	29.7
Detected cancers, per 1,000	3.3[‡, §]	9.2[#]	5.8[#]
Interval cancers, per 1,000	1.8[‡, ǁ]	2.2[**]	1.8[**]
Reader characteristics			
Readers, *n*	33	206	120
Average true-positive rate, %	63.0	78.2	70.2
Average false-positive rate, %	3.5	4.1	5.1

Table 3.1: Continued

	Sweden	England	Norway
Pair characteristics			
Pairs, *n*	58	193	223
Average true-positive rate, %	70.3	83.0	81.6
Average false-positive rate, %	5.0	5.5	8.0
Disagreement between readers in a pair, %	2.3	3.0	5.6

IQR = interquartile range. * = Age was missing for 4 screening examinations. † = Women aged 49 years and older were invited every 24 months, while younger women were invited every 18 months. ‡ = The final screening assessment was missing for 16 screening examinations. § = Recall and breast cancer detection within 12 months after screening. ǁ = Women who did not have a screen-detected cancer but had a breast cancer diagnosed within 18 to 24 months after screening. # = Breast cancer detected before the next screening examination as a result of recall at screening. ** = Women who did not have a screen-detected cancer but had a breast cancer diagnosed before the next screening round.

3.3.2 Model Fit

The single-reader performance from the model showed good agreement with the Agresti-Coull adjusted observed data (Figure 3.4, dark-blue triangles). The individual TPR values showed Pearson correlation coefficients of 0.969, 0.973, and 0.988 for the Swedish, English, and Norwegian study sample, respectively. The individual FPR values had Pearson correlation coefficients of 0.978, 0.990 and 0.996 for the Swedish, English, and Norwegian study samples.

The light-blue dots in Figure 3.4 show the modeled disagreement rates compared against the observed data. For this comparison only, the pairs observed in the actual screening data set could be used (Sweden: 58 of the 528 possible pairs; England: 193 of the 21,115 possible pairs; Norway: 223 of the 7,140 possible pairs). Disagreement rates for positive cases showed lower levels of correlation with the Swedish, English, and Norwegian study samples, having Pearson correlation coefficients of 0.727, 0.658, and 0.532, respectively. As expected, pairs that read a relatively small number of examinations had large error bars for the estimates of positive pair disagreement. Disagreement rates for negative cases showed higher correlations with coefficients of 0.709, 0.846, and 0.923 for the Swedish, English, and Norwegian study samples, respectively.

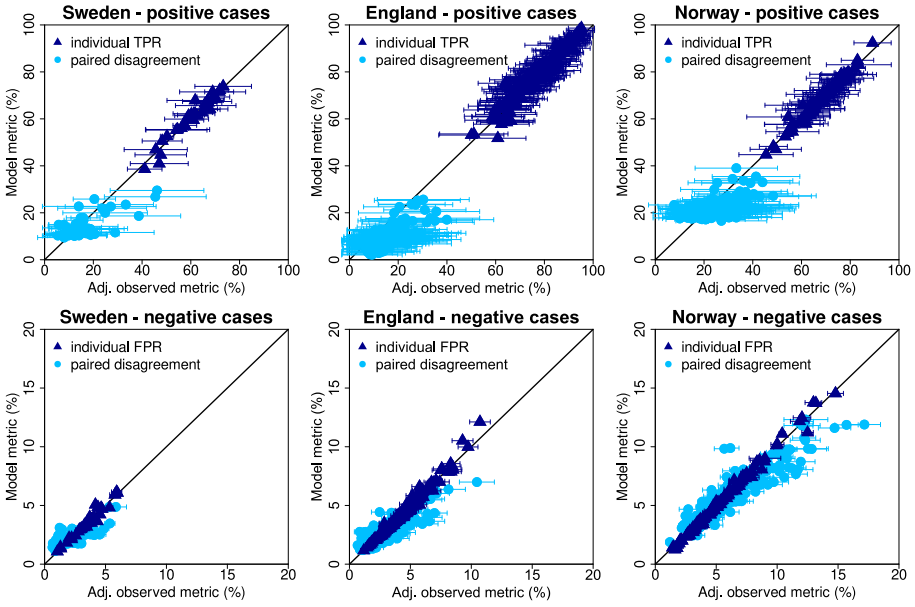


Figure 3.4: Comparison of modeled and observed single-reader performance and paired disagreement rates. The scatterplots show single-reader performance and paired disagreement rates for the Swedish, English, and Norwegian study sample. The dark-blue triangles represent single-reader performance (top: true-positive rates, bottom: false-positive rates). The light-blue dots represent the disagreement rates for pairs of readers (top: positive cases, bottom: negative cases). The diagonal represents perfect agreement between the modeled and observed data. The observed proportions were adjusted using the Agresti-Coull procedure for 95% confidence intervals. FPR = false-positive rate, TPR = true-positive rate.

The modeled individual TPR and FPR showed the commonly found association between TPR and FPR, with high TPR readers tending to have higher FPR (Figure 3.5).

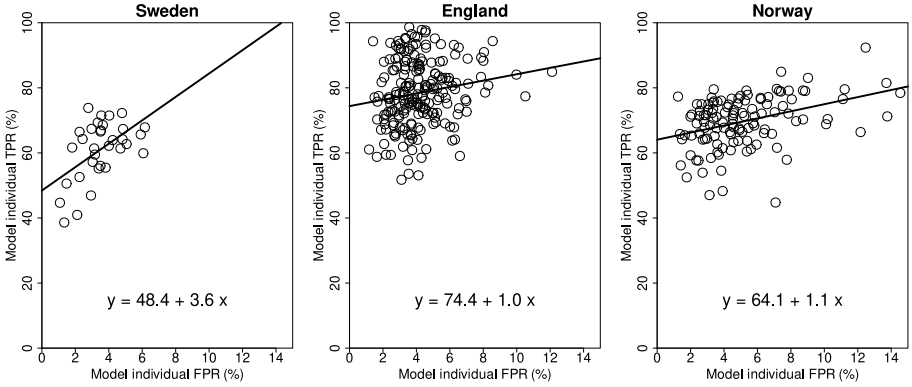


Figure 3.5: Modeled TPR and FPR of the individual readers. The scatterplots show true-positive and false-positive rates for the Swedish, English, and Norwegian study sample. The linear regression line shows the association between true-positive rate (TPR) and false-positive rate (FPR).

3.3.3 Screening Simulation

Double-reading strategies.—The grouped TPR and FPR of the 7 modeled pairing strategies, based on the modeled individual TPR and FPR from Figure 3.5, are shown in Figure 3.6 and Table 3.2. The pattern for all 3 study samples looks similar. According to our pairing rule, pairing readers who are similar in a certain performance metric resulted in a lower value for the paired outcome of that metric. Pairing strategies involving readers with similar TPR (TPR Sweden: 64.72% [95% confidence interval {CI}: 63.36%–66.07%], England: 81.23% [95% CI: 80.52%–81.95%], Norway: 78.83% [95% CI: 78.06%–79.60%]) resulted in a statistically significant reduction of the TPR when compared with random pairing strategies (Sweden: 67.95% [95% CI: 66.64%–69.25%], England: 85.38% [95% CI: 84.73%–86.03%], Norway: 80.66% [95% CI: 79.93%–81.39%]). In addition, the FPR of pairing strategies involving 2 similar FPR readers (FPR Sweden: 4.50% [95% CI: 4.46%–4.54%], England: 5.51% [95% CI: 5.47%–5.56%], Norway: 8.03% [95% CI: 7.99%–8.07%]) was statistically significantly lower compared with the FPR of the random pairing strategies (Sweden: 4.74% [95% CI: 4.70%–4.78%], England: 5.76% [95% CI: 5.71%–5.80%], Norway: 8.30% [95% CI: 8.26%–8.34%]). Conversely, pairing opposite readers increased the value of the paired outcome of what they were opposite in. However, none of the pairing strategies resulted in a statistically significant increased TPR compared with the random pairing strategy (Figure 3.6, Table 3.2). To extrapolate these findings to the context of screening programs, the resulting recall rate (RR) and CDR of the different modeled pairing strategies were calculated (Figure 3.S1). For all 3 data sets, pairing readers with similar FPR resulted in a significantly lower RR, while CDR did not significantly change.

Figure 3.7 shows the difference in the absolute number of TP and FP between the similar/opposite pairing strategies and the random pairing strategy, according to our pairing rule. The pairing strategy with similar FPR readers resulted in the best outcome, reducing the number of FPs with reductions of −3,605 (68,866 → 65,261; −5.23%), −3,447 (83,073 → 79,626; −4.15%), and −3,475 (120,146 → 116,671; −2.89%) FP screening examinations for Sweden, England, and Norway, respectively. The other pairing strategies resulted in similar or worse group outcomes. Compared with the random pairing strategy, the pairing strategy with similar TPR readers reduced the number of TP the most with −225 (5,084 → 4,859; −4.43%), −681 (14,364 → 13,683; −4.74%), and −192 (9,000 → 8,808; −2.13%) TP screening examinations for Sweden, England, and Norway, respectively. For the opposite pairing strategies, the number of FPs increased compared with random pairing, especially for pairs with opposite FPR characteristics readers (FP Sweden: [+739, 68,866 → 69,605; +1.07%],

England: [+1,291, 83,073 → 84,364; +1.55%], Norway: [+1,627, 120,146 → 121,773; +1.35%]). The TPs for the opposite pairing strategies remained essentially similar.

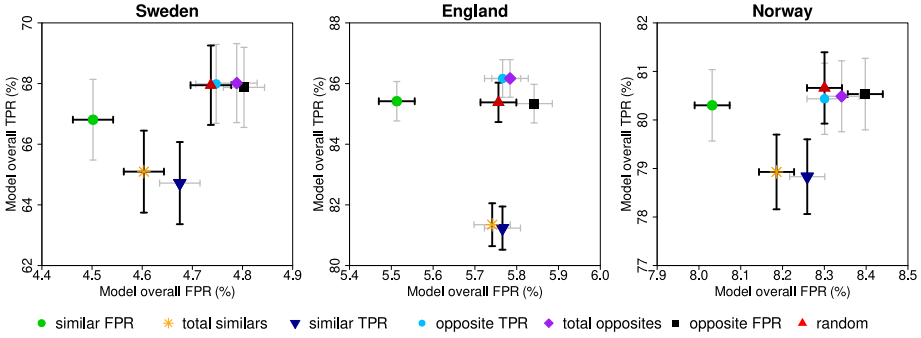


Figure 3.6: Screening performance for the different pairing strategies based on paired assessment. The scatterplots show overall true-positive and false-positive rates of the different pairing strategies for the Swedish, English, and Norwegian study samples. The colors and symbols represent the different pairing strategies. Error bars are 95% confidence intervals, obtained by bootstrap resampling ($n = 1{,}000$) and adjusted for 6 comparisons. Bold error bars indicate statistical significance. Please note that the axes are different, due to differences in true-positive rate (TPR) and false-positive rate (FPR) between the study samples.

Table 3.2: Screening performance for the different pairing strategies.

	Sweden		**England**		**Norway**	
	TPR (95% CI)	**FPR (95% CI)**	**TPR (95% CI)**	**FPR (95% CI)**	**TPR (95% CI)**	**FPR (95% CI)**
Similar FPR	66.81 (65.48, 68.14)	4.50 (4.46, 4.54)*	85.42 (84.77, 86.07)	5.51 (5.47, 5.56)*	80.30 (79.57, 81.04)	8.03 (7.99, 8.07)*
Total similars	65.10 (63.74, 66.45)*	4.60 (4.56, 4.64)*	81.35 (80.64, 82.05)*	5.74 (5.70, 5.78)	78.93 (78.16, 79.70)*	8.19 (8.14, 8.23)*
Similar TPR	64.72 (63.36, 66.07)*	4.67 (4.63, 4.72)	81.23 (80.52, 81.95)*	5.77 (5.72, 5.81)	78.83 (78.06, 79.60)*	8.26 (8.22, 8.30)
Opposite TPR	67.99 (66.69, 69.29)	4.75 (4.71, 4.79)	86.17 (85.54, 86.79)	5.77 (5.72, 5.81)	80.44 (79.70, 81.17)	8.30 (8.26, 8.34)
Total opposites	68.01 (66.71, 69.32)	4.79 (4.75, 4.83)	86.17 (85.55, 86.79)	5.78 (5.74, 5.83)	80.49 (79.76, 81.22)	8.34 (8.30, 8.38)
Opposite FPR	67.87 (66.55, 69.19)	4.80 (4.76, 4.84)	85.34 (84.70, 85.97)	5.84 (5.80, 5.89)	80.53 (79.80, 81.27)	8.40 (8.36, 8.44)*
Random (reference)	67.95 (66.64, 69.25)	4.74 (4.70, 4.78)	85.38 (84.73, 86.03)	5.76 (5.71, 5.80)	80.66 (79.93, 81.39)	8.30 (8.26, 8.34)

Note.—True-positive rate (TPR) and false-positive rate (FPR) are percentages and 95% CI are Bonferroni adjusted (P values < .05/6) confidence intervals, obtained by bootstrap resampling (n = 1,000). Nonoverlapping confidence intervals with the random pairing strategy were regarded as statistically significant, as indicated by the asterisk.

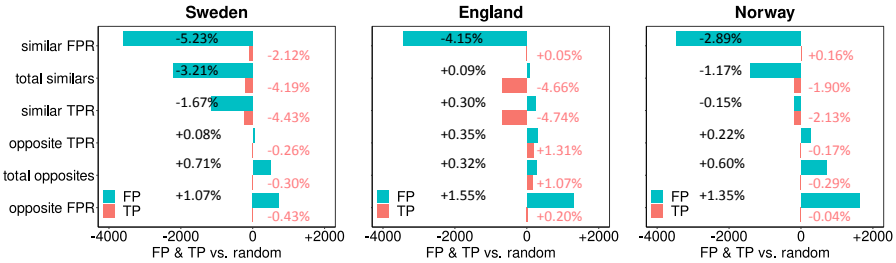


Figure 3.7: Change in the number of paired true positives (pink) and false positives (blue) for the different similar and opposite pairing strategies compared to the random pairing strategy. This simulation assumed a population size of 1,460,000 examinations. Positive assessments are determined by our pairing rule, which defines an examination as positive if any of the readers flags it as abnormal. Percentages indicate the change in the number of true positive (TP) or false positive (FP) compared with random pairing and vary as random pairing results in different numbers for TP and FP across the 3 different study samples.

Individual reading.—The performance measures of the individual reading simulation were compared against the double-reading performance measures for the different pairing strategies (Figure 3.8, Table 3.S1). The TPR and FPR of the individual simulation (Sweden: TPR: 60.46% [95% CI: 59.08%–61.83%], FPR: 3.50% [95% CI: 3.47%–3.54%], England: TPR: 78.72% [95% CI: 77.97%–79.48%], FPR: 4.17% [95% CI: 4.13%–4.21%], and Norway: TPR: 69.53% [95% CI: 68.68%–70.39%], FPR: 5.10% [95% CI: 5.06%–5.13%]) were statistically significantly lower compared with the TPR and FPR of all double-reading pairing strategies in all 3 data sets (Sweden: TPR ≥ 64.72%, FPR: ≥ 4.50%, England: TPR ≥ 81.23%, FPR: ≥ 5.51%, Norway: TPR ≥ 78.83%, FPR: ≥ 8.03%).

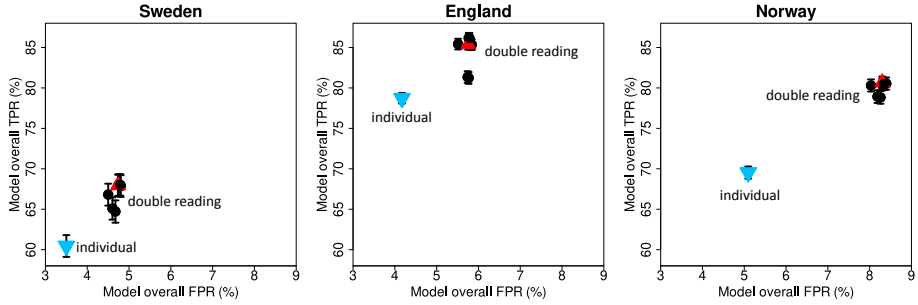


Figure 3.8: Screening performance for individual reading and different double-reading strategies. The scatterplots show the overall true-positive and false-positive rates of the different pairing strategies for the Swedish, English, and Norwegian study samples. The red triangles represent the average screening performance for the random pairing strategy, the black dots represent the specific pairing strategies, and the blue triangle with point down represents the performance for individual reading. Error bars are 95% confidence intervals, obtained by bootstrap resampling ($n = 1{,}000$) and adjusted for 7 comparisons. The true-positive rate (TPR) and false-positive rate (FPR) of the individual simulation were statistically significantly lower compared with the TPR and FPR of the double-reading pairing strategies in all 3 data sets.

3.4 Discussion

As demonstrated in this study, modeling can overcome limitations inherent to using real-world data directly. Our logistic regression models allowed for the exploration of both individual and paired reader performance during screening mammogram interpretation exhaustively, evaluating the latter even for case readings that did not take place in the real world. The strong Pearson correlations between the model-predicted values and observed data suggest that these models can effectively capture single-reader performance and disagreement rates between paired readers. While our study focused on pairing strategies in the context of breast cancer screening, the approach may have potential for broader applications across diverse research fields. The ability to model and simulate scenarios opens opportunities to explore real-world data sets with more statistical power.

There might be some discrepancies between the predicted performance and actual outcomes if this were to be implemented in actual screening practice. However, we anticipate that the modeled performance in this study will closely approximate actual performance since individual TPR and FPR and case-negative disagreement rates showed robust associations (Pearson correlation > 0.7) between model-predicted and observed data. Case-positive disagreement rates resulted in lower levels of association. This can be attributed to the relatively low number of positive cases per pair in the observed data, highlighting the challenge of assessing reader performance in the context of rare events. Nevertheless, the derived models allowed for the exploration of different pairing strategies, something that would not have been possible with sufficient statistical power in actual screening reading data sets. Of particular interest was the pairing strategy that emphasized the similarity in FPR between readers. This strategy, when compared with random pairing, consistently yielded a reduction in the number of FPs for all 3 study samples. This can be explained by the fact that similar FPR readers had a higher degree of concordance in negative assessments, leading to lower paired FPR values and thus lower grouped FPR, whereas random or opposite characteristic readers in a pair disagree more, which, in turn, led to higher FPR values based on our pairing rule in which discrepant readings were classified as positive readings. Constructing pairs of radiologists with similar FPR characteristics did not show a significant change in overall TPR. This suggests that strategically pairing radiologists with similar FPR characteristics may increase reading performance. Additional analyses (Figure 3.S1) also showed that pairing readers with similar FPR resulted in a significantly lower RR, while CDR did not significantly change. Therefore, pairing readers with similar FPR characteristics, thereby reducing FPR/RR without significantly changing TPR/CDR,

may offer a practical solution to reduce the number of unnecessary examinations forwarded to consensus or arbitration, most probably without negatively affecting CDRs in actual practice. Such a reduction will reduce the workload on health care professionals and the number of potential false-positive recalls, alleviating the burden on both screened women and the health care system.

The differences in performance introduced by the different pairing strategies are overall small, as could be expected [54], but for a large screening program with a single picture archiving and communication systems (PACS) system, so might be the costs, if any, of implementing a specific paired-reading strategy. Such a PACS system, from which readers can read mammograms from any location, provides opportunities for strategic pairing by implementing automated pairing algorithms. The automatic pairing algorithms could be based on the available radiologists and their regularly benchmarked performance metrics. Peer review and feedback sessions of discrepant examinations may help radiologists improve their screening performance and bring the number of discrepancies in their FP down, naturally employing our pairing strategy of pairing readers with similar FPR characteristics. However, before implementing and to ascertain cost-effectiveness, thorough analyses should be conducted.

Further studies should show how radiologists that are relatively similar in FPR but still somewhat different in TPR can be distinguished. This is crucial to avoid missing more cases of cancers. Further studies are also needed to assess the impact of pairing similar FPR readers on final screening outcomes after consensus/arbitration. In addition, it should be noted that pairing readers with similar characteristics results in higher variation in performance among different pairs. In contrast, pairing opposite characteristic readers results in more consistent performance measures across pairs, consequently leading to more consistent results among women. There may also be another optimal pairing strategy that further maximizes paired reading performance based on other factors. A previously published study by Gandomkar et al. [55] proposed pairing radiologists with different cognitive eye-tracking metrics to optimize double reading. However, this study was performed with an enriched test set and investigated the performance of different pairs but not the group performance of all radiologists together. Furthermore, eye trackers are not, perhaps yet, used in screening programs. Eventually, prospective screening studies should be considered before implementing any pairing strategy in practice.

Our findings also raise possibilities for other future applications, like lung cancer screening. Even in the nonmedical environment, similar models could be used, for

instance if double reading were ever used in airport bag screening. Also, in breast cancer screening programs with single reading, a new potential use of an optimal pairing strategy may be if the program incorporates AI as a second stand-alone reader. This would allow for the AI settings to be adjusted to the performance of the paired human reader to optimize this hybrid reading screening performance. This is especially important since our findings do suggest that double reading is an important mechanism to improve early detection of breast cancer (Figure 3.8), as previously shown [45,50,102]. However, the individual modeling in this study is based on readers who in screening practice performed double reading. Readers who know they will be reading on their own will probably behave differently, thus warranting further research. In programs already using double reading, AI might eventually replace one of the readers. The human-AI combination would then again allow for the AI settings to be adjusted to the performance of the paired human reader to optimize the combination of the human reader and AI. Finally, using AI for decision support for radiologists may help to align the FPR of the readers, enhancing our pairing strategy, especially when both readers rely on AI in a similar manner. However, such an approach warrants thorough additional investigation and is deemed beyond the scope of our current study objectives.

Our study has several strengths and limitations. A major strength of this study is the inclusion of 3 different data sets, which allowed us to evaluate the effect of the pairing strategies in 3 different countries with different screening practices. The fact that the results for all 3 data sets appear similar strengthens our conclusions. Furthermore, the models of radiologists' assessments developed in this study showed strong associations between the modeled and observed data and may thus be helpful for future modeling studies that include actual screening interpretation, in which each examination is read by either 1 or 2 radiologists. A limitation of our study was that the pairing strategies were based on the difference in performance between readers, but many different combinations of pairing strategies exist, and we did not maximize the pairing strategies to be as similar or opposite as possible as a group. Furthermore, consensus/arbitration assessments could not be modeled as we did not have information on the consensus/arbitration readers to develop a model of that screening interpretation stage. We also did not have information on which readers were blinded, so we could not control for these differences. Further research with a large radiologists' cohort that includes more information on lesions, the radiologists' experience levels, and characteristics of consensus and/or arbitration readers is needed to possibly identify other prospective selection criteria for optimally pairing readers.

In conclusion, this study showed the potential of logistic regression models to predict individual and paired reader performance during screening mammogram interpretation. With it, it was shown that strategic pairing of radiologists with similar FPR characteristics demonstrates promise in reducing false positives while preserving overall TPR. Telemedicine provides opportunities for strategic pairing, but the implementation of automated pairing algorithms and exploration of additional prospective selection criteria for double reading are essential steps in translating these findings into practice. Future research should also focus on the impact of consensus and arbitration decisions on reader pairing strategies.

Acknowledgments

The CSAW data set was obtained under the authorization of the regional Ethical Review Board in Stockholm. The CO-OPS data set received funding through an NIHR Postdoctoral Fellowship and an NIHR Career Development Fellowship (CDF-2016-09-018). The data set from BreastScreen Norway was obtained in compliance with the Cancer Registry of Norway Regulations. The simulated data sets generated and/or analyzed as part of the current study are available from the corresponding author upon reasonable request.

Supplementary Material

To extrapolate the findings of Figure 3.6 to the context of screening programs, the resulting recall rate (RR) and cancer detection rate (CDR) of the different modeled pairing strategies were calculated. Figure 3.S1 shows that the pattern of the RR and CDR graph looks similar to the pattern of the FPR and TPR graph (Figure 3.6). The confidence intervals indicate that pairing readers with similar FPR characteristics can significantly decrease RR without significantly affecting the CDR.

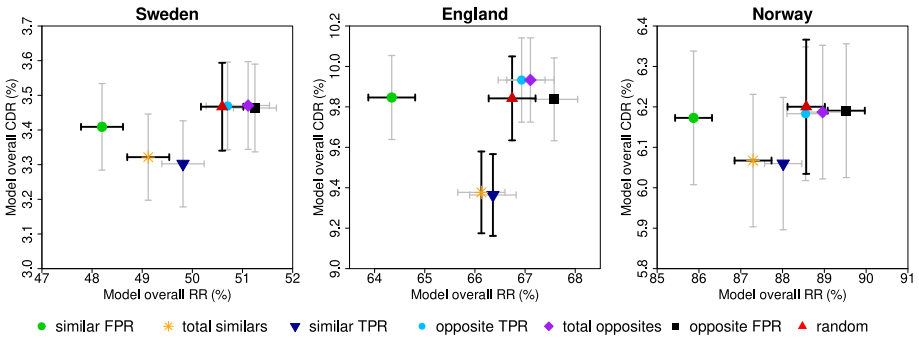


Figure 3.S1: Screening performance for the different pairing strategies based on paired assessment. The scatterplots show overall cancer-detection and recall rates of the different pairing strategies for the Swedish, English, and Norwegian study samples. The colors and symbols represent the different pairing strategies. Error bars are 95% confidence intervals, obtained by bootstrap resampling (n = 1,000) and adjusted for 6 comparisons. Bold error bars indicate statistical significance. Please note that the axes are different, due to differences in cancer detection rate (CDR) and recall rate (RR) between the study samples.

As a secondary analysis, the performance of individual reading was compared against that of double reading (Table 3.S1). Instead of comparing the 6 double-reading pairing strategies against the random pairing strategy (Bonferroni adjustment .05/6), we now compared individual reading against all 7 double-reading strategies (Bonferroni adjustment .05/7). Therefore, the confidence intervals in Table 3.S1 are slightly wider than those in Table 3.2 from the manuscript.

Table 3.S1: Screening performance for individual reading and double-reading pairing strategies.

	Sweden		England		Norway	
	TPR (95% CI)	**FPR (95% CI)**	**TPR (95% CI)**	**FPR (95% CI)**	**TPR (95% CI)**	**FPR (95% CI)**
Similar FPR	66.81 (65.45, 68.16)*	4.50 (4.46, 4.54)*	85.42 (84.76, 86.08)*	5.51 (5.47, 5.56)*	80.30 (79.55, 81.05)*	8.03 (7.99, 8.07)*
Total similars	65.10 (63.72, 66.47)*	4.60 (4.56, 4.64)*	81.35 (80.63, 82.07)*	5.74 (5.70, 5.79)*	78.93 (78.14, 79.71)*	8.19 (8.14, 8.23)*
Similar TPR	64.72 (63.34, 66.10)*	4.67 (4.63, 4.72)*	81.23 (80.51, 81.96)*	5.77 (5.72, 5.81)*	78.83 (78.05, 79.62)*	8.26 (8.22, 8.30)*
Opposite TPR	67.99 (66.66, 69.31)*	4.75 (4.71, 4.79)*	86.17 (85.53, 86.80)*	5.77 (5.72, 5.81)*	80.44 (79.69, 81.18)*	8.30 (8.26, 8.34)*
Total opposites	68.01 (66.69, 69.34)*	4.79 (4.75, 4.83)*	86.17 (85.54, 86.80)*	5.78 (5.74, 5.83)*	80.49 (79.74, 81.23)*	8.34 (8.30, 8.38)*
Opposite FPR	67.87 (66.53, 69.22)*	4.80 (4.76, 4.84)*	85.34 (84.69, 85.99)*	5.84 (5.80, 5.89)*	80.53 (79.78, 81.28)*	8.40 (8.36, 8.44)*
Random	67.95 (66.61, 69.28)*	4.74 (4.70, 4.78)*	85.38 (84.72, 86.04)*	5.76 (5.71, 5.80)*	80.66 (79.91, 81.41)*	8.30 (8.26, 8.34)*
Individual (reference)	60.46 (59.08, 61.83)	3.50 (3.47, 3.54)	78.72 (77.97, 79.48)	4.17 (4.13, 4.21)	69.53 (68.68, 70.39)	5.10 (5.06, 5.13)

Note.—True-positive rate (TPR) and false-positive rate (FPR) are percentages and 95% CI are Bonferroni adjusted (*P* values < .05/7) confidence intervals, obtained by bootstrap resampling (*n* = 1,000). Non-overlapping confidence intervals with the individual reading strategy were regarded as statistically significant, as indicated by the asterisk.

Chapter 4

Enhancing Radiologist Reading Performance by Ordering Screening Mammograms Based on Characteristics That Promote Visual Adaptation

Jessie J.J. Gommers, Sarah D. Verboom, Katya M. Duvivier, Jan-Kees van Rooden, A. Fleur van Raamt, Janneke B. Houwers, Dick B. Naafs, Lucien E.M. Duijm, Craig K. Abbey, Michael A. Webster, Mireille J.M. Broeders, Ioannis Sechopoulos

Published in Radiology 2024; 313(1):e240237.
DOI: 10.1148/radiol.240237

Abstract

Background

Mammographic background characteristics may stimulate human visual adaptation, allowing radiologists to detect abnormalities more effectively. However, it is unclear whether density, or another image characteristic, drives visual adaptation.

Purpose

To investigate whether screening performance improves when screening mammography examinations are ordered for batch reading according to mammographic characteristics that may promote visual adaptation.

Materials and Methods

This retrospective multireader multicase study was performed with mammograms obtained between September 2016 and May 2019. The screening examinations, each consisting of four mammograms, were interpreted by 13 radiologists in three distinct orders: randomly, by increasing volumetric breast density (VBD), and based on a self-supervised learning (SSL) encoding (examinations automatically grouped as "looking similar"). An eye tracker recorded radiologists' eye movements during interpretation. The area under the receiver operating characteristic curve (AUC), sensitivity, and specificity of random-ordered readings were compared with those of VBD- and SSL-ordered readings using mixed-model analysis of variance. Reading time, fixation metrics, and perceived density were compared using Wilcoxon signed-rank tests.

Results

Mammography examinations (75 with breast cancer, 75 without breast cancer) from 150 women (median age, 55 years [IQR, 50–63]) were read. The examinations ordered by increasing VBD versus randomly had an increased AUC (0.93 [95% CI: 0.91, 0.96] vs 0.92 [95% CI: 0.89, 0.95]; $P = .009$), without evidence of a difference in specificity (89% [871 of 975] vs 86% [837 of 975], $P = .04$) and sensitivity (both 81% [794 of 975 vs 788 of 975], $P = .78$), and a reduced reading time (24.3 vs 27.9 seconds, $P < .001$), fixation count (47 vs 52, $P < .001$), and fixation time in malignant regions (3.7 vs 4.6 seconds, $P < .001$). For SSL-ordered readings, there was no evidence of differences in AUC (0.92 [95% CI: 0.89, 0.95]; $P = .70$), specificity (84% [820 of 975], $P = .37$), sensitivity (80% [784 of 975], $P = .79$), fixation count (54, $P = .05$), or fixation time in malignant regions (4.6 seconds, $P > .99$) compared with random-ordered readings. Reading times were significantly higher for SSL-ordered readings compared with random-ordered readings (28.4 seconds, $P = .02$).

Conclusion

Screening mammography examinations ordered from low to high VBD improved screening performance while reducing reading and fixation times.

4.1 Introduction

Breast cancer screening has significantly improved the early detection of malignancies, thereby reducing breast cancer mortality [96,127,128]. However, given the challenging task of identifying cancers in complex mammographic backgrounds, the high volume of screening examinations, and the low prevalence of breast cancers at screening, there is a need to find strategies to increase the efficiency and accuracy of screening interpretation [35,62,129]. Retrospective studies revealed that 20%–50% of the interval and screen-detected cancers were visible on prior screening mammograms [28,30,130–133], suggesting that mammography alone could have been sufficient for detection of these cancers. These findings emphasize the potential to improve the interpretation of mammograms, hopefully resulting in earlier detection and improved prognosis.

Visual judgments of screening radiologists are affected by processes of human perception, including cognitive, perceptual, and decision-making factors, many of which have been studied in the context of mammography [66–68,134–136]. One such perceptual process is visual adaptation, which refers to changes in the sensitivity and operating characteristics of the visual system in response to the current stimuli the observer is viewing [137–139]. Visual adaptation can have a large influence on the appearance of images, known as visual aftereffects [140], and can also influence visual performance by altering sensitivity or the detectability of stimuli. For example, adaptation may serve to discount the expected properties of a visual environment, thereby enhancing the visual salience of unexpected information [141,142].

Few studies have considered adaptation in the context of radiologic assessments. However, Kompaniez-Dunigan et al. reported that exposure to mammogram patches altered how dense, fatty [66], or blurry [68] the images appeared. They also found that search times for simulated targets in fatty or dense mammogram patches were faster when the nonradiologist observers were adapted to images of the background type within which they were searching [67]. More recently, Parthasarathy et al. [69] reported similar adaptation aftereffects in digital breast tomosynthesis images. Although these results suggest that adaptation could significantly impact perception and performance during radiologic readings, a limitation was that these studies were conducted with nonradiologists under controlled conditions. Thus, the question remains whether similar adaptation effects manifest in screening radiologists in an actual screening reading environment.

Leveraging visual adaptation for batch reading screening mammograms requires the interpreting radiologist to be adapted to the characteristics of the mammogram

they are reading. To do this, the mammograms for interpretation need to be ordered so that images with characteristics that promote similar visual adaptation are grouped together. It is not clear, however, whether mammographic density, or another more complex image characteristic, is the main driver of visual adaptation for the interpreting radiologist. To our knowledge, there have been no previous attempts to order screening mammograms for individual batch reading based on image properties.

Therefore, the aim of this study was to determine whether the screening performance of radiologists during batch reading of screening mammograms can be improved when the screening examinations are ordered according to characteristics that may promote visual adaptation.

4.2 Materials and Methods

The need for ethical approval for this retrospective multireader multicase study was waived by the Research Ethics Committee of Radboud university medical center (registration number 2021–13186).

4.2.1 Study Sample

The screening examinations consisted of anonymized two-view (craniocaudal and mediolateral oblique) bilateral mammograms from the Dutch National Breast Cancer Screening Program. The mammograms were acquired between September 2016 and May 2019 using digital mammography systems (Lorad Selenia and Selenia Dimensions; Hologic) in accordance with Dutch guidelines, which involves the use of the automatic exposure control to set the technique for each image acquisition.

Screening examinations were randomly selected to achieve the target area under the receiver operating characteristic (ROC) curve (AUC) as described in the statistical analyses section. Mammograms of women with breast implants and examinations with missing images or ground truth data were excluded. Ground truth data included biopsy-proven cancers and noncancers with at least 2 years of follow-up. Findings of all noncancer examinations were classified as either true negative or false positive, depending on whether the women were recalled. Data to distinguish between normal (healthy breast tissue) and benign (nonmalignant lesions) findings were not available, and information about histopathologic tumor types, tumor sizes, and lesion types was retrieved through linkage with the national cancer registry.

4.2.2 Mammogram Ordering

Three approaches for ordering mammography examinations for readings were tested.

In the first approach, screening examinations were ordered by increasing volumetric breast density (VBD) percentage (VolparaDensity, version 1.5.4.0; Volpara Health Technologies) based on the average VBD of the four mammograms within each examination. Increasing VBD, as opposed to decreasing, was chosen to allow radiologists to start their sessions with low-density mammograms that represent an easy reading task, with a slow transition toward more difficult-to-interpret, higher-density mammograms.

As stated earlier, another more complex image characteristic might drive visual adaptation for radiologists. Therefore, the second approach was based on a deep learning network that was trained with self-supervised learning (SSL). Such a network is trained to summarize the information contained in an image by identifying several features. Although the features are not predefined by human input, the deep learning model was informed that features of different views (i.e., craniocaudal vs mediolateral oblique), from rotated mammograms, and from mammograms of both breasts of the same woman should be similar, while features from mammograms of different women should be different. With this method, mammograms that look similar should also have similar features. For the SSL order, any screening examination could serve as the initial examination. In this study, the examination with the most extreme features identified was selected first, followed by the next examination with features most similar to those of the just-selected examination. In that way, two consecutive screening examinations, each with its four mammograms, have minimal differences between their features. A detailed explanation of this encoding method can be found in Appendix 4.S1.

In the third approach, the order of screening examinations was randomly generated and was used as a reference for current screening practice.

4.2.3 Observer Study

A fully crossed multireader multicase study was performed with three reading sessions, each separated by at least 2 weeks. Each session included one of three different orders of mammograms: VBD, SSL, or random order.

Thirteen screening-certified breast radiologists (including K.M.D., J.K.v.R., A.F.v.R., J.B.H., D.B.N.) from different reading units in the Netherlands participated. The 13 involved breast cancer screening radiologists had 4–32 years of experience with

screening mammography (median, 12 years). The radiologists, on average, spend 1.4 days per week on screening, and the median volume of screening examinations read per year was 11,000 (range, 10,000–22,000). All radiologists performed reading in all three sessions, with the order of the first, second, and third sessions randomized for each radiologist.

The eye movements of the radiologists were recorded with a remote eye tracker (Tobii Pro Spark; Tobii). Radiologists were pretrained with a set of 10–20 screening examinations to gain familiarity with the workstation. Radiologists did not have access to previous screening mammograms or any other personal information about the women and were informed about the study aim but were blinded to the specific orders being tested. Radiologists were able to zoom and switch between the mediolateral oblique and craniocaudal views, but they were not allowed to pan or adjust the window and level settings.

The 150 screening examinations were split into three reading batches of 50 examinations each and were read in the same day. For each screening examination, radiologists provided the following: *(a)* a case-based probability of a malignancy score ranging from 0 to 100 (100 indicating the highest suspicion of malignancy), *(b)* a Breast Imaging Reporting and Data System (BI-RADS), fifth edition, density category ranging from fatty (BI-RADS category a) to dense (BI-RADS category d) [10], and *(c)* a recall decision (yes or no). If radiologists decided to recall, they were instructed to mark the location for which they recalled the woman. Figure 4.S1 shows the reader study interface. Reading times were recorded from the moment the screening examination first appeared until the radiologist moved to the next examination.

For all three ordering conditions, radiologists started reading the first screening examination without prior visual adaptation. To maintain visual adaptation after recalibration or a break, a readaptation movie was presented. This movie showed the five previous screening examinations from the preceding batch that day, each was displayed for 3 seconds in an order specific to the ordering condition of that day. More detailed information about the observer study can be found in Appendix 4.S2.

4.2.4 Eye-Tracker Data

Raw eye-tracker data were classified into eye fixations, which are brief pauses during which the eyes cease movement and focus on specific points in the visual field. The classification of eye fixations included different data processing steps inspired by the Tobii I-VT filter [143,144], as detailed in Appendix 4.S3.

The total number of fixations per screening examination, combining fixations from both the craniocaudal and mediolateral oblique views (e.g., yellow dots in Figure 4.1), was tracked. Additionally, the time spent fixating within a 20-mm radius around the lesion center (defined as the lesion region, red circle in Figure 4.1) was measured. Finally, the time to first fixation (TTFF) in that lesion region was recorded, representing the time interval for a radiologist to notice the lesion from the moment the examination first appeared.

4.2.5 Statistical Analyses

To achieve a study power of 0.8 for the detection of a difference of at least 0.05 in the AUC, a minimum of 124 examinations were required for interpreting by the 13 readers [145]. Therefore, for this study, a total of 150 screening examinations were randomly selected, including 75 with breast cancer and 75 with no cancer.

AUC analysis.—As the primary outcome measure, the AUC for the random-ordered reading was compared with the AUC for the VBD- and SSL-ordered readings. For the multireader multicase analysis of variance, the ROC curves were plotted and their AUCs and 95% CIs were calculated using the radiologists' probability of malignancy scores (ranging from 0 to 100) and iMRMC software package (version 4.0.3; Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration). Additional details about this analysis can be found in Appendix 4.S4. *P* values were based on a *t* test, assuming that the test statistic followed a Student *t* distribution [146].

Specificity and sensitivity.—As secondary outcome measures, reader-averaged specificity and sensitivity for the random-ordered readings were compared with those of the VBD- and SSL-ordered readings using the radiologists' recall decisions. This analysis was performed following the same statistics as described for the AUC analysis.

Reading times.—The median reading times per examination across all 13 radiologists, which are less susceptible to being influenced by extreme values, were compared for the random-ordered readings against the VBD- and SSL-ordered readings. For this comparison, the Wilcoxon signed-rank test was used, accounting for the dependency of the paired examinations. Additional analyses on reading time were reported separately for Volpara Density Grades (categories a, b, c, and d), which correlate with visual BI-RADS, fifth edition, density categories [10].

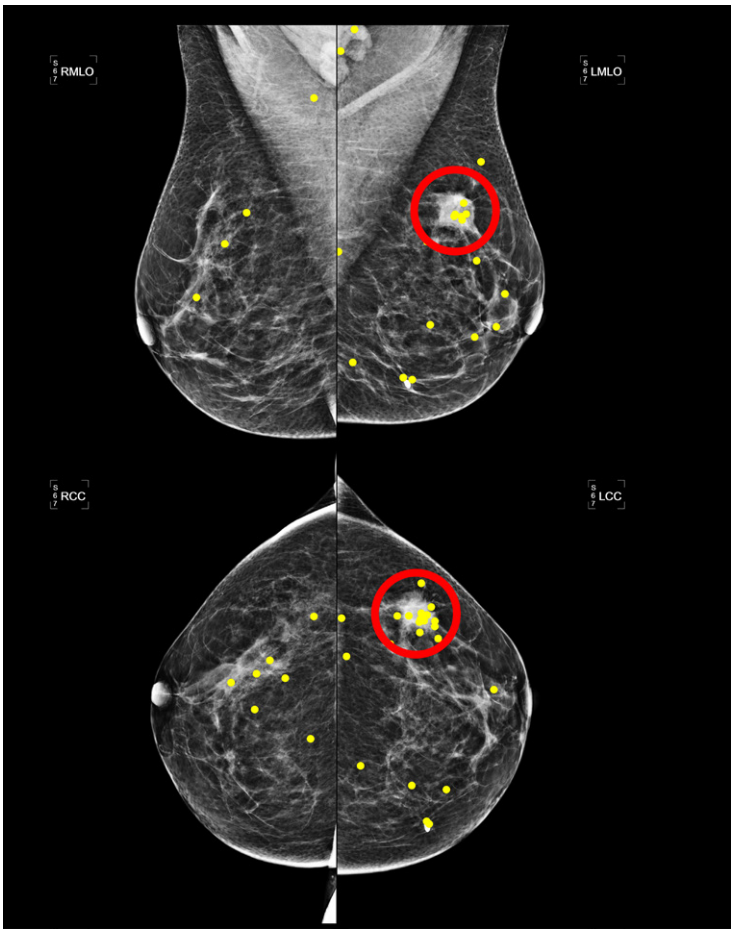


Figure 4.1: Mammograms in a 54-year-old woman with an invasive ductal carcinoma in the lateral upper quadrant of the left breast. The yellow dots represent eye fixations, and the red circle represents the lesion region. LCC = left craniocaudal, LMLO = left mediolateral oblique, RCC = right craniocaudal, RMLO = right mediolateral oblique.

Fixations.—Eye-tracking data were also compared using the Wilcoxon signed-rank test, again taking paired screening examinations into account. For the examination-based comparisons, median values were calculated for the number of fixations, fixation duration on malignant lesion regions, and TTFF at those regions, using data from all 13 radiologists for each examination. When the eye tracker did not register fixations in the lesion region for a specific reader in a specific examination, TTFF outcomes for that reader and examination were excluded for all three orders to ensure that the median TTFF calculation for each reading order was based on data from the same readers.

Additional analyses on reading time and fixation outcomes were reported separately for cancer and noncancer examinations. These included examinations with unanimous radiologist agreement on the correct recall decision for all three orders (consensus) and those with disagreement (nonconsensus). The consensus examinations were assumed to be relatively easy and/or obvious. The nonconsensus examinations were assumed to be more difficult.

Density categories.—The Wilcoxon signed-rank test was used to compare the perceived BI-RADS density categories (categories, a, b, c, and d) [10] of the radiologists across the three different reading orders. Each radiologist-examination pair was analyzed by comparing the density categories of the different reading orders (150 examinations × 13 radiologists = 1950 radiologist-examination pairs per reading order).

Statistical significance for each endpoint was defined as a $P < .025$, which included the Bonferroni correction for the comparison of the random order against the VBD and SSL orders.

4.3 Results

4.3.1 Study Sample Characteristics

The study sample comprised screening examinations of 150 women (median age, 55 years [IQR, 50–63]) (Table 4.1). The SSL order resulted in more occurrences of consecutive cancer and noncancer examinations than did the random and VBD orders (random: 2.3 ± 1.6, VBD: 2.1 ± 1.5, SSL: 6.0 ± 6.3) (Figure 4.S2).

Table 4.1: Characteristics of the selected study group.

Characteristic	Value
No. of women	150
Median age (years)*	55 (49–72, 50–63)
Median breast volume (cm^3)†	785.8 (511.4–1168.9)
Median compressed breast thickness (mm)†	58.5 (48.3–67.9)
Median percentage VBD†	7.6 (5.0–11.8)
Screening outcome‡	
True positive	64 (43)
False negative	11 (7)
True negative	69 (47)
False positive	4 (3)

Table 4.1: Continued

Characteristic	Value
Outcome	
Not malignant[§]	75 (50)
Malignant	75 (50)
Malignant lesion type[‖,#]	
Mass	43 (56)
Calcifications	22 (29)
Mass + calcifications	6 (8)
Architectural distortion	1 (1)
Architectural distortion + calcifications	2 (3)
Asymmetric density	3 (4)
Malignant histopathologic type[‖,**]	
Invasive NST	48 (62)
DCIS	17 (22)
ILC	7 (9)
Mixed NST/ILC	5 (6)
Other invasive	1 (1)
Median size of malignant lesions (mm)[†,††]	16.0 (10.3–21.0)

Note. – Unless otherwise indicated, data in parentheses are percentages. DCIS = ductal carcinoma in situ, ILC = invasive lobular carcinoma, NST = no special type, VBD = volumetric breast density from VolparaDensity, version 1.5.4.0; Volpara Health Technologies. [*] = The first set of numbers in parentheses is the age range, and the second set is the IQR. [†] = Data in parentheses are IQRs. [‡] = Screening outcome not available for two examinations. [§] = Data to distinguish between normal and benign findings were not available. [‖] = Some screening examinations included multiple lesions, which resulted in more lesions than examinations. [#] = Lesion type was not available for two malignant lesions. [**] = The type of histopathologic finding was not available for one malignant lesion. [††] = Tumor diameter was not available for 21 malignant lesions, but the reason for this is unknown.

4.3.2 Performance Analysis

Compared with reading mammography examinations in random order, radiologists improved their screening performance when reading examinations by increasing VBD, with the reader-averaged AUC increasing from 0.92 (95% CI: 0.89, 0.95) to 0.93 (95% CI: 0.91, 0.96) (P = .009) (Figure 4.2). This increased AUC was consistent for 12 of the 13 radiologists (Table 4.S1). There was no evidence of a difference in the reader-averaged AUC between the random- and SSL-ordered readings (both 0.92 [95% CI: 0.89, 0.95]; P = .70). An ROC curve plot with 95% CIs is presented in Figure 4.S3.

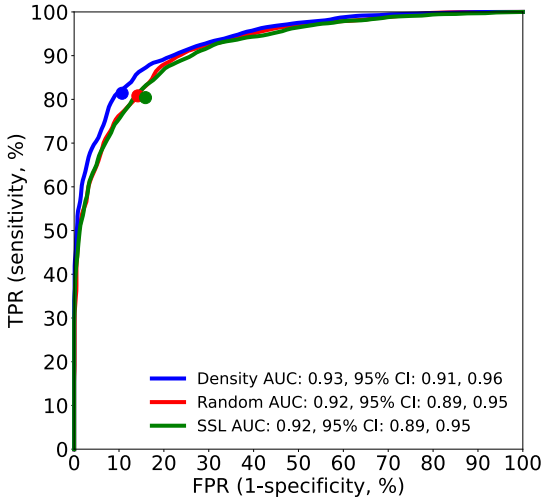


Figure 4.2: Reader-averaged receiver operating characteristic (ROC) curves for the radiologists reading screening mammography examinations that were ordered by increasing volumetric breast density (blue), randomly (red), and by self-supervised learning (SSL) (green). The dots indicate radiologists' operating points based on their recall decisions. AUC = area under the ROC curve, FPR = false-positive rate, TPR = true-positive rate.

4.3.3 Specificity and Sensitivity

On average, with the Bonferroni correction, there was no evidence of a difference in reader-averaged specificity when comparing the random order with the increasing VBD and SSL order (random vs VBD: 837 of 975 [86%] vs 871 of 975 [89%], $P = .04$; random vs SSL: 837 of 975 [86%] vs 820 of 975 [84%], $P = .37$) (Figure 4.2, Table 4.S2). Also, for radiologist-averaged sensitivities there was no evidence of a difference between the random-ordered readings and the VBD- and SSL-ordered readings (random vs VBD: 788 of 975 vs 794 of 975 [both 81%], $P = .78$; random vs SSL: 788 of 975 [81%] vs 784 of 975 [80%], $P = .79$). This finding was consistent across histopathologic tumor types, tumor sizes, and lesion types (Table 4.S3).

4.3.4 Reading Times

The median reading time per screening examination was shorter when radiologists read examinations from low to high VBD than when they read examinations in random order (24.3 seconds vs 27.9 seconds, $P < .001$) (Figure 4.3). This effect was especially pronounced in Volpara Density Grades b, c, and d (Figure 4.S4). Only for the obvious cancer examinations, where all radiologists agreed to recall the woman in all three reading orders, there was no evidence of a difference between the reading times of the VBD- and random-ordered readings ($P = .16$) (Figure 4.S5).

The median reading times were longer for the SSL-ordered readings than for the random-ordered readings (28.4 seconds vs 27.9 seconds, $P = .02$). Individual reading times are shown in Table 4.S4.

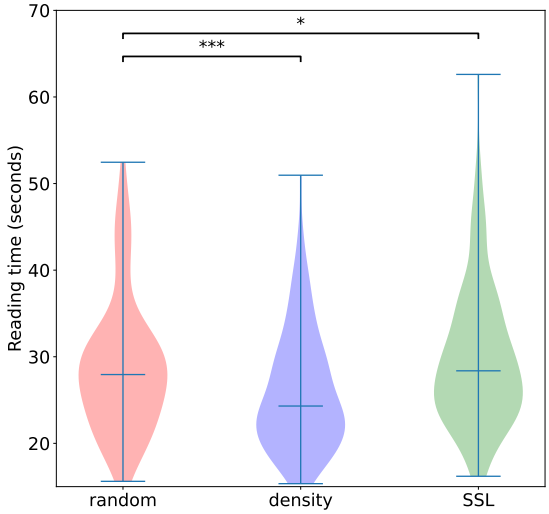


Figure 4.3: Violin plot shows median reading times of the radiologists reading screening mammography examinations that were ordered randomly (red), by increasing volumetric breast density (blue), and by self-supervised learning (SSL) (green). The random reading order was compared against the density and SSL order, and the asterisks indicate statistical significance: *** $P < .0005$, * $P < .025$.

4.3.5 Fixations

The median number of fixations and the time fixated on the malignant lesion regions were lower for the VBD-ordered reading than for the random-ordered reading (random vs VBD: 52 fixations vs 47 fixations, $P < .001$; 4.6 seconds vs 3.7 seconds in the lesion region, $P < .001$) (Figure 4.4, Table 4.S4). There was no evidence of a difference between the random- and SSL-ordered readings for median number of fixations and fixation time within the lesion region (random vs SSL: 52 fixations vs 54 fixations, $P = .05$; both 4.6 seconds in the lesion region, $P > .99$). Figure 4.S6 and 4.S7 contain information about fixation outcomes categorized into consensus and nonconsensus examinations.

TTFF at the lesion region was faster for the true-positive examinations with consensus among all radiologists than for the nonconsensus examinations (Figure 4.S8). However, there was no evidence of a difference in median TTFF for the reading orders (random vs VBD: 1.1 seconds vs 1.0 second, $P = .20$; random vs SSL: 1.1 seconds vs 1.0 second, $P = .34$) (Figure 4.4, Table 4.S4).

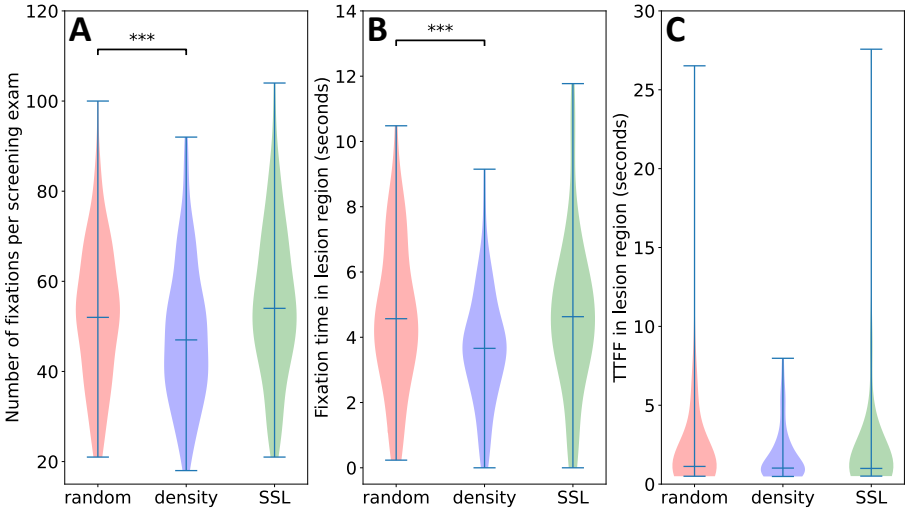


Figure 4.4: Violin plots show fixation outcomes for the radiologists reading screening mammography examinations that were ordered randomly (red), by increasing volumetric breast density (blue), and by self-supervised learning (SSL) (green). (A) The total number of fixations for the screening examinations (including fixations in both craniocaudal and mediolateral oblique views). (B) Fixation time in the 20-mm lesion region for the 75 screening examinations with malignant lesions. (C) Time to first fixation (TTFF) on the 20-mm radius lesion region in the 75 screening examinations with malignant lesions. The random reading order was compared against the density and SSL order, and the asterisks indicate statistical significance: *** $P < .0005$.

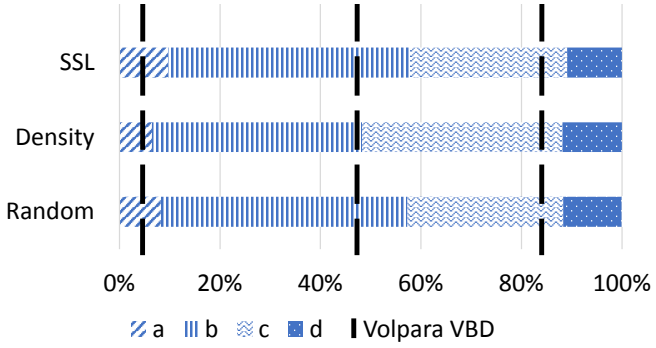


Figure 4.5: Radiologists' perceived Breast Imaging Reporting and Data System categories a–d. Outcomes from all radiologists and examinations were aggregated. The vertical dashed lines represent Volpara Density Grade percentage volumetric breast density (VBD) thresholds, which correlate with visual Breast Imaging Reporting and Data System, fifth edition, density categories (a < 3.5%, 3.5% ≤ b < 7.5%, 7.5% ≤ c < 15.5%, d ≥ 15.5%). SSL = self-supervised learning.

4.3.6 Density Category Assessments

Figure 4.5 shows how adaptation affected the mammographic density perceived by radiologists. The perceived density categories were different for the VBD- and random-ordered readings ($P < .001$) (Figure 4.S9). Within the VBD order, radiologists classified more examinations as BI-RADS c and d and fewer as BI-RADS a and b. Differences in perceived density categories were also observed between the random- and SSL-ordered readings ($P = .02$), with both increasing and decreasing radiologist-perceived density.

4.4 Discussion

The results of our study show that reading screening mammography examinations ordered from low to high volumetric breast density (VBD), compared with reading in a random order, consistently improved radiologists' screening performance, as measured with the area under the receiver operating characteristic curve (0.93 vs 0.92, $P = .009$), while reducing reading time (24.3 seconds vs 27.9 seconds, $P < .001$) and lesion-fixation time (3.7 seconds vs 4.6 seconds, $P < .001$). By integrating quantitative density metrics into the screening workflow and implementing automated sorting algorithms in the screening worklist, organizing batch readings of screening examinations from low to high VBD would be feasible to implement and may serve as a valuable strategy to optimize breast cancer screening.

The improved screening performance and reduced reading time (especially in denser mammograms) when batch reading of mammograms by increasing VBD aligns with the adaptation hypothesis [67,68,147]. This hypothesis suggests that radiologists gradually adjust to the characteristic structure of mammograms, focusing on more unexpected properties and thereby improving performance and reducing reading time. Consistent with the reduced reading time, the number of fixations per screening examination and the fixation duration in the lesion region were also lower. This finding may indicate a more confident decision of the radiologist, causing them to have fewer fixations in the density-ordered reading. Alternatively, adaptation could indirectly affect fixations by increasing the effective salience of the targets. McDermott et al. [141] also found changes in the number and duration of fixations following adaptation to different color distributions in a color search task, where observers located a color singleton on a dense background of distractors, but argued that this reflected changes in the relative stimulus contrast rather than in the way the images were sampled.

There was no evidence of a difference in diagnostic performance or fixation outcomes between the SSL-ordered reading and the random-ordered reading. However, the reading time was longer for SSL-ordered readings than for random-ordered readings. Interestingly, SSL-ordered reading resulted in numerous consecutive cancers and noncancers, for example, starting with 22 cancer examinations. This ordering may have influenced the radiologists by increasing their uncertainty and potentially affecting their decision making, which resulted in longer reading times. Despite this potential psychological effect, radiologists demonstrated comparable performance with SSL-ordered and random-ordered readings. The clustering effect of cancers will be less pronounced in a dataset with a cancer prevalence typical of screening. Hence, it is possible that SSL ordering will result in better outcomes in a screening setting where this clustering effect is less prevalent.

There was no evidence of differences in TTFF in the lesion region across different reading orders, indicating that radiologists did not detect malignant lesions faster in any particular order. The enhanced performance with specific ordering may relate more to improved lesion characterization rather than faster detection. Visual adaptation may play a role in differentiating between normal versus unexpected characteristics in images [137–139]. However, lesion characterization involves more complex cognitive processes, such as pattern recognition, knowledge of lesion characteristics, and clinical experience, as well as metacognition or confidence in judgments, which may be influenced by adaptation [148,149]. Further research is needed to fully understand the influence of visual adaptation on radiologic assessments.

Radiologists batch reading the screening examinations by increasing VBD perceived the mammograms to be denser than when reading the mammograms in the SSL or random order. A previous study by Kompaniez et al. [66] revealed that prior exposure to fatty mammogram patches caused an intermediate texture to appear denser and vice versa. This finding may explain why gradually increasing VBD caused the mammograms to look denser in the VBD-ordered reading.

It would be interesting to also investigate what happens if the density order is reversed—ordering by decreasing VBD, and thus starting with the highest-density mammograms, which are likely more difficult to interpret. Due to limitations in the number of ordering conditions that the readers could be asked to perform, we were not able to investigate this in our study. Radiologist preferences regarding the reading order and considerations about fatigue may ultimately also influence the most effective reading order.

Our study had limitations. First, the study was performed in an experimental setting with a cancer-enriched case set and no prior screening mammograms were available. High cancer prevalence can lead to higher sensitivity and lower specificity among radiologists, as found in prior research [35,92]. In addition, radiologists were aware of the enriched nature of the case set, which may have introduced a "laboratory effect" [35,150]. However, any such effect would have similarly affected all three reading conditions, thereby minimizing its impact on the ordering comparison. Future research should investigate ordering strategies in a real screening population, preferably through a prospective study. Second, our study was performed with screening radiologists from the Netherlands; however, screening practices around the world vary. Therefore, our results may not be generalizable to all other screening programs. Third, gaze point measurements are subject to error, potentially leading to variability in the accuracy of fixation outcomes. We tried to minimize those effects by applying different processing steps based on the principles of the Tobii I-VT filter [143,144]. Finally, for some examinations, we relied on visual inspection by screening radiologists for the lesion locations, as biopsy-proven locations were unavailable. The radiologists who provided these locations were not involved as readers in our study. While the absence of biopsy-proven locations may have affected lesion region fixation analyses, most reader localizations in this study aligned with our ground truth, minimizing the potential impact.

In conclusion, the findings of our study indicate that ordering mammography examinations from low to high volumetric breast density (VBD) improved screening performance while reducing reading and fixation times. Although larger studies are needed to validate these findings, our results suggest that ordering screening mammograms for reading by increasing VBD may be feasible. The integration of automated density algorithms could facilitate the organization of the screening program worklist. Assessing this ordered list may improve radiologists' ability to detect and characterize suspicious findings, potentially leading to better outcomes for screened women and increased efficiency. However, future research should assess what other order criteria may further optimize radiologists' screening performance (e.g., reading by decreasing VBD or ordering by increasing probability of malignancy scores via artificial intelligence). If proven successful, future studies should also assess the effectiveness of these ordering strategies prospectively in actual screening settings.

Acknowledgments

The authors thank the additional breast screening radiologists who participated in the study: H.G. de Bruin, MD, W. Dinkelaar, MD, A.M.B. Fauquenot-Nollen, MD, M.P.M.

Gielens, MD, M.W. Imhof-Tas, F.H. Jansen, MD, P.E.J.M. Sallevelt, MD, PhD, M.M. Snoeren, MD. The authors would also like to thank M.P. Eckstein, PhD, and D.S. Klein, PhD, for providing valuable insights into eye tracking during a visit to their Vision and Image Understanding Lab at the University of California, Santa Barbara.

Supplementary Material

Appendix 4.S1– Self-supervised Learning Encoding

The self-supervised learning order (SSL) aims to group similar looking mammography examinations to minimize the change from one examination to the next. The definition of "looking similar" is not based on hand-crafted features, but features learned by a network that was trained with an SSL framework. For this, a Resnet-50 was trained with a SimCLR framework to create a 2048-feature encoding of a mammogram cropping [151]. SimCLR is a contrastive SSL framework for visual representations. In contrastive learning, the distance between encodings of positive pairs of images is minimized while the distance between encodings of negative pairs is maximized. Positive pairs are usually defined by different transforms of the same image, for example a rotation or crop. Transforms are chosen in such a way that an image after that transform should still receive the same encoding. For mammogram ordering, in this study the following transforms were selected: selection of a random crop of 35 x 35 mm out of any of the views, random flip and/or rotation, random color jitter, and random gaussian blur. Negative pairs were chosen as mammograms of different women. Using these definitions for positive and negative pairs will ensure that encodings of different views in the same examination will receive a similar encoding, while views from examinations of different women will receive a different encoding. The Resnet-50 was trained with a dataset of 43,443 Dutch screening examinations (both normal and cancer cases) of unique women consisting of four views each.

The SSL order of the 150 screening examinations in this study was defined in several steps. First, each mammographic view was segmented with a U-Net [152] to remove the background and pectoral muscle and the largest possible rectangle was fitted inside the breast. The resulting rectangles were used as input to the model, to avoid any influence from the shape and background. The output feature vectors were averaged over each rectangle of each of the four views to result in one feature vector for each screening examination. The examination with the maximum feature norm was selected as the starting examination. The next examination was the examination that minimized the Euclidean distance of the encodings with the previous just-selected one. This process was repeated until all examinations were selected in order.

Appendix 4.S2– Observer Study Details

The reader study was performed in a controlled setting with ambient room lighting of approximately 45 lux. A twelve-megapixel monitor certified for breast imaging (Coronis Uniti, MDMC12133 Barco) and calibrated to the DICOM grayscale standard display function was used.

The Tobii eye tracker, which was used during the study, was placed below the viewing monitor, and the radiologists' eye movements were recorded with a sampling frequency of 60 Hz without requiring the radiologists to wear any equipment or restricting their head movement. Before each session, a 9-point calibration process was performed to ensure accurate tracking of eye movements. A fixation cross was used before each examination to ensure eye tracker calibration. If needed, radiologists repeated the 9-point calibration step before proceeding to the next examination.

Reading times were also recorded. Any time spent recalibrating, taking breaks, or watching the readaptation movie was excluded from the recorded reading times. To prevent outliers in reading time due to interruptions, precautions were taken. Prior to the start of the study, the readers were instructed to use the restroom if necessary and to stow away their phones to minimize interruptions. Additionally, one of the authors remained present in the same room as the readers to ensure an uninterrupted study. Consequently, extreme reading times were not excluded, as it was believed that these deviations were justifiable and not attributable to interruptions.

Appendix 4.S3– Eye-tracking data

Raw eye-tracking data were classified into eye fixations using different data processing steps and default values inspired by the Tobii I-VT Filter [143,144]. First, missing eye-tracking data shorter than 75 msec, the default value chosen by Tobii, was linearly interpolated. Second, the average position of where the left and the right eye were looking, was calculated. If only one eye was detected by the eye tracker, the data from that eye was used. Next, noise was reduced by taking the moving median of three consecutive data points, the default value used by Tobii [144]. Lastly, fixations were defined as consecutive gaze points with a velocity below 359 mm per second on the screen, considering the Tobii default velocity threshold of 30°/second [144,153] and assuming a distance of 67 cm between the reader and the screen. To prevent long fixations being split up by measurement errors, neighboring fixations that were within 75 msec and 6 mm of each other [143,144,154] and fell on the same image were merged. Fixations shorter than 60 msec were discarded as very short fixations are not meaningful when studying radiologist behavior where the eye and brain require some time to register and process visual information [143,144].

Appendix 4.S4– Statistical Analysis

The area under the receiver operating characteristic (ROC) curve (AUC) in the random order was compared with those in the density and SSL orders. ROC curves were plotted, and their AUC values and 95% confidence intervals were calculated using the radiologists' probability of malignancy scores (ranging from 0 to 100) in a multireader multicase analysis of variance, using the iMRMC software package (version 4.0.3, Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration). In the analysis of variance model, three factors were included: modality (reading order), reader, and case (screening examination); these three factors are considered standard effects in multireader multicase studies [155–157]. Modality was a fixed effect, as the three reading orders selected were included in this study. Readers and cases were treated as cross-correlated random effects because of the interest in generalizing to the population of readers and cases for the specific comparisons that were indicated [157–159]. The software estimated all components of variance (variances and covariances) using a "one-shot" nonparametric estimate based on the properties of U statistics, providing unbiased estimates of variance components, accounting for variability and correlations between readers and cases.

Additional Tables

Table 4.S1: Area under the curve of the receiver operating characteristic curves for each of the radiologists reading screening mammography examinations that were ordered randomly, by increasing volumetric breast density, or by SSL.

Reader	Random	Density	SSL
R1	0.94	0.92	0.95
R2	0.92	0.93	0.89
R3	0.87	0.92	0.91
R4	0.92	0.94	0.91
R5	0.94	0.95	0.95
R6	0.92	0.93	0.92
R7	0.93	0.95	0.85
R8	0.91	0.92	0.93
R9	0.92	0.95	0.93
R10	0.94	0.96	0.91
R11	0.92	0.94	0.89
R12	0.89	0.90	0.95
R13	0.94	0.95	0.94
Average	0.92 (0.89, 0.95)	0.93 (0.91, 0.96)	0.92 (0.89, 0.95)
P value		.009	.70

Note.—95% CIs are shown within parentheses. SSL = self-supervised learning.

Table 4.S2: Specificity & sensitivity for each of the radiologists reading screening mammography examinations that were ordered randomly, by increasing volumetric breast density, or by SSL.

Reader	Random	Density	SSL
Specificity			
R1	87 (65/75)	81 (61/75)	77 (58/75)
R2	93 (70/75)	91 (68/75)	88 (66/75)
R3	84 (63/75)	84 (63/75)	88 (66/75)
R4	85 (64/75)	91 (68/75)	88 (66/75)
R5	83 (62/75)	99 (74/75)	91 (68/75)
R6	87 (65/75)	88 (66/75)	81 (61/75)
R7	93 (70/75)	93 (70/75)	79 (59/75)
R8	72 (54/75)	76 (57/75)	71 (53/75)
R9	87 (65/75)	88 (66/75)	85 (64/75)
R10	81 (61/75)	91 (68/75)	87 (65/75)
R11	87 (65/75)	92 (69/75)	81 (61/75)
R12	91 (68/75)	99 (74/75)	97 (73/75)
R13	87 (65/75)	89 (67/75)	80 (60/75)
Average	86 (81, 91)	89 (84, 94)	84 (78, 90)
P value		.04	.37
Sensitivity			
R1	87 (65/75)	91 (68/75)	89 (67/75)
R2	65 (49/75)	83 (62/75)	76 (57/75)
R3	76 (57/75)	81 (61/75)	73 (55/75)
R4	80 (60/75)	76 (57/75)	81 (61/75)
R5	87 (65/75)	76 (57/75)	84 (63/75)
R6	85 (64/75)	84 (63/75)	83 (62/75)
R7	79 (59/75)	79 (59/75)	80 (60/75)
R8	89 (67/75)	88 (66/75)	88 (66/75)
R9	79 (59/75)	89 (67/75)	84 (63/75)
R10	85 (64/75)	85 (64/75)	76 (57/75)
R11	83 (62/75)	76 (57/75)	81 (61/75)
R12	69 (52/75)	65 (49/75)	63 (47/75)
R13	87 (65/75)	85 (64/75)	87 (65/75)
Average	81 (74, 88)	81 (74, 88)	80 (74, 87)
P value		.78	.79

Note.—Specificity & sensitivity are shown in percentages with the numerator and denominator shown in parentheses. For average values, 95% CIs are shown in parentheses. SSL = self-supervised learning.

Table 4.S3: Radiologist-averaged sensitivities across various histopathologic tumor types, tumor sizes, and lesion types when reading screening mammography examinations that were ordered randomly, by increasing volumetric breast density, or by SSL.

Subanalyses	Random	Density	SSL
Histopathologic tumor type			
Invasive NST	87 (521/598)	86 (515/598)	87 (523/598)
DCIS	68 (141/208)	71 (147/208)	65 (135/208)
ILC	86 (78/91)	87 (79/91)	79 (72/91)
Mixed NST/ILC	74 (48/65)	82 (53/65)	82 (53/65)
Other invasive	0 (0/13)	0 (0/13)	8 (1/13)
Tumor size			
0–10 mm	88 (161/182)	87 (159/182)	87 (158/182)
10–20 mm	88 (276/312)	89 (276/312)	89 (277/312)
20–50 mm	75 (166/221)	76 (169/221)	76 (169/221)
≥ 50 mm	73 (19/26)	85 (22/26)	69 (18/26)
Missing	71 (166/234)	72 (168/234)	69 (162/234)
Lesion type			
Mass	80 (438/546)	80 (439/546)	81 (444/546)
Calcifications	79 (217/273)	83 (227/273)	78 (212/273)
Mass + calcifications	95 (74/78)	91 (71/78)	95 (74/78)
Architectural distortion	100 (13/13)	85 (11/13)	85 (11/13)
Architectural distortion + calcifications	81 (21/26)	69 (18/26)	69 (18/26)
Asymmetric density	64 (25/39)	72 (28/39)	64 (25/39)

Note.—Data are radiologist-averaged sensitivity percentages with the numerator and denominator shown in parentheses. DCIS = ductal carcinoma in situ, ILC = invasive lobular carcinoma, NST = no special type, SSL = self-supervised learning.

Table 4.S4: Median outcomes per examination for each of the radiologists reading screening mammography examinations that were ordered randomly, by increasing volumetric breast density, or by SSL.

Reader	Random	Density	SSL
Reading time			
R1	27.7	18.4	19.3
R2	36.1	29.7	34.9
R3	18.4	16.0	14.8
R4	34.8	35.4	35.8
R5	36.2	41.9	35.5
R6	15.9	15.2	17.5
R7	22.3	21.9	27.4
R8	24.7	25.5	24.6
R9	28.6	34.8	39.8
R10	26.0	16.6	23.9
R11	26.1	28.1	31.9
R12	36.6	24.7	28.3
R13	27.5	26.0	33.3
Average	27.8	25.7	28.2
Number of fixations			
R1	54	33	35
R2	82	75	74
R3	30	23	20
R4	47	72	57
R5	73	75	65
R6	27	25	28
R7	35	34	47
R8	50	46	47
R9	62	75	86
R10	47	25	39
R11	46	56	74
R12	87	55	63
R13	54	46	68
Average	53	49	54
Fixation time in lesion region			
R1	5.0	2.9	3.1
R2	6.7	5.9	6.9
R3	3.6	3.0	3.0

Table 4.S4: Continued

Reader	Random	Density	SSL
R4	3.4	5.3	5.7
R5	7.2	6.2	8.2
R6	2.9	2.3	2.4
R7	4.7	4.0	4.7
R8	3.3	3.3	4.1
R9	5.8	5.8	7.1
R10	3.4	2.4	3.0
R11	3.9	3.7	4.9
R12	4.9	3.0	3.6
R13	4.7	4.7	6.1
Average	4.6	4.0	4.8
TTFF in lesion region			
R1	0.8	0.8	0.7
R2	1.0	0.8	0.8
R3	0.8	0.9	0.9
R4	1.5	1.1	1.0
R5	1.3	1.1	1.3
R6	0.7	0.9	1.1
R7	0.7	0.8	0.7
R8	1.3	1.1	0.9
R9	2.8	2.7	2.5
R10	1.1	1.0	1.0
R11	1.7	1.6	1.0
R12	4.8	5.8	6.4
R13	0.9	1.0	0.8
Average	1.5	1.5	1.5

Note.—Reading times, time in the lesion region, and time to first fixation (TTFF) in the lesion region are shown in seconds. SSL = self-supervised learning.

Additional Figures

Figure 4.S1: Illustration of the reader study interface. The radiologists were asked three questions: "Give your suspicion of malignancy" (ranging from normal to unsure to radiologically malignant), "What density do these images have?," and "Would you recall this woman?" (with possible answers: no and yes). The questions were originally presented in Dutch.

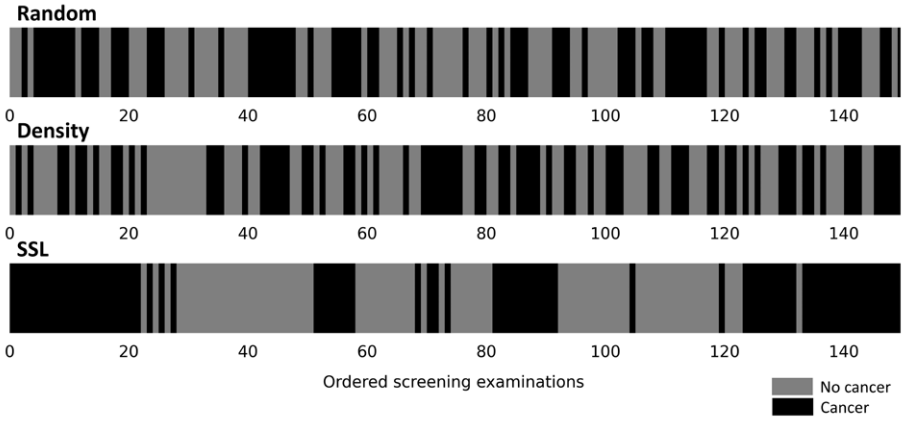


Figure 4.S2: Plots indicating the frequency of consecutive cancer and noncancer examinations for the three different reading orders. SSL = self-supervised learning.

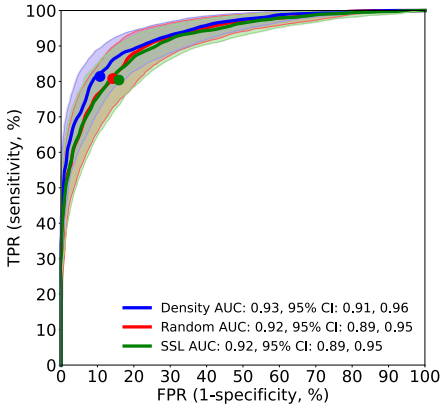


Figure 4.S3: Reader-averaged receiver operating characteristic (ROC) curves for the radiologists reading screening mammography examinations that were ordered by increasing volumetric breast density (blue), randomly (red), and by self-supervised learning (SSL) (green). The dots indicate radiologists' operating points based on their recall decisions. The area under the ROC curves are shown together with 95% CIs. It should be noted, however, that the CIs are considerably wider than the differences between the ROC curves. The errors characterized by the bands are representative of each curve individually, but they are not representative of the differences between the curves. The differences between the ROC curves are smaller because of correlations due to common readers and cases (in the same way that a paired comparison may be able to detect a small difference between two variables with large variances). Our hypotheses are not focused on the individual ROC curves (or the absolute value of the area under the ROC), but rather on the differences between them. And thus, these CIs are not directly relevant to our hypotheses. AUC = area under the ROC curve, FPR = false-positive rate, TPR = true-positive rate.

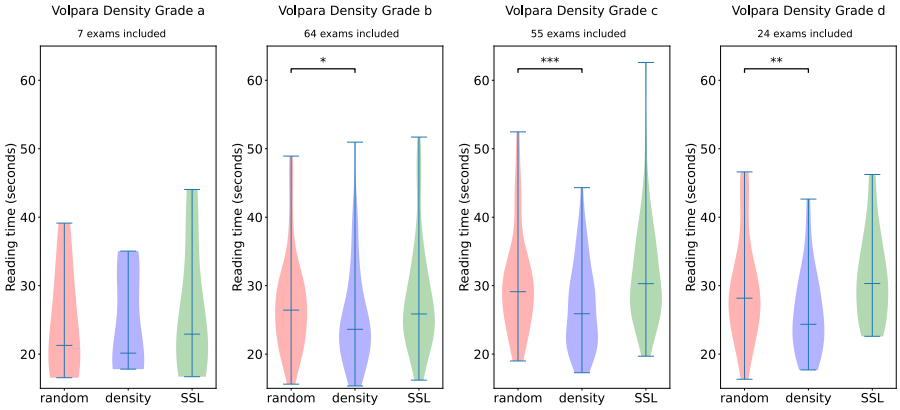


Figure 4.S4: Violin plots show median reading times of the radiologists reading screening mammography examinations that were ordered randomly (red), by increasing volumetric breast density (blue), and by self-supervised learning (SSL) (green), categorized by Volpara Density Grades (categories a, b, c, and d). The Volpara Density Grades correlate with visual BI-RADS 5th edition density categories (a < 3.5%, 3.5% ≤ b < 7.5%, 7.5% ≤ c < 15.5%, d ≥ 15.5%). The random reading order was compared against the density and SSL order, and the asterisks indicate statistical significance: *** $P < .000125$, ** $P < .00125$, * $P < .00625$.

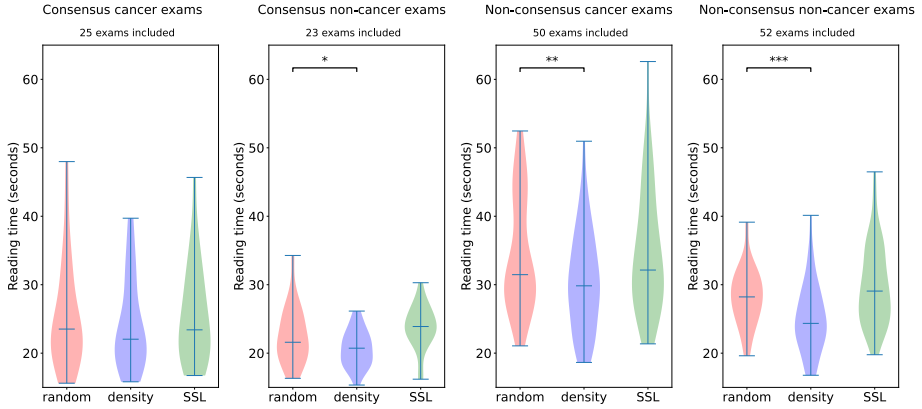


Figure 4.S5: Violin plots show median reading times of the radiologists reading screening mammography examinations that were ordered randomly (red), by increasing volumetric breast density (blue), and by self-supervised learning (SSL) (green), categorized by consensus and nonconsensus cancer and noncancers examinations. For the consensus examinations all radiologists agreed on the recall decision for all three orders. For the nonconsensus examinations some radiologists had different recall decisions in some orders. The random reading order was compared against the density and SSL order, and the asterisks indicate statistical significance: *** $P < .000125$, ** $P < .00125$, *$P < .00625$.

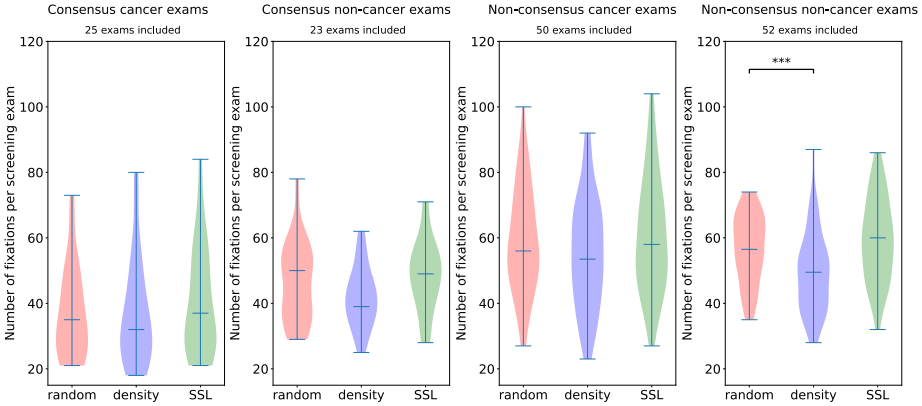


Figure 4.S6: Violin plots show total number of fixations (including fixations in both craniocaudal and mediolateral oblique views) for the radiologists reading screening mammography examinations that were ordered randomly (red), by increasing volumetric breast density (blue), and by self-supervised learning (SSL) (green), categorized by consensus and nonconsensus cancer and noncancers examinations. For the consensus examinations all radiologists agreed on the recall decision for all three orders. For the nonconsensus examinations some radiologists had different recall decisions in some orders. The random reading order was compared against the density and SSL order, and the asterisks indicate statistical significance: *** $P < .000125$.

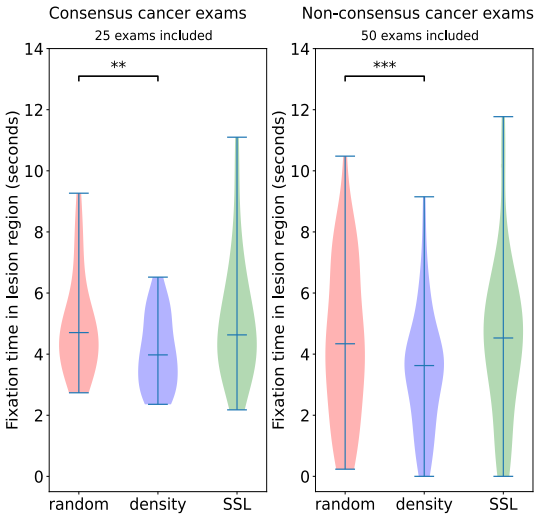


Figure 4.S7: Violin plots show fixation time in the 20 mm radius malignant lesion regions for the radiologists reading screening mammography examinations that were ordered randomly (red), by increasing volumetric breast density (blue), and by self-supervised learning (SSL) (green), categorized by consensus and nonconsensus cancer and noncancers examinations. For the consensus examinations all radiologists agreed on the recall decision for all three orders. For the nonconsensus examinations some radiologists had different recall decisions in some orders. The random reading order was compared against the density and SSL order, and the asterisks indicate statistical significance: *** $P < .00025$, ** $P < .0025$.

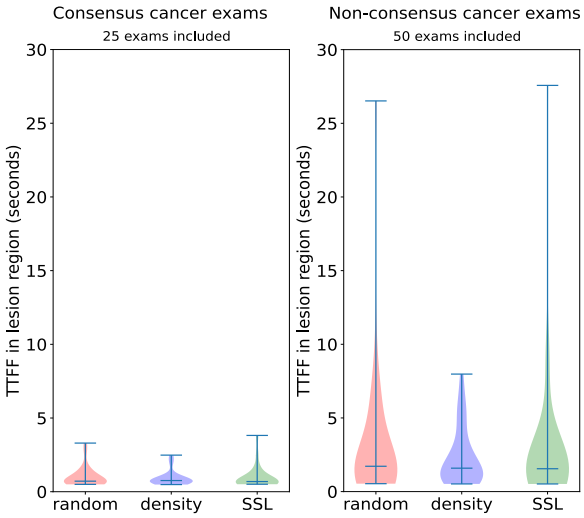


Figure 4.S8: Violin plots show time to first fixation (TTFF) at the 20 mm radius malignant lesion region for the radiologists reading screening mammography examinations that were ordered randomly (red), by increasing volumetric breast density (blue), and by self-supervised learning (SSL) (green), categorized by consensus and nonconsensus cancer and noncancers examinations. For the consensus examinations all radiologists agreed on the recall decision for all three orders. For the nonconsensus examinations some radiologists had different recall decisions in some orders. The random reading order was compared against the density and SSL order.

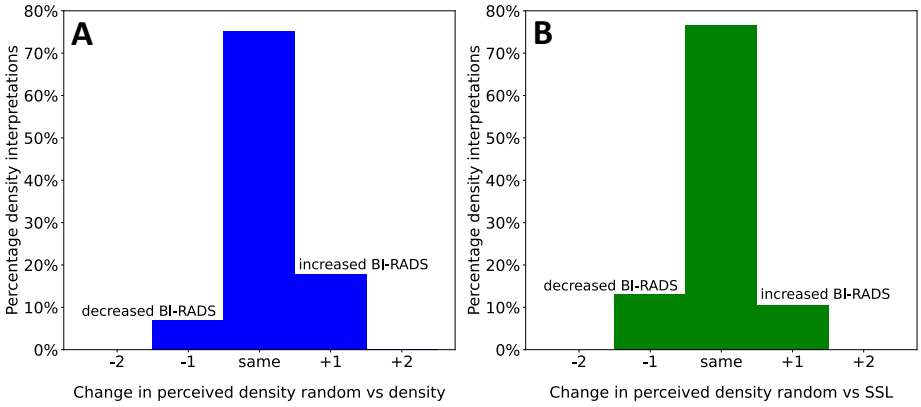


Figure 4.S9: Change in radiologists' perceived Breast Imaging Reporting and Data System density categories a-d. (A) Density- versus random-ordered readings. (B) Self-supervised learning- (SSL) versus random-ordered readings. The figure aggregates the comparisons of all radiologist-examination pairs.

Chapter 5

Influence of AI Decision Support on Radiologists' Performance and Visual Search in Screening Mammography

Jessie J.J. Gommers, Sarah D. Verboom, Katya M. Duvivier, Jan-Kees van Rooden, A. Fleur van Raamt, Janneke B. Houwers, Dick B. Naafs, Lucien E.M. Duijm, Miguel P. Eckstein, Craig K. Abbey, Mireille J.M. Broeders, Ioannis Sechopoulos

Published in Radiology 2025; 316(1):e243688.
DOI: 10.1148/radiol.243688

Abstract

Background

Artificial intelligence (AI) decision support may improve radiologist performance during screening mammography interpretation, but its effect on radiologists' visual search behavior remains unclear.

Purpose

To compare radiologist performance and visual search patterns when reading screening mammograms with and without an AI decision support system.

Materials and Methods

In this retrospective multireader multicase study, 12 breast screening radiologists with 4–32 years of experience (median, 12 years) from 10 institutions evaluated screening mammograms acquired between September 2016 and May 2019. Assessments were conducted unaided and with a Food and Drug Administration–approved, European Commission–marked AI decision support system, which assigns a region suspicion score from 1 to 100, with 100 indicating the highest malignancy likelihood. An eye tracker monitored readers' eye movements. Area under the receiver operating characteristic curve (AUC), sensitivity, and specificity between unaided and AI-assisted reading were compared using multireader multicase analysis software. Reading times, breast fixation coverage (percentage breast covered by fixations within 2.5° visual angle radius), fixation time, and time to first fixation within the lesion region were compared using bootstrap resampling (n = 20,000).

Results

Mammography examinations (75 with breast cancer, 75 without breast cancer) from 150 women (median age, 55 years [IQR, 50–63 years]; age range, 49–72 years) were read. The mean AUC was higher with AI support versus unaided reading (unaided, 0.93 [95% CI: 0.91, 0.96]; AI-supported, 0.97 [95% CI: 0.95, 0.98]; $P < .001$). There was no evidence of a difference in mean sensitivity (81.7% [735 of 900 readings] vs 87.2% [785 of 900]; $P = .06$), specificity (89.0% [801 of 900] vs 91.1% [820 of 900]; $P = .46$), or reading time (29.4 vs 30.8 seconds; $P = .33$). Breast fixation coverage was lower with AI support (11.1% vs 9.5% of breast area; $P = .004$), while fixation time in the lesion region was higher (4.4 vs 5.4 seconds; $P = .006$). There was no evidence of a difference in time to first fixation within the lesion region (3.4 vs 3.8 seconds; $P = .13$).

Conclusion
Radiologists improved their breast cancer detection accuracy when reading mammography with AI support, spending more fixation time on suspicious areas and less on the rest of the breast, indicating a more efficient search.

5.1 Introduction

Interest is growing in deep learning–based artificial intelligence (AI) systems aiming to enhance breast cancer screening performance, address the shortage of radiologists, and improve efficiency [81,160–162]. An initial step toward broader adoption is to provide radiologists with an AI decision support system. Previous studies have demonstrated that AI for decision support improves radiologist performance by increasing sensitivity, maintaining or increasing specificity, and not extending reading time [82–85,88]. However, the impact of AI on radiologists' visual search patterns remains underexplored.

Understanding how AI influences visual search is crucial, since AI recommendations could direct radiologists' eye fixations to areas identified as suspicious. This can potentially make the search more efficient but also raises concerns about automation bias, where radiologists may overly rely on AI suggestions, even if they are incorrect [92–94]. Eye tracking offers the possibility to study these interactions, providing insights into potential changes in search behavior.

Previous eye-tracking studies have examined interactions with earlier versions of computer-aided detection (CAD) tools. For example, Drew and colleagues [90] observed that CAD users had a more restricted search area compared with non–CAD users. Similarly, Klein et al. [91] found that CAD support improved performance, but CAD also led to reduced eye movements and shorter search times among observers. However, these studies primarily focused on older CAD systems or involved nonradiologist observers. Thus, the question remains whether similar effects would be observed in screening radiologists using AI decision support.

This study aims to fill that gap by comparing radiologist performance and visual search patterns when reading screening mammograms with and without an AI decision support system, using screening performance and eye-tracking metrics. By investigating how AI influences radiologists' visual search through eye tracking, the aim of this study was to gain insights that may help optimize AI integration into radiologic workflows and improve diagnostic accuracy.

5.2 Materials and Methods

This retrospective study used the same screening examinations previously used in another study [163]. In that earlier study, 13 radiologists interpreted the same examinations in different orders to investigate the effect of case ordering. In the current study, 12 of those radiologists participated again, allowing for direct comparison between their earlier interpretations and those from the new AI-assisted session in this study. The research ethics committee of Radboud university medical center waived the need for ethical approval (no. 2021–13186).

5.2.1 Study Design and Sample

In this fully crossed multireader multicase study, multiple readers evaluate multiple cases, with the analysis accounting for both reader and case variability. Bilateral two-view (craniocaudal and mediolateral oblique) screening mammograms from the Dutch National Breast Cancer Screening Program were included. Mammograms were acquired in multiple screening regions between September 2016 and May 2019 with use of digital systems (Lorad Selenia and Selenia Dimensions; Hologic). Data from women presenting for screening at the participating sites in this period were eligible for inclusion and randomly selected. Data from women with incomplete information or breast implants were excluded.

5.2.2 AI System

The AI system used for this study was Transpara (version 2.1.0; ScreenPoint Medical), which is Food and Drug Administration approved and European Commission marked. It uses deep convolutional neural networks to detect lesions suspicious for breast cancer on digital mammograms and breast tomosynthesis images. The system has been trained and tested with databases from multiple institutions with varied populations, equipment vendors, and screening programs [80,88,161,164,165]. None of the data used in this study were used to train the AI system.

Transpara identifies soft tissue lesions and calcifications, assigning a region suspicion score ranging from 1 to 100, with 100 indicating the highest likelihood of malignancy. The lesion with the maximum region score determines the overall examination risk category. Examinations were considered to be at low (maximum region score, < 40), intermediate (score, ≥ 40 and < 60), medium-high (score, ≥ 60 and < 80), or very high (score, ≥ 80) risk of having a malignant neoplasm present.

With AI support, radiologists were immediately shown the AI risk category upon opening the case. Additionally, they could click a button to display AI markers for

regions detected as suspicious for calcification or soft tissue lesions, along with the corresponding region suspicion scores.

5.2.3 Reader Sessions

Twelve Dutch screening radiologists (including K.M.D., C.J.v.R., A.F.v.R., J.B.H., and D.B.N.) participated in this study (Appendix 5.S1). Readers' years of experience with screening mammography ranged from 4 to 32 years (median, 12 years), with a median annual reading volume of 11,000 examinations (range, 10,000–22,000 examinations). The study was conducted at a single site, although the readers were from 10 different institutions. Each reader first performed the unaided reading session and then the one with AI support. Randomizing the session order across readers was not possible because the unaided sessions were conducted as part of a previous study [163]. There was a minimum period of 52 days between sessions for the same radiologist. In that previous study, it was found that ordering examinations by increasing volumetric breast density improved performance [163]. Consequently, the AI-supported session in this study followed the same approach, with examinations ordered by increasing volumetric breast density. Readers were informed about the enriched cancer prevalence in the study set but were unaware of specific proportions and did not receive feedback during their reading task. For each examination, readers provided a level of suspicion score using a slider ranging from normal to unsure to radiologically malignant. The slider corresponded to an underlying 0–100 scale that was not visible to the readers. Independent from this score, readers made a binary recall decision (yes or no). If a recall was made, readers were required to annotate the identified suspicious region or regions. Reading times were automatically recorded from the moment the examination was first displayed until the reader continued to the next examination.

5.2.4 Eye Tracking

The eye movements of the readers were recorded using an eye tracker (Pro Spark; Tobii Technology) at a sampling rate of 60 Hz, using a nine-point calibration protocol. The eye tracker registered *x* and *y* coordinates of the readers' gaze points on the screen, allowing for the calculation of fixations using the standard velocity threshold of 30° per second [143,144]. Every examination started with a center fixation cross, where the radiologists had to fixate on the cross and could recalibrate if necessary.

To evaluate how thoroughly the radiologists examined the breast area, the proportion of breast area covered with eye fixations was calculated. Mammographic views were segmented with a U-Net [152] to remove the background. Circles (radius, 2.5° [166]) were placed at each (*x, y*) fixation point, and fixation coverage was calculated as the

proportion of breast covered by circles and averaged across views and breasts for each examination and reader. For cancer examinations, analysis focused on fixations within the lesion region, defined as an area extending 3.0° of visual angle beyond the lesion radius to account for measurement errors of the eye tracker and uncertainties in the tumor radius estimates. Outcomes included total fixation time and time to first fixation within the lesion region. For each examination, the fixation times across all lesion regions from all views (craniocaudal and mediolateral oblique, both breasts) were summed, and the fastest time to first fixation from the first lesion region fixated on was used. Missed cancers were categorized into search, recognition, or decision errors. Search errors occurred when the radiologist failed to fixate within the lesion region. Recognition and decision errors occurred when the lesion region was fixated but the radiologist decided not to recall the case, with fixation times of 1,000 msec or less [166–168] considered recognition errors and longer fixation times classified as decision errors. For examinations with multiple lesions or a single lesion visible in both views, the lesion region with the longest lesion fixation time determined the error classification.

5.2.5 Statistical Analysis

Sample size was determined through a prior reader study [163], using the approach proposed by Hillis et al. [145]. To achieve a study power greater than 0.8 with an equal ratio of normal to abnormal examinations, a sample size of 150 was estimated to be sufficient to detect a minimum performance change resulting from AI support of 0.05 in the area under the receiver operating characteristic curve (AUC).

Breast cancer detection accuracy.—AUCs, sensitivity, and specificity with and without AI support were compared using mixed-model analysis of variance and the *t* test from multireader multicase analysis software (iMRMC, version 4.0.3; Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, Food and Drug Administration). Additionally, 95% CIs were calculated using the same software. The software treats readers and cases as cross-correlated random effects and the reading condition as a fixed effect, taking into account readers and cases being correlated and paired across reading conditions [157,158]. Receiver operating characteristic curves and their AUCs were computed using the suspicion score (scale, 0–100). Sensitivity and specificity were computed using the binary recall decisions. Cancer cases included biopsy-proven malignant lesions, and noncancer cases included benign and normal findings, as information to differentiate between normal and benign findings was not available. Several secondary AUC analyses were performed by analyzing subgroups according to *(a)* mammographic

lesion type (soft tissue, calcifications, or both); *(b)* histologic type; *(c)* tumor size (pathologic and/or mammographic); *(d)* breast density dichotomized into two categories: low density (Volpara Density Grade scores a and b) and high density (Volpara Density Grade scores c and d), determined by VolparaDensity (version 1.5.4.0; Volpara Health) [10,169]; *(e)* radiologist experience (< 10 years or ≥ 10 years), and *(f)* correct lesion localization. Correct lesion localization was defined as the reader having annotated the lesion location within 7 mm outside of the lesion radius, based on the median lesion size across all lesions. If the distance exceeded this threshold for all annotations in the examination, the reader's level of suspicion score was adjusted to the lowest score for that reader in that session. Standalone AI performance was compared with that of the readers by using the continuous AI examination scores. Sensitivity and specificity were also compared per AI category.

Reading time.—Reading times between reading conditions were compared using nonparametric bootstrap resampling. Examinations and readers were sampled with replacement 20,000 times to obtain differences in mean reading time when reading without versus with AI support. The number of bootstrapped differences in reading time either greater than 0 or less than 0 was counted, and the smaller of the two counts was divided by 20,000 and multiplied by 2 to obtain a two-tailed *P* value. Reading times were also compared per AI category using the same bootstrap procedure.

Eye tracker outcomes.—The proportion of the breast area covered by fixations (within 2.5° visual angle) was compared between readings without and with AI support using the aforementioned bootstrapping procedure. For the AI-supported sessions, fixation coverage was also compared before and after activating the AI markers. Additionally, readers' fixation time within the lesion region or regions, the time to first fixation in the lesion region, and the proportion of search, recognition, and decision errors were compared between readings with and without AI support using the same bootstrapping procedure.

$P < .05$ was considered indicative of a statistically significant difference for all analyses.

5.3 Results

5.3.1 Study Sample Characteristics

The final screening examinations from 150 women (median age, 55 years [IQR, 50–63 years]; age range, 49–72 years) included 75 biopsy-proven malignant cases and 75 noncancer cases with at least 2-year follow-up (Table 5.1).

Table 5.1: Characteristics of the selected study group.

Characteristic	Value
No. of women	150
Age at screening (years)	
Mean ± SD	57 ± 7
Median*	55 (50–63)
Range	49–72
Breast thickness (mm)*	58 (48–68)
Glandular radiation dose (mGy)*	1.9 (1.4–2.4)
Breast density†, ‡	
a	7 (4.7)
b	64 (43)
c	55 (37)
d	24 (16)
Screening outcome†, §	
True positive	64 (43)
True negative	69 (46)
False positive	4 (2.7)
False negative	11 (7.3)
Mammographic lesion type‖	
Mass	45/89 (51)
Calcification	32/89 (36)
Mass and calcification	6/89 (6.7)
Asymmetry	3/89 (3.4)
Architectural distortion	1/89 (1.1)
Architectural distortion and calcifications	2/89 (2.2)
No. of lesions per woman†	
One	63 (84)
Two	10 (13)
Three	2 (2.7)
Histologic type‖, #	
Invasive carcinoma of no special type	50/86 (58)
Ductal carcinoma in situ	21/86 (24)
Invasive lobular carcinoma	9/86 (10)
Mixed	5/86 (5.8)
Other invasive	1/86 (1.2)
Tumor size (mm)*, **	15.0 (9.2–21.0)

Note.—The study group was a randomly selected, cancer-enriched sample. * = Data are medians, with IQRs in parentheses. † = Data are numbers of examinations (n = 150), with percentages in parentheses. Percentages may not add up exactly due to rounding. ‡ = Volpara Density Grades scores (determined by VolparaDensity [version 1.5.4.0, Volpara Health] [10,169]]), which correlate with visual Breast Imaging Reporting and Data System 5th edition density categories [10]. § = Screening outcome was missing for two noncancer screening examinations due to missing recall data. ‖ = Data are numbers of lesions, with percentages in parentheses. Percentages may not add up exactly due to rounding. # = Histologic type was missing for three lesions. ** = Tumor sizes were primarily determined from pathology reports, and when unavailable, they were estimated from mammographic images.

5.3.2 Breast Cancer Detection Accuracy

Reading with AI support resulted in a higher reading performance than when reading unaided (AUC, 0.97 [95% CI: 0.95, 0.98] vs 0.93 [95% CI: 0.91, 0.96]; $P < .001$) (Figure 5.1A), with all radiologists showing an improvement (Figure 5.1B, Table 5.S1). Subgroup analysis showed no decrease in AUC with AI support across all cancer subgroups, mammographic lesion types, breast density categories, and radiologist experience levels (Table 5.2). The AI standalone AUC was 0.99 (95% CI: 0.97, 1.00) (Figure 5.1A, 5.S1).

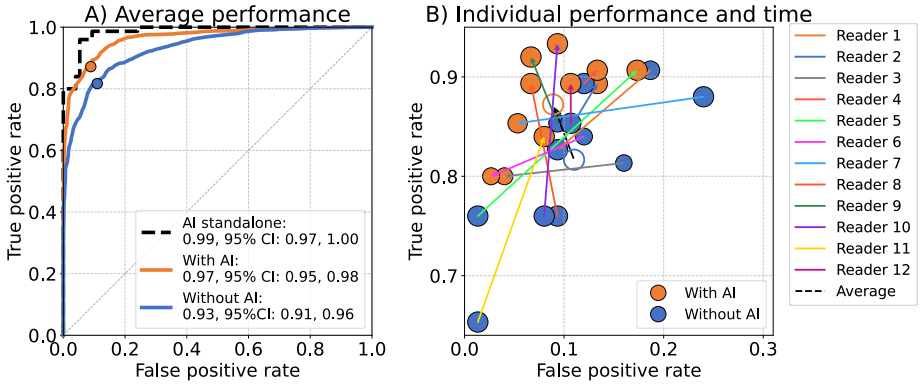


Figure 5.1: (A) Reader-averaged receiver operating characteristic curves of radiologists reading with and without artificial intelligence (AI) support and the standalone receiver operating characteristic curve of the AI system. Colored dots indicate the mean true- and false-positive rates of the radiologists. Areas under the receiver operating characteristic curves are reported together with their 95% CIs, which were derived using mixed-model analysis of variance. (B) Bubble chart shows changes in true-positive rate, false-positive rate, and reading time for each radiologist. The size of the colored disks is proportional to the mean reading time of each radiologist. Arrows point from the performance when reading without AI support to reading with AI support. Open circles represent the mean across all radiologists. Reading with AI support resulted in a higher reading performance than when reading unaided.

There was no evidence of a difference in mean sensitivity (87.2% [785 of 900; 95% CI: 81.4, 93.1] with AI vs 81.7% [735 of 900; 95% CI: 74.6, 88.7] without AI; $P = .06$) and specificity (91.1% [820 of 900; 95% CI: 86.8, 95.4] vs 89.0% [801 of 900; 95% CI: 83.7, 94.3]; $P = .46$) between the two reading conditions (Table 5.3, Figure 5.1A). Mean recall agreement was 86.9% for cancers and 89.7% for noncancers (Table 5.S2). Across AI risk categories, there was a trend suggesting that radiologists with AI support showed higher sensitivity for examinations categorized as very high risk by the AI tool (677 of 708 [95.6%] vs 634 of 708 [89.5%]) and higher specificity for low-risk examinations (554 of 564 [98.2%] vs 529 of 564 [93.8%]) compared with reading without AI support (Table 5.3). Figure 5.2 shows examples where AI support changed the number of recalls.

Table 5.2: Mean AUC overall and for subgroups read without and with AI support.

Subgroup	AUC without AI	AUC with AI	AUC difference*
All examinations	0.93 (0.91, 0.96)	0.97 (0.95, 0.98)	+0.03 (+0.02, +0.05)
Mammographic lesion type			
Soft tissue	0.93	0.95	+0.03
Calcifications	0.93	0.97	+0.04
Soft tissue and calcifications	0.97	0.99	+0.03
Histologic type			
Invasive carcinoma of no special type	0.95	0.97	+0.02
Ductal carcinoma in situ	0.89	0.96	+0.06
Invasive lobular carcinoma	0.96	0.99	+0.02
Mixed and other invasive†	0.86	0.91	+0.05
Tumor size			
0–10 mm	0.91	0.96	+0.04
10–20 mm	0.95	0.98	+0.03
≥ 20 mm	0.92	0.96	+0.03
Breast density‡			
Low	0.94	0.97	+0.03
High	0.92	0.96	+0.03
Experience level			
Least (< 10 years)	0.93	0.97	+0.03
Most (≥ 10 years)	0.93	0.96	+0.03
Location specific§	0.90	0.93	+0.03

Note.—Data in parentheses are 95% CIs, which were not calculated for subgroups because small sample sizes limited statistical testing. Areas under the receiver operating characteristic curves (AUCs) were higher when examinations were read with artificial intelligence (AI) support compared with no AI support. AUCs, without and with AI support, were compared using mixed-model analysis of variance from a multireader multicase analysis software (iMRMC, version 4.0.3). * = The differences in this table are calculated using precise values before rounding. Therefore, the displayed differences may not exactly match the result of subtracting the rounded AUC without AI from the rounded AUC with AI. † = Groups were combined due to limited sample sizes. ‡ = Breast density was dichotomized into two categories: low density (Volpara Density Grade scores a and b) and high density (Volpara Density Grade scores c and d), determined by VolparaDensity (version 1.5.4.0, Volpara Health) [10,169]. § = To account for incorrectly localized lesions, true-positive localizations were considered those that were within 7 mm outside of the reference radius of the lesion. Sensitivity analyses using alternative thresholds of 4, 10, and 20 mm gave similar results.

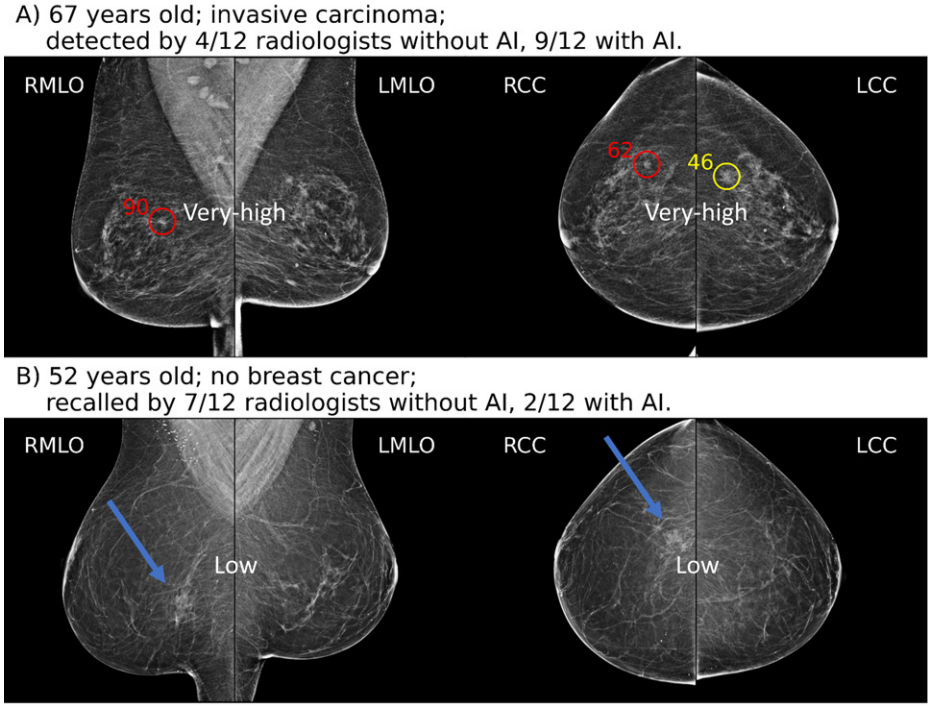


Figure 5.2: (A) Screening mammograms in a 67-year-old woman with a mass in the upper outer quadrant of the right breast. Four of the 12 radiologists detected this invasive carcinoma of no special type when reading without artificial intelligence (AI) support, while the same radiologists, along with five others, detected the cancer with AI support. The AI tool classified this examination as very high risk, with a maximum region score of 80 or higher at the cancer location. (B) Screening mammograms in a 52-year-old woman without breast cancer who was recalled by seven of the 12 radiologists when reading without AI support, compared with only two radiologists when reading with AI support, as five radiologists switched to no recall with AI support. The arrows indicate the locations annotated by the radiologists. The AI tool classified this examination as low risk, with a maximum region score under 40.

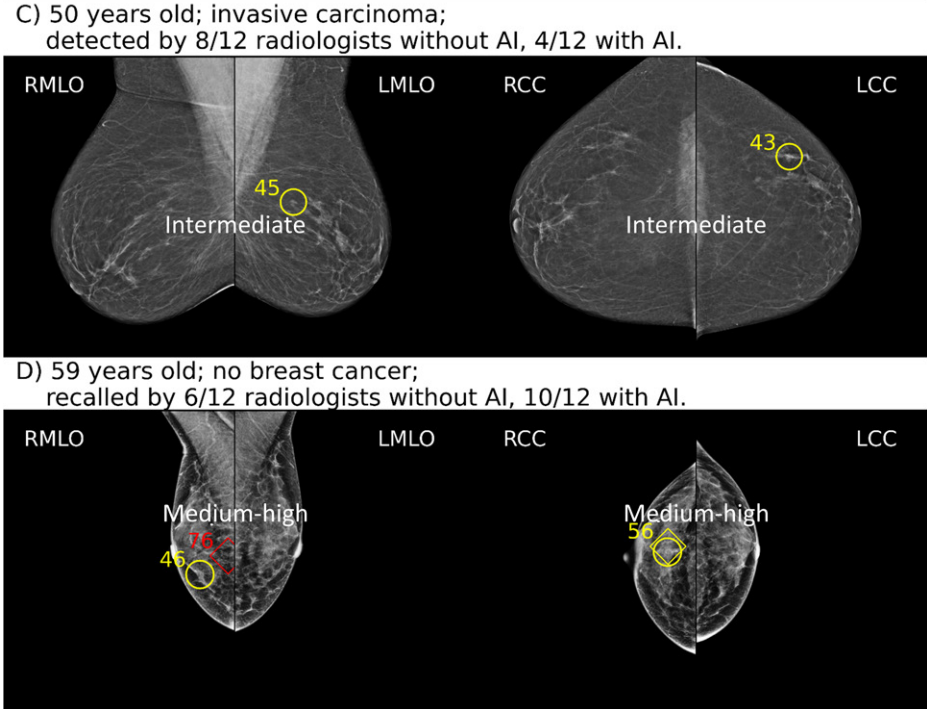


Figure 5.2: Continued

(C) Screening mammograms in a 50-year-old woman with an ill-defined mass in the upper outer quadrant of the left breast, diagnosed as invasive mixed ductal/lobular carcinoma. Eight of the 12 radiologists recalled this woman when reading without AI support, while only four radiologists recalled this woman when reading with AI support, as five radiologists switched to no recall and one radiologist switched to recall. The AI tool classified this examination as intermediate risk, with a maximum region score between 40 and 59. (D) Screening mammograms in a 59-year-old woman without breast cancer who was recalled by six of the 12 radiologists without AI support and by 10 radiologists when reading with AI support, with five radiologists switching to recall and one radiologist switching to no recall. The AI tool classified this examination as medium-high risk, with a maximum region score between 60 and 79. The AI examination category and AI-marked regions with region scores are shown as in the reader study, with diamonds indicating calcifications and circles indicating soft tissue lesions. LCC = craniocaudal image of the left breast, LMLO = mediolateral oblique image of the left breast, RCC = craniocaudal image of the right breast, RMLO = mediolateral oblique image of the right breast.

Table 5.3: Mean sensitivity and specificity of radiologists reading without and with AI support, categorized per AI risk category.

	Sensitivity (%)			Specificity (%)		
	Without AI	**With AI**	**Difference**	**Without AI**	**With AI**	**Difference**
All examinations	81.7 (735/900) [74.6, 88.7]	87.2 (785/900) [81.4, 93.1]	+5.6 (50/900) [-0.3, +11.4]	89.0 (801/900) [83.7, 94.3]	91.1 (820/900) [86.8, 95.4]	+2.1 (19/900) [-3.9, +8.1]
AI risk category*						
Low	NA†	NA†	NA†	93.8 (529/564)	98.2 (554/564)	+4.4 (25/564)
Intermediate	46 (22/48)	40 (19/48)	-6.3 (3/48)	86.8 (250/288)	85.8 (247/288)	-1.0 (3/288)
Medium-high	54.9 (79/144)	61.8 (89/144)	+6.9 (10/144)	46 (22/48)	40 (19/48)	-6.3 (3/48)
Very high	89.5 (634/708)	95.6 (677/708)	+6.1 (43/708)	NA‡	NA‡	NA‡

Note.—Data in parentheses are numbers of readings, and data in brackets are 95% CIs, which were not calculated for subgroups because small sample sizes limited statistical testing. With artificial intelligence (AI) support, sensitivity trended towards an increase for examinations categorized as very high risk by the AI tool, while specificity trended toward an improvement for low-risk category examinations. Sensitivity and specificity percentages, without and with AI support, were analyzed using mixed-model analysis of variance from a multireader multicase analysis software (iMRMC, version 4.0.3). NA = not applicable. * = Examinations were classified into AI risk categories based on maximum region scores: low risk (score, < 40), intermediate risk (score, ≥ 40 and < 60), medium-high risk (score, ≥ 60 and < 80), and very high risk (score, ≥ 80) for the presence of malignancy. † = No cancer examinations were classified as low risk by the AI tool. ‡ = No noncancer examinations were classified as very high risk by the AI tool.

5.3.3 Reading Time

There was no evidence of a difference in mean reading time with and without AI support (30.8 vs 29.4 seconds; $P = .33$) (Figure 5.3A, Table 5.4). Changes in reading time varied by radiologist (Figure 5.1B, Table 5.S3), assessment (Appendix 5.S2, Figure 5.S2), and AI category (Figure 5.3B). The mean reading time was reduced for low-risk AI category examinations (20.1 vs 25.1 seconds; $P < .001$) and increased for medium-high-risk AI category examinations (47.2 vs 36.7 seconds; $P = .006$) when radiologists used AI support (Figure 5.3B).

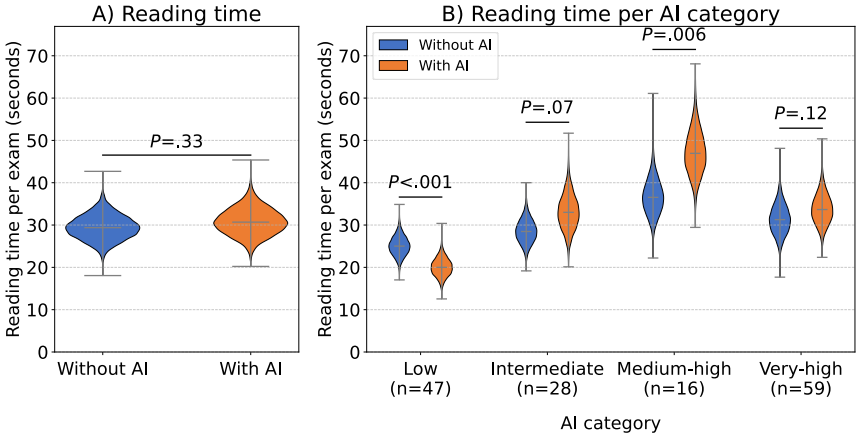


Figure 5.3: Violin plots show mean reading times per screening examination for radiologists reading without and with artificial intelligence (AI) support (A) for all examinations and (B) per AI risk category. The horizontal line within each violin represents the median. Examinations and readers were sampled with replacement 20,000 times to obtain differences in mean reading times when reading without versus with AI support. The number of bootstrapped differences either greater than 0 or less than 0 was counted, and the smaller of the two counts was divided by 20,000 and multiplied by 2 to obtain two-tailed P values. Examinations were classified into AI risk categories based on maximum region scores: low risk (score, < 40), intermediate risk (score, ≥ 40 and < 60), medium-high risk (score, ≥ 60 and < 80), and very high risk (score, ≥ 80) for the presence of malignancy. Overall reading time did not differ between readings with and without AI support but was reduced for examinations classified as low risk by the AI tool.

Table 5.4: Time, coverage, and false-negative errors for radiologists reading without and with AI support.

Parameter	Without AI	With AI	Difference	*P* value
Reading time (sec)	29.4	30.8	+1.3	.33
Breast fixation coverage (%)	11.1	9.5	-1.5	.004
Lesion fixation time (sec)	4.4	5.4	+0.9	.006
Time to first fixation (sec)*	3.4	3.8	+0.4	.13
No. false-negative readings	165	115		
Search error (%)	12.7	8.7	-3.9	.46
Recognition error (%)	30.1	20.5	-9.6	.16
Decision error (%)	57.2	70.8	+13.6	.10

Note.—Examinations and readers were sampled with replacement 20,000 times to obtain differences in mean outcomes when reading without versus with artificial intelligence (AI) support. For each outcome, the number of bootstrapped differences either greater than 0 or less than 0 was counted, and the smaller of the two counts was divided by 20,000 and multiplied by 2 to obtain two-tailed *P* values. With AI support, breast fixation coverage decreased, while lesion fixation time increased. Reading time, time to first fixation in the lesion region, and the distribution of different types of false-negative errors remained unchanged. * = In lesion region.

5.3.4 Breast Fixation Coverage

Reading with AI support resulted in a lower percentage of the breast area covered by fixations compared with unaided reading (9.5% vs 11.1%; *P* = .004) (Figure 5.4A, Table 5.4), with the majority of the radiologists (10 of 12) showing this reduction (Table 5.S4). Figure 5.5 shows examples of screening examinations where breast coverage was lower when reading with AI support. With AI support, radiologists covered more of the breast area before activating the AI markers than after they were activated (7.5% vs 4.7%; *P* < .001) (Figure 5.4B).

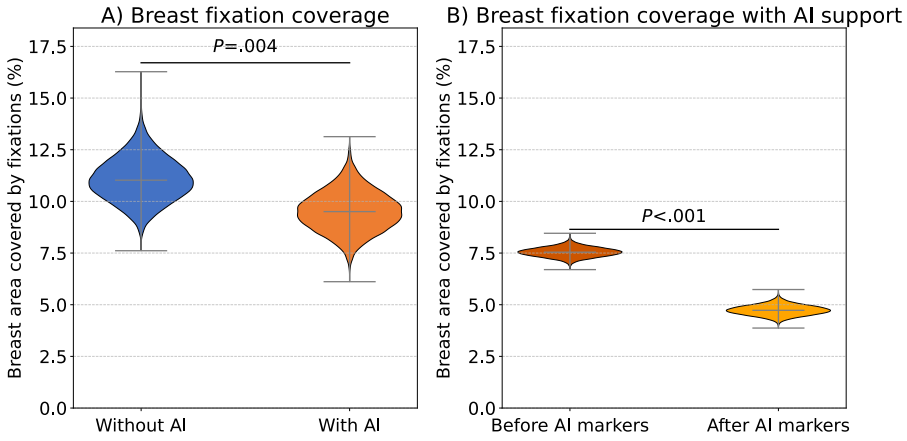


Figure 5.4: Violin plots show mean percentage of the breast area covered by fixations with the standard useful field of view (2.5° of visual angle) (A) for radiologists reading without and with artificial intelligence (AI) support and (B) for radiologists using AI support, before and after activating the AI markers. The horizontal line within each violin represents the median. Examinations and readers were sampled with replacement 20,000 times to obtain differences in mean breast coverage. The number of bootstrapped differences either greater than 0 or less than 0 was counted, and the smaller of the two counts was divided by 20,000 and multiplied by 2 to obtain two-tailed P values. The proportion of the breast area covered by fixations decreased with AI support, particularly after the AI markers were activated.

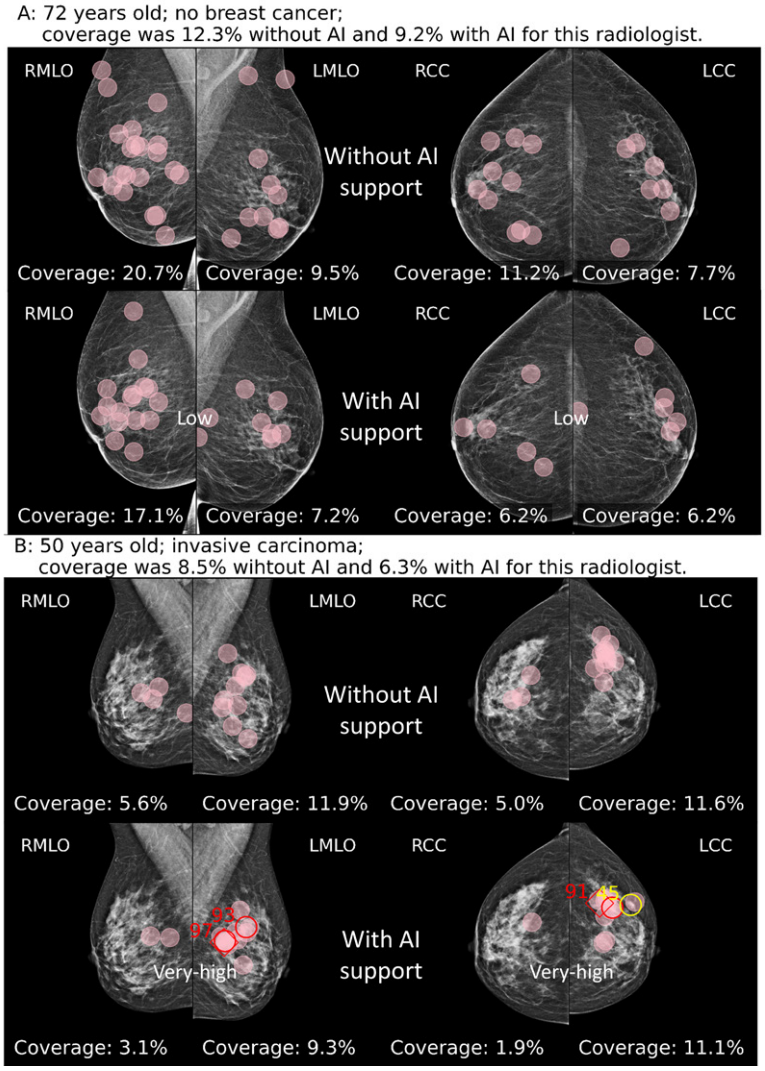


Figure 5.5: (A) Fixations (pink circles) of a radiologist while reading without and with artificial intelligence (AI) support on a screening mammogram in a 72-year-old woman without breast cancer. The radiologists did not recall the woman in either reading condition. The mean breast fixation coverage for this radiologist was 12.3% without AI support and 9.2% with AI support. The AI tool classified this examination as low risk, with a maximum region score under 40. (B) Fixations of a radiologist while reading without and with AI support on a screening mammogram in a 50-year-old woman with invasive carcinoma of no special type located dorsolaterally in the left breast. The radiologist recalled this woman in both reading conditions, with a mean breast fixation coverage of 8.5% without AI support and 6.3% with AI support. The AI tool classified this examination as very high risk, with a maximum region score of 80 or higher (red numbers; yellow number represents intermediate risk). Diamonds indicate calcifications, and circles denote soft tissue lesions. LCC = craniocaudal image of the left breast, LMLO = mediolateral oblique image of the left breast, RCC = craniocaudal image of the right breast, RMLO = mediolateral oblique image of the right breast.

5.3.5 Fixation Outcomes for Cancer Examinations

The mean fixation time within the lesion regions was higher when radiologists read with AI support (5.4 vs 4.4 seconds; *P* = .006) (Figure 5.6A; Tables 5.4, 5.S5). There was no evidence of a difference in the mean time to first fixation in the lesion region (3.8 vs 3.4 seconds; *P* = .13) or in the mean search, recognition, and decision error rates between AI-supported and unaided reading (search: 8.7% vs 12.7% [*P* = .46]; recognition: 20.5% vs 30.1% [*P* = .16]; decision: 70.8% vs 57.2% [*P* = .10]) (Figure 5.6B, 6C; Tables 5.4, 5.S5). Additional analyses using alternative thresholds (Figure 5.S3) to distinguish between recognition and decision errors, as well as a larger lesion region (extending 4.0° of visual angle beyond the lesion radius instead of 3.0°), resulted in similar conclusions, with no evidence of significant differences, except for lesion fixation time within the lesion region, which was consistent with the original analyses.

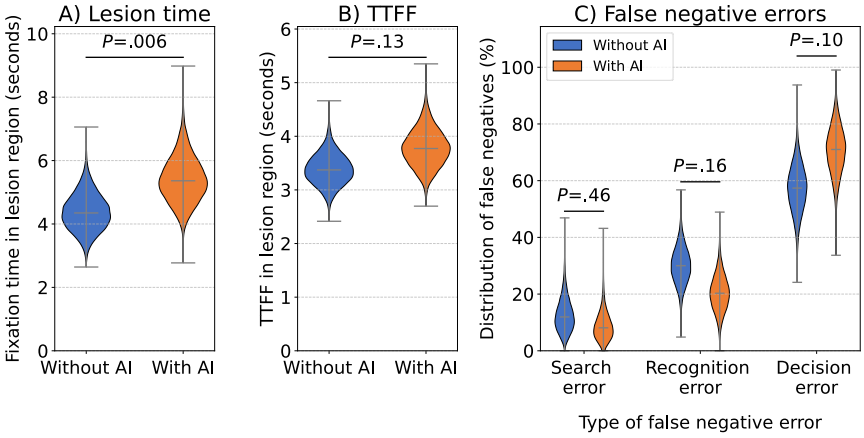


Figure 5.6: Violin plots show (A) mean fixation time within the lesion region (area extending 3.0° of visual angle beyond the lesion radius) for radiologists reading without and with artificial intelligence (AI) support; (B) mean time to first fixation (TTFF) within the lesion region compared for radiologists reading without and with AI support; and (C) mean search, recognition, and decision error rates among false-negative cases compared for radiologists reading without and with AI support. The horizontal line within each violin represents the median. Examinations and readers were sampled with replacement 20,000 times to obtain differences in mean outcomes when reading without versus with AI support. For each outcome, the number of bootstrapped differences either greater than 0 or less than 0 was counted, and the smaller of the two counts was divided by 20,000 and multiplied by 2 to obtain two-tailed *P* values. With AI support, fixation time on the lesion increased, while the time to first fixation within the lesion region and the distribution of false-negative error types remained unchanged.

After the observer studies were concluded, it was determined that due to a technical error, the AI markers of two normal examinations categorized as intermediate risk by the AI tool were not displayed, and that five other examinations assessed as low risk by the AI tool were displayed to the observers as intermediate risk (with no

markers). All analyses were repeated excluding these seven cases. The results were similar to those from the original 150 examinations (data not shown).

5.4 Discussion

Although artificial intelligence (AI) decision support may improve radiologist performance during screening mammography interpretation, its effect on radiologists' visual search behavior during mammography interpretation remains unclear. In this enriched multireader multicase study of 150 screening mammograms, AI support improved radiologists' screening performance, as indicated by a higher area under the receiver operating characteristic curve (0.97 vs 0.93; $P < .001$), without changing reading time (30.8 vs 29.4 seconds; $P = .33$). Eye-tracking data provided insights into the mechanism behind this improvement, showing that with AI support, radiologists spent more time fixating on lesion-containing regions (5.4 vs 4.4 seconds; $P = .006$) rather than on normal areas (as total breast fixation coverage decreased: 9.5% vs 11.1%; $P = .004$), adopting a more targeted search strategy. This change likely helped radiologists detect suspicious areas (addressing search errors) more effectively and better recognize these areas as requiring recall (addressing recognition errors).

The improvement in AUC aligns with previous studies on AI decision support in screening mammography [82–85]. Our results showed that high-performing radiologists improved even further with AI support, demonstrating that AI can enhance accuracy even for experienced radiologists. Across AI risk categories, we observed a trend suggesting that radiologists with AI support showed higher sensitivity in examinations categorized as higher risk by the AI tool and improved specificity for lower-risk examinations. While small sample sizes limited statistical testing, these trends suggest that AI positively influences recall decisions when its risk scoring is accurate.

There was no evidence of a difference in reading time between reading unaided and with AI support. Nevertheless, with AI support, radiologists spent less time on AI-assessed low-risk examinations and more time on medium-high risk examinations, suggesting that AI encourages radiologists to allocate more time to the latter, consistent with previous studies [82,84,170]. Considering that actual screening practice typically involves more low-risk examinations, AI decision support might actually reduce overall reading time and improve efficiency.

Despite no observed difference in overall reading time, eye-tracking data revealed that radiologists examined the images less comprehensively when assisted by AI.

Breast fixation coverage was lower, while fixation time in lesion regions was higher, suggesting that AI helped radiologists spend more time examining diagnostically relevant areas. This aligns with findings from CAD studies, where observers assisted by CAD focused more on areas marked by CAD and covered less of the search area [90,91]. Although the conclusions were similar, coverage percentages in these earlier studies were higher than those in this study, likely because those studies used smaller monitors, while in this study we used a larger standard mammography monitor (708 × 472 mm) displaying both breasts and views, as per standard screening practice. The observed reduction in breast coverage with AI support, compared with no AI support, may suggest a more efficient search, similar to how experienced radiologists cover smaller image areas but have better diagnostic performance [168,171]. However, reduced coverage could also increase the risk of missing cancers not detected by AI, which underscores the need for AI to be highly accurate and for radiologists to feel accountable for their decisions.

Interestingly, although radiologists spent more time examining diagnostically relevant areas with AI support, there was no evidence of a difference in time to first fixation in the lesion region between reading with and without AI support, suggesting that radiologists did not detect lesions faster with AI. This may be because, with AI support, radiologists initially followed their usual search strategy while considering the AI-assigned examination category before activating the AI markers. Only after clicking the AI marker button did their attention shift primarily to the AI-flagged regions while also reviewing some additional areas. This pattern of starting with comprehensive breast coverage followed by a more targeted focus on AI-identified regions suggests that our strategy of initially presenting only the AI risk category and allowing radiologists to decide when to view the AI markers seems effective in maintaining radiologists' own search strategies and potentially minimizing automation bias.

Notably, AI alone outperformed the combined efforts of readers and AI. However, the focus of this study was on the combination of readers and AI, rather than AI as a stand-alone tool. While AI advancements hold promise, its real-world standalone screening performance remains uncertain, and relying solely on AI currently raises unanswered questions about legal liability, safety, trust, transparency, and governance [172].

Our study has limitations. First, the cancer-enriched case set introduced a laboratory effect [150]. Future prospective studies should investigate the impact of AI support in an actual screening set. Second, radiologists did not have access to previous screening examinations or any other information, making assessments more difficult than in real screening practice. Third, this study used images from a single mammography vendor

and assessed only one AI system. Fourth, the order of the reading sessions was not randomized due to pragmatic constraints that required us to reuse the existing data from a previous reader study for the non–AI-aided reading [163]. Although the AI-supported session was performed at least 7 weeks later to minimize learning effects, implicit memory effects, such as contextual cuing [173], cannot be completely ruled out, highlighting a potential confounding factor that should be considered in future studies. Nevertheless, previous research by Gur et al. [174] concluded that memory effect did not significantly impact performance, suggesting no clear learning bias. In addition, no feedback was provided during or between the sessions, meaning that if radiologists recognized the examinations, they had no knowledge of the final diagnosis. Furthermore, the improvements observed with AI align with findings from previous studies with a counterbalanced design [82–84], supporting the validity of our results. Fifth, radiologists could only annotate one suspicion per view, possibly leading to a satisfaction of search effect [175], but this was true across conditions, reducing its influence. Sixth, lesion locations were determined by radiologists not participating in the study and could not be confirmed against biopsy-proven locations. While this may have affected lesion fixation analyses, the consistency across both reading conditions mitigates its impact. Seventh, the use of the density-based ordering may have influenced the results, but this ordering was consistent across both reading conditions, limiting its impact on the overall conclusions. Finally, as this study was performed with screening examinations and radiologists from the Netherlands only, the results may not be generalizable to other countries.

In conclusion, our study suggests that incorporating artificial intelligence (AI) as a decision support tool can improve radiologists' performance and efficiency by directing their attention to the most relevant examinations and regions within them. This approach likely improves lesion detection and recognition without increasing reading time and may even help reduce it. Further research using representative screening prevalence datasets is needed to confirm these potential efficiency gains. Future research should also explore alternative strategies for radiologists using AI decision support, such as showing all AI information upfront or providing it all only upon request, to determine the most effective approach.

Acknowledgments

The authors thank the additional breast screening radiologists who participated in the study: W. Dinkelaar, MD; A.M.B. Fauquenot-Nollen, MD; M.P.M. Gielens, MD; M.W. Imhof-Tas, MD; F.H. Jansen, MD; P.E.J.M. Sallevelt, MD, PhD; and M.M. Snoeren, MD. The authors also thank ScreenPoint Medical for providing the artificial intelligence system for this study.

Supplementary Material

Appendix 5.S1– Reader Sessions

Twelve Dutch screening radiologists (including K.M.D., J.K.v.R., A.F.v.R., J.B.H., D.B.N.) participated in this study. To mitigate potential memory effects, none of the readers were involved in the case selection process. To familiarize themselves with the AI system, readers underwent training involving 10 to 20 examinations, which were not part of the study. Readers had access to both views of each breast but were not provided with any other information about the case nor prior screening examinations. While they could zoom in and out and adjust the zoom location by repositioning the mouse, they were not allowed to adjust window width/level settings to keep reading time low. The reader study was performed with a 12-megapixel mammography-certified display (Coronis Uniti, Barco).

Additional Tables

Table 5.S1: AUC for each radiologist reading without and with AI support.

Reader	Years of experience with screening mammography	AUC without AI	AUC with AI	AUC difference*
R1	4	0.92	0.96	+0.05
R2	18	0.93	0.96	+0.03
R3	11	0.92	0.95	+0.03
R4	26	0.94	0.98	+0.05
R5	32	0.95	0.95	0.00
R6	12	0.93	0.97	+0.03
R7	25	0.92	0.96	+0.04
R8	5.5	0.95	0.97	+0.02
R9	6	0.96	0.96	+0.01
R10	13	0.94	0.97	+0.04
R11	4	0.90	0.97	+0.07
R12	6	0.95	0.97	+0.02
Mean	13.5	0.93	0.97	+0.03
95% CI mean		0.91, 0.96	0.95, 0.98	+0.02, +0.05

Note.—Areas under the receiver operating characteristic curves (AUCs) were higher when examinations were read with artificial intelligence (AI) support compared with no AI support. * = The differences in this table are calculated using precise values before rounding. Therefore, the displayed differences may not exactly match the result of subtracting the rounded AUC without AI from the rounded AUC with AI.

Table 5.S2: Sensitivity and specificity for each radiologist reading without and with AI support.

Reader		Sensitivity (%)				Specificity (%)			
	Years of experience with screening mammography	Without AI	With AI	Difference*	Agreement†	Without AI	With AI	Difference*	Agreement†
R1	4	91 (68/75)	83 (62/75)	-8 (6/75)	92 (69/75)	81 (61/75)	91 (68/75)	+9 (7/75)	85 (64/75)
R2	18	83 (62/75)	89 (67/75)	+7 (5/75)	91 (68/75)	91 (68/75)	87 (65/75)	-4 (3/75)	93 (70/75)
R3	11	81 (61/75)	80 (60/75)	-1 (1/75)	85 (64/75)	84 (63/75)	96 (72/75)	+12 (9/75)	88 (66/75)
R4	26	76 (57/75)	89 (67/75)	+13 (10/75)	84 (63/75)	91 (68/75)	93 (70/75)	+3 (2/75)	92 (69/75)
R5	32	76 (57/75)	91 (68/75)	+15 (11/75)	85 (64/75)	99 (74/75)	83 (62/75)	-16 (12/75)	84 (63/75)
R6	12	84 (63/75)	80 (60/75)	-4 (3/75)	83 (62/75)	88 (66/75)	97 (73/75)	+9 (7/75)	88 (66/75)
R7	25	88 (66/75)	85 (64/75)	-3 (2/75)	92 (69/75)	76 (57/75)	95 (71/75)	+19 (14/75)	81 (61/75)
R8	5.5	89 (67/75)	91 (68/75)	+1(1/75)	93 (70/75)	88 (66/75)	87 (65/75)	-1 (1/75)	91 (68/75)
R9	6	85 (64/75)	92 (69/75)	+7 (5/75)	85 (64/75)	91 (68/75)	93 (70/75)	+3 (2/75)	92 (69/75)
R10	13	76 (57/75)	93 (70/75)	+17 (13/75)	83 (62/75)	92 (69/75)	91 (68/75)	-1 (1/75)	93 (70/75)
R11	4	65 (49/75)	84 (63/75)	+19 (14/75)	81 (61/75)	99 (74/75)	92 (69/75)	-7 (5/75)	93 (70/75)
R12	6	85 (64/75)	89 (67/75)	+4. (3/75)	88 (66/75)	89 (67/75)	89 (67/75)	0 (0/75)	95 (71/75)
Mean	13.5	81.7	87.2	+5.6	86.9	89.0	91.1	+2.1	89.7
95% CI mean		74.6, 88.7	81.4, 93.1	-0.3, +11.4		83.7, 94.3	86.8, 95.4	-3.9, +8.1	

Note.—Data in parentheses are numbers of readings. There was no evidence of a difference in mean sensitivity and specificity between readings without and with artificial intelligence (AI) support. * = The differences in this table are calculated using precise values before rounding. Therefore, the displayed differences may not exactly match the result of subtracting the rounded value without AI from the rounded value with AI. † = Agreement refers to the number of examinations for which each reader provided the same recall decision across both reading sessions (without and with AI support): true positives or false negatives for cancer cases (sensitivity), and true negatives or false positives for noncancer cases (specificity).

Table 5.S3: Mean reading time per screening examination for each radiologist reading without and with AI support.

Reader	Years of experience with screening mammography	Reading time without AI	Reading time with AI	Reading time difference*
R1	4	21.1	19.9	-1.2
R2	18	34.6	41.9	+7.4
R3	11	16.6	17.5	+0.9
R4	26	41.7	41.3	-0.5
R5	32	47.3	48.6	+1.3
R6	12	17.3	17.6	+0.3
R7	25	29.1	22.9	-6.3
R8	5.5	39.7	42.9	+3.2
R9	6	18.5	27.1	+8.6
R10	13	30.2	27.7	-2.6
R11	4	27.1	32.1	+5.0
R12	6	30.0	30.2	+0.2
Mean	13.5	29.4	30.8	+1.4

Note.—Reading times are shown in seconds. There was no evidence of a difference in mean reading time when reading without and with artificial intelligence (AI) support. * = The differences in this table are calculated using precise values before rounding. Therefore, the displayed differences may not exactly match the result of subtracting the rounded time without AI from the rounded time with AI.

Table 5.S4:. Mean percentage of the breast area covered by fixations with the standard useful field of view (2.5° of visual angle) for each radiologist reading without and with AI support

Reader	Years of experience with screening mammography	Fixation coverage without AI	Fixation coverage with AI	Fixation coverage difference*
R1	4	8.9	6.4	-2.6
R2	18	10.4	12.2	+1.7
R3	11	7.0	6.2	-0.8
R4	26	16.3	11.1	-5.2
R5	32	18.1	15.2	-2.8
R6	12	8.4	6.2	-2.2
R7	25	8.2	5.6	-2.6
R8	5.5	15.6	13.7	-1.9
R9	6	7.3	7.2	-0.1
R10	13	11.7	9.8	-1.9
R11	4	11.8	12.6	+0.8
R12	6	9.1	8.1	-1.1
Mean	13.5	11.1	9.5	-1.5

Note.—Fixation coverages are shown in percentages. With artificial intelligence (AI) support, the mean percentage of the breast area covered by fixations was reduced. * = The differences in this table are calculated using precise values before rounding. Therefore, the displayed differences may not exactly match the result of subtracting the rounded fixation coverage without AI from the rounded fixation coverage with AI.

Table 5.S5: Mean fixation time and TTFF within the lesion region for each radiologist reading without and with AI support.

		Fixation time in lesion region			TTFF in lesion region		
Reader	**Years of experience with screening mammography**	**Without AI**	**With AI**	**Difference***	**Without AI**	**With AI**	**Difference***
R1	4	3.2	4.2	+1.0	2.2	3.1	+0.9
R2	18	5.8	6.1	+0.3	3.2	3.8	+0.6
R3	11	3.4	4.1	+0.7	2.0	2.0	0.0
R4	26	6.1	9.9	+3.8	4.6	3.5	-1.1
R5	32	7.4	9.3	+2.0	3.4	3.2	-0.2
R6	12	2.7	4.1	+1.4	1.6	1.7	+0.1
R7	25	2.7	3.8	+1.1	4.5	4.2	-0.3
R8	5.5	7.3	6.4	-0.9	4.6	6.3	+1.8
R9	6	2.4	3.8	+1.3	1.9	3.6	+1.7
R10	13	3.4	3.4	0.0	3.4	2.6	-0.7
R11	4	3.2	4.4	+1.2	7.0	7.9	+0.9
R12	6	4.9	5.3	+0.4	2.4	3.7	+1.3
Mean	13.5	4.4	5.4	+1.0	3.4	3.8	+0.4

Note.—Fixation times and time to first fixation (TTFF) are shown in seconds. With artificial intelligence (AI) support, the mean fixation time within the lesion region increased, while the time to first fixation in this region remained unchanged. * = The differences in this table are calculated using precise values before rounding. Therefore, the displayed differences may not exactly match the result of subtracting the rounded value without AI from the rounded value with AI.

Additional Figures

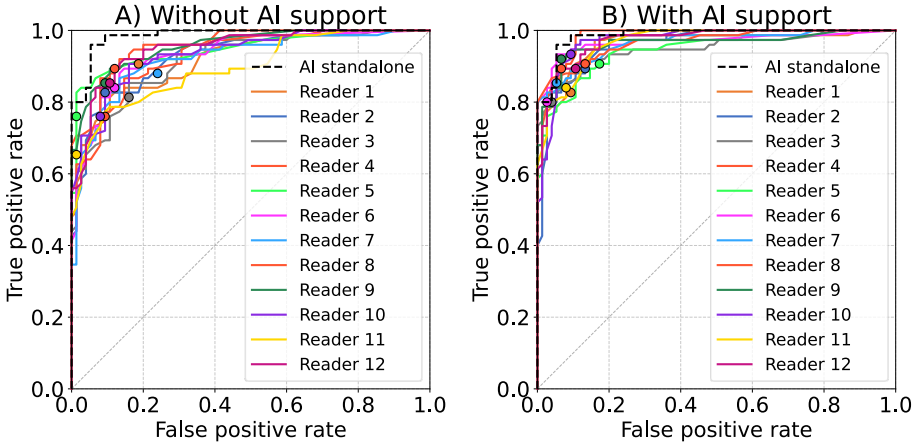


Figure 5.S1: Comparison between the standalone receiver operating characteristic curve of the artificial intelligence (AI) system and the receiver operating characteristic curves of the individual radiologists reading (A) without AI support and (B) with AI support. Curves are derived using mixed-model analysis of variance. Colored dots indicate the true- and false-positive rates of the individual radiologists. Radiologists improved screening performance when reading with AI support.

Appendix 5.S2– Reading Time per Radiologist Assessment

Methods

As a secondary analysis, reading times per screening examination were evaluated based on the radiologist assessment without and with AI support. The assessments were classified into four categories based on the radiologist's recall decision: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The analysis grouped the results as follows, with the first assessment being performed without AI and the second with AI:

- Cancers: TP → TP, FN → TP, TP → FN, FN → FN
- Non-cancers: TN → TN, FP → TN, TN → FP, FP → FP
- Reading times without and with AI were compared within each assessment combination (e.g., TP → TP without vs with AI). These comparisons were limited to within the same category (e.g., TP → FN vs. FN → TP comparisons were not made) to ensure that the effects observed were due to AI support, rather than differences in individual reading times.

Results

Reading times varied with diagnostic outcomes (Figure 5.S2). Reading times generally increased when radiologists changed a negative recall decision without AI support to a positive one with AI support. In contrast, reading times either decreased or showed no significant difference when switching from a positive decision to a negative one. For cancer cases that were correctly identified as TP when reading without AI, reading times remained unchanged when reading with AI support. However, for cancer cases that were missed (FN) when reading without AI, reading times increased when reading with AI support, likely due to AI's high-risk scores prompting more careful re-evaluation, often resulting in correct identification of the cancer. For non-cancer cases, the AI mostly assigned low-risk scores, with some exceptions categorized as intermediate or medium-high-risk. TNs with AI support had lower reading times compared to without AI, likely because AI's low-risk scores confirmed the initial low suspicion, enhancing efficiency. Conversely, cases that became FPs with AI support had increased reading times, probably due to AI's higher suspicion scores causing radiologists to spend more time on flagged regions and sometimes making incorrect recall decisions.

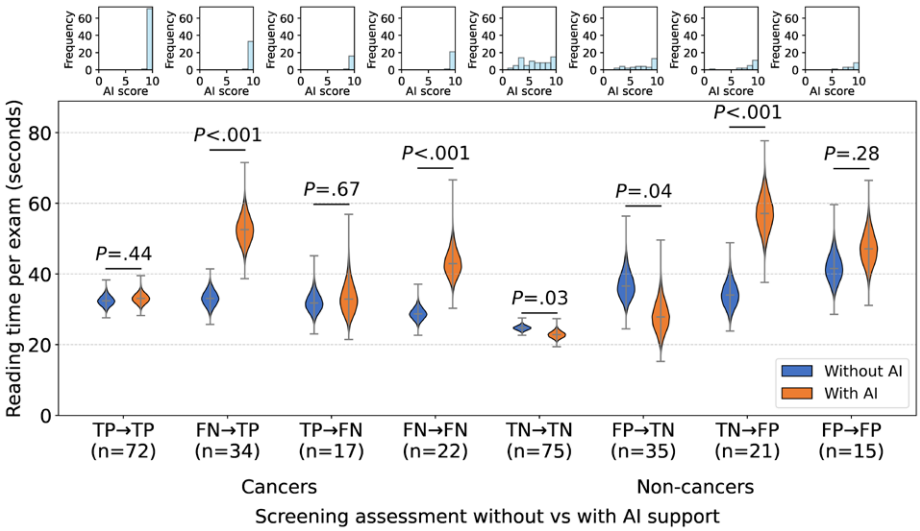


Figure 5.S2: Violin plots show mean reading times per screening examination compared per diagnostic outcome, with the first outcome being for reading without artificial intelligence (AI) and the second outcome for reading with AI support. The horizontal line within each violin represents the median. Examinations and readers were sampled with replacement 20,000 times to obtain differences in mean reading times when reading without versus with AI support. The number of bootstrapped differences either greater than 0 or less than 0 was counted, and the smaller of the two counts was divided by 20,000 and multiplied by 2 to obtain two-tailed *P* values. Reading times varied with diagnostic outcomes. FN = false negative, FP = false positive, TN = true negative, TP = true positive.

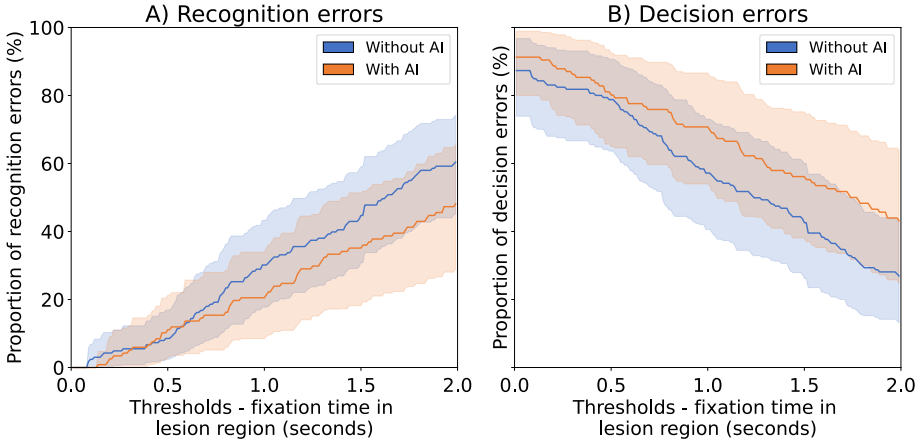


Figure 5.S3: Proportion of (A) recognition and (B) decision errors for a range of thresholds (0-2 seconds) when reading without and with artificial intelligence (AI) support. One graph is the complement of the other, although they start at different points due to variations in the proportion of search errors. Examinations and readers were sampled with replacement 20,000 times to obtain

95% CIs. There was no significant difference in the proportion of recognition and decision errors between readings without and with AI support, although a trend was observed toward a lower proportion of recognition errors and a higher proportion of decision errors with AI support.

Chapter 6

Automation Bias and Timing of Artificial Intelligence Decision Support in Screening Mammography

Jessie J.J. Gommers, Sarah D. Verboom, Mireille J.M. Broeders, Ioannis Sechopoulos

Published in SPIE Proceedings Volume 13928, Medical Imaging 2026:
Image Perception, Observer Performance, and Technology Assessment
DOI: 10.1117/12.3086110

Abstract

Purpose

To investigate how the timing of AI decision support influences radiologist performance, and to explore the role of automation bias in their decision making.

Materials and Methods

A reader study was performed using 150 screening mammograms (75 cancers: 75 normal) from the Dutch Screening Program (median age, 61 years [IQR, 55–67 years]). Examinations were analyzed using an FDA-approved commercial artificial intelligence (AI) system. The dataset was enriched with examinations where AI performance was suboptimal, defined as probability-of-malignancy (PoM) scores < 43 (out of 100) for cancer cases or PoM ≥ 43 for noncancer cases. Nine Dutch breast cancer screening radiologists read all examinations with (i) no AI support, (ii) on-request AI support (AI examination-level category displayed immediately but AI-detected suspicious region markers displayed only after a button click), and (iii) full concurrent AI support (all AI findings shown immediately). Radiologists recorded their recall decision and PoM score (1–100), and completed a post-study survey. Area under the receiver operating characteristic curve (AUC), sensitivity, and specificity were compared between the three reading conditions and across optimal and suboptimal AI results using analysis of variance (MRMCaov). Reading time was compared using bootstrap resampling.

Results

For examinations with optimal AI results, AUC was higher when reading with AI support than when reading without AI support (on-request: 0.967 vs 0.899; $P < .001$, full concurrent: 0.970 vs 0.899, $P < .001$). Conversely, for examinations with suboptimal AI results, AUC was lower with AI support than without it (on-request: 0.230 vs 0.383; $P = .006$, full concurrent: 0.213 vs 0.383, $P = .002$). Overall, there was no significant difference in AUC between the two AI-supported reading methods (on-request: 0.830, full concurrent: 0.829; $P = .89$). For most reading conditions, sensitivity, specificity, and reading time did not significantly differ. All radiologists preferred to use AI support, and seven preferred it to be on request.

Conclusion

This study suggests that reading with AI support can improve radiologist performance when AI results are accurate, but may reduce performance when AI results are suboptimal, independent of when AI region marks are displayed.

6.1 Introduction

Artificial intelligence (AI) is increasingly being used or considered as decision support for radiologists during screening mammography interpretation, showing positive results in reader studies [82–85,170,176–178]. Recently, prospective screening trials in nationwide screening programs have also implemented AI decision support [86–88]. However, there is a risk that radiologists may stop critically engaging with the AI results and become overly reliant on the system, known as automation bias [92,93]. As AI becomes more widely integrated into clinical practice, it is crucial to understand how and to what extent radiologists are influenced by automation bias, and how the timing of AI result presentation may affect their diagnostic interpretations. To the best of our knowledge, no prior studies have examined the timing of AI decision support. Therefore, the aim of this study is to investigate how the timing of AI decision support influences radiologist performance, and to explore the role of automation bias in their decision making. More knowledge about how radiologists interact with AI decision support can help to optimize its implementation in terms of both performance and timing.

6.2 Materials and Methods

This fully-crossed multireader multicase (MRMC) study was waived from ethical approval by the research ethics committee of Radboud university medical center.

6.2.1 Screening Examinations

The study was performed with 150 two-view (craniocaudal and mediolateral oblique view) bilateral digital screening mammograms from asymptomatic women (median age, 61 years [IQR, 55–67 years]) participating in the Dutch Breast Cancer Screening Program, obtained as part of the Personalised RISk-based MAmmography screening (PRISMA) study [179]. Exclusion criteria were missing images, missing ground truth outcomes, and breast implants. Examinations were selected with a 1:1 ratio of cancer and noncancer cases, including both screen-detected and interval cancers. The dataset was purposely selected to overrepresent examinations where the AI had low scores for cancers and high scores for noncancers.

6.2.2 AI System

Examinations were analyzed using a commercially available AI system (Transpara, version 2.1.0-A, Screenpoint) that has received Conformité Européenne mark approval and clearance from the U.S. Food and Drug Administration. The system has been

trained and validated on large, multi-institutional datasets collected using equipment from various vendors and representing diverse populations across multiple screening programs [80,88,161,164,165]. The system evaluates mammograms for suspicious findings with a score ranging from 1 to 100. The system highlights suspicious calcifications and soft tissue lesions with region marks if the score is ≥ 43. Based on the maximum region score, this version of the AI system categorizes the screening examination-level suspicion as *low* without any region marks (maximum region score < 43), *intermediate* (maximum region score between 43 and 74), or *elevated* (maximum region score ≥ 75). For the purposes of this study, scores were considered suboptimal if they were < 43 for cancer cases or ≥ 43 for noncancer cases. All other scores were considered optimal.

6.2.3 Reader Sessions

Nine Dutch breast cancer screening radiologists participated in the reader study. Eight radiologists had over 10 years of experience in breast imaging and five had more than 10 years of experience in breast cancer screening. The median screening reading volume was 10,000 examinations per year (IQR 9,000-14,000). Radiologists read the included examinations three times, once under each condition: (i) no AI support, where the examination was shown without any AI information, (ii) on-request AI support, where the AI examination-level category was displayed immediately upon opening the screening examination, but the AI region marks and their corresponding scores were only revealed if and when a radiologist clicked a button, and (iii) full concurrent AI support, where both the AI examination-level category and all AI region marks with their scores were shown immediately upon opening the screening examination. The order of the sessions was randomized across readers and separated by at least 18 days. The examination order was randomized across AI conditions but kept the same across all sessions. Reading was performed on a calibrated breast-imaging monitor (Barco Coronis Uniti) and radiologists were blinded to clinical history and the proportion of cancers included. An eye tracker (Tobii Pro Spark) kept track of the eyes of the radiologists on the screen. Radiologists were asked to fill in their recall decision (yes or no) and probability of malignancy score (ranging from 1 to 100). After completing the final reading session, radiologists completed a survey regarding their preferences for AI decision support. Reading time was automatically tracked by the software, starting from the moment the radiologist opened the examination until proceeding to the next examination.

6.2.4 Statistical Analysis

The primary outcome measures were area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, and reading time across the three reading conditions. Secondary outcomes included responses from a post-study survey assessing radiologist preferences for AI decision support.

AUC, sensitivity, and specificity were compared between the three reading conditions using MRMC analysis of variance (MRMCaov, version 0.3.0) [180]. Reading time was compared using bootstrap resampling, with examinations and readers sampled with replacement across 20,000 iterations. Stratified analyses were performed based on AI performance, distinguishing between optimal and suboptimal AI results and evaluated using bootstrap resampling. *P* values were Bonferroni corrected and considered statistically significant below .05/3.

6.3 Results

6.3.1 Study Sample Characteristics

The study sample characteristics are shown in Table 6.1. Breast density was predominantly category b (48%) or c (37%), and the most frequent histologic subtype was invasive carcinoma of no special type (58%).

Table 6.1: Characteristics of the selected study group.

Characteristic	Value
No. of women	150
Age at screening (years)	
Mean ± SD	61 ± 7
Median*	61 (55–67)
Range	49–75
Breast density†, ‡	
a	15 (10)
b	72 (48)
c	55 (37)
d	8 (5)
Screening outcome†	
True positive	67 (45)
True negative	67 (45)
False positive	8 (5)
False negative	8 (5)

Table 6.1: Continued

Characteristic	Value
No. of lesions per woman[†]	
One	72 (96)
Two	3 (4)
Histologic type[§]	
Invasive carcinoma of no special type	45/78 (58)
Ductal carcinoma in situ	19/78 (24)
Invasive lobular carcinoma	10/78 (13)
Other invasive	4/78 (5)
Tumor size (mm) [‖]	12 (7–17)

Note.—The study group was a randomly selected, cancer-enriched sample. [*] = Data are medians, with IQRs in parentheses. [†] Data are numbers of examinations (n = 150), with percentages in parentheses. Percentages may not add up exactly due to rounding. [‡] = Transpara Density Grades scores (determined by Transpara [version 2.1.0-A, Screenpoint Medical] [181]), which correlate with visual Breast Imaging Reporting and Data System 5th edition density categories [10]. [§] = Data are numbers of lesions, with percentages in parentheses. Percentages may not add up exactly due to rounding. [‖] = Tumor size was missing for four breast cancers.

6.3.2 Diagnostic Performance

Figure 6.1 and Table 6.2 display the diagnostic performance measures for the three reading conditions. Across all examinations, reading with AI support trended toward higher AUC compared to reading without AI support, although differences did not reach statistical significance (on request: P = .04; full concurrent: P = .06). For examinations with optimal AI results (n = 102), both on-request and fully concurrent AI support resulted in a significantly higher AUC compared to reading without AI support (P < .001 for both AI reading conditions). Conversely, for examinations with suboptimal AI results (n = 48), average radiologist performance was lower, and use of AI support resulted in a significantly lower AUC compared to readings without AI support (on request: P = .006; full concurrent: P = .002). No statistically significant difference in AUC was observed between the two AI-supported reading conditions (on request vs full concurrent). For most conditions, reader-averaged sensitivity and specificity did not significantly differ.

Table 6.2: Reader-averaged AUC, sensitivity, and specificity for the three reading conditions and split for optimal and suboptimal AI results.

	No AI support	On request AI support	Full concurrent AI support	No vs on request AI	No vs full concurrent AI	On request vs full concurrent AI
				P value	*P* value	*P* value
All examinations (*n* = 150)						
AUC	0.800 (0.746, 0.854)	0.830 (0.776, 0.885)	0.829 (0.773, 0.885)	.04	.06	.89
Sensitivity (%)	73.3 (64.4, 82.3)	76.1 (67.4, 84.9)	72.6 (61.7, 83.5)	.28	.86	.13
Specificity (%)	73.6 (64.6, 82.6)	74.8 (64.0, 85.6)	75.1 (62.4, 87.8)	.73	.80	.93
Optimal AI results (*n* = 102)						
AUC	0.899 (0.859, 0.938)	0.967 (0.946, 0.987)	0.970 (0.952, 0.987)	< .001*	< .001*	.60
Sensitivity (%)	79.6 (71.2, 88.1)	83.8 (76.1, 91.5)	80.8 (70.1, 91.5)	.08	.78	.25
Specificity (%)	88.3 (81.6, 94.9)	94.4 (88.5, 100.0)	94.4 (89.3, 99.6)	.01*	.04	1.00
Suboptimal AI results (*n* = 48)						
AUC	0.383 (0.221, 0.546)	0.230 (0.076, 0.383)	0.213 (0.075, 0.351)	.006*	.002*	.72
Sensitivity (%)	27.2 (5.0, 49.4)	19.8 (0.0, 40.7)	12.3 (0.0, 27.6)	.29	.03	.058
Specificity (%)	60.1 (47.7, 72.6)	56.7 (41.0, 72.4)	57.3 (37.1, 77.5)	.53	.76	.92

Note.—Data in parentheses are 95% CIs. Suboptimal artificial intelligence (AI) results were defined as scores < 43 for cancers or ≥ 43 for noncancers, all others were optimal. AUC = area under the receiver operating characteristic curve. * = denotes statistical significance.

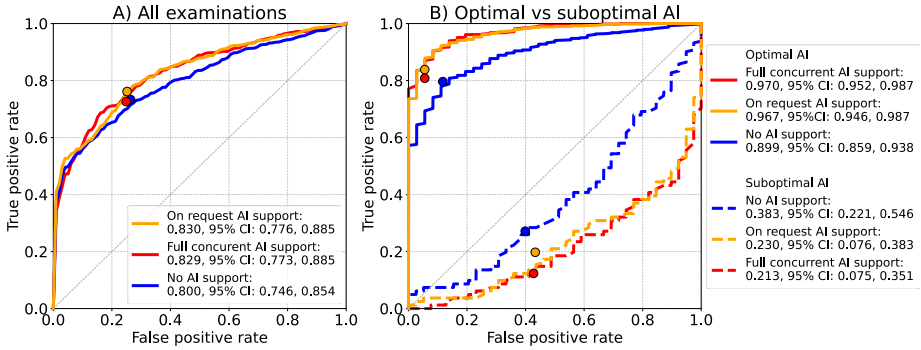


Figure 6.1: Reader-averaged receiver operating characteristic curves for the three reading conditions (A) for all examinations and (B) split for examinations with optimal and suboptimal artificial intelligence (AI) results. Suboptimal AI results were defined as scores < 43 for cancers or ≥ 43 for noncancers, all others were optimal. Colored dots indicate the reader-averaged operating point based on binary recall decisions. Areas under the receiver operating characteristic curves are reported together with their 95% CIs, which were derived using mixed-model analysis of variance.

6.3.3 Reading Time

Without AI support, the average reading time per examination was 32 seconds, compared to 35 seconds for both AI-supported methods. However, differences between reading conditions were not statistically significant. When distinguishing between optimal and suboptimal AI results no significant differences were found either (Figure 6.2).

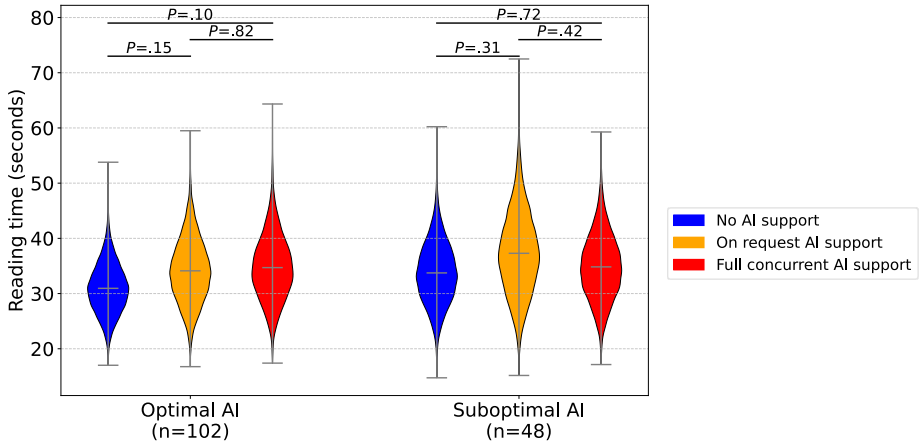


Figure 6.2: Violin plots show mean reading times per screening examination for the three reading conditions, stratified for optimal and suboptimal AI results. Suboptimal artificial intelligence (AI) results were defined as scores < 43 for cancers or ≥ 43 for noncancers, all others were optimal.

6.3.4 Radiologist Preferences

All nine radiologists preferred to use AI for decision support during their reading. Two radiologists preferred full concurrent AI support, with both the AI examination-level category and AI-detected regions displayed immediately upon opening the examination. The other seven radiologists preferred to see the AI-detected regions only upon request. Among these, four preferred to see the AI examination-level category immediately, while the other three would have preferred to access the category information also upon request.

6.4 Conclusion

In conclusion, this study shows that AI support can improve radiologist performance when the AI provides accurate (optimal) results, but may reduce performance when the AI output is suboptimal. Although no significant differences were observed in diagnostic accuracy or reading time between the two AI-supported methods, radiologists preferred to read with AI support on request rather than having all AI information displayed immediately. These findings suggest that radiologists are prone to automation bias regardless of when AI region marks are presented (immediately upon opening a case vs on request). More research is needed to investigate the impact of not showing the AI examination-level category immediately, an approach that was not tested in our study. Our next analyses will use the eye-tracking data to investigate how timing and accuracy of AI decision support influence radiologists' visual search patterns and interpretation strategies.

Acknowledgments

The authors thank the breast screening radiologists who participated in the study. The authors also thank ScreenPoint Medical for providing the artificial intelligence system for this study.

Chapter 7

General Discussion

This thesis aimed to improve radiologists' interpretation performance in screening mammography. Accurate mammography interpretation requires balancing the detection of clinically relevant breast cancers while minimizing false alarms. The findings from the studies presented in this thesis highlight how small adjustments in the existing screening infrastructure can improve radiologists' interpretation, creating opportunities for *real impact within reach*. It is expected that artificial intelligence (AI) will help with the interpretation of screening mammography in the near future. An important advantage of AI is that it may improve screening outcomes while also reducing radiologists' workload. In this final chapter, I will reflect on the findings from this thesis and provide more insight into the possible future directions for the use of AI in screening mammography. First, I will address the question of **who** should interpret the screening examinations. Second, I will discuss **how** the reader should carry out the interpretation. I will specifically focus on improving the interpretation of screening mammography in the Dutch Breast Cancer Screening Program, but the suggestions likely apply to other breast cancer screening programs as well.

Currently, all mammograms acquired in the Dutch Breast Cancer Screening Program are interpreted by two available radiologists, usually in the order in which the images were acquired, and without the involvement of AI. With approximately 4,000 screening examinations acquired on a single day and about 30 radiologists available for the reading, this means radiologists interpret almost 270 examinations each. However, there is room for improvement, both in enhancing screening outcomes and reducing radiologists' workload.

7.1 Who should interpret?

Until now, the answer would be two radiologists, with a consensus discussion or a third radiologist in case of disagreement. However, with the growing evidence that AI can detect breast cancer with accuracy similar to or exceeding that of individual radiologists [80,81], this answer may no longer be correct. There may be no single correct answer to this question, since the optimal reader combination may vary per screening examination and screening outcomes may benefit from specific reader pairings. Instead, a better question to answer could be: *who should interpret what and with whom?*

7.1.1 Who should interpret with whom?

Since double reading is the current practice, let's start by answering the question: *who should interpret with whom?*

7.1.1.1 Two radiologists

To date, radiologists that double read the same examinations are paired based on their availability. In **Chapters 2 and 3** of this thesis, we investigated whether strategically pairing radiologists, based on their individual performance characteristics, could improve the overall accuracy of double reading in breast cancer screening. Specifically, we tested whether pairing radiologists with either similar or opposite performance characteristics would lead to better group outcomes. Although substantial variation in individual radiologist performance was observed, the performance-based pairing strategies investigated in **Chapter 2** were not found to significantly improve overall screening performance compared to pairing radiologists at random. This may suggest that, based on individual performance metrics, it does not matter who reads with whom, or that the effects of performance-based pairing were too small to detect with the data available. To address this limitation, we used a different approach in **Chapter 3** of this thesis. This new approach included statistical modeling to simulate radiologist decisions and allowed for the comparison of many different reader pairs. Simulations showed that pairing radiologists with similar false-positive rates showed promise, reducing over 3,400 false alarms per year without significantly affecting cancer detection. On the other hand, it was found that pairing radiologists with similar true-positive rates decreased overall cancer detection and, hence, should be avoided. However, no pairing strategy consistently outperformed random pairing in both improving cancer detection and reducing false alarms. Therefore, while the findings indicate differences in outcomes between pairing strategies, they are not yet strong enough to advise on who should interpret with whom. Further research is needed to evaluate the impact of different pairing strategies and should include the impact on consensus meetings or arbitration decisions, accounting for important later steps in the screening workflow. Besides exploring pairing strategies based on individual performance characteristics, other promising pairing strategies are worth investigating. For instance, pairing experienced with inexperienced radiologists, as prior research has shown that they search images differently [168,182], and pairs of radiologists with different visual search patterns have been found to achieve better diagnostic performance [55].

7.1.1.2 AI as an independent reader

Perhaps a more exciting direction to determine who should interpret with whom in the future is the integration of AI into the double-reading process. With AI showing comparable performance to radiologists [80,81], as confirmed by the results in **Chapter 5**, different ways of implementing AI as an independent reader have been explored. Strategies include using AI as an additional reader or as a replacement for one of the two radiologists.

7.1.1.2.1 Two radiologists and AI as an independent reader

Using AI as an additional reader builds directly on the existing double-reading workflow and maintains the current roles of the two radiologists while adding an additional reader. Two previous prospective studies have investigated this triple reading [183,184]. In the Swedish ScreenTrustCAD study, two blinded radiologists and an AI system independently assessed all examinations, with any examination flagged by one of the three readers subsequently discussed in a consensus meeting. Results from this study showed that triple reading led to a nearly 7% increase in cancer detection, but at the expense of 46% more consensus discussions, and a 5% higher recall rate [183]. In another study, AI was implemented in breast cancer screening institutions in Hungary and used as a safety net that prompted arbitration when AI-flagged examinations received a negative assessment during normal double reading. Findings from this study indicated that triple reading increased the cancer detection rate by almost 5%, but at the same time, more than doubled the arbitration rate and increased the recall rate by about 4% [184]. Results from these two studies demonstrate that adding AI as an additional reader shows promise to increase cancer detection, but also results in increased workload and false positives. Therefore, triple reading does not strike the right balance in screening outcomes. Furthermore, the increased workload will be problematic for countries with a shortage of radiologists [162,185]. Hence, I do not think the question of who should interpret with whom will be answered with two radiologists and AI as an independent reader.

7.1.1.2.2 One radiologist and AI as an independent reader

A more plausible answer would be: one radiologist and AI as an independent reader. Several retrospective studies have simulated the use of AI as a replacement for the first or second reader, using historical reads from double-reading screening programs [104,106,107,164,186–191]. These studies reported encouraging findings, being able to reduce the radiologist workload, often with a tradeoff between sensitivity and specificity [104,106,107,164,186–189,191], but in most cases with similar or improved diagnostic accuracy compared with that of standard double reading [106,164,186–188,190,191]. However, retrospective studies cannot fully replicate the real-world screening scenario, particularly how AI-detected breast cancer signs influence radiologists and how they are managed in subsequent diagnostic steps, including arbitration and consensus reads. The ScreenTrustCAD study design allowed comparison of different reader combinations and is, to date, the only prospective study that reported on the effect of AI as an independent second reader in an actual screening workflow [183]. The research group reported that replacing one radiologist with AI increased cancer detection by 4%, and after consensus review, reduced recall rates by 4% compared with standard double reading. These

findings show that replacing one radiologist with AI could improve breast cancer detection while also reducing workload, suggesting this may be a good alternative strategy in double-reading programs. However, increased detection with AI included both in situ and invasive cancers, and follow-up research is needed to determine whether the improved detection leads to a reduction in interval cancers without adding to the overdiagnosis burden of screening. Unfortunately, the single-arm paired design of the ScreenTrustCAD study does not allow for future comparisons of interval cancers and, therefore, randomized controlled trials are needed. Another prospective paired-reader study in France is currently recruiting participants to investigate the effect of AI as a second reader [192]. Although not randomized, the study does compare the AI-assisted scenario against the standard of care. It will be interesting to compare their results with those from the ScreenTrustCAD study. Similar to exploring which radiologists should read together, as described in **Chapters 2 and 3**, it would be interesting to investigate if the human-AI reader combination can be optimized. Since radiologists vary in their performance metrics, it would, for instance, be worth exploring if the AI's operating point can be adjusted to complement the specific radiologists' strengths and weaknesses. If this proves to be effective, it could potentially improve screening performance and, hence, make it more certain that a radiologist should interpret with an independent AI reader.

7.1.2 Who should interpret what?

It is possible that the question of who interprets with whom cannot fully be answered without knowing what examinations need to be interpreted by how many radiologists. Mammograms are currently added to the worklists of radiologists without considering *who should interpret what*. Ideally, an examination is read by the reader, or pair of readers, that is most likely to interpret the case accurately. However, precise methods that characterize an examination and assign it to the "best" reader(s) for that case, to the best of my knowledge, do not yet exist. However, with the rise of AI algorithms that assess breast density, analyze mammograms for signs of breast cancer, and estimate the certainty of its prediction, a triage strategy that determines which examinations should be interpreted by whom, and by how many readers could be feasible. I can, for instance, imagine that mammograms that are estimated to have a high breast density, and are therefore generally more difficult to interpret, could benefit from being interpreted by two radiologists rather than one, and should be assigned to higher-performing radiologists. Additionally, I anticipate that an examination with AI-detected abnormalities may particularly benefit from an interpreting radiologist who excels in characterizing abnormalities of that type (e.g., calcifications). Alternatively, if AI were involved as an independent reader, the examination could potentially benefit from another radiologist who is particularly

skilled at detecting and characterizing abnormalities different from those detected by AI, thereby providing complementary value. However, evidence about which specific radiologists to select for the interpretation of specific examinations is lacking.

7.1.2.1 Single or double reading

While it remains uncertain who should interpret what examinations, it may be feasible to determine the optimal number of readers per examination. There is growing interest in a more personalized approach in which examinations are assigned to a different number of readers based on AI-generated probability of malignancy (PoM) scores. Previous studies have demonstrated that AI can distinguish between examinations with a high and low PoM, with substantially more breast cancers in examinations with a high PoM score and only a minority of cancers in those with a low PoM score [165,193–196]. This stratification allows for multiple possible triaging strategies, in which examinations are read by two, one, or no human reader(s). Examples of such approaches include assigning low-PoM examinations to a single human reader or standalone AI review (i.e., no recall), as well as directing high-PoM examinations to double reading or direct recall based on the AI-assigned score. Retrospective simulation studies have explored various triaging strategies with different AI vendors and thresholds [106,165,186,189,190,193,194,196–199]. Overall, results show promise to substantially reduce workload while maintaining or improving screening performance, laying the foundation for follow-up prospective studies.

So far, two prospective studies have reported on the effect of AI triaging, with low-PoM examinations assigned to single reading and high-PoM examinations assigned to double reading [161,200,201]. The Mammography Screening with Artificial Intelligence trial (MASAI) in Sweden is a randomized controlled trial that compared standard double reading against AI-supported reading, with AI used for triage and decision support [200]. In this trial, AI-supported reading was found to reduce the workload by 44% while improving screening performance. More specifically, trial results showed that compared with standard double reading, AI-supported reading resulted in a significant 29% increase in cancer detection, mostly of small, lymph node-negative, invasive cancers, without increasing the harm of false positives and the overdiagnosis of low-grade in situ cancer. The study by Lauritzen and colleagues compared screening outcomes before and after implementing an AI system for triage and decision support in the Danish Breast Cancer Screening Program [201]. They used a newer version of the AI system used in the MASAI trial, but triaged a smaller proportion of examinations to single reading (i.e., applied a lower threshold). Lauritzen et al. reported that after the implementation of AI-

supported reading, cancer detection was significantly increased by 17%, with an increase in the rate of small invasive cancers (≤1 cm), while the false-positive rate and screen-reading workload were reduced by 32% and 34%, respectively. The results from these two studies indicate that triaging examinations to single and double reading may be a plausible strategy to improve cancer detection and screening efficiency in a double-reading setting. Findings on interval cancers, screening sensitivity, and screening specificity are expected to be published soon and will provide further insights into the prognostic implications of using AI for triaging examinations to one or two radiologists. With the emerging evidence from prospective studies, additional trials are being prepared and conducted. In Norway, a prospective randomized controlled trial, named Artificial Intelligence in Mammography Screening (AIMS), is currently recruiting participants to investigate the effect of AI for triaging to single and double reading [202]. In the Netherlands, a similar prospective randomized controlled trial with two rounds of screening is being discussed. This trial, named Accurate and Intelligent Reading for EArlier breast cancer Detection II (aiREAD II), is a continuation of the project described in this thesis, and would make it possible to determine whether the positive findings obtained during the past five years will translate into real-world screening.

Although these findings about the opportunities of single and double reading do not directly provide an answer to which specific radiologists should interpret what examinations, it opens new questions about who should interpret what. Examinations likely to be normal and assigned to single reading may benefit from other readers than examinations with a higher PoM assigned to double reading. In the MASAI trial, separate worklists were used for single and double reading, and single-reading worklists were only assigned to experienced radiologists. Therefore, depending on whether an examination was interpreted by one or two readers, the interpreting radiologist could differ. In addition to radiologist experience, individual performance characteristics, like those investigated in **Chapters 2 and 3**, could be used for assigning specific readers to specific cases. I can imagine that low-PoM examinations benefit most from radiologists who have a high specificity, whereas high-PoM examinations may gain more from radiologists with a high sensitivity. In this way, AI may help to optimize the workflow by matching examinations to readers whose strengths complement the case characteristics. However, more research is needed, and any strategy would need to account for practical constraints. Availability of radiologists and additions and departures of radiologists would, for instance, need to be considered to ensure feasible implementation. Additionally, any system for assigning readers would need to be regularly recalibrated to reflect changes in radiologists' performance over time.

7.1.2.2 AI as a standalone reader

Beyond the question of which radiologists should interpret what examinations, a growing question concerns whether radiologists should interpret all examinations. Prospective studies conducted so far have investigated the effect of AI with at least one human reader involved in the interpretation of each examination. To my knowledge, no prospective study has yet investigated the effect of excluding human interpretation from all or some of the examinations. Ethical, legal, and social implications most likely hindered the use of AI as the sole reader in actual screening settings [203]. Surveys among women have also shown that, despite a positive attitude toward the use of AI, most women still prefer to have at least one human reader involved in the interpretation of their mammograms [204–206]. However, with continuous improvements in AI, increasing trust, and new developments in regulatory frameworks, I believe AI will eventually be used as a standalone tool for at least some of the screening examinations.

Since the majority of examinations (~99%) in screening do not result in a breast cancer diagnosis, AI could potentially have a substantial impact if used as a standalone reader to identify these cases. Previous retrospective studies have shown that AI is capable of identifying normal examinations and could thus pre-select these to be excluded from human interpretation [106,189,194,196–198]. The findings after pre-selection showed improvements in specificity [189,194,197] and reductions in recall rates [106,198], without substantially affecting cancer detection. However, retrospective results cannot be directly translated to real screening practice, particularly since studies of this type do not assess how radiologists are influenced by interpreting fewer examinations, with a higher expected cancer prevalence, knowing the AI marked them with a higher PoM. To know the true effect of excluding likely normal examinations from human interpretation, prospective trials in actual screening settings are needed. In the Netherlands, a large prospective randomized controlled screening trial, named imPROved breasT cancEr sCreening with sTandlonE anD responsible AI (PROTECTED-AI), is being discussed. PROTECTED-AI aims to compare the screening outcomes of standard double reading with those of AI-supported reading, using AI to identify examinations that can be assessed as normal and therefore excluded from human interpretation, and to aid radiologists in interpreting the rest of the examinations. To minimize the potential risk of harm, the plan is to obtain standard radiologist interpretations outside the trial for the AI-assessed normal examinations. These readings will determine the actual screening outcome.

In addition to excluding AI-determined normal examinations from human interpretation and to potentially improve cancer detection, another approach could be

to automatically recall AI-determined high-PoM examinations. Although prospective evidence is lacking, retrospective studies have shown the potential to improve sensitivity and/or specificity when excluding examinations with low and high AI PoM scores from human interpretation [165,186,190,193,198]. Nevertheless, only a small portion of the screening examinations will be assessed with a high PoM, and therefore, excluding these from human interpretation will have only a minor effect on workload reduction.

However, just like radiologists, AI is not perfect, with its performance suffering in certain cases, resulting in false-negative and false-positive interpretations. Ideally, AI should not be used when its predictions are inaccurate, and, therefore, it is important to be able to assess for which examinations the AI interpretation is likely to be correct and for which it is not. As part of aiREAD, the project to which this thesis contributes, Sarah Verboom et al. developed and tested an algorithm that predicts the certainty of the AI prediction per examination [207]. Results showed that a hybrid reading strategy, in which recall decisions were made by the AI when its predictions were likely to be accurate (i.e., certain) and by radiologist double reading when the AI's predictions were uncertain, was able to reduce the radiologist reading workload by nearly 40% while maintaining recall and cancer detection rates similar to standard double reading. These findings indicate that AI can be safely used as a standalone reader for mammography interpretation when its predictions are certain. Incorporating AI certainty metrics can thus help ensure safer use of AI. Using AI certainty metrics may also help to build trust in AI due to improved transparency with AI acknowledging its strengths and limitations [208].

7.1.3 My perspective on who should interpret

Given these findings, the question of who interprets screening examinations may soon be redefined. I believe it is increasingly likely that breast cancer screening in the near future will include a combination of different triaging strategies, excluding some examinations from human interpretation and triaging other examinations to single or double human reading. To be able to answer the question "*Who should interpret?*" and more specifically "*Who should interpret what and with whom*", I believe all examinations should be processed by an AI algorithm that assesses both the PoM and the certainty of its prediction, with the outputs used to determine which readers should be involved for what examinations. Figure 7.1 illustrates what I think the interpretation strategy should look like in the future. It shows that examinations assessed by AI as normal with high confidence (i.e., low AI PoM and high certainty) can be excluded from radiologist interpretation, whereas the remaining examinations should be single- or double-read by radiologists, depending on the AI certainty.

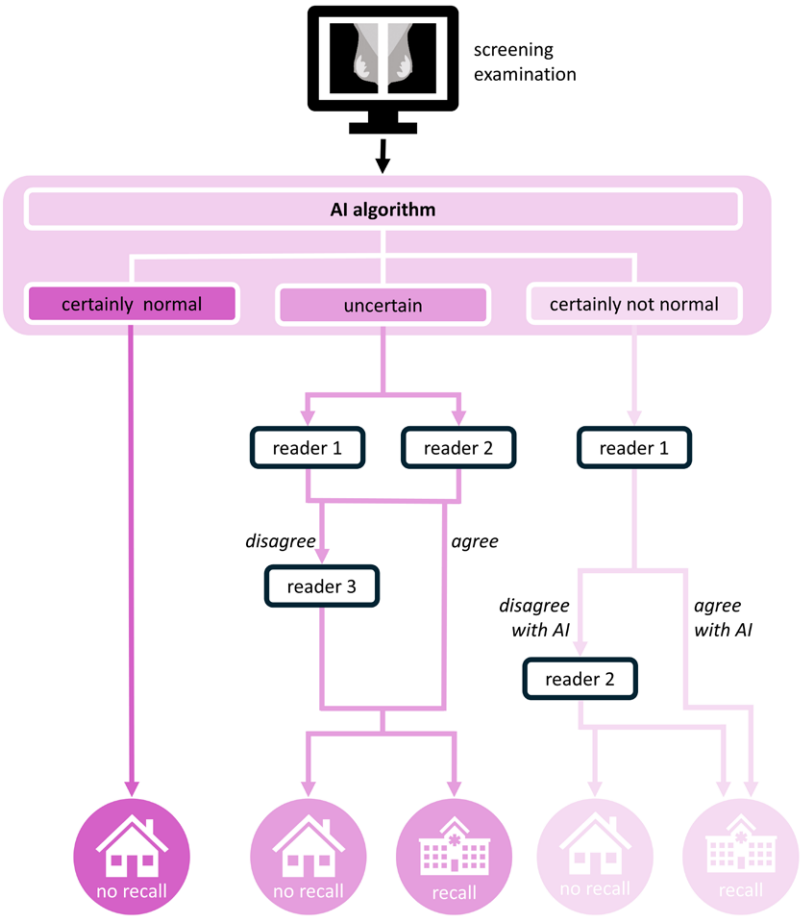


Figure 7.1: Illustration of my recommended future approach for the interpretation of mammography screening examinations, combining the interpretation of radiologists and artificial intelligence (AI).

7.2 How should the reader interpret?

Once it is established who interprets the screening examination, the next question to consider is: *how should the reader interpret?* Clearly, the answer depends on who is being considered the designated reader: the AI or one or two radiologists.

How AI interprets is largely shaped during the development and training phases, during which choices made about the training data, the model architecture, and the optimization methods shape how the AI will process and understand information when used. The focus of this thesis, however, was not on AI development or on

understanding the internal interpretation mechanisms of AI models. AI algorithms are known for being "black boxes", where information goes in and decisions come out, but it is not always clear how these decisions are derived. Recent AI developments that incorporate temporal information from multiple prior mammograms of the same woman may provide additional insight into how AI interprets images. Early results suggest that the temporal-comparison approach can improve AI's performance [209], similar to how reviewing previous examinations benefits radiologists during interpretation [210]. However, if AI is giving accurate results, do we really need to understand how it arrived at its conclusion? It may be more important to know that its conclusion is accurate. Since multiple commercial AI systems for mammography already exist, a more important question to consider might be which AI system should be used for interpretation. To be able to select the AI system that works best for its intended purpose, more studies comparing different AI algorithms are needed.

While an AI's interpretation is fixed after training, radiologists' interpretation can still be influenced by factors such as fatigue, distractions, or workload, affecting interpretation accuracy [33]. To help radiologists during mammography interpretation, multiple strategies exist. Examples include optimizing the reading environment, incorporating breaks, and providing feedback [33,211]. In this thesis, the focus was on using AI for decision support and the ordering of screening examinations to improve interpretation. I will discuss what I have learned about this in the next paragraphs.

7.2.1 AI for decision support

AI for decision support can be one of the most promising ways to help radiologists in interpreting screening mammograms. In **Chapter 5** of this thesis, the effect of AI decision support on radiologists' diagnostic accuracy and visual search was investigated. Similar to previously published findings [82–88], our results showed that AI decision support improved radiologists' performance. Radiologists spent more time and detected more cancers in examinations classified as suspicious by the AI, while they spent less time and had fewer unnecessary recalls in examinations that were assigned a low PoM score by the AI. In addition, our study included eye-tracking data, which provided new insights into the impact of AI decision support on radiologists' visual search patterns, a topic that, to the best of my knowledge, has not been investigated before. We found that with AI decision support radiologists adopted a more targeted search strategy, covering a smaller proportion of the breast with eye movements and spending more time examining regions that contained true lesions. This suggests that AI decision support can help radiologists prioritize their attention, enabling them to focus more on the examinations and regions where it matters most.

Despite promising results, adoption of AI for decision support has been slow due to concerns about overreliance on AI outcomes [92–95]. The more targeted search strategy observed with AI decision support in our reader study can make search more efficient (i.e., more focus on suspicious regions and reassurance for less suspicious areas), but also raises concerns about automation bias if the AI makes mistakes and the radiologists rely on its outputs. Findings from **Chapter 6** of this thesis indicate that radiologists are prone to automation bias when using AI decision support, with improved screening performance when the AI provided accurate results, but reduced performance when the AI output was suboptimal.

Nevertheless, some regional screening programs have already implemented AI for decision support and have shown promising results with improved cancer detection [86,88,201]. While automation bias is a valid concern, AI decision support may still provide greater benefits than the risk it poses, especially since AI is evolving quickly and standalone performance continues to improve. In an effort to balance the risks and benefits, each implementation applied its own approach. In the regional Córdoba Breast Cancer Screening Program in Spain, AI was temporarily (March 2021–March 2022) used for concurrent decision support, with radiologists interpreting examinations while simultaneously viewing the AI findings [88]. The limited duration of this implementation was due to the subsequent testing of a more radical AI triaging strategy [212]. In the Capital Region of Denmark, AI has been used to triage examinations to single and double reading and for decision support since November 2021 [201]. In this implementation, decision support is only used for examinations deemed suspicious by the AI, not for examinations with low PoM scores, with the aim of minimizing false positives. Instead of simultaneously viewing AI results, radiologists in Denmark are instructed to consider the AI lesion markings only after they have evaluated the examination themselves. In Germany, twelve screening units were involved in the PRAIM (PRospective multicenter observational study of an integrated AI system with live Monitoring) implementation study [86]. In this study (July 2021–February 2023), AI decision support was primarily used to label confident negative AI predictions, aiming to reduce the recall rate. AI suspicious findings were only shown when examinations identified as suspicious by the AI were not recalled by the radiologists. Interestingly, this implementation allowed radiologists to decide per examination whether they wanted to read with or without AI decision support. Although the described implementations are different, they all show promising results.

To further improve the benefits of AI decision support and reduce the risks of automation bias, AI certainty estimations could be used to select which AI results

should and should not be shown to the radiologists. By displaying AI outcomes only when they are certain and likely to provide meaningful benefit, and avoiding its use when outputs are uncertain, the risk of misleading radiologists would be reduced. Additionally, the timing of AI decision support may influence radiologists' performance and the degree of automation bias. However, to the best of my knowledge, no studies have yet suggested when the overall AI examination score and AI region markers should be made available to the radiologists. Displaying AI outputs before or immediately after opening an examination will likely influence radiologists' search and interpretation strategies the most, which is helpful if the AI is correct, but can bring risks if the AI is incorrect. In **Chapter 6**, we explored the difference between displaying AI region markers immediately upon opening an examination or showing them only when requested by the radiologist, with the overall AI score displayed as soon as the examination was opened in both strategies. The study showed that the two decision-support strategies did not result in significant differences in radiologists' performance and reading time. Most radiologists preferred to view the AI region markers upon request rather than immediately upon opening the examination. In fact, some radiologists would have preferred for the overall AI score to also appear only upon request, but we did not evaluate this option. More research is needed to investigate how displaying both the overall AI score and region markers only when requested affects interpretation. Other options for AI decision support include displaying AI results only when they conflict with radiologists' interpretations (i.e., serving as a safety net) or using AI solely for arbitration or consensus discussions. Ultimately, the ideal strategy would maximize radiologists' performance while minimizing reading time.

7.2.2 Ordering mammograms

To date, there are no guidelines on the order in which mammograms should be read, and therefore, most radiologists read the mammograms in the order they were acquired. However, studies demonstrating that exposure to mammograms can induce visual adaptation [66–69] prompted us to investigate whether ordering mammograms based on textural features could help radiologists during interpretation. Findings from the reader study described in **Chapter 4** of this thesis show that interpreting mammograms by increasing volumetric breast density can improve radiologists' diagnostic accuracy and reduce reading time. Even though the difference in screening performance between reading in a random order and by increasing density was small, it might have a substantial impact when multiplied by the large number of examinations interpreted every day. This finding suggests that ordering screening examinations in the radiologists' worklists is an important yet previously unrecognized factor that could improve radiologists' interpretation of screening mammography.

Although reading mammograms by increasing volumetric breast density showed promise, multiple other ways for ordering can be considered. Focusing on breast density, it would be interesting to investigate whether starting a batch with dense breasts and ordering by decreasing density instead of increasing density would induce similar effects. Ordering by increasing or decreasing density will most probably result in similar adaptation effects, and therefore, perhaps the performance improvement would be the same. However, in addition to visual adaptation, other factors, such as fatigue and vigilance, may also affect reader performance and thus influence what reading order may be most suitable. There is evidence of a vigilance decrement developing over time in various repetitive visual tasks, including the assessment of medical images [213]. Vigilance decrement is defined as a decline in the ability to detect targets as time on task increases [214]. This may suggest that radiologists perform better when the more difficult examinations (i.e., the densest ones) are presented first rather than last, and might thus be a motivation to explore the effects of reading mammography examinations in order of decreasing density. The CO-OPS trial in England, however, found no evidence of a vigilance decrement [62]. In this trial, standard double reading, where both readers interpreted a batch of mammograms in the same order, was compared with reversed double reading, where the second reader reviewed the same batch of mammograms in reverse order from the first reader. It was expected that if vigilance decrement occurred, cancer detection rates would decline over the course of a batch in the standard double-reading group, but not in the reversed reading group, as readers in this group would experience vigilance decrements at different points within the batch. However, no overall difference in cancer detection was found between the two reading groups. Instead, the authors found an increase in efficiency over time, with reduced recall rates, false positives, and reading times. In contrast to the vigilance decrement effect, these findings align with the idea that readers become better adapted to images as they move through a batch. Another study with digital breast tomosynthesis examinations found similar results regarding performance changes with time on task, including reduced recall rates and interpretation times, without a corresponding change in sensitivity [61]. Instead of reading the more difficult examinations in the beginning, this may thus be a motivation to interpret the more difficult cases at the end of a batch, when radiologist performance may be higher (i.e., ordering by increasing density as described in **Chapter 4**). However, the median batch size in the CO-OPS trial and tomosynthesis study by Abbey et al. were 35 and 7, respectively, which may have been too small for the vigilance decrement to occur. There was, however, some evidence of vigilance decrement shown in an observational study in the Norwegian Breast Cancer Screening Program [60]. This study included larger batch sizes and found not only reduced false-positive

recalls and reading times, but also a significant reduction in true-positive (cancer detected) assessments with time on task. In this Norwegian study, cancer detection decreased from 4.0 per 1,000 readings at image position 10 to 3.7 per 1,000 at image position 150. Observational results from the Dutch Program (average 125 examinations per batch) showed similar findings, with reduced suspicion scores and reading time over the course of a batch, but the study did not report on cancer detection specifically [59]. These results suggest that readers get less suspicious over a batch, resulting in a reduced sensitivity, consistent with vigilance decrement, potentially driven by reduced time spent per examination. Another explanation for the decreased sensitivity and a corresponding increased specificity over time could be the prevalence effect theory described by Evans and colleagues [35]. This theory states that the proportion of cancers in a set influences the performance of radiologists, with higher sensitivity in high-prevalence settings compared to low-prevalence settings. In case of extremely rare occurrences, it is usually wise to be conservative — *did I just see a dancing cactus? I do not think so*. In screening-reading, this prevalence effect develops throughout a reading session. That is, as readers interpret more examinations, their expectations shift. If few abnormalities were detected in the examinations they already interpreted, radiologists expect few abnormalities in the remaining examinations as well. This expectation increases their threshold for recalling a woman and decreases the time they spend reading each case. The resulting lower recall rates and shorter reading times over the course of a reading session, together with the decreased cancer detection in larger batch sizes, support the prevalence effect theory. These findings may provide a rationale for combining the increasing and decreasing density orders, with the densest cases (and therefore probably the most difficult ones) in the middle of a reading session, when perhaps the radiologists are at their peak performance and the balance in recall and cancer detection is optimal. However, more research in screening settings is needed to determine if this, or another reading order, is most effective.

In addition to breast density metrics, AI-generated PoM scores could also be used for the ordering of screening examinations. As explained by Evans et al., cancer prevalence affects radiologists' performance [35]. Grouping examinations with high AI PoM scores will result in subsets of examinations with an increased expected cancer prevalence and may thus improve radiologists' sensitivity for the examinations where it matters most. In contrast, clustering examinations with lower AI PoM scores may help radiologists improve their specificity and reduce reading time. Concerns about vigilance decrement might motivate ordering from high to low AI PoM scores, placing most cancers at the beginning of a reading session, before any decline in reader vigilance occurs, and leaving examinations with low AI PoM scores for later

in the batch, when recall rates decline. However, to the best of my knowledge, the effect of ordering examinations based on AI-generated PoM scores has not yet been investigated.

Nevertheless, the likely future implementation of AI offers a straightforward way to use breast density metrics or PoM scores for the ordering of screening examinations. Large, controlled studies in screening prevalence settings are needed to further investigate the sequential effects during screen-reading and to help identify the optimal reading order that can help radiologists during mammography interpretation.

7.2.3 My perspective on how the reader should interpret

Taken together, how screening examinations should be interpreted depends on who is interpreting them. The way an AI system interprets screening examinations cannot be altered once it has been trained, but its operating point can be set. Extensive testing is therefore essential to determine the optimal threshold and ensure performance comparable to or superior to that of radiologists. The choice of this threshold should depend on its intended mode of implementation (e.g., triage, standalone reader, decision support) and the screening setting where it will be used, since population characteristics and screening practices vary across the world. Additionally, I think AI certainty metrics should be incorporated to determine, on an examination-by-examination basis, whether the AI results should be used, either standalone and/or for decision support, or should be ignored due to its certainty being low. Focusing on radiologists as readers, preparing this thesis taught me that AI decision support and case ordering by increasing breast density can help radiologists improve their screening performance. Findings from this thesis also showed that radiologists' screening performance varies considerably. Given that no optimal pairing strategy was identified and AI is expected to play an increasingly prominent role in screening, I think future studies should examine whether adapting the AI's operating point based on the performance of the paired radiologist can enhance the combined effectiveness of AI and human interpretation.

7.3 Conclusion

This thesis explored immediate and accessible opportunities to improve the interpretation of screening mammography. Throughout the preceding chapters, we learned that optimizing who reads which examinations, in what order, and with which AI information could improve screening outcomes. These findings show that small adjustments in the reading workflow can lead to meaningful outcome improvements, demonstrating that *real impact is within reach*. While integrating AI may not be considered as immediate as the other strategies, its implementation is expected to become increasingly feasible, with the introduction in the Dutch Screening Program scheduled for 2028. To fully evaluate the potential of the interpretation strategies investigated in this thesis, prospective studies should be performed within actual screening practice, where radiologists' decisions influence examination outcomes, and reading batches are of the expected size and prevalence. I hope that aiREAD II, a proposed randomized controlled trial building on the work described in this thesis, may eventually translate the interpretation strategies into screening practice, providing valuable insights into the real-world impact and potentially contributing to improved outcomes for women undergoing breast cancer screening.

Bibliography

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229–63. doi:10.3322/caac.21834
2. NKR. NKR cijfers [Internet]. 2025 Jan [cited 2025 Mar 22]. Available from: https://nkr-cijfers.iknl.nl/
3. Rijksinstituut voor Volksgezondheid en Milieu. Factsheet 2024 Bevolkingsonderzoek Borstkanker [Internet]. 2024 Dec [cited 2025 Mar 22]. Available from: https://www.rivm.nl/documenten/factsheet-bevolkingsonderzoek-borstkanker
4. van der Waal D, Verbeek ALM, den Heeten GJ, Ripping TM, Tjan-Heijnen VCG, Broeders MJM. Breast cancer diagnosis and death in the Netherlands: a changing burden. Eur J Public Health. 2015;25:320–4. doi:10.1093/eurpub/cku088
5. Zielonke N, Gini A, Jansen EEL, Anttila A, Segnan N, Ponti A, et al. Evidence for reducing cancer-specific mortality due to screening for breast cancer in Europe: A systematic review. Eur J Cancer. 2020;127:191–206. doi:10.1016/j.ejca.2019.12.010
6. Erasmus University Medical Center & RIVM. Landelijke monitor bevolkingsonderzoek borstkanker 2024. 2025 Nov. p. 1–11.
7. PDQ® Screening and Prevention Editorial Board, National Cancer Institute. PDQ Breast Cancer Screening. [Internet]. 2025 Apr [cited 2026 Jan 3]. Available from: https://www.cancer.gov/types/breast/hp/breast-screening-pdq
8. Bassett LW, Kim CH. Breast Imaging: Mammography and Ultrasonography. Magn Reson Imaging Clin N Am. 2001;9:251–71. doi:10.1016/S1064-9689(21)00072-6
9. 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;72:214–25. doi:10.1016/j.semcancer.2020.06.002
10. Sickles EA, D'Orsi CJ, Bassett LW, Appleton CM, Berg WA, Burnside ES. ACR BI-RADS® mammography. Am Coll Radiol. 2013.
11. Marcon M, Fuchsjäger MH, Clauser P, Mann RM. ESR Essentials: screening for breast cancer - general recommendations by EUSOBI. Eur Radiol. 2024;34:6348–57. doi:10.1007/s00330-024-10740-5
12. Ren W, Chen M, Qiao Y, Zhao F. Global guidelines for breast cancer screening: A systematic review. The Breast. 2022;64:85–99. doi:10.1016/j.breast.2022.04.003
13. National Health Service. When you'll be invited for breast screening and who should go [Internet]. 2024 Oct [cited 2025 Mar 28]. Available from: https://www.nhs.uk/conditions/breast-screening-mammogram/when-youll-be-invited-and-who-should-go/
14. Hofvind S, Bennett RL, Brisson J, Lee W, Pelletier E, Flugelman A, et al. Audit feedback on reading performance of screening mammograms: An international comparison. J Med Screen. 2016;23:150–9. doi:10.1177/0969141315610790
15. European Commission. European guidelines on breast cancer screening and diagnosis - Screening ages and frequencies [Internet]. 2024 Jul. Available from: https://cancer-screening-and-care.jrc.ec.europa.eu/en/ecibc/european-breast-cancer-guidelines
16. Nickson C, Velentzis LS, Brennan P, Mann GB, Houssami N. Improving breast cancer screening in Australia: a public health perspective. Public Health Res Pract. 2019;29:2921911. doi:10.17061/phrp2921911
17. Oeffinger KC, Fontham ETH, Etzioni R, Herzig A, Michaelson JS, Shih YCT, et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA. 2015;314:1599. doi:10.1001/jama.2015.12783

18. Rijksinstituut voor Volksgezondheid en Milieu. Tijdelijke verlenging uitnodigingsinterval gemiddeld 29 maanden [Internet]. 2023 Jan [cited 2025 Mar 23]. Available from: https://www.rivm.nl/bevolkingsonderzoek-borstkanker/mammografie/later-uitgenodigd

19. Rijksinstituut voor Volksgezondheid en Milieu. Beoordeling borstfoto's [Internet]. 2025 Jun [cited 2025 Mar 23]. Available from: https://www.rivm.nl/bevolkingsonderzoek-borstkanker/mammografie/beoordeling-borstfotos

20. Rijksinstituut voor Volksgezondheid en Milieu. Uitslag mammografie [Internet]. 2023 Aug [cited 2025 Mar 23]. Available from: https://www.rivm.nl/bevolkingsonderzoek-borstkanker/uitslag

21. Yankaskas BC, Klabunde CN, Ancelle-Park R, Renner G, Wang H, Fracheboud J, et al. International comparison of performance measures for screening mammography: can it be done? J Med Screen. 2004;11:187–93. doi:10.1258/0969141042467430

22. Breast Cancer Surveillance Consortium. Performance measures for 2,301,766 Screening Mammography Examinations from 2011 - 2018 [Internet]. 2023 Jun [cited 2025 Apr 7]. Available from: https://www.bcsc-research.org/index.php/statistics/screening-performance-benchmarks/performance-measures

23. Netherlands comprehensive cancer organisation. National monitoring of the breast cancer screening programme in the Netherlands 2019 [Internet]. 2021 Sep [cited 2025 Apr 5]. Available from: https://iknl.nl/en/screening

24. JDM Otten, Harry de Koning, Nicolien van Ravesteyn, Lindy Kregting, Ruud M. Pijnappel, ALM Verbeek, AE de Bruijn, MJM Broeders. Landelijk Evaluatie Team voor bevolkingsonderzoek naar Borstkanker. Rapport I Uitkomstmaten bij borstkankerscreening – kwaliteit van leven [Internet]. 2020 Apr [cited 2025 Mar 29]. Available from: https://pure.eur.nl/en/publications/landelijk-evaluatie-team-voor-bevolkingsonderzoek-naar-borstkanke-2

25. van Ravesteyn NT, Kregting LM, Heijnsdijk EAM, de Bruijn AE, de Koning HJ, Broeders MJM, Otten JDM, Verbeek ALM, Pijnappel RM. Landelijke evaluatie van bevolkingsonderzoek naar borstkanker in Nederland, LETB XV [Internet]. 2023 Mar [cited 2025 Mar 29]. Available from: https://www.rivm.nl/documenten/letb-resultaten-bevolkingsonderzoek-borstkanker-2023-rapport-xv

26. Canelo-Aybar C, Ferreira DS, Ballesteros M, Posso M, Montero N, Solà I, et al. Benefits and harms of breast cancer mammography screening for women at average risk of breast cancer: A systematic review for the European Commission Initiative on Breast Cancer. J Med Screen. 2021;28:389–404. doi:10.1177/0969141321993866

27. de Munck L, Fracheboud J, de Bock GH, den Heeten GJ, Siesling S, Broeders MJM. Is the incidence of advanced-stage breast cancer affected by whether women attend a steady-state screening program? Int J Cancer. 2018;143:842–50. doi:10.1002/ijc.31388

28. Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. Npj Breast Cancer. 2017;3:1–13. doi:10.1038/s41523-017-0014-x

29. Weber RJP, van Bommel RMG, 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;158:471–83. doi:10.1007/s10549-016-3882-0

30. Hovda T, Hoff SR, Larsen M, Romundstad L, Sahlberg KK, Hofvind S. True and Missed Interval Cancer in Organized Mammographic Screening: A Retrospective Review Study of Diagnostic and Prior Screening Mammograms. Acad Radiol. 2022;29:S180–91. doi:10.1016/j.acra.2021.03.022

31. Magdy O, Magdy M, Hussein D, Fahim M, Faker E, Kamal M. Interval breast cancer: radiological surveillance in screening Egyptian population. Egypt J Radiol Nucl Med. 2024;55. doi:10.1186/s43055-024-01193-3

32. Klompenhouwer EG, Duijm LEM, 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;24:1097–104. doi:10.1007/s00330-014-3102-4

33. Ekpo EU, Alakhras M, Brennan P. Errors in Mammography Cannot be Solved Through Technology Alone. Asian Pac J Cancer Prev. 2018;19:291–301. doi:10.22034/APJCP.2018.19.2.291

34. Michalopoulou E, Darker I, Iotti V, Slonim E, de Koning HJ, Souza RA, et al. Breast imaging readers' performance in the PERFORMS test-set based assessment scheme within the MyPeBS international randomised study. Eur J Radiol. 2025;183:111938. doi:10.1016/j.ejrad.2025.111938

35. Evans KK, Birdwell RL, Wolfe JM. If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening. PLoS ONE. 2013;8:e64366. doi:10.1371/journal.pone.0064366

36. Wallis MG. How do we manage overdiagnosis/overtreatment in breast screening? Clin Radiol. 2018;73:372–80. doi:10.1016/j.crad.2017.09.016

37. Puliti D, Duffy SW, Miccinesi G, de Koning H, Lynge E, Zappa M, et al. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. J Med Screen. 2012;19 Suppl 1:42–56. doi:10.1258/jms.2012.012082

38. Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, et al. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. J Med Screen. 2012;19 Suppl 1:57–66. doi:10.1258/jms.2012.012083

39. Voogd AC, Molnar Z, Nederend J, Schipper RJ, Strobbe LJA, Duijm LEM. Predictors of re-attendance at biennial screening mammography following a false positive referral: A study among women in the south of the Netherlands. Breast. 2024;74:103702. doi:10.1016/j.breast.2024.103702

40. Long H, Brooks JM, Harvie M, Maxwell A, French DP. How do women experience a false-positive test result from breast screening? A systematic review and thematic synthesis of qualitative studies. Br J Cancer. 2019;121:351–8. doi:10.1038/s41416-019-0524-4

41. Larsen M, Moshina N, Holen ÅS, Bergan MB, Hofvind S. Re-attendance to mammographic screening after a false positive screening result. J Med Screen. 2025;9691413251329671. doi:10.1177/09691413251329671

42. Mao X, He W, Humphreys K, Eriksson M, Holowko N, Yang H, et al. Breast Cancer Incidence After a False-Positive Mammography Result. JAMA Oncol. 2024;10:63–70. doi:10.1001/jamaoncol.2023.4519

43. Román M, Castells X, Hofvind S, von Euler-Chelpin M. Risk of breast cancer after false-positive results in mammographic screening. Cancer Med. 2016;5:1298–306. doi:10.1002/cam4.646

44. Taylor-Phillips S, Stinton C. Double reading in breast cancer screening: considerations for policy-making. Br J Radiol. 2020;93:20190610. doi:10.1259/bjr.20190610

45. Taylor P, Potts HWW. 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;44:798–807. doi:10.1016/j.ejca.2008.02.016

46. Hautus MJ, Macmillan NA, Creelman CD. Detection Theory: A User's Guide. 3rd ed. New York: Routledge; 2021. 452 p. doi:10.4324/9781003203636

47. Camilla Flåt Aglen, Elin Wølner Bjørnson, Anne Kathrin Ertzaas, Cecilie Løbak Hestman, Solveig Hofvind, Åsne Sørlien Holen, et al. BreastScreen Norway: 25 years of organized mammographic screening [Internet]. 2022 Jun. Available from: https://www.kreftregisteret.no/globalassets/mammografiprogrammet/rapporter-og-publikasjoner/2022-25-arsrapport_webversjon.pdf

48. 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;287:749–57. doi:10.1148/radiol.2018171010

49. Euler-Chelpin M von, Lillholm M, Napolitano G, Vejborg I, Nielsen M, Lynge E. Screening mammography: benefit of double reading by breast density. Breast Cancer Res Treat. 2018;171:767–76. doi:10.1007/s10549-018-4864-1

50. 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;119:503–7. doi:10.1038/s41416-018-0195-6

51. Geertse TD, Setz-Pels W, van der Waal D, Nederend J, Korte B, Tetteroo E, et al. Added Value of Prereading Screening Mammograms for Breast Cancer by Radiologic Technologists on Early Screening Outcomes. Radiology. 2022;302:276–83. doi:10.1148/radiol.2021210746

52. European Commission. European guidelines on breast cancer screening and diagnosis - Organising screening programs [Internet]. 2017 Nov. Available from: https://cancer-screening-and-care.jrc.ec.europa.eu/en/ecibc/european-breast-cancer-guidelines?topic=62&usertype=60&filter_1=68&filter_2=70&updatef2=0

53. Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition--summary document. Ann Oncol. 2008;19:614–22. doi:10.1093/annonc/mdm481

54. Brennan PC, Ganesan A, Eckstein MP, Ekpo E, 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;26:717–23. doi:10.1016/j.acra.2018.06.017

55. 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;45:4844–56. doi:10.1002/mp.13161

56. Kurvers RHJM, Herzog SM, Hertwig R, Krause J, Carney PA, Bogart A, et al. Boosting medical diagnostics by pooling independent judgments. Proc Natl Acad Sci U S A. 2016;113:8777–82. doi:10.1073/pnas.1601827113

57. Cohen EO, Lesslie M, Weaver O, Phalak K, Tso H, Perry R, et al. Batch Reading and Interrupted Interpretation of Digital Screening Mammograms Without and With Tomosynthesis. J Am Coll Radiol. 2021;18:280–93. doi:10.1016/j.jacr.2020.07.033

58. Burnside ES, Park JM, Fine JP, Sisney GA. The Use of Batch Reading to Improve the Performance of Screening Mammography. Am J Roentgenol. 2005;185:790–6. doi:10.2214/ajr.185.3.01850790

59. Abbey CK, Webster MA, Geertse T, Waal D van der, Tetteroo E, Pijnappel R, et al. Sequential reading effects in Dutch screening mammography. In: Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment [Internet]. SPIE; 2020 [cited 2025 Aug 17]. p. 66–70. Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11316/113160G/Sequential-reading-effects-in-Dutch-screening-mammography/10.1117/12.2549320.full doi:10.1117/12.2549320

60. Backmann HA, Larsen M, Danielsen AS, Hofvind S. Does it matter for the radiologists' performance whether they read short or long batches in organized mammographic screening? Eur Radiol. 2021;31:9548–55. doi:10.1007/s00330-021-08010-9

61. Abbey CK, Bandos AI, Parthasarathy MK, Webster MA, Zuley ML. Changes in Reader Performance During Sequential Reading of Breast Cancer Screening Digital Breast Tomosynthesis Examinations. Radiology. 2024;313:e232885. doi:10.1148/radiol.232885

62. 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;315:1956–65. doi:10.1001/jama.2016.5257

63. Iima M, Satake H. Sequential Reading Effects in Digital Breast Tomosynthesis: Improving False-Positive Rates Without Compromising Cancer Detection. Radiology. 2024;313:e242642. doi:10.1148/radiol.242642

64. Webster MA. Visual Adaptation. Annu Rev Vis Sci. 2015;1:547–67. doi:10.1146/annurev-vision-082114-035509

65. Philip Nelson. From photon to neuron: Light, imaging, vision. Princeton University Press; 2017.

66. Kompaniez E, Abbey CK, Boone JM, Webster MA. Adaptation Aftereffects in the Perception of Radiological Images. PLoS ONE. 2013;8:e76175. doi:10.1371/journal.pone.0076175

67. Kompaniez-Dunigan E, Abbey CK, Boone JM, Webster MA. Adaptation and visual search in mammographic images. Atten Percept Psychophys. 2015;77:1081–7. doi:10.3758/s13414-015-0841-5

68. Kompaniez-Dunigan E, Abbey CK, Boone JM, Webster MA. Visual adaptation and the amplitude spectra of radiological images. Cogn Res Princ Implic. 2018;3:3. doi:10.1186/s41235-018-0089-4

69. Parthasarathy MK, Zuley ML, Bandos AI, Abbey CK, Webster MA. Visual adaptation to medical images: a comparison of digital mammography and tomosynthesis. J Med Imaging Bellingham Wash. 2023;10:S11909. doi:10.1117/1.JMI.10.S1.S11909

70. Giger ML, Chan HP, Boone J. Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. Med Phys. 2008;35:5799–820. doi:10.1118/1.3013555

71. Gilbert FJ, Astley SM, Gillan MGC, Agbaje OF, Wallis MG, James J, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med. 2008;359:1675–84. doi:10.1056/NEJMoa0803545

72. James JJ, Gilbert FJ, Wallis MG, Gillan MGC, Astley SM, Boggis CRM, et al. Mammographic Features of Breast Cancers at Single Reading with Computer-aided Detection and at Double Reading in a Large Multicenter Prospective Trial of Computer-aided Detection: CADET II. Radiology. 2010;256:379–86. doi:10.1148/radiol.10091899

73. Gromet M. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. AJR Am J Roentgenol. 2008;190:854–9. doi:10.2214/AJR.07.2812

74. Bargalló X, Santamaría G, Del Amo M, Arguis P, Ríos J, Grau J, et al. Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. Eur J Radiol. 2014;83:2019–23. doi:10.1016/j.ejrad.2014.08.010

75. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med. 2007;356:1399–409. doi:10.1056/NEJMoa066099

76. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL. Diagnostic Accuracy of Digital Screening Mammography with and without Computer-aided Detection. JAMA Intern Med. 2015;175:1828–37. doi:10.1001/jamainternmed.2015.5231

77. Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D'Orsi C, et al. Effectiveness of computer-aided detection in community mammography practice. J Natl Cancer Inst. 2011;103:1152–61. doi:10.1093/jnci/djr206

78. Mahoney MC, Meganathan K. False Positive Marks on Unsuspicious Screening Mammography with Computer-Aided Detection. J Digit Imaging. 2011;24:772–7. doi:10.1007/s10278-011-9389-7

79. Tchou PM, Haygood TM, Atkinson EN, Stephens TW, Davis PL, Arribas EM, et al. Interpretation time of computer-aided detection at screening mammography. Radiology. 2010;257:40–6. doi:10.1148/radiol.10092170

80. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111:916–22. doi:10.1093/jnci/djy222

81. Yoon JH, Strand F, Baltzer PAT, Conant EF, Gilbert FJ, Lehman CD, et al. Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis. Radiology. 2023;307:e222639. doi:10.1148/radiol.222639

82. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology. 2019;290:305–14. doi:10.1148/radiol.2018181371

83. Lee JH, Kim KH, Lee EH, Ahn JS, Ryu JK, Park YM, et al. Improving the Performance of Radiologists Using Artificial Intelligence-Based Detection Support Software for Mammography: A Multi-Reader Study. Korean J Radiol. 2022;23:505–16. doi:10.3348/kjr.2021.0476

84. Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool. Radiol Artif Intell. 2020;2:e190208. doi:10.1148/ryai.2020190208

85. Kim HE, Kim HH, Han BK, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health. 2020;2:e138–48. doi:10.1016/S2589-7500(20)30003-0

86. Eisemann N, Bunk S, Mukama T, Baltus H, Elsner SA, Gomille T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. Nat Med. 2025;1–8. doi:10.1038/s41591-024-03408-6

87. Chang YW, Ryu JK, An JK, Choi N, Park YM, Ko KH, et al. Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study. Nat Commun. 2025;16:2248. doi:10.1038/s41467-025-57469-3

88. Elías-Cabot E, Romero-Martín S, Raya-Povedano JL, Brehl AK, Álvarez-Benito M. Impact of real-life use of artificial intelligence as support for human reading in a population-based breast cancer screening program with mammography and tomosynthesis. Eur Radiol. 2024;34:3958–66. doi:10.1007/s00330-023-10426-4

89. Gandomkar Z, Mello-Thoms C. Visual search in breast imaging. Br J Radiol. 2019;92:20190057. doi:10.1259/bjr.20190057

90. Drew T, Cunningham C, Wolfe JM. When and Why Might a Computer-aided Detection (CAD) System Interfere with Visual Search? An Eye-tracking Study. Acad Radiol. 2012;19:1260–7. doi:10.1016/j.acra.2012.05.013

91. Klein DS, Karmakar S, Jonnalagadda A, Abbey CK, Eckstein MP. Greater benefits of deep learning-based computer-aided detection systems for finding small signals in 3D volumetric medical images. J Med Imaging. 2024;11:045501. doi:10.1117/1.JMI.11.4.045501

92. Al-Bazzaz H, Janicijevic M, Strand F. Reader bias in breast cancer screening related to cancer prevalence and artificial intelligence decision support—a reader study. Eur Radiol. 2024;34:5415–24. doi:10.1007/s00330-023-10514-5

93. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. Radiology. 2023;307:e222176. doi:10.1148/radiol.222176

94. Rezazade Mehrizi MH, Mol F, Peter M, Ranschaert E, Dos Santos DP, Shahidi R, et al. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. Sci Rep. 2023;13:9230. doi:10.1038/s41598-023-36435-3

95. Baltzer PAT. Automation Bias in Breast AI. Radiology. 2023;307:e230770. doi:10.1148/radiol.230770

96. 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. doi:10.1258/jms.2012.012078

97. 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;314:1615–34. doi:10.1001/jama.2015.13183

98. 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;71:209–49. doi:10.3322/caac.21660

99. Broeders MJM, Onland-Moret NC, Rijken HJTM, Hendriks JHCL, Verbeek ALM, 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;39:1770–5. doi:10.1016/s0959-8049(03)00311-3

100. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. Radiographics. 2003;23:881–95. doi:10.1148/rg.234025083

101. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. Ann Intern Med. 2007;146:502–10. doi:10.7326/0003-4819-146-7-200704030-00006

102. 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;180:1461–7. doi:10.2214/ajr.180.5.1801461

103. European Commission. European Commission Initiatives on Breast and Colorectal Cancer - Use of double reading in mammography screening. [Internet]. 2019. Available from: https://healthcare-quality.jrc.ec.europa.eu/ecibc/european-breast-cancer-guidelines/organisation-of-screening-programme/double-reading-in-mammography-screening

104. Marinovich ML, Wylie E, Lotter W, Lund H, Waddell A, Madeley C, et al. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. EBioMedicine. 2023;90:104498. doi:10.1016/j.ebiom.2023.104498

105. 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 [Internet]. medRxiv; 2022. p. 2021.02.26.21252537. Available from: https://www.medrxiv.org/content/10.1101/2021.02.26.21252537v2 doi:10.1101/2021.02.26.21252537

106. 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;32:8238–46. doi:10.1007/s00330-022-08909-x

107. 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;6:1581–8. doi:10.1001/jamaoncol.2020.3321

108. 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). J Digit Imaging. 2020;33:408–13. doi:10.1007/s10278-019-00278-0

109. 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. doi:10.1186/1745-6215-15-17

110. 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;297:33–9. doi:10.1148/radiol.2020192212

111. 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:e468–74. doi:10.1016/S2589-7500(20)30185-0

112. 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;294:265–72. doi:10.1148/radiol.2019190872

113. 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;32:602–12. doi:10.1007/s00330-021-07965-z

114. 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:e89418. doi:10.1371/journal.pone.0089418

115. Paci E, Broeders M, Hofvind S, Puliti D, Duffy SW, EUROSCREEN Working Group. European breast cancer service screening outcomes: a first balance sheet of the benefits and harms. Cancer Epidemiol Biomarkers Prev. 2014;23:1159–63. doi:10.1158/1055-9965.EPI-13-0320

116. Lauby-Secretan B, Scoccianti C, Loomis D, Benbrahim-Tallaa L, Bouvard V, Bianchini F, et al. Breast-cancer screening--viewpoint of the IARC Working Group. N Engl J Med. 2015;372:2353–8. doi:10.1056/NEJMsr1504363

117. Martiniussen MA, Sagstad S, Larsen M, Larsen ASF, Hovda T, Lee CI, 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;32:5974–85. doi:10.1007/s00330-022-08711-9

118. 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:e054005. doi:10.1136/bmjopen-2021-054005

119. Gommers JJJ, Abbey CK, Strand F, Taylor-Phillips S, Jenkinson DJ, Larsen M, et al. Optimizing the Pairs of Radiologists That Double Read Screening Mammograms. Radiology. 2023;309:e222691. doi:10.1148/radiol.222691

120. Bouwman RW, Mackenzie A, van Engen RE, Broeders MJM, Young KC, Dance DR, et al. Toward image quality assessment in mammography using model observers: Detection of a calcification-like object. Med Phys. 2017;44:5726–39. doi:10.1002/mp.12532

121. He X, Park S. Model observers in medical imaging research. Theranostics. 2013;3:774–86. doi:10.7150/thno.5138

122. Gandomkar Z, Lewis SJ, Li T, Ekpo EU, Brennan PC. A machine learning model based on readers' characteristics to predict their performances in reading screening mammograms. Breast Cancer Tokyo Jpn. 2022;29:589–98. doi:10.1007/s12282-022-01335-3

123. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology. 2009;253:641–51. doi:10.1148/radiol.2533082308

124. Birchenhall C. Numerical Recipes in C: The Art of Scientific Computing. Econ J. 1994;104:725–6.

125. Agresti A, Coull BA. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. Am Stat. 1998;52:119–26. doi:10.1080/00031305.1998.10480550

126. Groenewoud JH, Otten JDM, Fracheboud J, Draisma G, van Ineveld BM, Holland R, et al. Cost-effectiveness of different reading and referral strategies in mammography screening in the Netherlands. Breast Cancer Res Treat. 2007;102:211–8. doi:10.1007/s10549-006-9319-4

127. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. Br J Cancer. 2013;108:2205–40. doi:10.1038/bjc.2013.177

128. Nyström L, Andersson I, Bjurstam N, Frisell J, Nordenskjöld B, Rutqvist LE. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. The Lancet. 2002;359:909–19. doi:10.1016/S0140-6736(02)08020-0

129. Molins E, Macià F, Ferrer F, Maristany MT, Castells X. Association between Radiologists' Experience and Accuracy in Interpreting Screening Mammograms. BMC Health Serv Res. 2008;8:91. doi:10.1186/1472-6963-8-91

130. Hoff SR, Samset JH, Abrahamsen AL, Vigeland E, Klepp O, Hofvind S. Missed and true interval and screen-detected breast cancers in a population based screening program. Acad Radiol. 2011;18:454–60. doi:10.1016/j.acra.2010.11.014

131. Hovda T, Tsuruda K, Hoff SR, Sahlberg KK, Hofvind S. Radiological review of prior screening mammograms of screen-detected breast cancer. Eur Radiol. 2021;31:2568–79. doi:10.1007/s00330-020-07130-y

132. Vijayargahavan GR, Watkins J, Tyminski M, Venkataraman S, Amornsiripanitch N, Newburg A, et al. Audit of prior screening mammograms of screen-detected cancers: Implications for the delay in breast cancer detection. Semin Ultrasound CT MR. 2023;44:62–9. doi:10.1053/j.sult.2022.12.003

133. Hofvind S, Skaane P, Vitak B, Wang H, Thoresen S, Eriksen L, et al. Influence of Review Design on Percentages of Missed Interval Breast Cancers: Retrospective Study of Interval Cancers in a Population-based Screening Program. Radiology. 2005;237:437–43. doi:10.1148/radiol.2372041174

134. Brennan PC, Gandomkar Z, Ekpo EU, Tapia K, Trieu PD, Lewis SJ, et al. Radiologists can detect the 'gist'of breast cancer before any overt signs of cancer appear. Sci Rep. 2018;8:8717. doi:10.1038/s41598-018-26100-5

135. Wolfe JM, Wu CC, Li J, Suresh SB. What do experts look at and what do experts find when reading mammograms? J Med Imaging. 2021;8:045501. doi:10.1117/1.JMI.8.4.045501

136. Kundel HL, Nodine CF, Krupinski EA, Mello-Thoms C. Using Gaze-tracking Data and Mixture Distribution Analysis to Support a Holistic Model for the Detection of Cancers on Mammograms. Acad Radiol. 2008;15:881–6. doi:10.1016/j.acra.2008.01.023

137. Kohn A. Visual Adaptation: Physiology, Mechanisms, and Functional Benefits. J Neurophysiol. 2007;97:3155–64. doi:10.1152/jn.00086.2007

138. Clifford CWG, Webster MA, Stanley GB, Stocker AA, Kohn A, Sharpee TO, et al. Visual adaptation: Neural, psychological and computational aspects. Vision Res. 2007;47:3125–31. doi:10.1016/j.visres.2007.08.023

139. Webster MA. Adaptation and visual coding. J Vis. 2011;11:3. doi:10.1167/11.5.3

140. Thompson P, Burr D. Visual aftereffects. Curr Biol. 2009;19:R11–4.

141. McDermott KC, Malkoc G, Mulligan JB, Webster MA. Adaptation and visual salience. J Vis. 2010;10:17. doi:10.1167/10.13.17

142. Wissig SC, Patterson CA, Kohn A. Adaptation improves performance on a visual search task. J Vis. 2013;13:6. doi:10.1167/13.2.6

143. Olsen A. Determining the Tobii I-VT Fixation Filter's Default Values. Tobii Technol. 2012.

144. Olsen A. The Tobii I-VT Fixation Filter. Tobii Technol. 2012.

145. Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods an updated and unified approach. Acad Radiol. 2011;18:129–42. doi:10.1016/j.acra.2010.09.007

146. Obuchowski NA, Gallas BD, Hillis SL. Multi-Reader ROC studies with Split-Plot Designs: A Comparison of Statistical Methods. Acad Radiol. 2012;19:1508–17. doi:10.1016/j.acra.2012.09.012

147. Webster MA. Probing the functions of contextual modulation by adapting images rather than observers. Vision Res. 2014;104:68–79. doi:10.1016/j.visres.2014.09.003

148. Arnold DH, Saurels BW, Anderson NL, Johnston A. An observer model of tilt perception, sensitivity and confidence. Proc R Soc B Biol Sci. 288:20211276. doi:10.1098/rspb.2021.1276

149. Gallagher RM, Suddendorf T, Arnold DH. Confidence as a diagnostic tool for perceptual aftereffects. Sci Rep. 2019;9:7124. doi:10.1038/s41598-019-43170-1

150. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The "Laboratory" Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations1. Radiology. 2008;249:47–53. doi:10.1148/radiol.2491072025

151. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In. PMLR; 2020. p. 1597–607.

152. Verboom SD, Caballo M, Peters J, Gommers J, van den Oever D, Broeders MJM, et al. Deep learning-based breast region segmentation in raw and processed digital mammograms: generalization across views and vendors. J Med Imaging Bellingham Wash. 2024;11:014001. doi:10.1117/1.JMI.11.1.014001

153. Olsen A, Matos R. Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In. 2012. p. 317–20.

154. Komogortsev OV, Gobert DV, Jayarathna S, Koh DH, Gowda S. Standardization of automated analyses of oculomotor fixation and saccadic behaviors. IEEE Trans Biomed Eng. 2010;57. doi:10.1109/TBME.2010.2057429

155. Dorfman DD, Berbaum KS, Metz CE. Receiver Operating Characteristic Rating Analysis: Generalization to the Population of Readers and Patients with the Jackknife Method. Invest Radiol. 1992;27:723.

156. Obuchowski NA, Beiden SV, Berbaum KS, Hillis SL, Ishwaran H, Song HH, et al. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. Acad Radiol. 2004;11:980–95. doi:10.1016/j.acra.2004.04.014

157. Gallas BD. One-Shot Estimate of MRMC Variance: AUC. Acad Radiol. 2006;13:353–62. doi:10.1016/j.acra.2005.11.030

158. Gallas BD, Bandos A, Samuelson FW, Wagner RF. A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators. Commun Stat - Theory Methods. 2009;38:2586–603. doi:10.1080/03610920802610084

159. Gallas BD, Brown DG. Reader studies for validation of CAD systems. Neural Netw. 2008;Advances in Neural Networks Research: IJCNN '0721:387–97. doi:10.1016/j.neunet.2007.12.013

160. Hickman SE, Woitek R, Le EPV, Im YR, Mouritsen Luxhøj C, Aviles-Rivero AI, et al. Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis. Radiology. 2022;302:88–104. doi:10.1148/radiol.2021210391

161. Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, et al. 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. 2023;24:936–44. doi:10.1016/S1470-2045(23)00298-X

162. Wing P, Langelier MH. Workforce shortages in breast imaging: impact on mammography utilization. AJR Am J Roentgenol. 2009;192:370–8. doi:10.2214/AJR.08.1665

163. Gommers JJJ, Verboom SD, Duvivier KM, van Rooden JK, van Raamt AF, Houwers JB, et al. Enhancing Radiologist Reading Performance by Ordering Screening Mammograms Based on Characteristics That Promote Visual Adaptation. Radiology. 2024;313:e240237. doi:10.1148/radiol.240237

164. Elhakim MT, Stougaard SW, Graumann O, Nielsen M, Lång K, Gerke O, et al. Breast cancer detection accuracy of AI in an entire screening population: a retrospective, multicentre study. Cancer Imaging. 2023;23:127. doi:10.1186/s40644-023-00643-x

165. Lauritzen AD, Rodríguez-Ruiz A, Von Euler-Chelpin MC, Lynge E, Vejborg I, Nielsen M, et al. An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload. Radiology. 2022;304:41–9. doi:10.1148/radiol.210948

166. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. Acad Radiol. 1996;3:137–44. doi:10.1016/s1076-6332(05)80381-2

167. Kundel HL, Nodine CF, Carmody D. Visual Scanning, Pattern Recognition and Decision-making in Pulmonary Nodule Detection. Invest Radiol. 1978;13:175.

168. Krupinski EA. Medical image perception: evaluating the role of experience. In: Human Vision and Electronic Imaging V. SPIE; 2000. p. 281–9. doi:10.1117/12.387164

169. Tromans C, Chan A, Johnston L, Highnam R. ECR 2016 EPOS [Text] [Internet]. European Congress of Radiology - ECR 2016; 2016 [cited 2024 Sep 9]. Optimizing VolparaDensity's volumetric breast density thresholds to align with the American College of Radiology BI-RADS 5th edition updates. Available from: https://epos.myesr.org/poster/esr/ecr2016/C-0793

170. Pinto MC, Rodriguez-Ruiz A, Pedersen K, Hofvind S, Wicklein J, Kappler S, et al. Impact of Artificial Intelligence Decision Support Using Deep Learning on Breast Cancer Screening Interpretation with Single-View Wide-Angle Digital Breast Tomosynthesis. Radiology. 2021;300:529–36. doi:10.1148/radiol.2021204432

171. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Acad Radiol. 1996;3:1000–6. doi:10.1016/S1076-6332(96)80032-8

172. Saw SN, Ng KH. Current challenges of implementing artificial intelligence in medical imaging. Phys Med. 2022;100:12–7. doi:10.1016/j.ejmp.2022.06.003

173. Chun MM, Jiang Y. Contextual cueing: implicit learning and memory of visual context guides spatial attention. Cognit Psychol. 1998;36:28–71. doi:10.1006/cogp.1998.0681

174. Gur D, Rockette HE, Armfield DR, Blachar A, Bogan JK, Brancatelli G, et al. Prevalence Effect in a Laboratory Environment. Radiology. 2003;228:10–4. doi:10.1148/radiol.2281020709

175. Berbaum KS, Franken EA, Dorfman DD, Rooholamini SA, Kathol MH, Barloon TJ, et al. Satisfaction of search in diagnostic radiology. Invest Radiol. 1990;25:133–40. doi:10.1097/00004424-199002000-00006

176. Gommers JJJ, Verboom SD, Duvivier KM, van Rooden CJ, van Raamt AF, Houwers JB, et al. Influence of AI Decision Support on Radiologists' Performance and Visual Search in Screening Mammography. Radiology. 2025;316:e243688. doi:10.1148/radiol.243688

177. van Winkel SL, Rodríguez-Ruiz A, Appelman L, Gubern-Mérida A, Karssemeijer N, Teuwen J, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. Eur Radiol. 2021;31:8682–91. doi:10.1007/s00330-021-07992-w

178. Conant EF, Toledano AY, Periaswamy S, Fotin SV, Go J, Boatsman JE, et al. Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. Radiol Artif Intell. 2019;1:e180096. doi:10.1148/ryai.2019180096

179. Broeders MJ. PRISMA: Informatie voor professionals [Internet]. Available from: https://www.prisma-studie.nl/over-de-studie/informatie-voor-professionals/

180. Smith BJ, Hillis SL, Pesce LL. MRMCaov: Multi-Reader Multi-Case Analysis of Variance [Internet]. 2023. (R package version 0.3.0). Available from: https://github.com/brian-j-smith/MRMCaov

181. ScreenPoint Medical. 510(k) Summary Transpara Density [Internet]. 2023 [cited 2025 Jul 25]. Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf23/K232096.pdf

182. Gandomkar Z, Tay K, Brennan PC, Mello-Thoms C. Recurrence quantification analysis of radiologists' scanpaths when interpreting mammograms. Med Phys. 2018;45:3052–62. doi:10.1002/mp.12935

183. Dembrower K, Crippa A, Colón E, Eklund M, Strand F. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health. 2023;5:e703–11. doi:10.1016/S2589-7500(23)00153-X

184. Ng AY, Oberije CJG, Ambrózay É, Szabó E, Serfőző O, Karpati E, et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. Nat Med. 2023;29:3044–9. doi:10.1038/s41591-023-02625-9

185. Britisch Society of Breast Radiology, NHS England, Public Health England, The Royal College of Radiologists. The breast imaging and diagnostic workforce in the United Kingdom: results of a survery of NHS Breast Cancer Screening Programme units and radiology departments. London: The Royal College of Radiologists; 2016.

186. Elhakim MT, Stougaard SW, Graumann O, Nielsen M, Gerke O, Larsen LB, et al. AI-integrated Screening to Replace Double Reading of Mammograms: A Population-wide Accuracy and Feasibility Study. Radiol Artif Intell. 2024;6:e230529. doi:10.1148/ryai.230529

187. Sharma N, Ng AY, James JJ, Khara G, Ambrózay É, Austin CC, et al. Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms. BMC Cancer. 2023;23:460. doi:10.1186/s12885-023-10890-7

188. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577:89–94. doi:10.1038/s41586-019-1799-6

189. Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. Lancet Digit Health. 2022;4:e507–19. doi:10.1016/S2589-7500(22)00070-X

190. Frazer HML, Peña-Solorzano CA, Kwok CF, Elliott MS, Chen Y, Wang C, et al. Comparison of AI-integrated pathways with human-AI interaction in population mammographic screening for breast cancer. Nat Commun. 2024;15:7525. doi:10.1038/s41467-024-51725-8

191. van Winkel SL, Peters J, Janssen N, Kroes J, Loehrer EA, Gommers J, et al. AI as an independent second reader in detection of clinically relevant breast cancers within a population-based screening programme in the Netherlands: a retrospective cohort study. Lancet Digit Health. 2025;100882. doi:10.1016/j.landig.2025.100882

192. Brigitte Seradour. Diagnostic Performance of Breast Cancer Screening Second Reading Process Assisted by AI (IMA-L2) [Internet]. 2024 Apr [cited 2025 Aug 25]. Available from: https://clinicaltrials.gov/study/NCT05800132

193. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation. Radiology. 2021;300:57–65. doi:10.1148/radiol.2021203555

194. Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. Eur Radiol. 2021;31:1687–92. doi:10.1007/s00330-020-07165-1

195. Larsen M, Aglen CF, Lee CI, Hoff SR, Lund-Hanssen H, Lång K, et al. Artificial Intelligence Evaluation of 122 969 Mammography Examinations from a Population-based Screening Program. Radiology. 2022;303:502–11. doi:10.1148/radiol.212381

196. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, 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;29:4825–32. doi:10.1007/s00330-019-06186-9

197. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A Deep Learning Model to Triage Screening Mammograms: A Simulation Study. Radiology. 2019;293:38–46. doi:10.1148/radiol.2019182908

198. Fisches ZV, Ball M, Mukama T, Štih V, Payne NR, Hickman SE, et al. Strategies for integrating artificial intelligence into mammography screening programmes: a retrospective simulation analysis. Lancet Digit Health. 2024;6:e803–14. doi:10.1016/S2589-7500(24)00173-0

199. Hickman SE, Payne NR, Black RT, Huang Y, Priest AN, Hudson S, et al. Mammography Breast Cancer Screening Triage Using Deep Learning: A UK Retrospective Study. Radiology. 2023;309:e231173. doi:10.1148/radiol.231173

200. Hernström V, Josefsson V, Sartor H, Schmidt D, Larsson AM, Hofvind S, et al. Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. Lancet Digit Health. 2025;7:e175–83. doi:10.1016/S2589-7500(24)00267-X

201. Lauritzen AD, Lillholm M, Lynge E, Nielsen M, Karssemeijer N, Vejborg I, et al. Early Indicators of the Impact of Using AI in Mammography Screening for Breast Cancer. Radiology. 2024;311:e232479. doi:10.1148/radiol.232479

202. Hofvind S. Artificial Intelligence in Mammography Screening in Norway (AIMS Norway) [Internet]. 2024 May [cited 2025 Aug 26]. Available from: https://clinicaltrials.gov/study/NCT06032390

203. Carter SM, Rogers W, Win KT, Frazer H, Richards B, Houssami N. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. The Breast. 2020;49:25–32. doi:10.1016/j.breast.2019.10.001

204. Holen ÅS, Martiniussen MA, Bergan MB, Moshina N, Hovda T, Hofvind S. Women's attitudes and perspectives on the use of artificial intelligence in the assessment of screening mammograms. Eur J Radiol. 2024;175:111431. doi:10.1016/j.ejrad.2024.111431

205. Carter SM, Popic D, Marinovich ML, Carolan L, Houssami N. Women's views on using artificial intelligence in breast cancer screening: A review and qualitative study to guide breast screening services. The Breast. 2024;77:103783. doi:10.1016/j.breast.2024.103783

206. Ozcan BB, Dogan BE, Xi Y, Knippa EE. Patient Perception of Artificial Intelligence Use in Interpretation of Screening Mammograms: A Survey Study. Radiol Imaging Cancer. 2025;7:e240290. doi:10.1148/rycan.240290

207. Verboom SD, Kroes J, Pires S, Broeders MJM, Sechopoulos I. AI Should Read Mammograms Only When Confident: A Hybrid Breast Cancer Screening Reading Strategy. Radiology. 2025;316:e242594. doi:10.1148/radiol.242594

208. Baltzer PAT. When AI Knows Its Limits. Radiology. 2025;316:e252198. doi:10.1148/radiol.252198

209. Rodriguez-Ruiz A. Using prior mammograms to improve specificity of a breast AI system: a large-scale retrospective multi-site validation. ECR. 2025; Vienna.

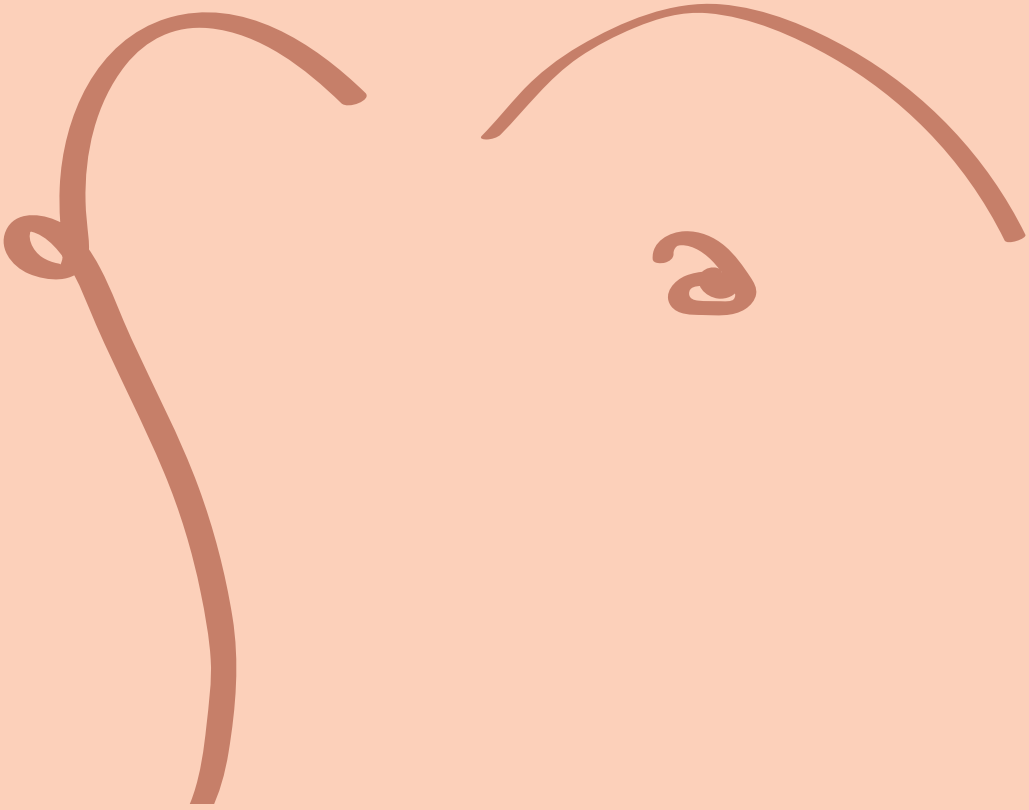
210. Akwo JD, Trieu PDY, Barron ML, Reynolds T, Lewis SJ. Access to prior screening mammograms affects the specificity but not sensitivity of radiologists' performance. Clin Radiol. 2024;79:e1549–56. doi:10.1016/j.crad.2024.09.007

211. Partridge GJW, Taib AG, Phillips P, James JJ, Satchithananda K, Sharma N, et al. Take a break: should breaks be enforced during digital breast tomosynthesis reading sessions? Eur Radiol. 2024;34:1388–98. doi:10.1007/s00330-023-10086-4

212. Elías-Cabot E. Artificial Intelligence in Breast Cancer Screening Programs (AITIC) [Internet]. 2025 Aug [cited 2025 Nov 27]. Available from: https://clinicaltrials.gov/study/NCT04949776

213. Taylor-Phillips S, Elze MC, Krupinski EA, Dennick K, Gale AG, Clarke A, et al. Retrospective review of the drop in observer detection performance over time in lesion-enriched experimental studies. J Digit Imaging. 2015;28:32–40. doi:10.1007/s10278-014-9717-9

214. Taylor-Phillips S, Jenkinson D, Stinton C, Kunar MA, Watson DG, Freeman K, et al. Fatigue and vigilance in medical experts detecting breast cancer. Proc Natl Acad Sci U S A. 2024;121:e2309576121. doi:10.1073/pnas.2309576121



Summaries

Summary – English

Breast cancer is the most commonly diagnosed cancer among women worldwide. In the Netherlands, approximately 1 in 7 women will develop breast cancer by the age of 80. Survival outcomes depend on the stage and characteristics of the cancer at the time of diagnosis, with earlier diagnosis being associated with better prognosis. Breast cancer screening aims to detect breast cancer at an early stage, allowing for more successful and less aggressive treatment.

Mammography is the most widely used imaging modality for breast cancer screening. This technique uses low-dose x rays to create two-dimensional images of the breast. Radiologists interpret these images to detect and characterize abnormalities suspicious for breast cancer and decide whether a woman should be recalled for further diagnostic workup. Screening mammography has been shown to reduce breast cancer-related mortality and improve patients' quality of life. However, some cancers are still missed, and some women are recalled despite not having breast cancer, highlighting the need for further improvements.

Limitations inherent to human interpretation of mammograms affect screening performance, contributing to both missed cancers and false alarms. This thesis investigates immediate and accessible strategies to improve mammogram interpretation within existing screening programs, aiming for *real impact within reach*. Specifically, this thesis focuses on optimizing the individual- and double-reading process and exploring the potential impact of artificial intelligence (AI) in improving screening outcomes.

Double reading in breast cancer screening is recommended by European guidelines, reflecting the principle that "*two heads are better than one*", with each mammogram interpreted independently by two radiologists. In **Chapters 2 and 3** of this thesis, we investigated whether strategically pairing radiologists, based on their individual performance characteristics (e.g., cancer detection and false-alarm rates), could improve the overall accuracy of double reading. Specifically, we evaluated whether pairing radiologists with either similar or opposite performance characteristics could improve group outcomes. In **Chapter 2**, exploratory analyses showed that different pairing strategies did not result in significant differences in overall grouped screening performance compared with random pairings. This may suggest that individual performance metrics are not the main factor to optimize double reading, or that the effects of performance-based pairing were too small to detect with the available data. To further investigate this, **Chapter 3** describes

a method where radiologist assessments were statistically modeled, providing greater power to explore different pairing strategies beyond those observed in practice. Simulations showed that pairing radiologists with similar false-positive rates could reduce false alarms without significantly affecting cancer detection overall. However, no pairing strategy consistently outperformed random pairing in both improving cancer detection and reducing false alarms. Therefore, while differences in group outcomes between pairing strategies were observed, they are not yet strong enough to advise on optimal pairing strategies. Further research is needed to evaluate the impact of different pairing strategies, taking into account downstream steps in the screening workflow (e.g., consensus meetings).

To date, no guidelines exist on the order in which mammograms should be read. Visual adaptation, the process by which the brain and eyes adjust to the visual environment, shapes perception based on previously viewed images. In mammography, this implies that the order in which images are interpreted could potentially affect diagnostic performance. **Chapter 4** describes a reader study in which we explored the concept of visual adaptation. In this study, radiologists read the same examinations in three different orders: random, by increasing breast density, and by an AI-based similarity metric (i.e., grouping similar-looking mammograms). Results showed that reading in the increasing breast density order improved diagnostic accuracy and reduced reading times. These findings suggest that ordering screening examinations by breast density may help radiologists during interpretation and improve screening performance.

The growing availability of AI for breast cancer detection presents new opportunities. **Chapter 5** describes a reader study in which we investigated the effect of AI decision support on radiologists' performance and visual search behavior. Radiologists showed improved diagnostic accuracy when reading with AI support compared to reading without it. We also observed changes in search patterns, with radiologists examining the images less comprehensively and spending more time looking at diagnostically relevant areas when assisted by AI. This suggests that AI can help radiologists prioritize their attention, making the search more efficient. However, reliance on AI also raises concerns about automation bias if the AI makes mistakes. **Chapter 6** describes this effect, showing that radiologists' screening performance improved when AI decision support provided accurate results, but suffered when the AI outputs were suboptimal. Together, these findings suggest that AI decision support can improve radiologist performance, but only when reliable, highlighting the importance of incorporating AI-certainty metrics.

The general discussion in **Chapter 7** reflects on the findings of this thesis and outlines perspectives on future directions in screening mammography. It describes a proposal for who should interpret which examinations and how this interpretation should be optimized.

Overall, this thesis shows that there are immediate and accessible opportunities to improve the interpretation of screening mammography, both by helping radiologists during interpretation and by incorporating AI. I hope the interpretation strategies investigated in this thesis will be tested in prospective studies within actual screening practice, providing insights into their impact and potentially improving outcomes for women undergoing breast cancer screening.

Samenvatting – Nederlands

Borstkanker is wereldwijd de meest voorkomende vorm van kanker bij vrouwen. In Nederland krijgt ongeveer 1 op de 7 vrouwen borstkanker voor haar 80ste levensjaar. Overlevingsuitkomsten zijn afhankelijk van het stadium en de eigenschappen van de kanker op het moment van diagnose. Als borstkanker vroeg wordt gevonden, is de kans op genezing groter. Het bevolkingsonderzoek borstkanker is bedoeld om de ziekte zo vroeg mogelijk op te sporen. Daardoor kan de behandeling minder zwaar zijn en is de kanker vaak beter te bestrijden.

In het bevolkingsonderzoek borstkanker wordt een mammografie gemaakt. Dit is een röntgenfoto van de borsten, gemaakt met een lage dosis straling. Twee radiologen bekijken de mammografieën en letten daarbij op afwijkingen die kunnen wijzen op borstkanker. Op basis daarvan wordt besloten of verder onderzoek nodig is in het ziekenhuis.

Het bevolkingsonderzoek met mammografie heeft ervoor gezorgd dat minder vrouwen overlijden aan borstkanker en dat de kwaliteit van leven van borstkankerpatiënten is verbeterd. Toch worden niet alle borstkankers ontdekt. Ook komt het voor dat vrouwen worden doorverwezen voor extra onderzoek, terwijl zij uiteindelijk geen borstkanker hebben (dit is een fout-positieve beoordeling). Er is dus ruimte om het bevolkingsonderzoek verder te verbeteren.

De manier waarop radiologen mammografieën beoordelen, heeft invloed op de kwaliteit van het bevolkingsonderzoek. Het beoordelen van mammografieën is lastig, omdat er maar weinig borstkankers voorkomen in de beelden die de radiologen beoordelen en de kenmerken vaak subtiel zijn. In dit proefschrift wordt onderzocht hoe de beoordeling van mammografieën binnen het bestaande bevolkingsonderzoek kan worden verbeterd. Het gaat daarbij om aanpassingen die relatief eenvoudig door te voeren zijn en waarmee impact binnen handbereik ligt. De focus ligt daarbij op het verbeteren van de beoordeling van de beelden door radiologen. Daarbij kijken we zowel naar de beoordeling door één radioloog als door twee radiologen. Ook onderzoeken we hoe kunstmatige intelligentie (AI) kan helpen bij het beoordelen van mammografieën.

In Europa wordt bij het bevolkingsonderzoek naar borstkanker aangeraden om mammografieën door twee radiologen te laten beoordelen. Dit heet dubbel lezen. Het idee hierachter is dat twee mensen samen meer zien dan één. Beide radiologen bekijken de beelden en beslissen onafhankelijk van elkaar of de vrouw moet worden

doorverwezen naar een ziekenhuis voor verder onderzoek. In **hoofdstuk 2 en 3** van dit proefschrift onderzochten we of uitkomsten van het bevolkingsonderzoek veranderen wanneer radiologen bewust aan elkaar worden gekoppeld op basis van hun individuele prestaties. Daarbij keken we naar hoe vaak radiologen kanker ontdekken en hoe vaak zij iemand onterecht doorsturen voor extra onderzoek. We onderzochten of het beter is om radiologen met ongeveer dezelfde prestaties te koppelen, of juist radiologen die van elkaar verschillen, en vergeleken dit met de huidige manier van willekeurig koppelen. In **hoofdstuk 2** zagen we geen duidelijke verschillen. De groepen van radiologen die waren gekoppeld op basis van hun individuele prestaties deden het ongeveer even goed als de groep met willekeurig gekoppelde radiologen. Dit kan betekenen dat het koppelen van radiologen op basis van hun individuele prestaties weinig effect heeft op hoe goed de groep als geheel presteert, of dat de effecten zo klein waren dat we ze in de data niet konden terugzien. In **hoofdstuk 3** hebben we het onderzoek verder uitgebreid door met behulp van statistische modellen op verschillende manieren radiologen aan elkaar te koppelen. Uit de resultaten bleek dat het koppelen van radiologen met een gelijk aantal fout-positieve beoordelingen resulteerde in minder vrouwen die onterecht werden doorverwezen, terwijl het aantal ontdekte kankers in de groep ongeveer gelijk bleef. Toch vonden we geen manier van koppelen die beter was dan willekeurig koppelen voor zowel het vinden van borstkanker als het verminderen van fout-positieve doorverwijzingen. Er is dus nog niet genoeg bewijs dat een bepaalde manier van koppelen betere groepsresultaten oplevert. Daarom is er meer onderzoek nodig waarbij ook andere stappen van het bevolkingsonderzoek, zoals overleg tussen radiologen over doorverwijzing, worden meegenomen.

Ons brein en onze ogen passen zich voortdurend aan de omgeving aan. Dit verschijnsel heet visuele adaptatie. Het bepaalt hoe we iets waarnemen op basis van wat we eerder hebben gezien. Mammografieën zien er allemaal anders uit. Een belangrijke reden hiervoor is de borstdensiteit, dat wil zeggen hoeveel klierweefsel er in de borst zit. Klierweefsel is wit op de beelden, waardoor borsten met meer klierweefsel er anders uitzien dan borsten met weinig klierweefsel. Bij het beoordelen van mammografieën kan dit betekenen dat de volgorde waarin radiologen de beelden bekijken, invloed heeft op hoe goed ze afwijkingen opmerken. Er zijn nog geen richtlijnen over de volgorde waarin mammogrammen gelezen moeten worden. In **hoofdstuk 4** beschrijven we een leesstudie waarin we visuele adaptatie hebben onderzocht. Radiologen bekeken dezelfde beelden op drie manieren: in willekeurige volgorde, van borsten met minder klierweefsel naar borsten met meer klierweefsel (borstdensiteit), en op basis van een volgorde waarbij beelden die veel op elkaar lijken achter elkaar werden gezet door AI. De

resultaten lieten zien dat radiologen beter en sneller presteerden wanneer ze de beelden van lage naar hoge borstdensiteit beoordeelden. Dit laat zien dat het op volgorde zetten van de beelden de resultaten van het bevolkingsonderzoek kan verbeteren.

AI biedt nieuwe mogelijkheden voor het opsporen van borstkanker. **Hoofdstuk 5** beschrijft een leesstudie waarin we onderzochten hoe AI-ondersteuning de prestaties en het zoekpatroon van radiologen beïnvloedt. AI-ondersteuning markeert plekken op een mammogram die volgens de AI verdacht zijn. Uit de studie bleek dat radiologen beter presteerden met hulp van AI-ondersteuning. Bovendien gingen ze gerichter zoeken: in plaats van overal in de borst te kijken, keken ze dankzij de AI-markeringen vooral naar de plekken waar de kanker zich bevond. Dit laat zien dat AI radiologen helpt om hun aandacht te richten op de belangrijke gebieden, waardoor ze sneller en gerichter kunnen zoeken. Toch moeten radiologen niet zomaar de AI-resultaten volgen. **Hoofdstuk 6** laat zien dat radiologen beter presteren als de AI juiste uitkomsten laat zien, maar slechter presteren wanneer de AI fouten maakt. AI kan radiologen dus helpen, maar alleen als de resultaten betrouwbaar zijn.

De algemene discussie in **hoofdstuk 7** blikt terug op de resultaten die in dit proefschrift zijn gepresenteerd en beschrijft mogelijke toekomstige verbeteringen voor het beoordelen van mammografieën. Er wordt beschreven wie welke mammogrammen zou moeten beoordelen en hoe deze beoordeling het beste kan worden gedaan.

Dit proefschrift laat zien dat er praktische en direct toepasbare mogelijkheden binnen handbereik zijn om de beoordeling van mammografieën binnen het bevolkingsonderzoek naar borstkanker te verbeteren. Ik hoop dat de nieuwe strategieën die in dit proefschrift worden beschreven, getest kunnen worden in studies binnen het bevolkingsonderzoek naar borstkanker. Op die manier kunnen we beter begrijpen wat het effect is en mogelijk bijdragen aan betere uitkomsten voor vrouwen die meedoen aan het bevolkingsonderzoek naar borstkanker.

Appendix

Research Data Management
PhD Portfolio
Publications
Curriculum Vitae
Dankwoord

Research Data Management

Ethics and privacy

All studies in this thesis were part of the aiREAD project: Accurate and Intelligent Reading for EArlier breast cancer Detection, supported by KWF Dutch Cancer Society and NWO Domain AES, as part of their joint strategic research program Technology for Oncology II (project number 17912). 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.

A statement confirming that all aiREAD studies were not subject to the Dutch Medical Research Involving Human Subjects Act (WMO) was obtained from the recognized Medical Ethics Review Committee 'METC Oost-Nederland' (registration number 2021–13186). All procedures involving human data were conducted in accordance with relevant national and international legislation and regulations, guidelines, codes of conduct, and Radboudumc policy.

All studies used existing imaging and clinical data collected from multiple breast cancer screening programs. All data reuse adhered to the informed-consent procedures of the original sources and complied with applicable laws, regulations, and the national Code of Conduct for Health Research. For most screening programs, an opt-out procedure was in place, and data from individuals who opted out were excluded. Prior to researcher access, datasets were pseudonymized, with re-identification keys securely maintained by the data providers and not disclosed to any project members.

Radiologists participating in the reader studies provided written informed consent for the collection and processing of their data within the aiREAD project. Consent for sharing the data after the project was not obtained to protect the personal and professional privacy of the participating radiologists.

Data collection and storage

All data was reused from existing data sources and included imaging and clinical data previously collected from various screening programs.

Data for **Chapters 2 and 3** were obtained from pre-existing datasets in Sweden, England, and Norway. Access to these datasets was facilitated through collaborations between Radboudumc and: (1) the Karolinska Institute, (2) the University of Warwick, and (3) Oslo University Hospital and the Cancer Registry of

Norway. The English dataset was securely stored and analyzed on a computer at the University of Warwick to ensure data protection. The Swedish and Norwegian data were securely stored and analyzed on our institutional servers, following Radboudumc's policies to safeguard the availability, integrity, and confidentiality of the data.

Data for **Chapters 4, 5, and 6** was obtained from the Preventicon and PRISMA dataset. The Preventicon dataset is stored on the Medical Imaging Department server, and the PRISMA dataset is stored on the IQ Health server, both at Radboudumc, in accordance with Radboudumc policies for ensuring data availability, integrity, and confidentiality. Reader studies were performed on computers from the Medical Imaging Department and secure online questionnaire data was obtained using Castor EDC. Paper informed consent forms from radiologists participating in the reader studies are securely stored in cabinets of the Medical Imaging Department at Radboudumc.

Data sharing according to the FAIR principles

Obtained results for **Chapters 2 and 3** cannot be made available for reuse, because of the conditions under which the underlying data for this was shared. The results reflect performance in a specific screening program, and disclosure compromises confidentiality and proprietary information. The datasets used for the analyses in these chapters remain the property of the original providers and have therefore not been deposited in the Radboud Data Repository. A part of the Swedish dataset is publicly accessible online (https://doi.org/10.5878/45vm-t798).

Data for **Chapters 4, 5, and 6** was obtained from the Preventicon and PRISMA dataset. Clinical data from the Preventicon data is archived in a Data Acquisition Collection in the Radboud Data Repository (https://doi.org/10.34973/f6ht-cq27). Due to the large size of the Preventicon dataset, images are stored on the Medical Imaging Department server at Radboudumc. The PRISMA dataset is archived in a Data Acquisition Collection in the Radboud Data Repository (https://doi.org/10.34973/p6hp-3n35). The obtained results from the reader studies in **Chapters 4, 5, and 6** cannot be made available for reuse, because no consent from the radiologists was obtained to share the results publicly, aiming to protect the personal and professional privacy of the participating radiologists. To support reproducibility and transparency, accompanying metadata of the reader studies

have been deposited as a Data Acquisition Collection in the Radboud Data Repository (https://doi.org/10.34973/r3qn-mn22).

All obtained results, except for those from the English dataset, are securely stored on secured department servers at the Radboudumc. Data will be archived for 15 years after termination of the study. The results were made reusable by adding sufficient documentation, including readme files, analysis scripts, and publication files.

Chapters 2, 4, and 5 are published hybrid open access in *Radiology*, which means that the papers are available for all RSNA members and publicly available after 12 months. **Chapter 3** is published open access in *Medical Decision Making*. **Chapter 6** is submitted and accepted by *SPIE Medical Imaging 2026: Image Perception, Observer Performance, and Technology Assessment* and will be published after the conference.

PhD Portfolio

Department:	Medical Imaging
PhD period:	17/08/2020 – 06/07/2026
PhD supervisors:	Prof. Dr. I. Sechopoulos, Prof. Dr. M.J.M. Broeders
PhD co-supervisors:	Dr. C.K. Abbey Dr. L.E.M Duijm

Training activities	Year(s)	Hours
Courses		
• RIHS – Introduction course for PhD candidates	2020	15.00
• Radboudumc – Introduction day	2021	6.00
• Deep Learning specialization	2021	128.00
• RU – Projectmanagement for PhD candidates	2021	52.00
• Radboudumc – Scientific integrity	2021	20.00
• RU – Design and Illustration	2022	26.00
• EUSOBI – Breast imaging research course	2024	8.00
• Radboudumc – Career development for PhD candidates & postdocs	2024	24.00
• RU – Art of Finishing Up	2024	10.00
Conferences		
• EUSOBI – Breast Imaging Annual Scientific Meeting*	2021	4.00
• RSNA – Radiology Annual Meeting†	2021	48.00
• RIHS – PhD Retreat	2022	16.00
• WEON – Epidemiology conference*	2022	16.00
• Radboudumc – Cancer Research Retreat Radboudumc*	2022	16.00
• ECR – European Congress of Radiology†	2022	32.00
• MIPS – Medical Image Perception Society congress†	2022	24.00
• EUSOBI – Breast Imaging Annual Scientific Meeting	2022	24.00
• SPIE – Medical Imaging conference	2023	40.00
• ECR – European Congress of Radiology†	2023	40.00
• RSNA – Radiology Annual Meeting*	2023	40.00
• SPIE – Medical Imaging conference†	2024	40.00
• ECR – European Congress of Radiology*	2024	32.00
• EBCC – European Breast Cancer Conference	2024	24.00
• RIHS – PhD retreat†	2024	16.00
• ECR – European Congress of Radiology†	2025	32.00
• Radboudumc – Cancer Research Retreat Radboudumc†	2025	8.00
• IBDW – International Breast Density & Cancer Risk Assessment Workshop†	2025	20.00
• NABIC – Nijmegen Advanced Breast Imaging Course†	2025	8.00
• EUCanScreen – Personalised Cancer Screening†	2025	8.00
• RSNA – Radiology Annual Meeting	2025	40.00

Training activities	Year(s)	Hours
Other		
• AXTI research meetings	2020–2026	200.00
• Monthly BIG-AXTI/Breast Impact meeting	2020–2026	50.00
• Monthly mamma research meeting	2020–2026	30.00
• Weekly department research meeting	2020–2026	80.00
• Journal club – epidemiology autumn	2020	28.00
• Journal club – epidemiology spring	2021	42.00
• Journal club – epidemiology autumn	2021	42.00
• Journal club – epidemiology spring	2022	42.00
• Journal club – epidemiology autumn	2022	42.00
• Journal club – epidemiology autumn	2023	42.00
• LRCB onderzoeksmiddag	2023	4.00
• LRCB onderzoeksmiddag	2024	4.00
• LRCB onderzoeksmiddag	2025	4.00
Teaching activities		
Lecturing		
• RU – Research Your Own Data	2022	28.00
• RU - Grant proposal	2022	28.00
• RU – Research Your Own Data	2023	28.00
Supervision		
• Supervision literature thesis BMS	2024–2025	10.00
• Co-supervision of master internship	2025	30.00
• Supervision of radiologist fellowship	2025	50.00
Total		**1601.00**

* = Poster presentation. † = Oral presentation.

Publications

Publications in this thesis

Gommers JJJ, Verboom SD, Broeders MJM, Sechopoulos I. Automation bias and timing of artificial intelligence decision support in screening mammography. SPIE Proc Med Imaging 2026 Image Percept Obs Perform Technol Assess. 2026;13928. doi:10.1117/12.3086110

Gommers JJJ, Verboom SD, Duvivier KM, van Rooden CJ, van Raamt AF, Houwers JB, et al. Influence of AI Decision Support on Radiologists' Performance and Visual Search in Screening Mammography. Radiology. 2025;316:e243688. doi:10.1148/radiol.243688

Gommers JJJ, Verboom SD, Duvivier KM, van Rooden JK, van Raamt AF, Houwers JB, et al. Enhancing Radiologist Reading Performance by Ordering Screening Mammograms Based on Characteristics That Promote Visual Adaptation. Radiology. 2024;313:e240237. doi:10.1148/radiol.240237

Gommers JJJ, Abbey CK, Strand F, Taylor-Phillips S, Jenkinson DJ, Larsen M, et al. Modeling Radiologists' Assessments to Explore Pairing Strategies for Optimized Double Reading of Screening Mammograms. Med Decis Making. 2024;44:828–42. doi:10.1177/0272989X241264572

Gommers JJJ, Abbey CK, Strand F, Taylor-Phillips S, Jenkinson DJ, Larsen M, et al. Optimizing the Pairs of Radiologists That Double Read Screening Mammograms. Radiology. 2023;309:e222691. doi:10.1148/radiol.222691

Other publications

van Niekerk CC, Verbeek ALM, Brummelkamp E, Otten JHDM, Groenewoud JHMM, van Rossum MM, **Gommers JJJ**. Trends in histological subtypes of cutaneous melanoma in the Netherlands from 1989 to 2023 - Influence of sunlight exposure and overdiagnosis. Tumori. 2026;112:51–8. doi:10.1177/03008916251388862

Gommers J, Hernström V, Josefsson V, Sartor H, Schmidt D, Hjelmgren A, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. Lancet. 2026;407:505–14. doi:10.1016/S0140-6736(25)02464-X

van Winkel SL, Peters J, Janssen N, Kroes J, Loehrer EA, **Gommers J**, et al. AI as an independent second reader in detection of clinically relevant breast cancers within a population-based screening programme in the Netherlands: a retrospective cohort study. Lancet Digit Health. 2025;100882. doi:10.1016/j.landig.2025.100882

Sliwicka O, Oostveen LJ, Swiderska Chadaj Z, van Everdingen WM, Michielsen K, **Gommers J**, et al. Radiation dose reduction of 50% in dynamic myocardial CT perfusion with skipped beat acquisition: a retrospective study. Acta Radiol. 2024;65:724–34. doi:10.1177/02841851241240446

Abbey CK, **Gommers JJJ**, Eckstein MP, Broeders MJM, Sechopoulos I. A bivariate binormal model for modelling double reading of screening mammograms. SPIE Proc Med Imaging 2024 Image Percept Obs Perform Technol Assess. 2024;12929. doi:10.1117/12.3008652

Verboom SD, Caballo M, Peters J, **Gommers J**, van den Oever D, Broeders MJM, et al. Deep learning-based breast region segmentation in raw and processed digital mammograms: generalization across views and vendors. J Med Imaging Bellingham Wash. 2024;11:014001. doi:10.1117/1.JMI.11.1.014001

de Vegt F, **Gommers JJJ**, Groenewoud H, Siersema PD, Verbeek ALM, Peters Y, et al. Trends and projections in the incidence of oesophageal cancer in the Netherlands: An age-period-cohort analysis from 1989 to 2041. Int J Cancer. 2022;150:420–30. doi:10.1002/ijc.33836

Gommers JJJ, Voogd AC, Broeders MJM, van Breest Smallenburg V, Strobbe LJA, Donkers - van Rossum AB, et al. Breast magnetic resonance imaging as a problem solving tool in women recalled at biennial screening mammography: A population-based study in the Netherlands. The Breast. 2021;60:279–86. doi:10.1016/j.breast.2021.11.014

Gommers JJJ, Duijm LEM, Bult P, Strobbe LJA, Kuipers TP, Hooijen MJH, et al. The Impact of Preoperative Breast MRI on Surgical Margin Status in Breast Cancer Patients Recalled at Biennial Screening Mammography: An Observational Cohort Study. Ann Surg Oncol. 2021;28:5929–38. doi:10.1245/s10434-021-09868-1

Curriculum Vitae

Jessie Joke José Gommers was born on May 27, 1997, in Eindhoven, The Netherlands. In 2015, she started her Bachelor's degree in Health Sciences at Maastricht University, specializing in Biology and Health. After obtaining her bachelor's degree in 2018, she started her Master's degree in Biomedical Sciences at Radboud University with a specialization in research and epidemiology. Her last master's internship focused on trends in the use of breast MRI in women recalled after breast cancer screening and investigated the association between preoperative MRI and surgical margin status. It was during this project that Jessie gained her first experience in breast cancer screening.

Following the completion of her master's degree in 2020, Jessie started working as a PhD candidate in the Advanced X-ray Tomographic Imaging (AXTI) group within the Department of Medical Imaging in Radboud university medical center in Nijmegen. During her PhD, she focused on improving the interpretation of screening mammography by combining the strengths of radiologists and artificial intelligence.

Currently, Jessie continues her academic career as a postdoctoral researcher at Radboud university medical center in Nijmegen and Lund University in Malmö, Sweden, and as project leader MRI-Dense at the Dutch Expert Center for Screening (LRCB). Her work remains focused on breast cancer screening. As postdoctoral researcher, she is now involved in two prospective screening trials, including the Screening Tomosynthesis trial with advanced REAding Methods (STREAM) and the Mammography Screening with Artificial Intelligence trial (MASAI). At LRCB, Jessie is involved in the MRI-Dense project, where she focuses on the quality assurance and optimization of supplemental MRI for women with extremely dense breasts within the Dutch Breast Cancer Screening Program. Additionally, Jessie joined the Radiological Society of North America (RSNA) in 2025 as a trainee member of the editorial board.

Dankwoord

De afgelopen jaren waren een periode van leren, lachen en groei, zowel op professioneel als op persoonlijk vlak, met fijne hoogtepunten én uitdagingen die mij sterker hebben gemaakt. Dit had ik niet alleen gekund. Daarom wil ik iedereen bedanken die hier, op welke manier dan ook, aan heeft bijgedragen. Een aantal personen wil ik graag persoonlijk bedanken.

To start with my supervision team: **Ioannis** and **Mireille**, you are the best duo of supervisors, with the perfect combination of skills that guided me through this journey. Meeting with both of you always left me reassured. I am grateful for your guidance, patience, support in shaping ideas, and the feedback that helped me move forward. **Ioannis**, thank you for always being full of great ideas, for being available to discuss anything no matter the time zone you are in, and for consistently looking after everyone's well-being, both at work and beyond. The way you lead and motivate people within the AXTI group is something I truly value and admire. Thank you for introducing me to so many people and the new possibilities that came from it. Also, thank you for making trips and conferences so much fun; I promise that next time I will make sure I am not too hungry, so we can save some margaritas ;). **Mireille**, bedankt dat je aan mij dacht bij de start van het aiREAD-project; zonder jou was ik waarschijnlijk niet aan dit avontuur begonnen, maar jij had vertrouwen in mij, en daar ben ik je heel dankbaar voor. Fijn dat je in mij bent blijven geloven en mij hebt geleerd knopen door te hakken en keuzes te maken in mijn analyses. Ik bewonder hoe jij altijd het overzicht weet te bewaren en oog hebt voor het grotere geheel; daar heb ik veel geleerd. To both of you, thank you for your trust, encouragement, and for opening doors to new opportunities. I am grateful to continue working with you.

To my co-promotor **Craig**, thank you for being such a great help throughout my PhD. Your mathematical and statistical expertise are truly impressive. I am very grateful for all that you have taught me and for the time you took to work through things with me, including running IDL scripts together. I also want to thank you for your warm welcome in Santa Barbara. I still remember how you and your wife, Shelly, picked me up at the bus station, invited me into your home for dinner, and showed me around the university and the lab. Thank you for all. Aan mijn andere co-promotor **Lucien**: ons avontuur begon eigenlijk al tijdens mijn masterstage, toen ik met jouw data aan de slag mocht gaan. Daar maakte ik voor het eerst echt kennis met de wereld van de borstkankerscreening, waar mijn interesse in dit vakgebied werd aangewakkerd. Ik bewonder je enthousiasme en de manier

waarop jij je klinische werk combineert met onderzoek. Ook waardeer ik hoe snel je altijd reageert op vragen, zelfs wanneer je aan de andere kant van de wereld bent. Bedankt voor je betrokkenheid en alles wat ik van je heb mogen leren. Ik hoop dat ik ooit een deel van de mooie reizen mag maken die jij hebt gemaakt. Je hebt zoveel landen bezocht dat ik ze onmogelijk allemaal kan evenaren ;).

Sarah, jij verdient als mede-PhD'er op het aiREAD-project een eigen paragraaf in dit hoofdstuk. Verschillende artikelen in dit proefschrift heb ik mede aan jou te danken. Ik heb ontzettend veel van je geleerd. Je wiskundige inzichten en programmeerskills hebben me enorm geholpen, en daar ben ik je heel dankbaar voor. Dank je wel voor al je goede ideeën, je geduld, de fijne gesprekken en natuurlijk de gezelligheid. Ik denk met veel plezier terug aan de vele congressen die we samen hebben bezocht.

To all collaborators who kindly provided data from their breast cancer screening programs for the radiologist pairing studies, I would like to sincerely thank you. In particular, I am grateful to **Fredrik Strand** for providing the data from Sweden, to **Sian Taylor-Phillips** and **David Jenkinson** for sharing data from England, and to **Solveig Hofvind** and **Marthe Larsen** for contributing the data from Norway. I truly appreciate your trust, the valuable meetings, discussions, and feedback along the way. This collaboration greatly expanded my understanding of breast cancer screening programs outside the Netherlands, and I am truly thankful for that.

I would also like to thank **Miguel Eckstein** and the members of his lab for welcoming me during my visit to the Vision and Image Understanding Lab at the University of California, Santa Barbara. Thank you for the wonderful Californian experience and for introducing me to the world of eye tracking. This experience led us to add eye tracking to the reader studies described in this thesis, which has proven to be highly valuable for the field.

Thank you, **Michael Webster**, for sharing your expertise on visual perception and adaptation, and for helping me better understand how perception dynamically changes with experience. Your insights have truly helped shape the design of our reader study. I am very grateful for the way visual adaptation and radiology could be brought together in this work, and I appreciate the insightful meetings and your valuable feedback throughout the project.

De leesstudies zouden niet mogelijk zijn geweest zonder de medewerking van de vele radiologen. Bedankt aan **Heleen de Bruin**, **Wouter Dinkelaar**, **Katya Duvivier**, **Janneke Faquenot-Nollen**, **Maaike Gielens**, **Janneke Houwers**, **Mechli**

Imhof-Tas, **Frits Jansen**, **Dick Naafs**, **Fleur van Raamt**, **Jan-Kees van Rooden**, **Peter Sallevelt**, **Miranda Snoeren**. Door met jullie mee te kijken tijdens het lezen van de mammogrammen heb ik heel veel geleerd. Dank voor jullie enthousiasme, interesse in de wetenschap, tijd en bereidheid om naar Nijmegen te komen.

To the members of the manuscript committee, **Prof. dr. I.D. Nagtegaal**, **Prof. dr. C.H. van Gils** en **dr. F. Kilburn-Toppin**, thank you for reading and assessing my thesis. I look forward to your questions during the defense. I would also like to thank the other opponents for their willingness to exchange ideas during my defense.

Next, I would like to thank all AXTI members who have been part of my journey. I am grateful for your "gezelligheid", support, and the coffee/tea breaks, which made my time within AXTI so much more enjoyable.

Marjolein, bedankt dat ik altijd met al mijn vragen bij jou terecht kan en dat je me zo goed helpt met allerlei organisatorische zaken. **Wendelien** and **Marco**, I still remember you showing me around the hospital when I was considering doing a PhD, and how welcome you made me feel. **Marco**, thank you for your early involvement in the aiREAD project, your thoughtful questions, and the coffee breaks. **Wendelien**, ook al ben je tegenwoordig meer betrokken bij de BIG groep, ik heb altijd heel erg genoten van je gezelligheid, de congressen die we samen hebben bezocht en je altijd snelle en behulpzame antwoorden op mijn vragen. **Marta** and **Olga**, you made my first conference abroad an amazing experience. Thank you, **Marta**, for being such a warm person and for bringing the group together by organizing fun group activities. **Olga**, thank you for all the fun nights and drinks we shared together. **Joana**, even though our time together in the lab was short, I still remember your contagious laugh – thank you for the nice time. **Koen**, bedankt voor al je hulp met computergerelateerde vragen. Ook al was je niet direct betrokken bij mijn project, wist je toch altijd de juiste vragen te stellen. **Luuk**, bedankt voor je behulpzaamheid en droge humor.

To the PhD's who started around the same time as I did: **Juan**, **Sjoerd**, **Mikhail**, **Franziska**, and **Noelia** thank you for making this journey a shared experience. It was nice to be in the same boat together, supporting one another and exchanging advice along the way. **Juan**, thank you for always bringing a smile to my face when you came into the office. I truly appreciate the many moments you sincerely asked how I was doing, our late-afternoon conversations about life, the finger-stretching sessions, and even the times you turned off my screen to make me stop working ;). **Sjoerd**, bedankt voor je gezelligheid, al vanaf de vroege uurtjes. Dank je wel voor je

betrokkenheid en voor het delen van de Limburgse hitjes en vlaai. **Mikhail**, I really admire your language skills, and it has been very nice to be able to speak Dutch together. Thank you for the good times and for always making sure the lab went for lunch. **Franziska**, being based in Germany did not take away from how much I valued the moments we shared. I really appreciated you joining us for special occasions, such as the AXTI weekends. **Noelia**, although you were working from abroad, it was always a pleasure to see you whenever you visited Nijmegen.

To the colleagues who joined my PhD journey later: **Daan**, bedankt voor je gezelligheid en adviezen. Het was altijd fijn dat je even langskwam voor een praatje, oog had voor het welzijn van iedereen in de groep en het kantoor gezelliger maakte met jouw initiatieven om samen leuke dingen te doen. **Lian**, fijn dat jij ook onderdeel bent van ons team en bedankt voor je bijdrage binnen de STREAM-studie. **Liselot**, het is altijd leuk om jouw praktische set-ups te zien en fijn om te merken dat er zoveel mooie samenwerkingen vanuit Twente mogelijk zijn. Dank je wel voor de fijne gesprekken en de gezellige tijd samen tijdens congressen. **Gustavo**, I am happy you joined AXTI and shared your amazing design skills. Thank you for being willing to be the photographer at my defense. **Martina**, thank you for always being in the office, for managing the servers, and for sharing your funny stories. Please do take care of yourself when going to the gym or riding your bike ;). **Raneim**, thank you for the good times together and for all the enjoyable conversations along the way. **Leo**, I was really glad you joined us. I enjoyed your thoughtful smiles and the positive energy you brought. **Maranda** en **Hanne**, het plastische chirurgie-duo, bedankt voor jullie gezelligheid en dat ik tijdens jullie presentaties standaard even met mijn pols draaide om te checken of alles nog goed zat ;). **Pim**, ik ben blij dat je na je sollicitatie hebt gekozen om bij AXTI te starten. Je hebt in een korte tijd al superveel bijgedragen en je bent echt een fijne aanvulling op het team. **Sven**, bedankt voor de goede tijd samen en voor het delen van je leuke sportieve verhalen. **Ou**, thank you for sharing your stories about China. I am happy to see someone is continuing with reader studies.

To the other colleagues in the BIG group, specifically **Suzanne** and **Alma**, thank you for collaborating between AXTI and BIG, for sharing your experiences, and for the nice times together. **Marialena** and **Nika**, thank you for the coffee corner breaks, the welcome distractions, and the good times at conferences (and in the toilets – you know what I mean, Marialena).

Jim en **Lindy**, fijn dat ik jullie ook heb leren kennen en bedankt voor de gezellige congresreisjes. **Jim**, ook al werken we niet direct samen, ik waardeer het heel erg

dat we af en toe kunnen overleggen en dat ik gebruik kon maken van jullie mooie PRISMA-dataset. **Lindy**, bedankt voor je gezelligheid, het warme welkom in het STREAM-team en de fijne samenwerking.

Aan het **LRCB** wil ik mijn dank uitspreken voor de waardevolle ondersteuning bij onze leesstudies. Bij jullie wordt op inspirerende wijze zichtbaar hoe de wetenschap wordt vertaald naar de dagelijkse praktijk. Dank jullie wel voor de mooie kans om bij jullie aan de slag te mogen gaan.

To my new colleagues in Sweden, thank you for making me feel so welcome. I am especially grateful to **Kristina** for making our collaboration possible and for giving me the chance to work with the inspiring MASAI data. I look forward to many more moments of working together.

Daarnaast wil ik ook graag iedereen buiten het werkveld bedanken. Lieve vrienden en (schoon)familie, bedankt voor de nodige afleiding en ontspanning. De vele borrels, weekendjes weg, festivals en gezellige etentjes hebben voor veel plezier en balans gezorgd naast het werk. Ook wil ik jullie bedanken voor jullie oprechte interesse in mijn werk en de steun die ik daarbij heb ervaren.

Tot slot wil ik graag de allerbelangrijkste mensen nog bij naam noemen.

Lieve **papa** en **mama**, hoe kan ik jullie ooit bedanken? Zonder jullie was ik nooit zover gekomen. Bedankt voor jullie onvoorwaardelijke liefde, hulp en steun, en voor alles wat jullie mij hebben meegegeven. Het was heel fijn om aan het begin van mijn PhD weer bij jullie te wonen. De ommetjes tijdens de lunchpauzes, de momenten waarop jullie mij bevestiging gaven als ik weer eens twijfelde, en de verwennerij hebben die periode extra waardevol gemaakt. Jullie kennen mij als geen ander, staan altijd voor mij klaar en weten precies wat ik nodig heb. Dankjulliewel voor alles! **Jeroen**, bedankt voor alle mooie momenten die we samen hebben gedeeld en voor alles wat jij mij in het begin van mijn PhD hebt bijgeleerd over programmeren. Ik ben je ontzettend dankbaar voor je geduld en de tijd die je daarvoor nam. Ook bedankt voor je nuchtere blik op het leven, die mij telkens weer hielp relativeren, en voor de fijne zwemavonden die een goede afleiding waren. **Rick**, ook jij bedankt voor je onvoorwaardelijke steun. Je vele telefoontjes onderweg naar Nijmegen maakten de files toch een stuk minder erg. Ik waardeer het hoe jij mij steeds weer laat zien dat het leven gevierd moet worden – daarin ben jij echt een groot voorbeeld voor mij!

Lieve **Bart**, waar moet ik beginnen? Ik wil jou voor heel veel bedanken. Bedankt dat je er altijd voor mij bent en met zoveel geduld naar mijn verhalen luistert. Ook dankjewel dat je nooit hebt geklaagd wanneer ik in de avonden en weekenden weer aan het werk was. Jij weet altijd precies wat je moet zeggen of doen om mij weer gerust te stellen. Bedankt voor je onvoorwaardelijke steun, liefde en de vele mooie momenten die we samen hebben mogen delen. Op naar nog veel meer tijd om samen te genieten. Ik hou van jou!

9 789465 152790 >

Radboudumc university medical center

IN DEI NOMINE FELICITER

Radboud University