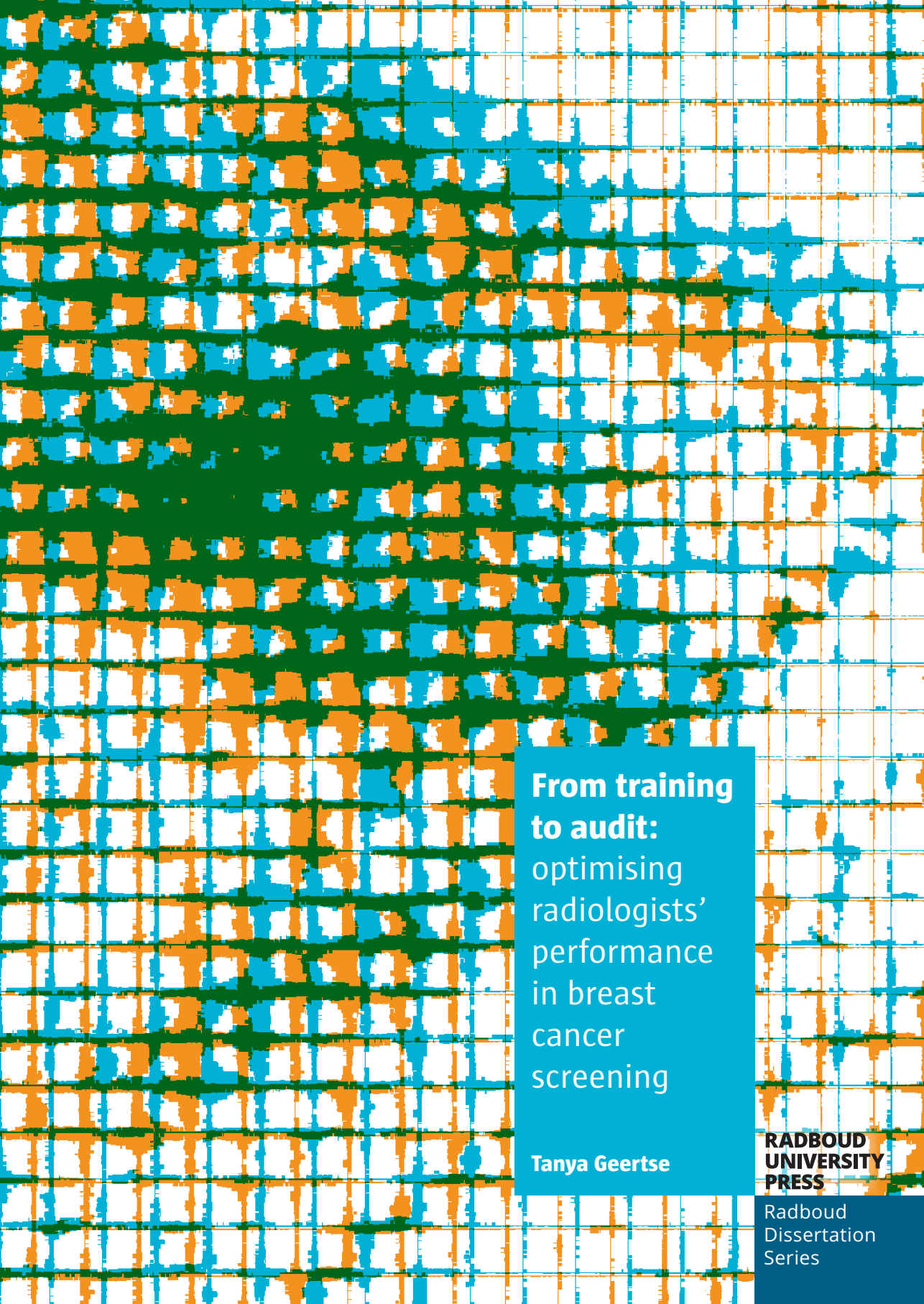


From training to audit: optimising radiologists' performance in breast cancer screening

Tanya Geertse

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

From training to audit

optimising radiologists' performance in breast cancer screening

Tanya Geertse

From training to audit: optimising radiologists' performance in breast cancer screening
Tanya Geertse

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: ANNIEK + design | Anniek van den Berge
Printing: DPN Rikken/Pumbo

ISBN: 9789465151960
DOI: 10.54195/9789465151960
Free download at: https://doi.org/10.54195/9789465151960

© 2026 Tanya Geertse

RADBOUD
UNIVERSITY
PRESS

This is an Open Access book published under the terms of Creative Commons Attribution-Noncommercial-NoDerivatives International license (CC BY-NC-ND 4.0). This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given to the creator, see http://creativecommons.org/licenses/by-nc-nd/4.0/.

From training to audit

optimising radiologists' performance in breast cancer screening

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

donderdag 9 april 2026
om 12.30 uur precies

door

Tanny Diana Geertse
geboren op 16 april 1972
te Reimerswaal

Promotoren:
Prof. dr. M.J.M. Broeders
Prof. dr. R.M. Pijnappel (UMC Utrecht)

Copromotoren:
Dr. D. van der Waal
Dr. L.E.M. Duijm (CWZ)

Manuscriptcommissie:
Prof. dr. W.J.J. Assendelft
Prof. dr. C. van Ongeval (UZ Leuven, België)
Prof. dr. S. Siesling (Universiteit Twente)

Table of contents

Radboud

Chapter 1

General Introduction

Breast cancer screening

Worldwide, breast cancer is the leading cause of female cancer deaths [1]. Screening enables earlier detection, which increases the likelihood of successful treatment and helps prevent breast cancer mortality. It also helps improve the chances of receiving conservative treatments, such as breast-conserving therapy rather than mastectomy. The most effective modality for breast cancer screening is still mammography [2-4].

While the benefits of early detection are well established, breast cancer screening also presents important challenges. Notably, some of the early detected breast cancers may be cancers that would never have been detected if the woman had not been screened and would never have caused her any harm during her lifetime – a phenomenon known as overdiagnosis [3-6]. Hence, the woman gains no benefit from the knowledge of a breast cancer diagnosis nor from the subsequent treatment. Unfortunately, it is currently not possible at the time of screening to determine whether the cancer will cause harm. Therefore, disregarding the detected cancer is not an option. Other major challenges in screening include false-negative and false-positive screening results. False negative results can lead to interval cancers – cancers that develop after a negative screening result and before the next scheduled examination. These cancers often have a poorer prognosis than screen-detected cancers and are more likely to have a higher histological grade, larger tumour size, and to be estrogen- and progesterone-receptor negative [7,8]. False positives, on the other hand, can cause anxiety, reduce future screening participation, and increase healthcare costs [9–12].

Based on a screening mammogram, the screening radiologists determine whether a woman should be recalled for further diagnostic assessment or not. The accuracy with which they make this decision – i.e., the performance of the screening radiologists – plays a crucial role in determining the overall balance between the benefits and harms of breast cancer screening.

Reading screening mammograms

In breast cancer screening programmes worldwide, all women in specific age groups are invited for a screening mammogram every 1-3 years, meaning millions of screening mammograms per year [13,14]. For example, in the Netherlands all women between 50-74 years are invited every 2 years, corresponding to around 1 million women per year [15]. As recommended, the reading of screening mammograms takes place in a setting with large batches of 100 or more [16-18].

Importantly, breast cancer screening typically involves imaging women who are assumed to be healthy (asymptomatic women). Therefore, the prevalence of detectable cancers in a screening population is very low compared to that in women who present at a hospital with breast-related symptoms (symptomatic women). As a result of the high volumes of mammograms and the low prevalence of breast cancers, subtle abnormalities and abnormalities obscured by glandular tissue of the breast are challenging to detect on a mammogram and are at risk of being missed by a screening radiologist.

The detection of abnormalities is the essential first step, followed by the challenge of interpreting whether the detected abnormality is suspicious for malignancy and should be recalled for further assessment. This interpretation is rarely a clear-cut, objective decision, but often a subjective judgement influenced by the radiologist's experience. Radiologists establish individual thresholds to distinguish between benign and potentially malignant findings. Reading mammograms in a screening setting constitutes risk management and therefore demands a different mindset compared with reading mammograms in a clinical setting, where the aim is to make a definitive diagnosis instead of differentiating between recall or no recall.

To effectively navigate this complex, risk based decision-making process, radiologists require targeted training and feedback on performance. Professional education is therefore crucial [19,20].

Quality Assurance and screening performance of radiologists

To support radiologists in achieving and maintaining high-quality performance, many breast cancer screening programmes incorporate a quality assurance (QA) programme. These QA programmes often include training and regular audits – independent evaluations that monitor screening performance and assess compliance with established requirements and target values [21,22]. QA programmes are particularly well developed in centrally organised, population-based screening programmes – typically coordinated at the national or regional level – where consistent monitoring and feedback mechanisms are embedded in the programme.

While the main goal of breast cancer screening is reducing mortality from this disease, the primary focus when monitoring the screening performance of the radiologists is on screening outcomes other than mortality itself. Among these

screening outcomes, there are three key metrics: the recall rate, the breast cancer detection rate and the positive predictive value of recall (PPV). The recall rate gives the proportion of screening examinations that are interpreted as abnormal and result in a recall for further clinical assessment. The cancer detection rate indicates the number of screening examinations at which breast cancer was detected, typically expressed per 1,000 screens. Finally, the PPV reflects the proportion of recalls at which breast cancer was detected.

Lowering the recall threshold results in a higher recall rate and may increase the cancer detection rate. However, it also typically leads to more false-positive screening results, thereby reducing the PPV. Conversely, raising the recall threshold may reduce false-positives but increases the risk of missing cancers (false-negative screening results or interval cancers). Based on feedback regarding their screening performance, radiologists must carefully consider this trade-off to find the most appropriate balance.

Thesis outline

This thesis aims to identify factors that support screening radiologists in improving their reading performance, ultimately optimising the benefit-harm balance of breast cancer screening. To this end, we examine various QA-tools and strategies implemented in breast cancer screening programmes, such as radiologist training, recall strategies, warning signals from radiographers to radiologists, and audits and feedback.

In several national quality assurance programmes, screening radiologists are required to obtain Continuing Medical Education (CME) credits to maintain and enhance their professional competence in interpreting screening mammograms [23-26]. To assess how this requirement contributes to improved performance, we conducted a systematic literature review on the utility of supplemental training in breast cancer screening. **Chapter 2** presents the results of this study. We evaluate different types of supplemental training for screening radiologists and examine whether performance improves and how such improvement is measured.

For screening radiologists, the many well-circumscribed masses occurring in a screening population often pose a dilemma when deciding whether or not to recall the lesion. Recalls due to well-circumscribed masses frequently lead to a false-positive screening result. In **chapter 3** we investigate the characteristics of

malignancies presenting as well-circumscribed masses on mammography as these characteristics provide insight into their clinical relevance. This, in turn, can help identify the potential improvements in the recall strategy. We present an overview of the literature and the results of an exploration of real-world screening practice.

In the Netherlands, the screening mammograms are acquired by specialised radiographers. Immediately after obtaining the mammogram, the radiographer checks the images for any abnormalities suspicious for cancer (prereading) and can set a warning signal for the radiologists. These warning signals may help prevent the radiologist from overlooking abnormalities. In **chapter 4**, we assess the added value of prereading by radiographers in relation to early screening outcomes.

In **chapter 5** we investigate the value of audits in breast cancer screening as a QA- tool. In the Netherlands, audits include not only the evaluation of screening performance indicators but also radiological reviews of mammograms. This chapter presents a retrospectively evaluation of four audit series performed in the Netherlands between 1996 and 2013.

Chapter 6 addresses a different aspect of audit data by retrospectively evaluating the utility of a specific QA-tool: the PPV-recall diagram. We investigate the suitability of this PPV-recall diagram for monitoring performance of screening radiologists in a breast cancer screening programme.

Finally, in the General Discussion, the main findings derived from this thesis are summarized and their implications and directions for further research are discussed.

References

1. World Health Organization. Global breast cancer initiative implementation framework: assessing, strengthening and scaling up of services for the early detection and management of breast cancer: executive summary. Geneva: WHO; 2023. Available from: https://www.who.int/publications/i/item/9789240067134. Accessed 12 Feb 2025.
2. Dibden A, Offman J, Duffy SW, Gabe R. Worldwide review and meta-analysis of cohort studies measuring the effect of mammography screening programmes on incidence-based breast cancer mortality. Cancers (Basel). 2020;12(4):976–991.
3. 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(11):2205–2240.
4. Marcon M, Fuchsjäger MH, Clauser P, Mann RM. ESR Essentials: screening for breast cancer – general recommendations by EUSOBI. Eur Radiol. 2024;34(10):6348–6357.
5. Barratt A. Overdiagnosis in mammography screening: a 45 year journey from shadowy idea to acknowledged reality. BMJ. 2015;350:h689.
6. Flemban AF. Overdiagnosis due to screening mammography for breast cancer among women aged 40 years and over: a systematic review and meta-analysis. J Pers Med. 2023;13(3):523.
7. Mills C, Sud A, Everall A, et al. Genetic landscape of interval and screen detected breast cancer. NPJ Precis Onc. 2024;8(1):1-9.
8. Niraula S, Biswanger N, Hu P, Lambert P, Decker K. Incidence, characteristics, and outcomes of interval breast cancers compared with screening-detected breast cancers. JAMA Netw Open. 2020;3(09):e2018179.
9. Bond M, Pavey T, Welch K, Cooper C, Garside R, Dean S, Hyde C. Systematic review of the psychological consequences of false-positive screening mammograms. Health Technol Assess. 2013;17(13):1–170.
10. Setz-Pels W, Duijm LE, Coebergh JW, Rutten M, Nederend J, Voogd AC. Re-attendance after false-positive screening mammography: a population-based study in the Netherlands. Br J Cancer. 2013;109(8):2044–2050.
11. 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(4):351–358. Erratum in: Br J Cancer. 2021;25(7):1031.
12. 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 Apr 10. Epub ahead of print. doi:10.1177/09691413251329671. PMID: 40207622.
13. OECD. Health at a Glance 2023: OECD Indicators. Paris: OECD Publishing; 2023. Available from: https://doi.org/10.1787/7a7afb35-en. Accessed 11 Sep 2025.
14. Katsika L, Boureka E, Kalogiannidis I, et al. Screening for breast cancer: a comparative review of guidelines. Life (Basel). 2024;14(6):777.
15. National Institute for Public Health and Environment. Monitor breast cancer screening 2023. Available from: https://www.rivm.nl/en/documenten/monitoring-breast-cancer-screening-2023. Accessed 10 Sep 2025.
16. Burnside ES, Park JM, Fine JP, Sisney GA. The use of batch reading to improve the performance of screening mammography. AJR Am J Roentgenol. 2005;185(3):790–796.

17. Cohen EO, Lesslie M, Weaver O, Phalak K, Tso H, Perry R, Leung JWT. Batch reading and interrupted interpretation of digital screening mammograms without and with tomosynthesis. J Am Coll Radiol. 2021;18(2):280–293.

18. 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(12):9548–9555.

19. Nodine CF, Kundel HL, Mello-Thoms C, et al. How experience and training influence mammography expertise. Acad Radiol. 1999;6(10):575–585.

20. European Commission, Joint Research Centre. European Commission Initiative on Breast Cancer – European Quality Assurance Scheme for Breast Cancer Services. Luxembourg: Publications Office of the European Union; 2025. Available from: https://data.europa.eu/doi/10.2760/7067385. Accessed 11 Sep 2025.

21. Hofvind S, Bennett RL, Brisson J, et al. Audit feedback on reading performance of screening mammograms: an international comparison. J Med Screen. 2016;23(3):150–159.

22. Qenam BA, Li T, Tapia K, Brennan PC. The roles of clinical audit and test sets in promoting the quality of breast screening: a scoping review. Clin Radiol. 2020;75(10):794.e1–794.e6.

23. Dutch Expert Centre for Screening. Kwaliteitsregister voor Screeningsradiologen in het Bevolkingsonderzoek op Borstkanker in Nederland. 2020. Available from: https://lrcb.nl/wp-content/uploads/2020/10/Reglement-Kwaliteitsregister-voor-screeningsradiologen_herziene-versie-1-okt-2020.pdf. Accessed 21 Jul 2025.

24. National Health Service England. Guidance for radiology and advanced radiographic practice in the NHS Breast Screening Programme. 2024. Available from: https://www.gov.uk/government/publications/breast-screening-quality-assurance-standards-in-radiology. Accessed 21 Jul 2025.

25. BreastCheck Ireland. Guidelines for Quality Assurance in Mammography Screening. 4th ed. 2015. Available from: https://www.breastcheck.ie/health-professionals.2721.html. Accessed 21 Jul 2025.

26. German Mammography Screening Program. Versorgung im Rahmen des Programms zur Früherkennung von Brustkrebs durch Mammographie-Screening, Anlage 9.2 BMV-Ärzte. 2024. Available from: https://fachservice.mammo-programm.de/de/qualitaetssicherung. Accessed 21 Jul 2025.

Radboud

Chapter 2

Utility of supplemental training to improve radiologist performance in breast cancer screening: a literature review

Geertse TD, Paap E, van der Waal D, Duijm LEM, Pijnappel RM, Broeders MJM

Journal of the American College of Radiology, 2019; 16(11):1528-1546

Abstract

Purpose

This literature review evaluates whether supplemental training for radiologists improves their breast screening performance and how this is measured.

Methods

A systematic search was conducted in PubMed on August 3, 2017. Articles were included if the study describes a supplemental training for radiologists reading mammograms to improve their breast screening performance and at least one outcome measure is reported. Quality of the study was assessed using the Medical Education Research Study Quality Instrument (MERSQI).

Results

Of 2199 identified articles, 18 studies were included, of which 17 showed an improvement on at least one of the outcome measures, for at least one training activity or subgroup. Two measurement approaches were found. For the first approach, measuring performance on test sets, sensitivity and specificity are the most reported outcomes (eight out of eleven studies). Recall rate is the most reported outcome (six out of seven studies) for the second approach, which measures performance in actual screening practice. The studies were mainly of moderate quality (MERSQI score: mean=11.7, SD=1.7), caused by small sample sizes and the lack of a control group.

Conclusions

Supplemental training helps radiologists improve their screening performance, despite the mainly moderate quality of the studies. There is a need for better designed studies. Future studies should focus on performance in actual screening practice and should look for methods to isolate the training effect. If test sets are used, focus should be on knowledge about correlation between performance on test sets and actual screening practice.

Introduction

Prevalence of disease is low in breast cancer screening. Only a small number of cancers can be detected in a large number of images. Reading mammograms in a screening setting therefore requires different skills compared to a clinical setting. In the Dutch national program, radiologists who want to interpret screening mammograms are obliged to participate in an 8-day starting course, as part of the quality assurance program. Once registered, screening radiologists have to read at least 3000 screening examinations per year, obtain at least 40 continuing medical education (CME) points over five years and participate in an audit once every three years [1]. Many countries require CME for radiologists in breast cancer screening, to maintain a high performance level, e.g., the UK [2], Australia, and the USA [3] also use these types of requirements.

Table 1: Search terms entered into PubMed database

Group	Search terms
Population	("Radiologists"[Mesh] OR "Radiology"[Mesh] OR radiolog*[tiab] OR Screen reader* [tiab] OR Observer* [tiab]) AND
Setting	(("Mass Screening"[Mesh] OR "Mammography"[Mesh] OR mammograph*[tiab] OR (breast[tiab] AND screen*[tiab]) OR mass screening*[tiab] OR mammogram*[tiab] OR breast imaging[tiab]) AND
Intervention	(Audit [tiab] OR "Feedback"[Mesh] OR Feedback* [tiab] OR "Education"[Mesh] OR education*[tiab] OR "Learning"[Mesh] OR learn*[tiab] OR workshop*[tiab] OR Module [tiab] OR Instruction* [tiab] OR Assessment* [tiab] OR Intervention* [tiab] OR Training* [tiab] OR Self-test [tiab] OR Lesson* [tiab] OR Evaluation [tiab] OR "Clinical Audit"[Mesh] OR Self-assessment [tiab] OR Educational measurement [tiab] OR "Self-Evaluation Programs"[Mesh] OR Teaching [tiab])

In general, the idea is that supplemental training helps radiologists improve screening performance. However, to our knowledge, no review of the literature has been performed to confirm this. We have therefore conducted a systematic literature search, with the aim to evaluate different types of supplemental training for screening radiologists, in order to gain insight into whether performance really improves and how improvement is measured.

Materials and Methods

Search Strategy

The search strategy was developed with the assistance of a librarian of the Radboud university medical center (Nijmegen, The Netherlands). To define search terms, the pearl growing technique [4] was used, where keywords from relevant studies were used as "pearls" [5-11]. The keywords were divided into three groups (Table 1): 1. The population (radiologists); 2. The setting (mammography screening); 3. The intervention (supplemental training activities to improve screening performance). We defined general terms, e.g., "education" and "assessment", but also specific terms such as "self-test" and "audit" based on the "pearls" [8,11]. The PubMed (Medline) search was performed on August 3, 2017. There were no limitations regarding type of journal or publication date. The references of all included articles were manually checked for additional relevant articles.

Study Selection

The search results were imported in the reference management software EndNote and duplicates were removed.

First, titles and abstracts were reviewed by one researcher (T.G.). The full texts of the selected articles were reviewed by two researchers independently (T.G., E.P.). Discrepancies were resolved by a third researcher (M.B.). Articles were included if they met the inclusion criteria: (a) the study population consists of radiologists (screening, possibly combined with clinical radiology, fellows or residents); (b) the study describes a type of supplemental training for radiologists reading mammograms to improve their screening performance; and (c) at least one outcome measure is reported, e.g., sensitivity, specificity, recall rate, positive predictive value (PPV) of recall or kappa value (measure of agreement with expert panel, e.g. in BI-RADS assessment).

Data Extraction and Quality Assessment

Table 2 shows the data extracted from the selected articles: study characteristics, outcome measures, and main conclusion. In addition, the quality of the studies was assessed using the Medical Education Research Study Quality Instrument (MERSQI), which was developed to appraise the methodologic quality of quantitative medical education studies [12] (see supplementary Appendix 1 for the score form). This instrument is based on six domains: study design, sampling of the population, type of data, validity of the evaluation instrument, data analysis, and outcomes. Within each domain a value between 0 and 3 or between 1 and 3 can be scored. The total

MERSQI score thus ranges from 5 to 18. Higher scores indicate a higher quality. No cutoff values have been defined.

Data Analysis

Due to the considerable variation in the study designs, type of training, and outcomes, we could not pool the studies quantitatively to estimate the effect of supplemental training on performance. Instead, we assessed how the effect of supplemental training was measured and qualitatively described the outcomes of the studies.

Table 2: Study characteristics, outcomes measures and main conclusions of the selected studies

Publication	Study design and aim of the study	Population and setting	Type of training
Adcock KA et al. 2004 [13]	• Retrospective cohort study • Performance outcomes from 1993 to 2002 • Evaluate changes in performance after specialization and self-learning were implemented in 1998.	• n=21 • USA • Screening	Radiologists chose to specialize in interpreting mammograms: access to specialized training and a high volume of readings. Self-study: 3-times-per-year mammogram interpretation self-assessment exercises for the subspecialists (2.5h per set of cases)
Berg WA et al. 2002 [14]	• Single group & non-equivalent control group; • Pre-test – post-test (54 cases) • Determine whether training in BI-RADS feature analysis would improve observer agreement with experienced breast imagers in mammographic lesion description and final assessments	• Intervention group: n=23 • Expert panel: n=3 • Control group of senior residents: n=6 • USA • Screening	Face-to-face training BI-RADS training (1d): • lectures of BI-RADS features (3.5h) • questions and answers (0.5h) • examples including discussion (2h)
Carney PA et al. 2012 [15]	• Randomized wait-list study • Recall rates during three time periods: 1) 9 mo before 2) 0-9 mo after (T1) 3) 9 -18 mo after (T2) • Describe the impact of a tailored web based educational program designed to reduce excessive recall.	• Intervention group: n=33 (n=23 completed the intervention) • Control group: n=21 (n=9 completed the intervention) • USA • Screening	Self-study, web-based (1h): • Audit data: feedback on screening performance • Breast cancer risk • Possible impact of unnecessary recalls • Knowledge questions were imbedded

Outcome measures of performance	Main conclusion	MERSQI
• Percentage stage 0 or 1 of detected cancers: Baseline: 1993= 84.5%; 1994=81.6%; 1995=85.5%; 1996=82.5%; 1997=82.8% After specialization and self-learning: 1998=88.8%; 1999=90.6%; 2000=88.0%; 2001=88.8%; 2002=90.2% • Sensitivity: Baseline: 1993-1995=0.704; 1996-1997=0.722 After specialization and self-learning: 1998-2002=0.801 • Callback rate: After specialization and self-learning: 1999=7.5%; 2000=7.6%; 2001=7.5%; 2002=7%	• By implementing a multifaceted initiative to improve interpretation of mammograms, the sensitivity increased substantially, as more cases of cancers at earlier stages were diagnosed without increasing the proportion of callbacks.	10.5
• The overall generalized κ value: - Pre-training κ=0.31, post-training κ=0.45 • The overall generalized κ value of sub-group (n=11): - Pre-training κ=0.37, immediately after training κ=0.53, 2-3 months after training κ=0.49 Accuracy measure: *Reference = expert panel:* • Biopsy sensitivity - Pre-training= 71%, post-training= 86%, ($p<0.05$) • Biopsy rate of benign lesions (false-positive rate) - Pre-training= 25%, post-training= 26%, ($p>0.05$) • A_z value - Pre-training= 0.78, post-training= 0.82, ($p<0.05$) *Reference = histopathologic truth:* • Biopsy sensitivity - Pre-training= 73%, post-training= 88%, ($p<0.05$) • Biopsy rate of benign lesions (false-positive rate) - Pre-training= 43%, post-training= 51%, ($p<0.05$) • A_z value - Pre-training= 0.70, post-training= 0.71, ($p>0.05$) • No significant differences for control group.	• Training in BI-RADS feature analysis and assessment resulted in improved consistency in lesion description. • The sensitivity of participants improved without a significant increase in false-positive results, as measured against the consensus of the experienced breast imagers. • The benefits of such training were retained after 2–3 months in a subgroup of 11 out of the 23 participants.	12
Recall rates: • Radiologists who completed the intervention: *Intervention group*: Baseline: 11.2%, T1: 10.8%, T2: 10.4% *Control group*: Baseline: 8.7 %, T1: 8.8%, T2: 9.2% Intervention vs. Control group: T1: OR=1.12, 95%CI=1.00-1.27, p=0.0569 T2: OR=1.10, 95% CI=0.96-1.25, p>0.05 • Radiologists who not completed the intervention: *Intervention group*: Baseline: 11.0%, T1: 9.4%, T2: 9.7% *Control group*: Baseline: 9.6%, T1: 9.0%, T2: 9.3%	• The study resulted in a null effect, which may indicate that a single 1-hour intervention is not adequate to change excessive recall. • It is likely that more complex approaches are needed to change radiologists practice patterns.	15

Table 2: Continued

Publication	Study design and aim of the study	Population and setting	Type of training
Ciatto S et al. 2006 [16]	• Single group • Pre-test – post-test (150 cases) • Evaluate the performance of a sample of radiologists undergoing a proficiency test of screening mammography	• 1st attempt: n=537 • 2nd attempt: n=146 • 3rd attempt: n= 22 • Italy • Screening	Self-study: Consulting a teaching DVD atlas, with several thousand mammography cases visible as an atlas or as an interactive test
Geertse TD et al. 2015 [17]	• Retrospective cohort study • Performance outcomes of four audit series • Evaluate the results of four series of audits, to investigate the value of these audits as a QA tool	• 26 RU, 28 RU, 29 RU and 17 RU (only group results) • Netherlands • Screening	Audit • Feedback on screening performance • Radiological review of screening examinations
Geller BM et al. 2014 [18]	• Randomized controlled trial • Pre-test – post-test (40 cases) • Evaluate the impact of two case-based educational interventions (self-paced DVD and a live expert-led seminar) designed to improve interpretive screening mammography performance.	• Intervention group DVD n=37 • Intervention group Live Seminar n=25 • Control group n=40 • USA • Screening	Teaching set: 40 cases: 18 cancer and 22 noncancer cases Teaching points for each case. 1) DVD with teaching set -> self-study 2) Live seminar (one-day) Teaching set on workstation + expert instructor + discussion

Outcome measures of performance	Main conclusion	MERSQI
Pass rate (test was passed when sensitivity > 80% and recall rate < 15%) : • 32.7% after one test • 66.7% after two (χ^2=15.1, p<0.0001) • 89.3% after three attempts (χ^2=10.2), p<0.01) Pass rate according to mammography experience (1st attempt): • <1,000/year : 22.7% passes • 1,000-2,000/year: 33.5% passes • >2,000/year: 54.8% passes (χ^2=4.17, p=0.04)	Diagnostic accuracy: • in a simulated scenario of screening mammography is poor; • depends on previous mammography experience; • improves with intensive training.	11
• Increase in recall rate from 0.66%, 1.07%, 1.22% to 1.58% (p<0.05). • Increase in detection rate from 3.3, 4.5, 4.8 to 5.4 per 1000 (p<0.01). • Decrease in PPV of recall from 51.9%, 43.5%, 41.5% to 35.5% (p<0.001). • Increase in sensitivity from 64.5%, 68.7%, 70.5% to 71.6% (p<0.05). • No trend was observed over time for the percentage of missed cancers for interval cancers (p= 0.460) and for screen-detected stage II cancers (p=0.323).	• For radiologists, an accurate understanding of their performance is essential to know which points are most in need of improvement. • It is recommended that in addition to bench-marking screening outcomes, a radiological review of screening examinations and immediate feedback should be part of an audit.	12.5
Effects of intervention (adjusted odds ratios) ** relative to expert recall / ** relative to cancer status* • Sensitivity: - Live vs. control OR=1.24 (p=0.190)*, OR=1.22 (p=0.384)**. - DVD vs. control OR=1.34 (p=0.050)*, OR=1.28 (p=0.237)**. • Specificity: - Live vs. control OR=0.8 (p=0.048)*, OR=0.79 (p=0.015)** . - DVD vs. control OR=0.9 (p=0.299)*, OR=0.92 (p=0.343)**. • PPV: - Live vs. control OR=1.13 (p=0.631)*, OR=1.11 (p=0.743)** . - DVD vs. control OR=1.94 (p=0.004)*, OR=1.81 (p=0.045)**. • NPV: - Live vs. control OR=1.08 (p=0.547)*, OR=1.06 (p=0.752)**. - DVD vs. control OR=0.96 (p=0.760)*, OR=0.96 (p=0.694)**.	• The DVD educational intervention resulted in a significant improvement in mammography interpretive screening performance on a test-set. • Results suggest that interpretive performance can be improved by educational interventions based on actual clinical cases with findings that are frequently misinterpreted.	14.5

Table 2: Continued

Publication	Study design and aim of the study	Population and setting	Type of training
Horst F van der et al. 2003 [19]	• Retrospective cohort study • Performance outcomes of two audit series • Describe the audit proces and evaluate audit results.	• 14 CU and 15 CU (group results) • Netherlands • Screening	Audit • Feedback on screening performance • Radiological review of screening examinations
Lee EH et al. 2014 [20]	• Single group & non-equivalent control group • Pre-test – post-test (25 cases) • Evaluate the efficacy of a mammography boot camp to improve performance in interpreting mammograms	• Intervention group: n=141 • Control group (experts): n=26 • Korea • Screening	Face-to-face training 3-day bootcamp consisting of lectures and group practice readings (250 cases, 200 cancer and 50 noncancer cases)
Lehman CD et al. 2001 [21]	• Single group • Pre-test – post-test (30 cases) • Evaluate the effect of training in the use of the lexicon on interpretive skills in screening mammography.	• n=14 • Ukraine • Screening	Face-to-face training 1-day ACR BI-RADS training consisting of lectures, focused on the use of the lexicon and assessment categories
Linver MN et al. 1992 [22]	• Retrospective cohort study • Audit results of two consecutive years • Analyse radiologist's performance (audit results) before and after attending mammography courses	• n=12 • USA • Screening + diagnostic	Face-to face training During the audit period the radiologists attended at least one dedicated 3- or 4 day basic mammography course. Some attended as many as four courses

Outcome measures of performance	Main conclusion	MERSQI
First audit series vs. second series: • Increase in recall rate from 0.68% to 0.92% • Increase in detection rate from 3.5 to 4.1 per 1000 • Decrease in PPV of recall from 50.6% to 44.8% • Interval cancers, significant lesion on previous screening mammogram: 31% vs. 18%.	• Audit offers tools to improve radiologist performance. • Reviewing screening films of referred and interval cancers with outcome results in hand boosts performance.	11.5
Average test scores (correct answers): • Intervention group: pre-camp 56.0±12.2%, post-camp 78.3±9.2% (p<0.001) • Control group: pre-camp 79.6±7.2% (no post-camp) • Pre-camp: Intervention Vs. Control Group p<0.01 • Post-camp: Intervention Vs. Control Group p=0.31 • Pre- and post-camp test scores were similar among the age groups and among the type of attending institution (p=0.38)	• The bootcamp improved the performance in Interpreting mammograms. • The bootcamp was equally effective for all age groups and all types of attending institution.	9.5
Mean values: • Sensitivity: - Pre-training = 50.5 (14.3-71.4)%, - post-training = 87.4 (66.7-100)% (p<0.0001) • Specificity: - Pre-training= 76.8 (52.2-91.3)%, - post-training= 88.6 (71.4-100)% (p<0.01) • PPV: - Pre-training= 43.4 (11.1-71.4)%, - post-training= 78.5 (57.1-100)% (p<0.0001)	• The training program was associated with significant improvement in interpretation of test mammograms.	10.5
Audit results1988 vs.1989: • Sensitivity: 80% vs. 87% (p=0.053) • PPV: 33% vs. 32% • No. of false-negative cases: 30 vs. 27. • Cancer detection rate (asymptomatic): • 0.48% vs. 0.62% • Cancer detection rate: increased from 7.35 per 1000 before attendance course to 11.26 per 1000 after attendance three or more courses (p=0.04) (n=2)	• Improvement in detection of breast cancers. • Although many factors have contributed, it is believed that the attendance at dedicated mammography courses was the major factor responsible for the improvement of mammography skills.	14

Table 2: Continued

Publication	Study design and aim of the study	Population and setting	Type of training
Luo P et al. 2005 [23]	• Single group • Pre-test – post-test (80 cases) • Investigate the effects of CAD training on performance in mammography interpretation	• n=3 • USA • Screening	Self-study: A four weeks' training on the knowledge of CAD, using a computer-based hypermedia approach, composed of a training set of 100 biopsy-proven cases.
Miglioretti DL et. al 2009 [24]	• Retrospective cohort study • Performance outcomes from 1996 to 2005 • Evaluate changes in screening mammogram interpretation for radiologists with and without fellowship training in breast imaging gained clinical experience	• n=321: n=214 without fellowship training and n=17 with fellowship training • USA • Screening	Fellowship training in breast imaging This study examined the effect of increasing years of experience of radiologists on their performance measures, results reported separately for radiologists with and those without fellowship training in breast imaging.
Mullen LA et al. 2017 [25]	• Retrospective cohort study • Performance outcomes during intervention period (7-months) compared with baseline (period of 3-years before intervention) • Determine the impact of interventions designed to reduce recall rates on screening performance metrics	• n=10 • USA • Screening	Face-to-face training 1st intervention (a 7-month period): Group discussion and personal review of all recalls and their outcomes (weekly basis) 2nd intervention (a 7-month period): Consensus double reading of all potential recalls.

Outcome measures of performance	Main conclusion	MERSQI
• Sensitivity: pre-training vs. post-training Observer 1: 0.51 vs. 0.72, Observer 2: 0.54 vs. 0.75 Observer 3: 0.56 vs. 0.77 • Specificity: pre-training vs. post-training Observer 1: 0.89 vs. 0.88, Observer 2: 0.86 vs. 0.91 Observer 3: 0.85 vs. 0.90 • ROC area (A_z): pre-training vs. post-training Observer 1: 0.752 vs. 0.861, Observer 2: 0.745 vs. 0.901, Observer 3: 0.749 vs. 0.904	• The study demonstrated increased performance with the aid of CAD after training and gaining experience with CAD. • CAD training for readers in mammography reading has effects on their image reading performance. • Sufficient training and experience with CAD is recommended before reader performance studies and in residency and technician training programs.	8.5
Radiologists without fellowship training: • Performance improved most during their 1st 3 years of clinical practice, when the odds of a false-positive reading dropped 11%-15% per year (p<.015) with no associated decrease in sensitivity (p>.89). • The number of women recalled per breast cancer detected decreased from 33 in their 1st year of practice, to 24 with 3 years of experience, to 19 with 20 years of experience. Radiologists with fellowship training: • They experienced no learning curve and reached desirable goals during their 1st year of practice. • Sensitivity significant increased with increasing years of experience (p=.043)	• Radiologists' interpretations of screening mammograms improve during their first few years of practice and continue to improve throughout much of their careers. • The additional training offered in breast imaging fellowship may better prepare radiologists for screening mammogram interpretation than the standard training offered to radiology residents. • Additional residency training and targeted continuing medical education may help reduce the number of work-ups of benign lesions while maintaining high cancer detection rates,	12.5
1st intervention vs. baseline, using FFDM: • Recall rate: 9.2% vs. 11.1% (p=0.0001) • Detection rate: 3.1 vs. 3.8 per 1000 (p>0.05) • PPV1: 3,1% vs. 3.4% (p>0.05) 1st intervention vs. baseline, using DBT: • Recall rate: 6.6% vs. 7.6% (p=0.018) • Detection rate: 6.2 vs. 4.8 per 1000 (p>0.05) • PPV1: 10.8% vs. 6.0% (p<0.0001) 2nd intervention vs. baseline using FFDM: • Recall rate: 9.9% vs. 11.1% (p=0.048) • Detection rate: 5.9 vs. 3.8 per 1000 (p>0.05) • PPV1: 5.7% vs. 3.4% (p<0.0001) 2nd intervention vs. baseline using DBT: • Recall rate 7.2% vs. 7.6% (p>0.05) • Detection rate: 5.7 vs. 4.8 per 1000 (p>0.05) • PPV1: 9.0% vs. 6.0% (p<0.0001)	• Although the study population is small, the data suggest that simple interventions can reduce recall rates, without compromising screening performance metrics. • These interventions could ultimately save resources and improve the patient experience.	12.5

Table 2: Continued

Publication	Study design and aim of the study	Population and setting	Type of training
Poot D et al. 2016 [9]	• Retrospective cohort study • Performance outcomes of module during three rotations (within 27 months) • Evaluate effectiveness of the simulation screening mammography module on interpretive skills	• Intervention group: n=39 • USA • Radiology residents	Face-to-face training Simulation module: • Interpret screening mammograms (212 1st rotation, 230 2nd rotation and 261 3rd rotation) • One-on-one feedback after 100 cases.
Raza S et al. 2016 [26]	• Single group • Pre-test – post-test • (200 cases) relative to: 1) Expert panel = truth 2) Quantitative assessment (Volpara Density) = truth • Evaluate accuracy of visual breast density categorization and to determine if accuracy of this assessment is affected by training	• Intervention group: n=20 • Expert panel: n=2 • USA • Screening	Face-to-face training Training module about density assessment: • a review of currently available methods; • limitations of the measurements; • a systematic visual technique.
Suleiman WI et al. 2016 [27]	• Retrospective cohort study • Performance outcomes of test-set in 3 consecutive years • Assess test-set performance of radiologists using the Breast Reader Assessment Strategy (BREAST) over a 3-year period.	• Intervention group: n=14 • Australia • Screening	Self-study Test-set (BREAST): • Reading test-set of 60 cases (20 cancer and 40 noncancer cases) on dedicated workstation • Decisions in online tool • Immediate feedback after test-set

Outcome measures of performance	Main conclusion	MERSQI
Performance outcomes three rotations: • Sensitivity: 1st rotation: 77.45% 2nd rotation: 79.41% 3rd rotation: 82.45%. • Specificity 1st rotation: 79.45% 2nd rotation: 82.93% 3rd rotation 86.01%. Overall performance vs. national benchmark: • Sensitivity: 84.5% vs. 84.9% • Specificity: 83.2% vs. 90.3%	• By interpreting an enriched data set of standardized screening mammograms with known outcomes, residents were given meaningful feedback on their interpretive skills of screening mammography compared to national benchmarks while in radiology residency training. • Using simulation modules for interpretation of screening mammograms is a promising method for training radiology residents to detect breast cancer and to help them achieve competence toward national benchmarks.	10
Proportion of correct density assessment (*expert panel = truth)*: • Pre-training= 65%, post-training= 72% (p=0.0011) Overall OR=1.41 (p<0.0001). • BI-RADS 1 OR=12.26 (p<0.0001) • BI-RADS 2 OR=0.66 (p=0.037) • BI-RADS 3 OR=0.4 (p<0.0001) • BI-RADS 4 OR=1.45 (p=0.10) Proportion of correct density assessment *(quantitative assessment = truth)*: • Pre-training= 63%, post-training= 65% (p=0.28) Overall OR=1.1 (p=0.26). Substantial agreement between expert panel and quantitative assessment (κ=0.78, 95%CI=0.72 to 0.83)	• Training resulted in improved overall accuracy of visual mammographic breast density assessment against the truth of the expert panel. • The educational training module enabled the reader to more precisely assess breast density by illustrating the boundaries between each of the BI-RADS density categories 1, 2, 3 and 4. • Substantial agreement between qualitative and quantitative breast density assessment exists.	10.5
Performance outcomes test-sets 2011 - 2012 – 2013: • Overall: -Sensitivity: 74%, 80%, 89% (p=0.002) -Specificity: 67%, 74%, 74% (p=0.003) -Location sensitivity: 51%, 73%, 79% (p=0.001) • More difficult cases: -Sensitivity: 66%, 77%, 86% (p=0.002) -Specificity: 62%, 82%, 84% (p=0.001) -Location sensitivity: 47%, 60%, 68% (p=0.02) • Less difficult cases: -Sensitivity: 78%, 97%, 98% (p=0.001) -Specificity: 61%, 78%, 84% (p=0.001) -Location sensitivity: 62%, 81%, 93% (p=0.001)	• Radiologists who undertake the BREAST programme demonstrate significant improvements in test-set performance during a 3-year period, highlighting the value of ongoing education through the use of test-set. • The results do not prove that radiologists' actual screening performance has improved.	13

Table 2: Continued

Publication	Study design and aim of the study	Population and setting	Type of training
Timmers JMH et al. 2012 [28]	• Single group • Pre-test – post-test (30 cases) • Study the impact of training on the inter-observer agreement in using the BI-RADS lexicon in the screening setting.	• Intervention group: n=55 (25 experienced and 30 new screening radiologists) • Expert panel: n=6 • Netherlands • Screening	Face-to-face training 2-day BI-RADS training: • 1-day theory • 1-day hands-on
Urban N et al. 2007 [29]	• Randomized, 2 groups • Each radiologist participated in 5 sessions. Performance outcomes of session 1 and 4 were compared (Set A and B, 45 cases per set, crossover design). • Test the hypothesis that providing low-volume radiologists (reading < 300 mammograms per year) with immediate feedback on their interpretations of difficult mammograms would improve their reading skills.	• Intervention group: n=35 (Group I: n=17, group II: n=18) • USA • Screening	Self-study CAMFP (Computer Assisted Mammography Feedback Program): • 5 sessions (of ca. 1 hour) over an 11 month period • Review / assess difficult mammograms (TIFF images) • Feedback at the end of the session

BI-RADS, Breast Imaging Reporting and Data System; κ, kappa (measure of agreement between participant of training and expert panel); A_z, area under the receiver operating characteristic (ROC) curve; χ^2, chi-square test; RU, reading unit; CU, central unit (=reading unit); PPV, positive predictive value; ACR, American College of Radiology; CAD, computer-aided detection; FFDM, full-field digital mammography; CME, continuing medical education.

Outcome measures of performance	Main conclusion	MERSQI
Average kappa agreement *(expert panel = truth)*: • All radiologists: Pre-training κ=0.44, Post-training κ=0.48 (p=0.15) • Experienced radiologists: Pre-training κ=0.48, Post-training κ=0.46 (p=0.60) • New screening radiologists: Pre-training κ=0.41, Post-training κ=0.50 (p=0.01)	• The training programme in the BI-RADS lexicon resulted in a significant improvement of agreement among new screening radiologists. • There was no difference in agreement among experienced screening radiologists. • Overall, the agreement among radiologists was moderate.	12
Session 1 vs. session 4: • Sensitivity: - Group I: 75% vs. 85% - Group II: 77% vs. 84% - Combined: 76% vs. 84% (p<0.001) • Specificity: - Group I: 82% vs. 76% - Group II: 81% vs. 76% - Combined: 81% vs. 76% (p=0.04) • Variability among radiologists decreased (p=0.035).	• The CAMFP intervention improved sensitivity and decreased variability among radiologist's interpretations. • Specificity decreased. • The program is potentially useful as a component of CME of radiologists.	11

Results

Included Studies

The search identified 2199 articles (Figure 1). After removing duplicates, 2139 articles were screened by assessing the title, and 395 articles were screened by evaluating the abstract as well. In total, 91 articles were selected for full text review, of which 75 were excluded: the study population did not consist of radiologists (n=10), no description of a supplemental training activity was provided (n=21), or no outcome measure was reported (n=42). Two articles had to be excluded because no European library could supply the full text.

After checking the references of the 16 included articles, two additional articles were considered to be eligible for inclusion. The final data set therefore consisted of 18 studies [9,13-29].

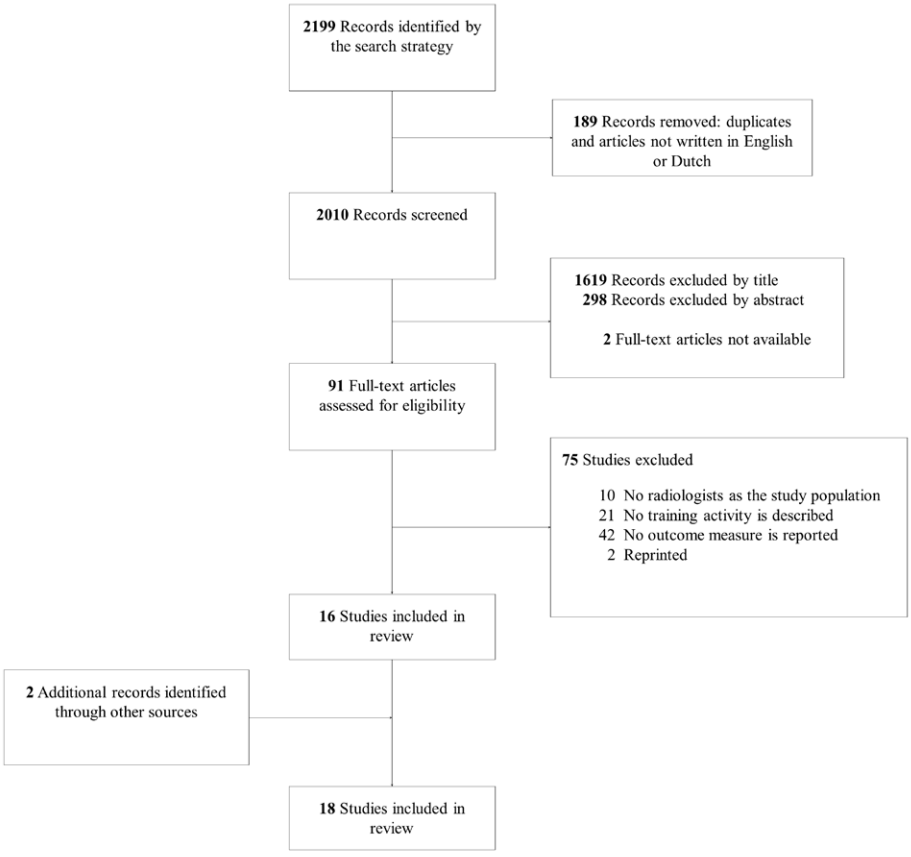


Figure 1: Flow diagram of the literature search and the inclusion of results

The included studies varied considerably in design and objectives as well as in the type of supplemental training. Nevertheless, we identified two approaches to measure potential improvements in screening performance as an effect of supplemental training: (1) interpretive performance on mammography test sets, where a test set is a pre-selected set of mammograms with a mix of normal and cancer cases (Table 3); (2) in actual screening practice (Table 4).

Performance on test sets

Table 3 gives an overview of the main results of the eleven studies [9,14,16,18,20, 21,23,26-29 that use performance on test sets to measure the effect of supplemental training. More details of these studies can be found in Table 2. All eleven studies showed an improvement of at least one outcome, which was statistically significant in ten studies. In one study [9], no statistical analysis was reported. Sensitivity (n=7 out of 11) and specificity (n=6 out of 11) are the most commonly reported outcomes.

Sensitivity improved in all seven studies. In five studies [16,18,21,27,29], this improvement was statistically significant ($0.0001<p<0.05$). In the study by Geller et al. [18], this was only true for the DVD educational intervention ($p=0.05$), but not for the live seminar ($p=0.190$).In two studies [9,23], no p-value was reported.

Specificity improved in four (out of six) studies. In two [21,27], this improvement was statistically significant ($p<0.01$ resp. $p=0.003$). In the other two [9,23], no statistical analysis was reported. Two studies [18,29] observed a statistically significantly decrease in specificity ($p=0.048$ resp. $p=0.04$). In the study by Geller et al. [18], this was only true for the live seminar, but not for the DVD educational intervention ($p=0.299$).

Two studies [18,21] reported the PPV, and both showed a statistically significant improvement ($p=0.004$ resp. $p<0.0001$). Similar to the sensitivity results, this improvement in PPV was limited to the DVD educational intervention in the study by Geller and colleagues [18]. Two studies [14,28] reported kappa values for the BI-RADS classification. In both studies, the kappa value improved. In the study by Timmers [28], this improvement was only statistically significant for new radiologists ($p=0.01$) and not for experienced radiologists ($p=0.60$). In the study by Berg et al. [14], no p-value was reported. The area under the receiver operating characteristic (ROC) curve was reported in two studies [14,23]. Both showed a statistically significant improvement ($p<0.05$). Recall rate, biopsy sensitivity, biopsy rate for benign lesions, percentage correct answers, and correct density assignment were all reported only once, and all showed statistically significant improvement ($0.0001<p<0.05$) [14,16,20,26].

Table 3 : Results of studies evaluating performance on test sets

Author	Type of training	Type of test set(s)	Subgroups
Ciatto [16]	S	One test, three attempts	
Luo [23]	S	Pre-test =Post-test (reordered)	
Suleiman [27]	S	Three test sets in three consecutive years	Overall More difficult cases Less difficult case
Urban [29]	S	Two tests, crossover design	
Geller [18]	S	Four pre-tests, one post-test	*Reference:* Expert panel Histopathologic
	F	Four pre-tests, one post-test	*Reference:* Expert panel Histopathologic
Berg [14]	F	Pre-test =Post-test	*Reference:* Expert panel Histopathologic
Lee [20]	F	Pre-test =Post-test	
Lehman [21]	F	One pre-test, one post-test	
Poot [9]	F	One test, three iterations	
Raza [26]	F	Pre-test =Post-test	
Timmers [28]	F	One pre-test, one post-test	Overall Experienced radiologists New radiologists

Az value, area under the receiver operating characteristic (ROC) curve; k value, kappa value=measure of agreement between participant of training and expert panel; DSF, digitized screen-film mammography; DM, digital mammography; N/A, not available; S, Self-study; F, face-to-face training; +, statistically significant increase; -, statistically significant decrease; =, no statistically significant difference.
* No statistical analysis is reported.

Number of test cases per test	Number of cancer cases per test	Type of mammography images	Sensitivity	Specificity	Recall rate	PPV	Biopsy sensitivity	Biopsy rate of benign lesions	κ value	Az value	Correct answers	Correct density assignments
150	17	DSF	+		-							
80	14	DSF	+*	+*						+		
60	20	DM	+	+								
			+	+								
			+	+								
45	25	DSF	+	-								
40	18	DSF										
			+	=		+						
			=	=		+						
40	18	DSF										
			=	-		=						
			=	-		=						
54	19	N/A							+*			
							+	=		+		
							+	+		=		
25	18	DM									+	
30	7	N/A	+	+		+						
266	65	DM	+*	+*								
200	N/A	DM										+
30	N/A	DM							=			
									=			
									+			

Table 4: Results of studies evaluating performance in actual screening practice

Author	Type of training	Period of outcomes	Subgroups	Type of mammography images	Sensitivity	Recall rate	Detection rate	PPV	False-positive rate	Number of false-negative cases	Percentage missed cancers for interval cancers	Percentage stage 0 and 1 of detected cancers
Adcock [13]	S	Ten consecutive years		SFM	+ *	= *						+ *
Carney [15]	S	Three periods: 1) 9 months before (baseline) 2) 0-9 months after (T1) 3) 9-18 months after (T2)		N/A		=						
Geertse [17]	F	Four periods: 1) 1990-1997 2) 1998-2003 3) 2001-2006 4) 2006-2011		 SFM SFM SFM SFM+DM	+	+	+	-			=	
Horst [19]	F	Two periods: 1)1992-1999 2)1997-2001		SFM		+ *	+ *	- *			- *	
Linver [22]	F	Two consecutive years		SFM	=		+	=		=		
Miglioretti [24]	F	Ten consecutive years	Fellowship NO YES	SFM		- =		+ =	- =			
Mullen [25]	F$	Two periods: 1) 3-years before (baseline) 2) 7-months during intervention		DM		-	=	=				
	F§	Two periods: 1) 3-years before (baseline) 2) 7-months during intervention		DM		-	=	+				

PPV, positive predictive value ; S, Self-study; F, face-to-face training; N/A, not available; SFM, screen-film mammography; DM, digital mammography; +, significant increase; -, significant decrease; =, no significant difference.

* No statistical analysis is reported.

$ 1st intervention

§ 2nd intervention

The number of included mammograms in the test sets used in the different studies varied from 25 to 266 cases. In all test sets, the number of cancers was higher than in a real screening situation. The ratio between cancer and normal cases varied from 1:8 to 1:0.4 (Table 3).
Seven studies [14,18,20,21,23,26,28] compared the results of a test set before (pre-test) and after (post-test) a training activity. In four of these studies [14,20,23,26], the pre-test and post-test consisted of the same mammograms.

Performance in actual screening practice

Table 4 gives an overview of the seven studies [13,15,17,19,22,24,25] that use performance in actual screening practice to measure the effect of supplemental training. This is done by comparing screening performance in different calendar periods or by evaluating trends over time. More details can be found in Table 2. Six of the seven studies showed an improvement of at least one outcome, which was statistically significant in four studies. In two [13,19], no statistical analysis was reported. Recall rate (n=6 out of 7), PPV (n=5 out of 7), detection rate (n=4 out of 7), and sensitivity (n=4 out of 7) are the most commonly reported outcomes.

Recall rate improved in five [15,17,19,24,25] out of six studies, which was statistically significant in three [17,24,25] ($0.0001<p<0.05$). In the study by Miglioretti et al. [24], this was only for radiologists without fellowship training ($p<0.001$). The radiologists with fellowship training showed no learning curve ($p=0.56$). No statistical analysis was reported in one study [19], and the decrease in recall rate was not statistically significant ($p>0.05$) in another study [15]. In one study [13], the recall rate remained the same.

Five studies reported the PPV, of which two showed an improvement [24,25]. In the study by Mullen et al. [25], this increase in PPV was only statistically significant ($p<0.0001$) for the second intervention (consensus double reading of all recalls) and not ($p>0.005$) for the first intervention (discussion of all recalls). In the study by Miglioretti et al. [24], this increase in PPV was not statistically significant ($p=0.066$ without fellowship training, $p=0.08$ with fellowship training). Two studies showed a decrease in PPV [17,19], which was statistically significant ($p<0.001$) in one study [17]. In the other study [19], no statistical analysis was reported. In one study [22], the PPV remained the same.

The four studies that looked at detection rate [17,19,22,25] all showed an improvement, which was statistically significant in two [17,22] ($p<0.01$ resp. $p=0.04$)

and not statistically significant ($p>0.05$) in one [25]. In one study [19], no statistical analysis was reported.

Four studies [13,17,22,24] reported the sensitivity. All of them showed an improvement, which was statistically significant ($p<0.05$) in two [17,24]. In the study by Miglioretti et al. [24], this was only for radiologists with fellowship training. The radiologists without fellowship training showed no significant trends ($p=0.15$). In one study [22], the improvement was not statistically significant ($p=0.053$), and another study [13] reported no statistical analysis.

Two studies [17,19] reported no change in the percentage of missed cancers among interval cancers from one audit to the next ($p>0.3$). False-positive rate, the number of false-negative cases, and the percentage of stage 0 and 1 of detected cancers (early-stage cancer detection) were all reported only once [13,22,24]. None of them showed a statistically significant improvement. None of the studies described the effect on specificity.

Quality Assessment

The MERSQI scores (Table 2) varied from 8.5 to 15 (mean=11.7, SD=1.7). Only two [15,18] were randomized controlled trials and scored the maximum 3 points for "Study design". None scored the maximum 3 points for "Sampling", because all were limited to a single institute and/or had response rates that were either low or not reported. In the "Type of data" domain, all studies included objectives measurements and therefore scored the maximum of 3 points. All studies scored low in the "Validity of evaluation instrument" domain, since they did not report on the reproducibility, consistency, and reliability of the method selected to measure the effect of training. The relationships of the selected outcomes to other variables were only reported by three studies [18,27,29]. Four studies [9,13,19,23] scored low for "Data analysis", because no statistical analyses were reported. In more than half of the studies (11 out of 18) the outcomes were measured using performance on test sets (Table 3). These studies are valued lower in the "Outcomes" domain than the studies using performance in actual screening practice (Table 4): 1.5 points compared to 3 points.

Discussion

Based on this review, we conclude that supplemental training appears to improve radiologists' performance. Seventeen of the eighteen studies showed an

improvement on at least one of the outcome measures, which was statistically significant in fourteen studies. Due to considerable heterogeneity in interventions used and effect measures reported, no summary effect could be estimated. Although we are therefore not able to draw conclusions on the extent of the effect and the type of supplemental training activities that are most effective, this review does increase our understanding of the importance of supplemental training to improve performance of breast screening radiologists and highlights the importance of well-designed studies.

Ideally, the effect of supplemental training on performance is determined in actual screening practice. This does involve several challenges. With a relatively low breast cancer incidence (e.g. seven per 1000 screened in the Netherlands [30]), most screening radiologists detect only a few cancers per month. To accurately measure a potential effect of training, especially on detection rate, data over a prolonged period of time are needed [31,32]. However, changes in performance over time can also be caused by factors other than the intervention, e.g., improvement in image quality due to improved skills of technologists or changes in the age distribution of the screened population [22]. Another challenge is the assessment of sensitivity and specificity [31]. These parameters can only be calculated once the exact number of interval cancers (i.e. breast cancers diagnosed after a negative screening examination and before the next examination) is known, which requires a follow-up period of at least one screening interval. In addition, a comprehensive cancer registry should be available to identify the interval cancers through linkage with the screening registry.

When measuring performance, it is important to look at the association between recall rate, detection rate, and PPV of recall. The international study by Elmore et al. [33] points out that factors such as national health policy issues (e.g., malpractice concerns) and population characteristics may strongly influence recall rates in different countries. In order to determine whether a decrease or increase of recall rate is the desired effect of supplemental training, it is important to know the baseline performance and the acceptable level of the performance indicators. The studies of Miglioretti [24] and Mullen [25] both showed a decrease in recall rate. Because of the high baseline recall rate (>10%), this result was favorable. On the other hand, the studies by Geertse et al. [17] and Van der Horst et al. [19] both showed an increase in recall rate. For these studies, the baseline recall rate of less than 1% was considered to be too low. Therefore, the increased recall rate was also interpreted as an improvement. The latter two studies [17,19] showed a decrease in

PPV of recall, which is a direct consequence of the desired effect on recall rate. On balance, the training effect is therefore still considered to be an improvement.

It is evident that using test sets to determine the effect of training offers many advantages: test sets could be timed in such a way that the intervention is the only likely explanation for measured changes in performance, test sets are usually heavily enriched and therefore have a much higher prevalence, cases with proven pathology can be used, and results are almost immediately available.

Several limitations have to be taken into account here as well. Performance is evaluated based on the gold standard set by expert panel review or based on histopathology. As noted by Onega et al. [32], one should realize that by using histopathology, no account is taken of false-positive findings that require recall. When pre- and post-test are the same, memory bias could occur. This could be addressed by including a control group, as described in the study by Berg et al. [14]. Luo et al. [23] attempted to avoid memory effects by reordering the cases for the post-test. When different test sets are used, the difficulty level could be different. Geller et al. [18] addressed this issue by adjusting for test set and case difficulty in the statistical analysis, and analyzing changes in performance of the intervention groups relative to the control group. Suleiman et al. [27] chose to calculate a difficulty index for each cancer case in three test sets to correct for varying difficulty. The number of normal and cancer cases per set may also influence the outcomes. Furthermore, using test sets may have a psychological effect. Because participants know they are being tested, they might lower their recall threshold in an attempt to maximize sensitivity [16]. This could explain why sensitivity improved in all test set studies and specificity did not.

In general, potential sources of bias exist for both approaches. Effects may be influenced by, e.g., differences in image quality and differences in experience level of the radiologists. This underlines the importance of including an appropriate control group.

All studies scored low on the "validity of the evaluation instrument", because the reproducibility, consistency, and reliability of the method to measure effect of training were not reported. None of the included studies that measured performance on test sets examined the correlation between test set performance and performance in screening practice. Instead, the authors usually referenced two other studies [34,35]. Both studies found a moderate correlation but differed in their findings on which performance measures were significantly correlated. Rutter

and Taplin [34] found no evidence of correlation for sensitivity, and a moderate correlation for specificity. In contrast, Soh et al. [35] found a stronger correlation for sensitivity than for specificity. Another limitation is that only four studies (22%) used a control group, and sample sizes were very small in general. In ten studies (56%), the number of participants was 35 or less.

This review shows that supplemental training activities for breast cancer screening radiologists help improve screening performance, despite the fact that the included studies were mainly of moderate quality, particularly in terms of sampling strategy and validity of evaluation instruments. Both approaches to measure changes in performance, i.e., test sets or actual screening practice, have their own strengths and limitations. For the future, there is a clear need for well-designed studies that focus on performance in screening practice and look for methods to isolate the effect of training. If test sets are used, information should be obtained on the correlation between test set performance and the performance in screening practice. Future studies should also address the point in time when offering supplemental training activities (e.g. in the first year or in the form of continuing education) to breast cancer screening radiologists is most effective.

References

1. Internal report. LRCB. Kwaliteitsregister voor Screeningsradiologen in het Bevolkingsonderzoek op Borstkanker in Nederland (Quality register for screening radiologists of the breast cancer screening in the Netherlands). Available at: http://www.lrcb.nl/resources/uploads/2017/10/Reglement-Kwaliteitsregister-voor-screeningsradiologen_1-jan-2018.pdf. Accessed January 2018.
2. Public Health England. Quality assurance guidelines for breast cancer screening radiology. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/470579/nhsbsp59_QA_radiology_uploaded_231015.pdf. Accessed March 2011.
3. Destouet JM, Bassett LW, Yaffe MJ, Butler PF, Wilcox PA. The ACR's Mammography Accreditation Program: ten years of experience since MQSA. J Am Coll Radiol 2005;2(7):585-94.
4. Schlosser RW, Wendt O, Bhavnani S, Nail-Chiwetalu B. Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive Pearl Growing. A review. Int J Lang Commun Disord 2006;41(5):567-82.
5. Carney PA, Bowles EJ, Sickles EA, et al. Using a tailored web-based intervention to set goals to reduce unnecessary recall. Acad Radiol 2011;18(4):495-503.
6. Geller BM, Ichikawa L, Miglioretti DL, Eastman D. Web-based mammography audit feedback. AJR Am J Roentgenol 2012;198(6):W562-7.
7. Hysong SJ, Teal CR, Khan MJ, Haidet P. Improving quality of care through improved audit and feedback. Implement Sci 2012;7:45.
8. Hofvind S, Bennett RL, Brisson J, et al. Audit feedback on reading performance of screening mammograms: An international comparison. J Med Screen 2016;23(3):150-9.
9. Poot JD, Chetlen AL. A Simulation Screening Mammography Module Created for Instruction and Assessment: Radiology Residents vs National Benchmarks. Acad Radiol 2016;23(11):1454-62.
10. Cook AJ, Elmore JG, Zhu W, et al. Mammographic interpretation: radiologists' ability to accurately estimate their performance and compare it with that of their peers. AJR Am J Roentgenol 2012;199(3):695-702.
11. Scott HJ, Gale AG. Breast screening: PERFORMS identifies key mammographic training needs. Br J Radiol 2006;79 Spec No 2:S127-33.
12. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. JAMA 2007;298(9):1002-9.
13. Adcock KA. Initiative to Improve Mammogram Interpretation. Perm J 2004;8(2):12-8.
14. Berg WA, D'Orsi CJ, Jackson VP, et al. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? Radiology 2002;224(3):871-80.
15. Carney PA, Abraham L, Cook A, et al. Impact of an educational intervention designed to reduce unnecessary recall during screening mammography. Acad Radiol 2012;19(9):1114-20.
16. Ciatto S, Ambrogetti D, Morrone D, Del Turco MR. Analysis of the results of a proficiency test in screening mammography at the CSPO of Florence: review of 705 tests. Radiol Med 2006;111(6):797-803.
17. Geertse TD, Holland R, Timmers JM, et al. Value of audits in breast cancer screening quality assurance programmes. Eur Radiol 2015;25(11):3338-47.

18. Geller BM, Bogart A, Carney PA, et al. Educational interventions to improve screening mammography interpretation: a randomized controlled trial. AJR Am J Roentgenol 2014;202(6):W586-96.
19. Van der Horst F, Hendriks JH, Rijken H. Breast cancer screening in The Netherlands: audit and training of radiologists. Semin Breast Dis 2003;6:114-22.
20. Lee EH, Jun JK, Jung SE, Kim YM, Choi N. The efficacy of mammography boot camp to improve the performance of radiologists. Korean J Radiol 2014;15(5):578-85.
21. Lehman CD, Miller L, Rutter CM, Tsu V. Effect of training with the american college of radiology breast imaging reporting and data system lexicon on mammographic interpretation skills in developing countries. Acad Radiol 2001;8(7):647-50.
22. Linver MN, Paster SB, Rosenberg RD, Key CR, Stidley CA, King WV. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. Radiology 1992;184(1):39-43.
23. Luo P, Qian W, Romilly P. CAD-aided mammogram training. Acad Radiol 2005;12(8):1039-48.
24. Miglioretti DL, Gard CC, Carney PA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology 2009;253(3):632-40.
25. Mullen LA, Panigrahi B, Hollada J, Panigrahi B, Falomo ET, Harvey SC. Strategies for Decreasing Screening Mammography Recall Rates While Maintaining Performance Metrics. Acad Radiol 2017.
26. Raza S, Mackesy MM, Winkler NS, Hurwitz S, Birdwell RL. Effect of Training on Qualitative Mammographic Density Assessment. J Am Coll Radiol 2016;13(3):310-5.
27. Suleiman WI, Rawashdeh MA, Lewis SJ, et al. Impact of Breast Reader Assessment Strategy on mammographic radiologists' test reading performance. J Med Imaging Radiat Oncol 2016;60(3):352-8.
28. Timmers JM, van Doorne-Nagtegaal HJ, Verbeek AL, den Heeten GJ, Broeders MJ. A dedicated BI-RADS training programme: effect on the inter-observer variation among screening radiologists. Eur J Radiol 2012;81(9):2184-8.
29. Urban N, Longton GM, Crowe AD, et al. Computer-assisted mammography feedback program (CAMFP) an electronic tool for continuing medical education. Acad Radiol 2007;14(9):1036-42.
30. Netherlands Comprehensive Cancer Organisation (IKNL). National evaluation of breast cancer screening in the Netherlands. Available at: https://www.iknl.nl/docs/default-source/PDF_Docs/breast_cancer_screening_in_the_netherlands_2015_uk.pdf?sfvrsn=2. Accessed July 2017.
31. Soh BP, Lee W, Kench PL, et al. Assessing reader performance in radiology, an imperfect science: lessons from breast screening. Clin Radiol 2012;67(7):623-8.
32. Onega T, Anderson ML, Miglioretti DL, et al. Establishing a gold standard for test sets: variation in interpretive agreement of expert mammographers. Acad Radiol 2013;20(6):731-9.
33. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammography interpretations in community-based programs. J Natl Cancer Inst 2003;95(18):1384-93.
34. Rutter CM, Taplin S. Assessing mammographers' accuracy. A comparison of clinical and test performance. J Clin Epidemiol 2000;53(5):443-50.
35. Soh BP, Lee W, McEntee MF, et al. Screening mammography: test set data can reasonably describe actual clinical reporting. Radiology 2013;268(1):46-53.

Supplementary material

Appendix 1: Score form Medical Education Research Study Quality Instrument (MERSQI)

Domain	MERSQI Item	Score	Max Score
Study design	Single group cross-sectional or single group posttest only	1	3
	Single group pretest & posttest	1.5	
	Nonrandomized, 2 groups	2	
	Randomized controlled trial	3	
Sampling	*Institutions studied:*		3
	1	0.5	
	2	1	
	3	1.5	
	Response rate, %:		
	Not applicable		
	<50 or not reported	0.5	
	50-74	1	
	≥75	1.5	
Type of data	Assessment by participants	1	3
	Objective measurement	3	
Validity of evaluation instrument	*Internal structure:*		3
	Not applicable		
	Not reported	0	
	Reported	1	
	Content:		
	Not applicable		
	Not reported	0	
	Reported	1	
	Relationships to other variables:		
	Not applicable		
	Not reported	0	
	Reported	1	
Data analysis	*Appropriateness of analysis:*		3
	Inappropriate for study design or type of data	0	
	Appropriate for study design, type of data	1	
	Complexity of analysis:		
	Descriptive analysis only	1	
	Beyond descriptive analysis	2	
Outcomes	Satisfaction, attitudes, perceptions, opinions, general facts	1	3
	Knowledge, skills	1.5	
	Behaviors	2	
	Patient/health care outcome	3	
Total possible score*			**18**

*Scores range from 5 to 18. Adapted from Reed DA et al. Association between funding and quality of published medical education research. JAMA 2007;298:1002–9.

Copyright JGME, 2016 JGME Workshop "Personal Trainer Approach to Writing for Education Journals: Ready, Set Go. Presented by: *J Graduate Medical Education* Managing Editor

Radboud

Chapter 3

The dilemma of recalling well-circumscribed masses in a screening population: a narrative literature review and exploration of Dutch screening practice

Geertse TD, van der Waal D, Vreuls W, Tetteroo E, Duijm LEM, Pijnappel RM, Broeders MJM

The Breast, 2023; (69) 431-440

Abstract

Background

In Dutch breast cancer screening, solitary, new or growing well-circumscribed masses should be recalled for further assessment. This results in cancers detected but also in false positive recalls, especially at initial screening. The aim of this study was to determine characteristics of well-circumscribed masses at mammography and identify potential methods to improve the recall strategy.

Methods

A systematic literature search was performed using PubMed. In addition, follow-up data were retrieved on all 8860 recalled women in a Dutch screening region from 2014 to 2019.

Results

Based on 15 articles identified in the literature search, we found that probably benign well-circumscribed masses that were kept under surveillance had a positive predictive value (PPV) of 0-2%. New or enlarging solitary well-circumscribed masses had a PPV of 10-12%. In general the detected carcinomas had a favorable prognosis. In our exploration of screening practice, 25% of recalls (2133/8860) were triggered by a well-circumscribed mass. Those recalls had a PPV of 2.0% for initial and 10.6% for subsequent screening. Most detected carcinomas had a favorable prognosis as well.

Conclusion

To recognize malignancies presenting as well-circumscribed masses, identifying solitary, new or growing lesions is key. This information is missing at initial screening since prior examinations are not available, leading to a low PPV. Access to prior clinical examinations may therefore improve this PPV. In addition, given the generally favorable prognosis of screen-detected malignant well-circumscribed masses, one may opt to recall these lesions at subsequent screening, if grown, rather than at initial screening.

Introduction

Mammographic breast cancer screening in combination with state-of-the art treatment is still the most effective strategy for a substantial reduction in mortality from this disease. Detection at an earlier stage results in less invasive treatment and improved survival [1, 2]. If an abnormality suspicious for cancer is seen on the screening mammogram, the woman is recalled for further assessment. In the Netherlands, the recall rate continuously increased over the years [3, 4]. This resulted in an increase in the cancer detection rate, but also in a disproportionate increase in false positive recalls, especially at initial (or first) screening examinations (see figure in Appendix A) [4]. False positive recalls cause anxiety, a lower re-attendance rate, and additional costs [5-7].

The result of a screening examination is classified using the Breast Imaging Reporting And Data System (BI-RADS) lexicon [8]. BI-RADS category 0, adapted for use in a screening setting, represents an abnormality with a low suspicion for cancer. It is assigned to recalls related to well-circumscribed masses, architectural distortions seen in one direction, and asymmetries [8]. Half of all recalls in the Dutch screening program are classified as BI-RADS 0 (2019: 12.6 per 1000 of 23.9 per 1000 recalls [52.7%]) [4]. For these BI-RADS 0 recalls, the positive predictive value of recall (PPV) was found to be 10%. Thus, in 90% of these women there was no cancer diagnosed after further assessment. Of these women with a false positive recall, 15% underwent a diagnostic biopsy [4]. Although the PPV of the separate radiological features classified as BI-RADS 0 is unknown, well-circumscribed masses are very common in a screening population (approximately 8% of all screening mammograms [9]). Well-circumscribed masses with a typically benign appearance, such as typical intramammary nodes, hamartomas, and oily cysts, are easily recognized by screening radiologists and do not have to be recalled for further assessment. But the vast majority of the well-circumscribed masses are not typically benign and fall into the "probably benign" category. In this category, it is harder for screening radiologists to decide whether further assessment is necessary, knowing that the PPV is very low (<2%) [10, 11].

The Dutch recall strategy indicates that solitary, new, or growing well-circumscribed masses should be recalled (BI-RADS 0) for further assessment. To determine whether the number of false positive recalls can be reduced by improving the recall strategy, we need to better understand the clinical relevance of well-circumscribed masses. Although several studies have been published on this topic, to our knowledge, no review of the literature has been performed to combine all available evidence.

The aim of this study is to investigate the characteristics of malignancies presenting as well-circumscribed masses on mammography, in order to identify the potential room for improvement in the recall strategy, particularly for initial screening. We report the results of a narrative literature review and the follow-up results of all screening examinations assigned a BI-RADS 0 in a Dutch screening region in the period 2014 to 2019.

Materials and methods

Literature review

A systematic search was performed in April 2021, using PubMed (Appendix B), and updated in April 2022. Key search terms included: "mammography", "well-circumscribed mass", and variations of these terms. There were no restrictions regarding the type of journal or publication date. Articles written in a language other than English or Dutch were excluded. Titles and abstracts were screened to determine relevance. We reviewed the references of all relevant articles (snowballing) for additional ones. Articles were included if they met the following inclusion criteria: (1) the study population consisted of women (with or without symptoms) undergoing periodic mammography examination (for screening or follow-up of probably benign lesions); and (2) the outcome of assessment of the well-circumscribed masses presenting on the mammography examinations was reported (at least cancer or no cancer). We excluded studies if they focused on women with a mutation in one of the breast cancer susceptibility genes.

Exploration of screening practice

The screening organization in the south of the Netherlands provided data on all women who participated in the Eindhoven region between January 1, 2014, and January 1, 2019, and who were recalled based on a lesion on their screening mammogram. The anonymized data included information on the screening outcome and the clinical assessment (radiology, pathology, and surgical procedures performed at a hospital after the recall). By participating in screening, women consent to their data being made available for evaluation purposes and research, unless they choose to opt out explicitly. We did not receive any data of women who objected to the use of their data. This study was performed under the national permit for breast cancer screening issued by the Ministry of Health, Welfare and Sports and did not require additional approval by a local institutional review board.

Details of the Dutch national breast cancer screening program have been described previously [12-14]. In short, participating women get a two-view full-field digital mammogram (Lorad Selenia, Hologic). All mammograms are performed by a radiographer specialized in mammography. Each mammogram is read by two certified screening radiologists independently. Only for subsequent screening examinations, prior examinations are available for comparison. Mammograms are classified according to the BI-RADS lexicon [8]. BI-RADS 1 or 2 implies no recall, whereas women with a BI-RADS 0, 4, or 5 are recalled for clinical assessment. The screening program does not allow a BI-RADS 3 since no short-term follow-up is available in the screening setting.

Only women with a BI-RADS 0 recall based on a well-circumscribed mass, according to the recall letter, were included in the analyses. The outcome of the clinical assessment of these women was evaluated. Multiple foci of cancer in one breast were counted as one cancer. Due to the exploratory nature of this study, only descriptive statistics are presented here.

Results

Literature review

The search identified 164 articles (Figure 1). After title and abstract screening and reference checking, a total of 15 articles was included in the review. The update in March 2023 did not yield any additional articles.

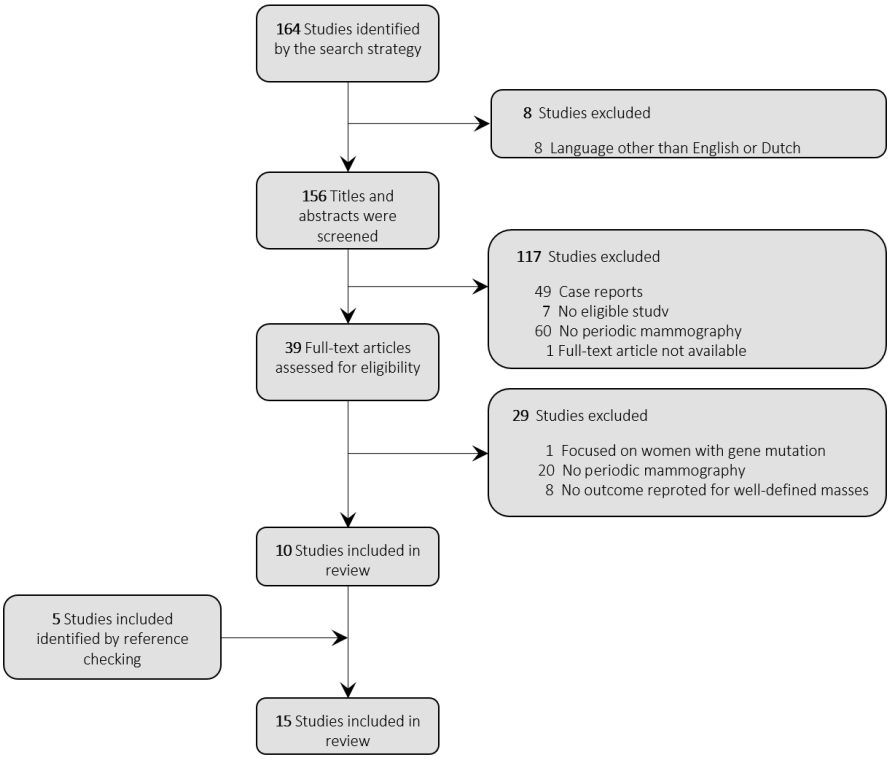


Figure 1: Flowchart of the literature search.

Table 1 shows the outcomes of the well-circumscribed masses described in the included articles. The study objectives of the selected articles were very heterogeneous. The number of cases related to well-circumscribed masses varied greatly, ranging from 24 to 1440. The number of cancers detected related to these well-circumscribed masses ranged from 0 to 91. The study populations consisted of women undergoing periodic mammography examination in the context of breast cancer screening (n=12 studies) [9, 17, 19-28] or surveillance for a probably benign lesion (n=3 studies) [15, 16, 18].

Four studies explicitly mentioned that all participants were asymptomatic [9, 21, 22, 26], four studies described in their method section that women with symptoms were included as well [15, 18 , 19, 24], and the other seven studies did not report information about symptoms [16, 17, 20, 23, 25, 27, 28].

In all studies, well-circumscribed masses generally had a benign outcome. The highest reported PPV was 17% [27]. This high PPV can probably be explained by the fact that only mammographically detected well-circumscribed masses that could also be detected on ultrasound or MRI were examined in this study. In the other studies, the PPV varied from 0% to 10.6%. To gain more insight into the clinical relevance of probably benign lesions, Sickles distinguished solitary from multiple well-circumscribed masses [15]. The PPV was 2.0% for solitary well-circumscribed masses and 0.4% for multiple masses. Later, Leung conducted a study together with Sickles to assess the need to recall women with multiple masses [21]. For this study they included a different population than in the previous study by Sickles [15]. They again found a PPV of 0.4%. Sickles [15] also described that the PPV increased to 11.5% when a well-circumscribed mass changed over time or became palpable. Opie et al. [17] reported that 10.6% of solid, enlarging well-circumscribed masses were malignant. Timmers et al. [25] reported similar findings: a PPV of 10% for well-circumscribed masses. This study was performed within the Dutch national breast cancer screening program and included only subsequent screening examinations. Most of the well-circumscribed masses in this study were therefore solitary, new or growing. In a study by Burrell et al. [19] the population was stratified into asymptomatic and symptomatic women. For the well-circumscribed masses, the PPV was 0% in asymptomatic women and 4.2% in symptomatic women. It should be noted that were only 24 mammography examinations showing a well-circumscribed mass.

Only eight studies [9, 18, 19, 21-23, 27, 28] reported the type of cancer related to the well-circumscribed masses. The study by Dhillon et al. [22] only included mucinous carcinomas. Due to the low prevalence, the number of cancers diagnosed among well-circumscribed masses in the smaller studies was very low. The larger studies [9, 18, 23, 27] reported 19, 91, 18, and 19 breast cancers, respectively, of which at least 50% were invasive breast carcinoma of no special type (IBC-NST) (16/19 [85%], 49/91 [54%], 9/18 [50%], and 10/19 [53%], respectively). Besides IBC-NST, several of these four larger studies also mentioned mucinous (0/19 [0%], 13/91 [14%], 5/18 [28%], and 1/19 [5%], respectively), papillary (0/19 [0%], 0/91 [0%], 1/18 [6%], and 3/19 [16%], respectively) and ductal carcinoma in situ (DCIS) (3/19 [16%], 25/91 [27%], 1/18 [6%], and 2/19 [11%], respectively) as cancer type. Farshid et al. [9] found that many of the DCIS cases had a papillary component.

Table 1: Characteristics of well-circumscribed masses of the selected articles

Reference	Study objective	Country	Population	Age	Sample size
Sickles, 1991 [15]	Establish the validity of managing probably benign lesions with periodic mammographic surveillance	USA	Women (asymptomatic or symptomatic) who underwent periodic mammographic surveillance for a probably benign lesion (during a 8.5-year period)	Range 28-96, Median 51	3,184 women
Datoc, 1991 [16]	Compare the efficacy of single-view and two-view examinations for the follow-up of mammographic findings associated with low suspicion for malignancy	USA	Women who underwent periodic mammographic surveillance for a probably benign lesion (6 months follow-up)	Range 26-81, Mean 53.7	498 women with a total of 666 mammographic abnormalities
Opie, 1993 [17]	Determine the yield of carcinoma in patients with a nonpalpable mammographic abnormality and identify which mammographic criteria will most likely yield a positive biopsy	USA	Women participating in screening who had a nonpalpable abnormality detected and biopsied	Range 24-86	295 women who underwent 332 biopsies
Sickles, 1994 [18]	Determine whether lesion size and patient age should prompt immediate biopsy of nonpalpable, circumscribed, noncalcified solid breast masses	USA	Women (asymptomatic or symptomatic) who underwent periodic mammographic surveillance for a probably benign lesion (during a 12.3-year[2] period)	Range 28-94, Median 50	58,415 mammograms (a woman can have more than one mammogram)
Burrell, 1996 [19]	Identify factors which may improve sensitivity and specificity of mammographic interpretation	UK	Women (asymptomatic or symptomatic) participating in screening who had a nonpalpable abnormality detected and biopsied	Range 30-75, Mean 55	416 women who underwent 425 biopsies (303 asymptomatic + 122 symptomatic)

1. PPV=11.5%, when a lesion changes on mammography or becomes palpable. No cancers were found among biopsy without mammographic change.
2. This 12.3-year period includes the 8.5-year period of the study of 1991.

Cases well-circumscribed masses	Number of breast cancers	Proportion of breast cancers	Type of breast cancer	Histological grade	Stage	Lymph node status	Receptor status
842 589 (one) 253 (multiple)	13 12 1	1.5%[1] 2.0% 0.4%	Only reported for all probably benign lesions combined	NR	Only reported for all probably benign lesions together	One demonstrated axillar lymph node metastasis	NR
314 masses of 666 abnormalities	0	0%	NA	NA	NA	NA	NA
47 masses	5	10.6%	Only reported for all lesions combined	NR	Only reported for all lesions together	NR	NR
1,403	19	1.4%	16 IBC-NST (84%) 3 DCIS (16%)	NR	3 stage 0 (16%) 14 stage I (74%) 2 stage II (11%)	One demonstrated axillar lymph node metastasis	NR
24	Asymptomatic 0 Symptomatic 1	Asymptomatic 0% Symptomatic 4.2%	1 Intracystic carcinoma (100%)	Only reported for all lesions together	Only reported for all lesions together	NR	NR

Table 1: Continued

Reference	Study objective	Country	Population	Age	Sample size
Hussain, 1999 [20]	Assess the nature of new densities and microcalcifications in the second round of breast screening	UK	Women participating in screening (2nd round), with abnormalities not present in 1st round	Range 50-64	311 lesions identified in 302 women
Leung, 2000 [21]	Assess the need for recalling women with multiple masses	USA	Women (asymptomatic) participating in screening with multiple bilateral masses	NR	84,615 examinations of 40,419 women
Dhillon, 2006 [22]	Describe the imaging features of 34 screen-detected mucinous carcinomas	Australia	Women (asymptomatic) participating in screening and with a screen-detected mucinous carcinoma	Range 48-82 Mean 65	214,507 women 2745 invasive cancers 45 mucinous cancers (11 mucinous cancers were excluded, 34 were described)
Farshid, 2008 [9]	Establish the reliability of FNAB as a first line diagnostic modality for assessment of category 3 screen-detected mass lesions	Australia	Women (asymptomatic) participating in screening and with a category 3B[3] solid circumscribed mass	Range 50-69	1,183 lesions (538 initial screening, 645 subsequent screening)
Bonetti, 2008 [23]	Confirm that FNAC is a reliable first diagnostic tool for the assessment of breast lesions	Italy	Women participating in screening and with a category 3B5 solid circumscribed mass	NR	388 lesions

3. According to the Tabar 5-tier grading scheme.
4. Not included here: 1 LCIS, 1 Leiomyosarcoma, 3 Lymphoma, 1 Metastasis.
5. In 2011 Farshid et al. published an extension of this study from 2008 [26], in which these percentages are mentioned.
6. Numbers are not reported.
7. Not included here: 1 LCIS
8. It was only reported that no highly aggressive tumors were observed in the series.

Cases well-circumscribed masses	Number of breast cancers	Proportion of breast cancers	Type of breast cancer	Histological grade	Stage	Lymph node status	Receptor status
53	2	3.8%	Only reported for all lesions together	Only reported for all lesions together	NR	Only reported for all lesions together	NR
1440 examinations with multiple masses among 907 women	4	0.4%	3 IBC-NST (75%) 1 Mucinous (25%)	2 Grade 1 (50%) 1 Grade 2 (25%) 1 Grade 3 (25%)	3 stage I (75%) 1 stage IIa (25%)	3 negative (75%) 1 positive (25%)	NR
30	30	NA	Mucinous	30 Grade 1 or 2 (100%)	NR	30 negative (100%)	NR
1,183	91[4]	7.7% (3% initial screening 13% subsequent screening)[5]	49 IBC-NST (54%) 13 Mucinous (14%) 2 Tubular (2%) 1 Medullary (1%) 1 Inv. Lobular (1%) 25 DCIS (27%)	Invasive[6]: 49.2% Grade 1 33.3% Grade 2 17.5% Grade 3 DCIS: 16.0% high grade	NR	NR	NR
388	18[7]	4.6%	9 IBC-NST (50%) 5 Mucinous (28%) 1 Medullary (6%) 1 Inv. Lobular (6%) 1 Inv. Papillary (6%) 1 DCIS (6%)	NR[8]	NR	NR	NR

Table 1: Continued

Reference	Study objective	Country	Population	Age	Sample size
Bandan, 2013 [24]	Evaluate BI-RADS as a predictive factor for suspicion of malignancy in breast lesions by correlating radiological findings with histological results in a breast cancer reference hospital	Brazil	Women (asymptomatic or symptomatic) participating in screening and recalled for FNAB, core biopsy or vacuum-assisted core biopsy	Range 16-84, Mean 49	580 women (recalls) 276 BI-RADS 3[9] 230 BI-RADS 4 74 BI-RADS 5
Timmers, 2013 [25]	Develop a prediction model for breast cancer based on common mammographic findings on screening mammograms, aiming to reduce reader variability in assigning BI-RADS	Netherlands	Women participating in subsequent screening who were recalled	Range 53-75 Mean 62	352 women (recalls) 120 BI-RADS 0 198 BI-RADS 4 34 BI-RADS 5
McDonald, 2017 [26]	Evaluate BI-RADS 3 assessment after recall from screening, before and after implementation of DBT Cohort 1: FFDM Cohort 2: DBT +FFDM	USA	Women (asymptomatic) participating in screening (without symptoms or physical examination findings and no prior history of breast cancer) and recalled	Range <40 - >69 App. 85% 40-69 Cohort 1: Mean 54.2 Cohort 2: Mean 53.8	Cohort 1: 184 BI-RADS 3 Cohort 2: 227 BI-RADS 3
Nakashima, 2017 [27]	Compare the visibility of circumscribed masses on DBT images and 2D mammograms and determine the usefulness of DBT for differentiation between benign and malignant circumscribed masses	Japan	Women participating in screening who were recalled	Malignant lesions: Mean 61 Benign lesions: Mean 53	1395 women (recalls)
Stepanek, 2019 [28]	Compare the utilization of BI-RADS 3 assessment after recall from screening before and after implementation of DBT Cohort 1: FFDM Cohort 2: DBT +FFDM	USA	Women participating in screening who were recalled	Range <40 - >69 App. 85% 40-69	BI-RADS 3 Cohort 1: 388 women (463 lesions) Cohort 2: 220 women (254 lesions)

Abbreviations: NR = not reported; NA = not available; IBC-NST = invasive breast carcinoma of no special type; DCIS = ductal carcinoma in situ; LCIS = lobular carcinoma in situ; Inv.= invasive; FNAB = fine needle aspiration biopsy; FNAC = fine needle aspiration cytology; BI-RADS = Breast Imaging Reporting And Data System; FFDM = full field digital mammography; DBT = digital breast tomosynthesis; ER=estrogen receptor; PR=progesterone receptor; HER2=human epidermal growth factor receptor 2

9. Nodules with circumscribed margins, clustered punctiform microcalcifications, and focal asymmetry without associated findings.

Cases well-circumscribed masses	Number of breast cancers	Proportion of breast cancers	Type of breast cancer	Histological grade	Stage	Lymph node status	Receptor status
248	2	0.80%	Only reported for all lesions together	NR	NR	NR	NR
60	6	10%	NR	NR	NR	NR	NR
Cohort 1: 41 Cohort 2: 61	Cohort 1: 1 Cohort 2: 0	Cohort 1: 2.4% Cohort 2: 0%	1 IBC-NST (100%)	NR	NR	NR	NR
115	19	17%	10 IBC-NST (53%) 2 DCIS (11%) 3 Papillary (16%) 2 Metaplastic (11%) 1 Mucinous (5%) 1 Phyllodes (5%)	NR	NR	NR	NR
Cohort 1: 83 Cohort 2: 47	Cohort 1: 2 Cohort 2: 1	Cohort 1: 2.4% Cohort 2: 2.1%	1 DCSI (50%) 1 IBC-NST (50%) 1 Papillary (100%)	NR	1 Stage 0 (50%) 1 Stage IA (50%) 1 Stage 0 (100%)	Negative Negative Negative	ER+PR+ ER-PR-HER2- ER+PR+

Only two of the larger studies [9, 18] reported prognostic factors: histological grade, tumor stage, lymph node status, or receptor status. In the study by Sickles [18], 88% (14/16) of invasive cancers were stage I, and an axillary lymph node metastasis was found in only one case (6%). In the study by Farshid et al. [9] grade 1 tumors were most common, accounting for 49.2% (31/63) of the cases. Furthermore, 33.3% (21/63) were grade 2, and only 17.5% (11/63) of the cancers were grade 3. Grade was not specified for three cases. Of the 25 DCIS cases that presented as a well-circumscribed mass, only four were of high grade. The 30 mucinous carcinomas described in the study by Dhillon et al. [22] were all grade 1 or 2 and all had a negative lymph node status.

Exploration of screening practice

Of the 8860 recalled women in the database we used to explore screening outcomes, 4574 (51.6%) were recalled for a BI-RADS 0 lesion (see Figure 2). Of these BI-RADS 0 lesions, 2133 (46.6%) presented as a well-circumscribed mass on the screening mammogram. A total of 179 (8.0%) women were diagnosed with breast cancer after recall for a well-circumscribed mass, 18 cancers at initial screening (PPV: 18/696 [2.6%]) and 161 cancers at subsequent screening (PPV: 161/1437 [11.2%]). In 13 of these 179 women, the cancer appeared to be unrelated to the well-circumscribed mass. Of these 13 women, eight women had a bilateral recall, with breast cancer diagnosed in the contralateral breast from where the well-circumscribed mass was detected. In the remaining five women, a malignancy was discovered elsewhere in the recalled breast (an incidental finding). These 13 women were excluded from our analyses. At subsequent screening, 152 breast cancers diagnoses were related to a well-circumscribed mass (PPV: 152/1437 [10.6%]). At initial screening, breast cancer related to a well-circumscribed mass was diagnosed in 14 of 696 screening examinations (PPV: 2.0%).

Table 2 shows the details of the 166 breast cancers related to a well-circumscribed mass. Most cancers were IBC-NST (initial screening: 10 of 14 cancers [71.4%]; subsequent screening: 107 of 152 cancers [70.1%]). We further observed invasive lobular carcinomas, but only in subsequent screening examinations (initial screening: 0 of 14 cancers [0%]; subsequent screening: 17 of 152 cancers [11.2%]). Of the rare subtypes of invasive cancers, mucinous (initial screening: 0 of 14 cancers [0%]; subsequent screening: 5 of 152 cancers [3.3%]) and tubular carcinomas (initial screening: 2 of 14 cancers [14.3%]; subsequent screening: 7 of 152 cancers [4.6%]) were most frequently encountered. DCIS was found in 10 women (initial screening: 2 of 14 women [14.3%]; subsequent screening: 8 of 152 women [5.3%]), 9 of these 10 women (90%) had a well-circumscribed mass such as fibroadenoma (n=2) or papillary lesion (n=7) with DCIS as an additional finding on biopsy.

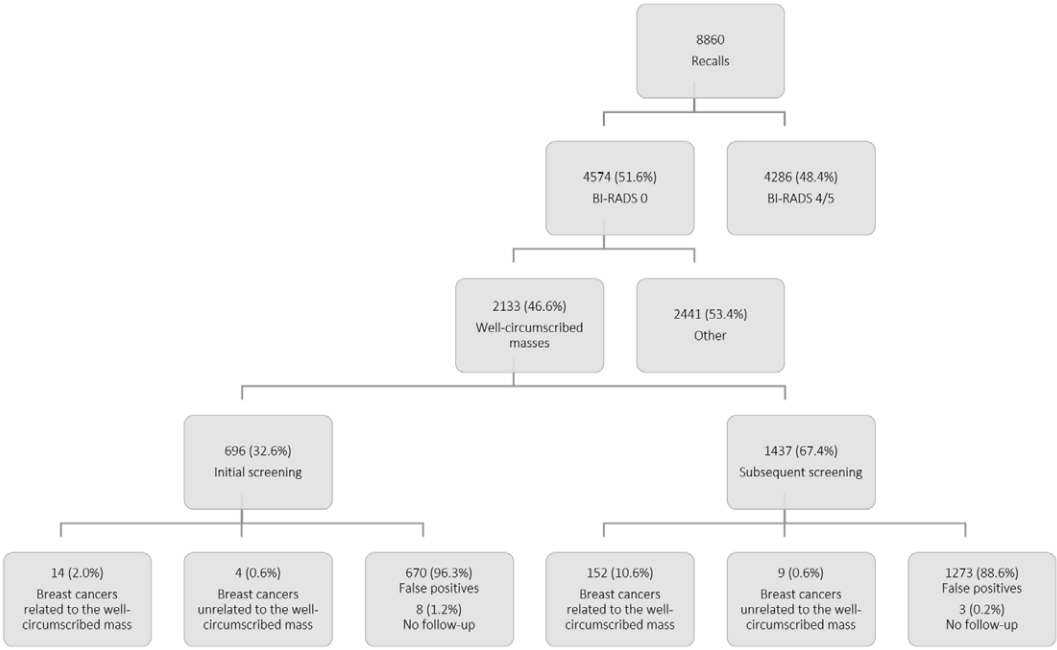


Figure 2: Screening mammography results of women recalled with BI-RADS 0 based on a well-circumscribed mass in a Dutch screening region in het period 2014-2019

The malignant well-circumscribed masses generally comprised cancers with a favorable prognosis. Most cancers were grade I or II (initial screening: 12 of 12 cancers [100%]; subsequent screening: 129 of 144 cancers [89.6%]) and had a negative lymph node status (initial screening: 8 of 12 cancers [66.7%]; subsequent screening: 115 of 144 cancers [79.9%]). The hormone receptor status of the invasive cancers was predominantly ER+, PR+ or - , and HER2- (initial screening: 11 of 12 cancers [92%]; subsequent screening: 128 of 144 cancers [88.9%]). HER2+ breast cancer was found in 1 of 12 women [8.3%] at initial screening and in 9 of 144 women [6.3%] at subsequent screening. Triple negative breast cancers were rare (initial screening: 0 of 12 women [0%]; subsequent screening: 7 of 144 women [4.9%]).

Table 2: Tumor characteristics of 166 screen-detected cancers presenting as a well-circumscribed mass at screening mammography

	Initial screening	Subsequent screening
Type of carcinomas	**n=14**	**n=152**
DCIS	2 (14.3%)	8 (5.3%)
IBC-NST	10 (71.4%)	107 (70.1%)
ILC	0 (0%)	17 (11.2%)
mixed IBC- NST/ILC	0 (0%)	5 (3.3%)
Inv tubular	2 (14.3%)	7 (4.6%)
Inv papillary	0 (0%)	1 (0.7%)
Inv mucinous	0 (0%)	5 (3.3%)
Inv mucinous / papillary	0 (0%)	1 (0.7%)
Inv neuroendocrine	0 (0%)	1 (0.7%)
Tumor size	**n=14**	**n=152**
Tis	2 (14.3%)	8 (5.3%)
T1A	2 (14.3%)	19 (12.5%)
T1B	7 (50.0%)	64 (42.1%)
T1C	2 (14.3%)	47 (30.9%)
T2	0 (0%)	11 (7.2%)
T3+	1 (7.1%)	3 (2.0%)
Grading		
DCIS:	**n=2**	**n=8**
Low	2 (100%)	4 (50.0%)
Intermediate	0 (0%)	3 (37.5%)
High	0 (0%)	1 (12.5%)
Bloom&Richardson:	**n=12**	**n=144**
I	7 (58.3%)	70 (48.6%)
II	5 (41.7%)	59 (41.0%)
III	0 (0%)	15 (10.4%)
Receptor status	**n=12**	**n=144**
ER+, PR+/-, HER2 -	11 (91.7%)	128 (88.9%)
ER+/-, PR+/-, HER2+	1 (8.3%)	9 (6.3%)
Triple negative	0 (0%)	7 (4.9%)
Lymph node status	**n=12**	**n=144**
Negative	8 (66.7%)	115 (79.9%)
Positive	3 (25.0%)	21 (14.6%)
Nx	1 (8.3%)	8 (5.6%)

Abbreviations: DCIS=ductal carcinoma in situ; IBC-NST= invasive breast carcinoma of no special type; ILC=invasive lobular carcinoma; Inv=invasive; ER=estrogen receptor; PR=progesterone receptor; HER2=human epidermal growth factor receptor 2; Nx=cannot be measured.

Discussion

The narrative literature review showed that probably benign well-circumscribed masses at mammography had a PPV of 0-2%. When limited to new or growing well-circumscribed masses, the PPV increased to 10-12%. In general, the cancers detected had a favorable prognosis. Our exploratory study showed that almost 25% of all recalls were triggered by a well-circumscribed mass on the screening mammogram. We found a PPV of 2.0% for initial screening examinations, 10.6% for subsequent screening examinations, and 8.0% for all screening examinations combined. Thus, the majority of well-circumscribed masses were benign, especially for initial screening examinations (98.0%). In addition, and in line with the literature review, most cancers detected had a favorable prognosis.

The Dutch study by Timmers et al. [25] reported a PPV of 10% for subsequent screening examinations. Of all other studies included in our literature review, the recall strategy in the study by Farshid et al. [9], performed within the Australian breast cancer screening program, most closely resembles the recall strategy of the Dutch breast cancer screening program. The authors reported an overall PPV of 8% for well-circumscribed masses. In an extension of this study, Farshid et al. [29] found that the PPV was 3% for initial screening examinations and 13% for subsequent screening examinations, which is quite similar to results from Dutch screening practice. Unlike the recall policy in the Netherlands, in the USA all probably benign well-circumscribed masses are recalled and assigned a BI-RADS 3, for which short-term surveillance is recommended. A PPV of 0-2% [15, 16, 18, 26, 28] has been reported for this setting. During surveillance, a morphological changes or an increase in size is an indication for needle biopsy, resulting in a PPV of biopsy of 10-12% [15, 17].

The difference in PPV between all probably benign well-circumscribed masses (0-2%) and those that are new or enlarging (10-12%) may at least partly explain the distinct difference between the PPV at initial (2.0%) versus subsequent (10.6%) screening examinations. The radiologists have no prior examinations to compare with during reading of initial screening examinations, resulting in the recall of more probably benign well-circumscribed masses. For subsequent screening examinations, the radiologists have prior examinations to compare with, which makes it possible to only recall new or enlarging well-circumscribed masses. This is true for the majority of the Dutch breast cancer screening population, because in the Netherlands the re-attendance rate is 91% [4].

The low PPV at initial screening examinations (2.0%) suggests that the balance between screen-detected cancers and false positive recalls is unfavorable. This balance could potentially be improved if a prior mammogram is available for comparison. Several studies have shown that, in breast cancer screening, the availability of prior mammograms for comparison reduces the false positive rate [30-34]. To our knowledge, no previous study has focused on initial screening. The extent to which prior mammograms could be made available at initial screening depends on how screening is organized. Most likely, women will have to give their consent for using their prior clinical mammograms for comparison. This can be facilitated by making women aware of the importance of providing their prior mammograms. A survey study by Horsley et al. reports that, even in a group of women who routinely underwent screening mammography, most women did not think that prior mammograms are important to decrease false positive recalls [35]. It is known that in the Netherlands a large proportion of women have had a mammogram in a clinical setting before reaching the screening starting age. Data from the Netherlands Institute for Health Services Research (NIVEL) show that yearly an estimated 1 in 50 women over the age of 25 have an appointment in the hospital because of fear of having breast cancer or breast problems [36, 37]. In the Netherlands further assessment of recalled participants is performed in a hospital and is not part of the screening program. Privacy legislation can therefore be an obstacle in retrieving both medical history and/or clinical mammograms for comparison in screening.

In general, the pathological characteristics of cancers found in an asymptomatic screening population differ from symptomatic and interval cancers [38, 39]. The poorer prognosis for interval cancers seem to be associated with their biological differences and more rapid tumor growth. The latter means that the preclinical detectable phase of high-grade carcinomas is often too short to be detected during screening. As a consequence, in screening in particular low-grade, slower growing carcinomas are detected, which could explain the mostly favorable prognosis of the screen-detected, malignant, well-circumscribed masses in our study. Given this generally favorable prognosis and the high re-attendance rate of 91% [4], it might be possible to wait and recall these lesions at subsequent screening, if grown, rather than recalling them at initial screening.

For the few rapidly growing and more aggressive carcinomas that present as a well-circumscribed mass at the time of screening, we need to find a mammographic feature that is able to identify these cancers and avoid a delay in detection. It is quite conceivable, that in the coming years new artificial intelligence algorithms will be developed which can help radiologists to identify these cancers.

Our study has several strengths and limitations. An important strength of this study is that it combines a literature review with an exploration of actual screening practice based on a large sample size. An important limitation of this study is that data on well-circumscribed masses was scarce and did not allow us to draw strong conclusions based on the few, mostly small, studies, identified by the literature search. In addition, for the exploratory study, the presence of a well-circumscribed mass was only based on the description in the recall letters, drafted by the screening radiologists, and could not be based on radiological review of the mammograms.

Conclusions

To recognize malignancies presenting as well-circumscribed masses, identifying solitary, new or growing lesions is key. This information is missing at initial screening since prior examinations are not available, leading to a low PPV. Access to prior clinical examinations may therefore improve this PPV. In addition, given the generally favorable prognosis of screen-detected, malignant, well-circumscribed masses, one may opt to recall these lesions at subsequent screening, if grown, rather than at initial screening.

Declaration of competing interest

The authors of this manuscript certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

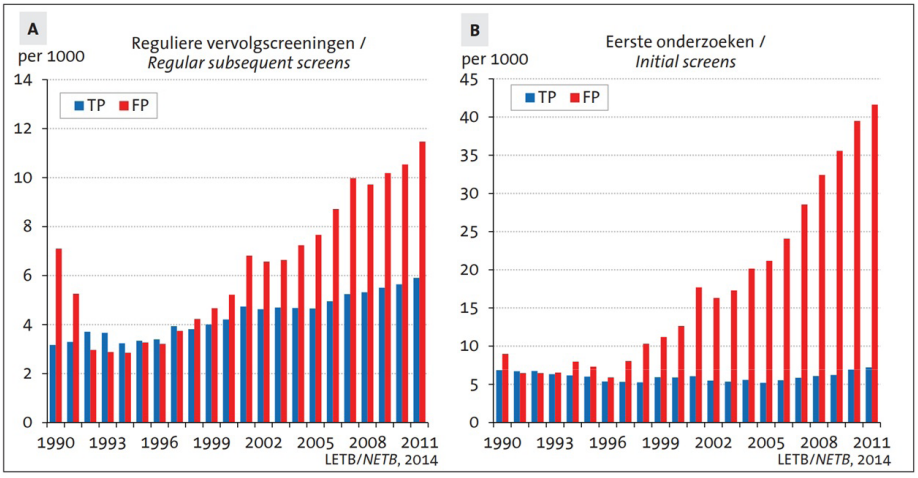
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

This study was performed under the national permit for breast cancer screening issued by the Ministry of Health, Welfare and Sports and did not require additional approval by a local institutional review board.

Supplementary material

Appendix A. True-positive (TP) and false-positive (FP) screen results per 1000 women screened from the Dutch breast cancer screening 1990-2011

Source: Figure 3.9 from the report "National evaluation of breast cancer screening in the Netherlands 1990 – 2011/2012: Thirteenth evaluation report", published by the National Evaluation Team for Breast cancer screening in April 2014



Figuur 3.9 Terecht-positieve (TP) en fout-positieve (FP) screeningsuitslagen per 1000 voor **A:** reguliere vervolgscreeningen, en **B:** eerste screeningsonderzoeken, 1990-2011

*Figure 3.9 True-positive (TP) and false-positive (FP) screen results per 1000 women screened for **A:** regular subsequent, and **B:** initial screens, 1990-2011*

Appendix B: Search Strategy

We performed the following search strategies in PubMed:
Component 1=breast cancer
Component 2=mammography
Component 3=well-defined mass

(1) (Breast Neoplasm*[tiab] OR Breast Tumo*[tiab] OR Breast Cancer*[tiab] OR Mammary Cancer*[tiab] OR Malignant Neoplasm of Breast[tiab] OR Breast Malignant Neoplasm*[tiab] OR Malignant Tumor of Breast[tiab] OR Breast Malignant Tumo*[tiab] OR Cancer of Breast[tiab] OR Cancer of the Breast[tiab] OR Mammary Carcinoma*[tiab] OR Mammary Neoplasm*[tiab] OR Breast Carcinoma*[tiab])

OR

(2) ("Mammography"[Mesh] OR Mammogra*[tiab] OR Digital Mammogra*[tiab])

AND

(3) (circumscribed mass*[tiab] OR well circumscribed mass*[tiab] OR well-circumscribed mass*[tiab] OR well defined mass*[tiab] OR well-defined mass*[tiab])

References

1. Dibden A, Offman J, Duffy SW, Gabe R. Worldwide Review and Meta-Analysis of Cohort Studies Measuring the Effect of Mammography Screening Programmes on Incidence-Based Breast Cancer Mortality. Cancers (Basel). 2020 Apr 15;12(4):976. https://doi.org/10.3390/cancers12040976.
2. 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 Jun 11;108(11):2205-40. https://doi.org/10.1038/bjc.2013.177.
3. National Evaluation Team for Breast cancer screening. National evaluation of breast cancer screening in the Netherlands 1990 – 2011/2012: Thirteenth evaluation report, https://www.rivm.nl/sites/default/files/2018-11/LETB%20XIII%20Definitief%20%28web%29.pdf; 2014 [accessed 7 December 2022]
4. IKNL Monitor 2019, published September 2021, Available from: https://iknl.nl/getmedia/03f70b9e-bc75-4e07-b099-2105741a0c8b/Monitor-bevolkingsonderzoek-borstkanker-2019.pdf [accessed 7 December 2022]
5. Bond M, Pavey T, Welch K, Cooper C, Garside R, Dean S, Hyde C. Systematic review of the psychological consequences of false-positive screening mammograms. Health Technol Assess. 2013 Mar;17(13):1-170, v-vi. https://doi.org/10.3310/hta17130.
6. Setz-Pels W, Duijm LE, Coebergh JW, Rutten M, Nederend J, Voogd AC. Re-attendance after false-positive screening mammography: a population-based study in the Netherlands. Br J Cancer. 2013 Oct 15;109(8):2044-50. https://doi.org/10.1038/bjc.2013.573.
7. 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 Aug;121(4):351-358. https://doi.org/10.1038/s41416-019-0524-4. Erratum in: Br J Cancer. 2021 Sep;125(7):1031.
8. Timmers JM, van Doorne-Nagtegaal HJ, Zonderland HM, van Tinteren H, Visser O, Verbeek AL, et al.. The Breast Imaging Reporting and Data System (BI-RADS) in the Dutch breast cancer screening programme: its role as an assessment and stratification tool. Eur Radiol. 2012 Aug;22(8):1717-23. https://doi.org/10.1007/s00330-012-2409-2.
9. Farshid G, Downey P, Gill P, Pieterse S. Assessment of 1183 screen-detected, category 3B, circumscribed masses by cytology and core biopsy with long-term follow up data. Br J Cancer. 2008 Apr 8;98(7):1182-90. https://doi.org/10.1038/sj.bjc.6604296.
10. Berment H, Becette V, Mohallem M, Ferreira F, Chérel P. Masses in mammography: what are the underlying anatomopathological lesions? Diagn Interv Imaging. 2014 Feb;95(2):124-33. https://doi.org/10.1016/j.diii.2013.12.010.
11. BI-RADS Committee. ACR BI-RADS atlas: Breast Imaging Reporting And Data System. 5th ed. Reston, Va: American College of Radiology, 2013.
12. Luiten JD, Voogd AC, Luiten EJT, Broeders MJM, Roes KCB, Tjan-Heijnen VCG, Duijm LEM. Recall and Outcome of Screen-detected Microcalcifications during 2 Decades of Mammography Screening in the Netherlands National Breast Screening Program. Radiology. 2020 Mar;294(3):528-537. https://doi.org/10.1148/radiol.2020191266.
13. Sankatsing VDV, van Ravesteyn NT, Heijnsdijk EAM, Looman CWN, van Luijt PA, Fracheboud J, et al. The effect of population-based mammography screening in Dutch municipalities on breast cancer mortality: 20 years of follow-up. Int J Cancer. 2017 Aug 15;141(4):671-677. https://doi.org/10.1002/ijc.30754.

14. Otten JDM, Fracheboud J, den Heeten GJ, Otto SJ, Holland R, de Koning HJ, et al. Likelihood of early detection of breast cancer in relation to false-positive risk in life-time mammographic screening: population-based cohort study. Ann Oncol. 2013 Oct;24(10):2501-2506. https://doi.org/10.1093/annonc/mdt227.

15. Sickles EA. Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. Radiology. 1991 May;179(2):463-8. https://doi.org/10.1148/radiology.179.2.2014293.

16. Datoc PD, Hayes CW, Conway WF, Bosch HA, Neal MP. Mammographic follow-up of nonpalpable low-suspicion breast abnormalities: one versus two views. Radiology. 1991 Aug;180(2):387-91. https://doi.org/10.1148/radiology.180.2.2068300.

17. Opie H, Estes N, Jewell W, Chang C, Thomas J, Estes M. Breast Biopsy for Nonpalpable Lesions: A Worthwhile Endeavor? Am Surg 1993;59:490–4.

18. Sickles EA. Nonpalpable, circumscribed, noncalcified solid breast masses: likelihood of malignancy based on lesion size and age of patient. Radiology. 1994 Aug;192(2):439-42. https://doi.org/10.1148/radiology.192.2.8029411.

19. Burrell HC, Pinder SE, Wilson AR, Evans AJ, Yeoman LJ, Elston CW, Ellis IO. The positive predictive value of mammographic signs: a review of 425 non-palpable breast lesions. Clin Radiol. 1996 Apr;51(4):277-81. https://doi.org/10.1016/s0009-9260(96)80346-1.

20. Hussain HK, Ng YY, Wells CA, Courts M, Nockler IB, Curling OM, Carpenter R, Perry NM. The significance of new densities and microcalcification in the second round of breast screening. Clin Radiol. 1999 Apr;54(4):243-7. https://doi.org/10.1016/s0009-9260(99)91159-5.

21. Leung JW, Sickles EA. Multiple bilateral masses detected on screening mammography: assessment of need for recall imaging. AJR Am J Roentgenol. 2000 Jul;175(1):23-9. https://doi.org/10.2214/ajr.175.1.1750023.

22. Dhillon R, Depree P, Metcalf C, Wylie E. Screen-detected mucinous breast carcinoma: potential for delayed diagnosis. Clin Radiol. 2006 May;61(5):423-30. https://doi.org/10.1016/j.crad.2005.10.008.

23. Bonetti F, Manfrin E. 3B circumscribed masses: to assess or not to assess? Br J Cancer. 2008 Aug 19;99(4):675-6; author reply 677-8. https://doi.org/10.1038/sj.bjc.6604500.

24. Badan GM, Júnior DR, Alberto C, Ferreira P, Augusto F, Ferreira T, et al. Positive predictive values of Breast Imaging Reporting and Data System (BI-RADS ®) categories 3, 4 and 5 in breast lesions submitted to percutaneous biopsy. Radiol Bras. 2013 Aug;46(4):209–13.

25. Timmers JM, Verbeek AL, In 't Hout J, Pijnappel RM, Broeders MJ, den Heeten GJ. Breast cancer risk prediction model: a nomogram based on common mammographic screening findings. Eur Radiol. 2013 Sep;23(9):2413-9. https://doi.org/10.1007/s00330-013-2836-8.

26. McDonald ES, McCarthy AM, Weinstein SP, Schnall MD, Conant EF. BI-RADS Category 3 Comparison: Probably Benign Category after Recall from Screening before and after Implementation of Digital Breast Tomosynthesis. Radiology. 2017 Dec;285(3):778-787. https://doi.org/10.1148/radiol.2017162837.

27. Nakashima K, Uematsu T, Itoh T, Takahashi K, Nishimura S, Hayashi T, Sugino T. Comparison of visibility of circumscribed masses on Digital Breast Tomosynthesis (DBT) and 2D mammography: are circumscribed masses better visualized and assured of being benign on DBT? Eur Radiol. 2017 Feb;27(2):570-577. https://doi.org/10.1007/s00330-016-4420-5.

28. Stepanek T, Constantinou N, Marshall H, Pham R, Thompson C, Dubchuk C, Plecha D. Changes in the Utilization of the BI-RADS Category 3 Assessment in Recalled Patients Before and After the Implementation of Screening Digital Breast Tomosynthesis. Acad Radiol. 2019 Nov;26(11):1515-1525. https://doi.org/10.1016/j.acra.2018.12.020.
29. Farshid G, Walker A, Battersby G, Sullivan T, Gill PG, Pieterse S, Downey P. Predictors of malignancy in screen-detected breast masses with indeterminate/equivocal (grade 3) imaging features. Breast. 2011 Feb;20(1):56-61. https://doi.org/10.1016/j.breast.2010.07.002.
30. Bassett LW, Shayestehfar B, Hirbawi I. Obtaining previous mammograms for comparison: usefulness and costs. AJR Am J Roentgenol. 1994 Nov;163(5):1083-6. https://doi.org/10.2214/ajr.163.5.7976879.
31. Callaway MP, Boggis CR, Astley SA, Hutt I. The influence of previous films on screening mammographic interpretation and detection of breast carcinoma. Clin Radiol. 1997 Jul;52(7):527-9. https://doi.org/10.1016/s0009-9260(97)80329-7.
32. Roelofs AA, Karssemeijer N, Wedekind N, Beck C, van Woudenberg S, Snoeren PR, et al. Importance of comparison of current and prior mammograms in breast cancer screening. Radiology. 2007 Jan;242(1):70-7. https://doi.org/10.1148/radiol.2421050684.
33. Nelson HD, O'Meara ES, Kerlikowske K, Balch S, Miglioretti D. Factors Associated With Rates of False-Positive and False-Negative Results From Digital Mammography Screening: An Analysis of Registry Data. Ann Intern Med. 2016 Feb 16;164(4):226-35. https://doi.org/10.7326/M15-0971.
34. Hardesty LA, Lind KE, Gutierrez EJ. Effect of Arrival of Prior Mammograms on Recall Negation for Screening Mammograms Performed With Digital Breast Tomosynthesis in a Clinical Setting. J Am Coll Radiol. 2018 Sep;15(9):1293-1299. https://doi.org/10.1016/j.jacr.2018.05.003.
35. Horsley RK, Kling JM, Vegunta S, Lorans R, Temkit H, Patel BK. Baseline Mammography: What Is It and Why Is It Important? A Cross-Sectional Survey of Women Undergoing Screening Mammography. J Am Coll Radiol. 2019 Feb;16(2):164-169. https://doi.org/10.1016/j.jacr.2018.07.002.
36. Donker GA. NIVEL Primary Care Database - Sentinel Practices 2014. Netherlands Institute for Health Services Research (NIVEL), https://www.nivel.nl/sites/default/files/bestanden/Peilstations_2014_Engels.pdf; 2016 [accessed 7 December 2022].
37. NIVEL Primary Care Database: Annual figures 2020 and trend figures 2016-2020 (in Dutch), https://www.nivel.nl/sites/default/files/bestanden/1004117_0.pdf; 2021 [accessed 7 December 2022].
38. Pálka I, Kelemen G, Ormándi K, Lázár G, Nyári T, Thurzó L, Kahán Z. Tumor characteristics in screen-detected and symptomatic breast cancers. Pathol Oncol Res. 2008 Jun;14(2):161-7. https://doi.org/10.1007/s12253-008-9010-7.
39. Gilliland FD, Joste N, Stauber PM, Hunt WC, Rosenberg R, Redlich G, Key CR. Biologic characteristics of interval and screen-detected breast cancers. J Natl Cancer Inst. 2000 May 3;92(9):743-9. https://doi.org/10.1093/jnci/92.9.743.

Radboud

Chapter 4

Added value of prereading screening mammograms for breast cancer by radiologic technologists on early screening outcomes

Geertse TD, Setz-Pels W, van der Waal D, Nederend J, Korte B, Tetteroo E, Pijnappel RM, Broeders MJM, Duijm LEM

Radiology, 2022; 302(2):276-283

Abstract

Background

In the Dutch breast cancer screening program, technologists pre-read the mammograms to identify possible abnormalities, leading to warning signals for radiologists. The best moment to present these warning signals is unknown.

Purpose

To determine the effect that blinding of technologists' warning signals has on radiologists' early screening outcome measures during interpretation of mammograms.

Materials and methods

In this prospective study running from September 2017 to May 2019, on alternating months, radiologists were either blinded or non-blinded to the warning signals of the technologist when interpreting screening mammograms for breast cancer. All discrepancies between radiologists and technologists were reviewed during quality assurance sessions every 6 weeks, which could result in secondary recalls. The outcome measures of this study were recall rate, cancer detection rate, and positive predictive value (PPV) of recall. A chi-square test was used to test for differences between the two groups.

Results

During the study period, 109596 women (mean age, 62 years ± 7 [standard deviation]) including 53291 in the blinded and 56305 in the non-blinded groups were included. The overall recall rate (including secondary recalls) was lower for women in the blinded group than in the non-blinded group (blinded: 1140 of 53291 [2.1%], non-blinded: 1372 of 56305 [2.4%], P=0.001). There was no evidence of cancer detection rate differences between the groups (blinded: 349 of 53291 [6.5 per 1000 screens], non-blinded: 360 of 56305 [6.4 per 1000 screens], P=0.75). The blinded group thus had a higher PPV of recall (blinded: 349 of 1140 [30.6%], non-blinded: 360 of 1372 [26.2%], P=0.02).

Conclusion

While interpreting screening mammograms for breast cancer, radiologists blinded to technologists' warning signals had lower recall rates with higher positive predictive values than non-blinded radiologists, yet cancer detection rates seemed to remain unchanged.

Introduction

In the Netherlands, a population-based breast cancer screening program was set up in 1989 to reduce breast cancer mortality [1,2]. Combined with state-of-the-art treatment, early detection of breast cancer by mammography is still the most effective strategy to achieve a substantial reduction in mortality from this disease [3,4].

Within the Dutch breast cancer screening program, the examinations are obtained by a technologist specializing in mammography. Dutch screening technologists are trained to assess mammograms for possible mammographic abnormalities (pre-reading) to provide screening radiologists with warning signals. At the introduction of pre-reading (in January 2003), the decision was taken to present these warning signals to the screening radiologist when opening a screening examination. Involving technologists in reading mammograms is expected to give them an additional challenge in their work, which increases motivation and makes them more aware of the importance of producing high-quality mammograms [5]. Furthermore, it is more likely that the technologist will take an additional image, to avoid an unnecessary recall, by looking at the mammograms from the perspective of a radiologist. To improve their skills in reading mammograms, technologists attend a quality assurance session every six weeks, in which they review a selection of screening examinations with a screening radiologist.

The performance of technologists in pre-reading or as a second reader in mammography screening has been investigated in several previous studies. These studies, mostly performed during the screen-film mammography era, provided evidence that technologists can learn to read mammograms when given adequate training [5-11]. In contrast, our study was designed to investigate the influence of pre-reading by technologists on the performance of radiologists and the best moment to present the warning signals. We hypothesized that blinding radiologists for warning signals during their interpretation of mammograms would lower the recall rate. To our knowledge, this topic has not previously been explored.

The aim of our study was to determine the effect that blinding of technologists' warning signals has on radiologist early screening outcome measures during interpretation of mammograms, including recall rate, cancer detection rate, and positive predictive value (PPV) of recall. We also investigated whether pre-reading by technologists influences the distribution of tumor characteristics and mammographic image features of screen-detected cancers.

Materials and Methods

Study participants

This prospective study was performed between September 2017 and May 2019 in four units (Eindhoven, Kempen, Den Bosch, Meierij), in a southern region of the Dutch nationwide breast cancer screening program. By participating in screening, women consent to making their data available for evaluation purposes and research unless they choose to opt out explicitly. We did not receive any data for women who objected to the use of their data. This study was performed within the national permit for breast cancer screening issued by the Ministry of Health, Welfare and Sports, and did not require additional approval by a local Institutional Review Board.

Standard mammography screening procedure

Details of the Dutch breast cancer screening program have been described previously [12,13]. In short, the Dutch screening program offers biennial mammography to women aged 50-75 who are invited to attend by a personal letter. Full-field digital mammographic examinations are obtained by a technologist specializing in mammography and consist of a medio-lateral-oblique and a cranio-caudal view of each breast. All examinations are acquired using a Lorad Selenia mammography system (Hologic Inc).

Immediately after obtaining the mammograms, the technologist checks the images for any abnormalities suspicious for cancer (pre-reading), in a reading room equipped with a dedicated workstation (Coronis 3MP, MDCG 3120-CB, Barco). Prior screens are available for comparison in the case of subsequent examinations. For each positive mammogram, the technologist annotates the location and type of abnormality. At the discretion of the technologist, extra views (e.g., Cleopatra view) or repeated views can be obtained.

Each mammogram is double read by two certified screening radiologists in a blinded fashion (i.e., the second reader is unaware of the first reader's decision). The mammograms are viewed on a SecurView mammography screening workstation (Hologic Inc) with 5-megapixel monitors (Coronis 5MP Mammo, MFGD 5621 HD, Barco). When opening a new exam, the radiologist receives an audible and visual warning signal if the technologist observed an abnormality suspicious for cancer. Like the technologists, the radiologists have prior screens available for comparison in the case of subsequent examinations. Radiologists classify each examination according to the Breast Imaging Reporting and Data System (BI-RADS) [14]. Women classified as BI-RADS 1 or 2, are not recalled. Women classified as BI-RADS 0, 4 or 5 are recalled to

a hospital for further assessment (referred to as "primary recall"). The program does not allow a BI-RADS 3 classification since there is no short-term follow up available in the screening setting [15]. In case of a discordant assessment, which occurs in about 2% of all double readings, arbitration by a third screening radiologist is applied. After arbitration, all examinations with a warning signal of the technologist and no primary recall are reviewed during 6-weekly quality assurance sessions supervised by a screening radiologist (for this study: J.N. or W.S-P). For each case, the supervising radiologist decides whether a secondary recall (BI-RADS 0, 4 or 5) is necessary.

Modified screening procedure for this study

The examinations were performed by 41 technologists. The median years of experience of the technologists was 14.6 years (range 0.5–25.6 years), and they perform and pre-read a median of 1926 screening mammograms per year. Technologists spend 4 hours on pre-reading in their initial 5-day screening course. Every 3 years, technologists follow at least 12 hours of supplemental training, including 4 hours on pre-reading of mammograms. In addition, they attend a quality assurance session every six weeks.

The examinations were read by 14 screening radiologists (including W.S-P, J.N., B.K., L.E.M.D). The median years of experience of the screening radiologists was 12 years (range 1-23 years), and they read a median of 9000 screening mammograms per year approximately. Radiologists are obliged to participate in an 8-day initial screening course, have to read a minimum of 3000 screening examinations per year, and obtain at least 40 continuing medical education (CME) points over five years.

During the initial reading, on alternating months, the radiologists were either blinded or non-blinded for the warning signals of the technologist. Non-blinded is the standard procedure. When blinded, the warning signals were not presented at all.

Assessment after recall and follow-up

Recalled women underwent further assessment in one of 15 hospitals with a specialized breast unit. During a one-year follow-up period, all radiology and pathology reports were collected (L.E.M.D.). Data were collected on diagnostic procedures, biopsy results, and breast cancer diagnoses, including Tumor-Node-Metastasis classification [16] and tumor characteristics.

Statistical analysis

The outcome measures were the recall rate (recalls per 100 screens), cancer detection rate (screen-detected cancers per 1000 screens) and positive predictive value (PPV)

of recall (screen-detected cancers per 100 recalls). Chi-square tests were used to test for differences between the blinded and non-blinded group in all outcome measures as well as differences in the distribution of tumor characteristics. $P<0.05$ indicated statistical significance. Statistical analyses were performed by one author (T.D.G.) using Statistical Package for Social Sciences (SPSS, version 27.0, IBM SPSS Statistics).

Results

Participant characteristics

A total of 109596 women (mean age, 62 years ± 7 [standard deviation]) who underwent screening examinations were included during the study period (Table 1). During reading, the radiologists were blinded to the warning signals of the technologists for 53291 screens and non-blinded for 56305 screens (Figure 1). For the examinations in the blinded group, 8.81% had extra views. The technologists gave 1339 warning signals (2.5%) that were not presented to the radiologists during reading (radiologists were blinded). The radiologist scored (after resolving of discordant readings) 566 BI-RADS 0 (with 161 not presented warning signals [28.4%]), 455 BI-RADS 4 (with 220 not presented warning signals [48.4%]), 68 BI-RADS 5 (with 58 not presented warning signals [85.3%]) and 52202 BI-RADS 1 or 2 (with 900 not presented warning signals [1.7%]). The 900 BI-RADS 1 or 2 scores with a not presented warning signal were reviewed during a quality assurance session. In the non-blinded group, 8.84% of the examinations had extra views (repeated views for technical reasons and additional views). Warning signals were presented to the radiologist for 1408 screens (2.5%). The radiologist scored (after resolving of discordant readings) 724 BI-RADS 0 (with 290 presented warning signals [40.1%]), 571 BI-RADS 4 (with 276 presented warning signals [48.3%]), 75 BI-RADS 5 (with 61 presented warning signals [81.3%]) and 54935 BI-RADS 1 or 2 (with 781 presented warning signals [1.4%]). The 781 BI-RADS 1 or 2 scores with a presented warning signal were reviewed during a quality assurance session.

Table 1: Participant characteristics

Characteristic	Screening Sample (n=109596)
Age (y)	
Mean*	62 ± 7
49-69	94800 (86.5)
>70	14769 (13.5)
Screening examination	
Initial	11238 (10.3)
Subsequent	98358 (89.7)

Note- Unless otherwise specified, data are numbers of women, with percentages in parentheses.
* Data are the mean ± standard deviation.

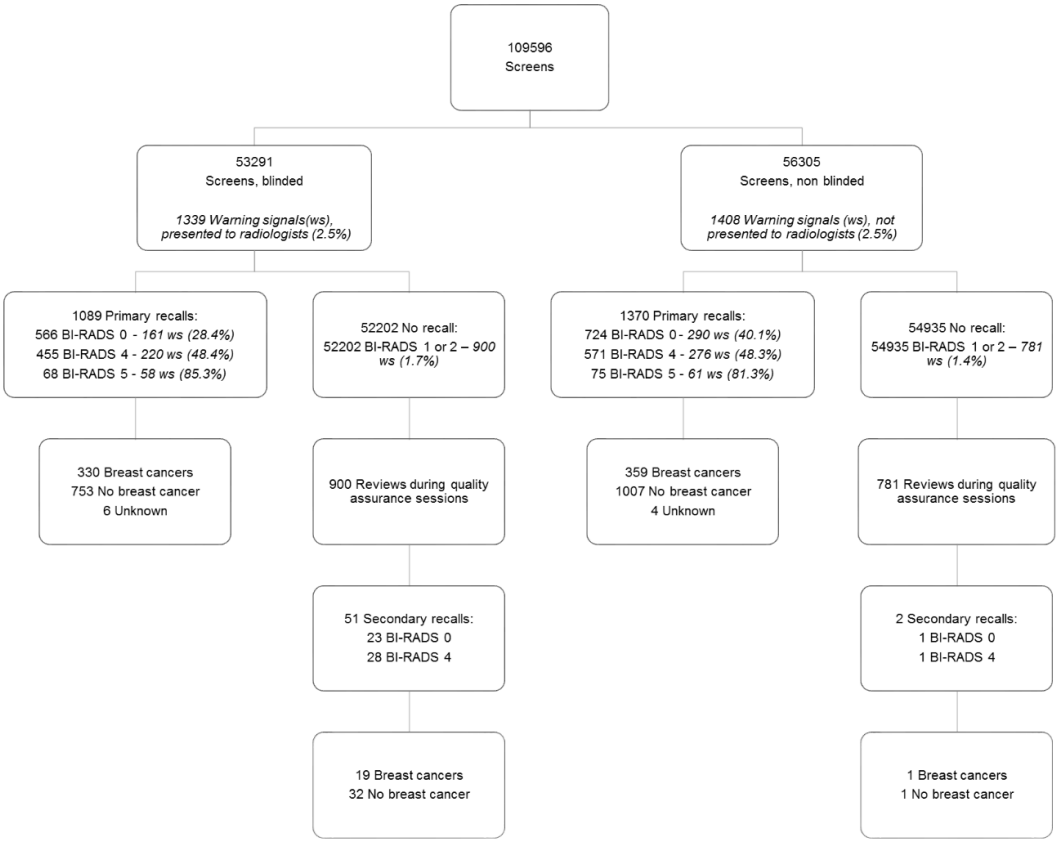


Figure 1: Flowchart of the study participants divided into a blinded group (no warning signal presented to the radiologist) and a non-blinded group (warning signal presented to the radiologist)

Screening results of the blinded group vs. the non-blinded group, excluding secondary recalls

The (primary) recall rate (blinded: 1089 of 53291 [2.0%], non-blinded 1370 of 56305 [2.4%], P<0.001) was lower for women in the blinded group than for those in the non-blinded group (Table 2). There was no evidence of cancer detection rate differences between the groups (blinded: 330 of 53291 [6.2 per 1000 screens], non-blinded: 359 of 56305 [6.4 per 1000 screens], P=0.70), yielding a higher PPV of recall for the blinded group (blinded: 330 of 1089 [30.3%], non-blinded 359 of 1370 [26.2%], P=0.03).

Table 2: The early outcome measures for the blinded and non-blinded groups, after primary recall by the radiologists (excluding the secondary recalls) and after the quality assurance session (overall recall, including secondary recalls)

Outcome Measure	Blinded group	Non-Blinded group	P-value
Screens, No. (%)	53291 (48.6)	56305 (51.4)	
Warning signals, No. (%)	1339 (2.5)	1408 (2.5)	
Reviews in quality assurance session, No. (%)	900 (1.7)	781 (1.4)	
Recall rate, No. (%)			
Primary recalls (excluding secondary recalls)	1089 (2.0)	1370 (2.4)	<0.001
Overall recalls (including secondary recalls)	1140 (2.1)	1372 (2.4)	0.001
Screen-detected cancer rate, No. (per 1000 screens)			
Primary recalls (excluding secondary recalls)	330 (6.2)	359 (6.4)	0.70
Overall recalls (including secondary recalls)	349 (6.5)	360 (6.4)	0.75
PPV of recall (%)			
Primary recalls (excluding secondary recalls)	30.3	26.2	0.03
Overall recalls (including secondary recalls)	30.6	26.2	0.02

PPV = positive predictive value

Screening results of the blinded group vs. the non-blinded group, including secondary recalls

At the quality assurance sessions where the technologists and one of the two supervising screening radiologists were present, more women were recalled at the second instance in the blinded group compared with the non-blinded group (51 secondary recalls in the blinded group versus 2 secondary recalls in the non-blinded group, Figure 1). Nevertheless, the overall recall rate including these secondary recalls (blinded: 1140 of 53291 [2.1%], non-blinded: 1372 of 56305 [2.4%, P=0.001) was lower in the blinded group than in the non-blinded group (Table 2). The secondary recalls resulted in the additional detection of 19 cancers in the blinded group and 1 cancer in the non-blinded group (Figure 1). However, we did

not find evidence that these additional screen-detected cancers resulted in a higher cancer detection rate, not even in the blinded group (blinded group, detection rate primary recalls: 330 of 53291 [6.2 per 1000 screens], blinded group, detection rate overall recalls: 349 of 53291 [6.5 per 1000 screens], P=0.46). We found no evidence of cancer detection rate differences, including the cancers detected at secondary recalls, between the blinded and non-blinded groups (blinded: 349 of 53291 [6.5 per 1000 screens], non-blinded: 360 of 56305 [6.4 per 1000 screens], P=0.75). The overall PPV of recall including the secondary recalls was higher for the blinded group (blinded: 349 of 1140 [30.6%], non-blinded: 360 of 1372 [26.2%], P=0.02). The PPV of recall of only the secondary recalls was 37.3% for the blinded group and 50% for the non-blinded group (blinded: 19 of 51 [37.3%], non-blinded: 1 of 2 [50%], P=0.62).

Tumor characteristics of the detected cancers in the blinded and non-blinded groups

Table 3 shows the tumor characteristics of all screen-detected breast cancers in both the blinded and non-blinded groups. We found no evidence of a difference in tumor characteristics between the blinded and non-blinded groups (p-value range, 0.25 – 0.93). In the blinded group, the proportion of ductal carcinoma in situ (DCIS) among detected cancers following secondary recalls was higher than that observed in primary recalls (blinded group, primary recalls: 61 of 330 [18.5%], blinded group, secondary recalls: 9 of 19 [47.4%], P=0.002).

Mammographic image features of the warning signals (pre-reading) of the technologists

In the blinded group, the technologists annotated a suspicious abnormality in 1339 out of 53291 screening examinations (2.5%) at pre-reading (Figs 2, 3). Table 4 presents the different features of those abnormalities. These were most often masses (767 of 1339 [57.3%]), followed by calcifications (237 of 1339 [17.7%]) and asymmetries (153 of 1339 [11.4%]). In the blinded group, the warning signals where not presented to the radiologists during interpreting the examinations, but where reviewed at the quality assurance sessions, which resulted in 51 secondary recalls. The majority of these secondary recalls were related to calcifications and asymmetries (both 18 of 51 [35.3%]). Of the 19 breast cancers detected through secondary recall, 11 showed suspicious calcifications at the screening mammogram (57.9%). The substantial number of warning signals for masses (767 of 1339 [57.3%]) resulted in only six secondary recalls, yielding one breast cancer. Most false positive secondary recalls were based on asymmetries (14 of 32 [43.8%]). The PPV of recall was highest for calcifications (11 of 18 [61.1%]), followed by architectural distortions (3 of 5 [60.0%]), and was lowest for masses with or without calcifications (0 of 4 [0%] resp. 1 of 6 [16.7%]).

Table 3: Tumor characteristics of cancers in the blinded and non-blinded groups, detected in response to primary recalls and secondary recalls[a]

	Blinded group			Non-blinded group			
Characteristic	**Primary recall**	**Secondary recall**	**Overall**	**Primary recall**	**Secondary recall**	**Overall**	**P-value[b]**
Detected cancers, No.	330	19	349	359	1	360	
Type of cancer, No. (%)							
DCIS	61 (18.5)	9 (47.4)	70 (20.1)	84 (23.4)	1 (100)	85 (23.6)	0.25
Invasive	269 (81.5)	10 (52.6)	279 (79.9)	275 (76.6)	0 (0)	275 (76.4)	
Type of invasive cancer, No. (%)							
NST	213 (79.2)	8 (80.0)	221 (79.2)	212 (77.1)	NA	212 (77.1)	0.93
Lobular	38 (14.1)	2 (20.0)	40 (14.3)	45 (16.4)		45 (16.4)	
Other	18 (6.7)	0 (0)	18 (6.5)	18 (6.5)		18 (6.5)	
Grade of DCIS, No. (%)							
1	11 (18.0)	2 (22.2)	13 (18.6)	15 (17.9)	1 (100)	16 (18.8)	0.57
2	27 (44.3)	4 (44.4)	31 (44.3)	31 (36.9)		31 (36.5)	
3	23 (37.7)	3 (33.3)	26 (37.1)	38 (45.2)		38 (44.7)	
Grade of invasive cancers, No. (%)							
1	117 (43.5)	2 (20.0)	119 (42.7)	117 (42.5)	NA	117 (42.5)	0.31
2	117 (43.5)	7 (70.0)	124 (44.4)	129 (46.9)		129 (46.9)	
3	34 (12.6)	1 (10.0)	35 (12.5)	25 (9.1)		25 (9.1)	
Unknown	1 (0.4)	0 (0)	1 (0.4)	4 (1.5)		4 (1.5)	
Tumor size of invasive cancers, No. (%)							
T1	229 (85.1)	8 (80.0)	237 (84.9)	225 (81.8)	NA	225 (81.8)	0.60
≥ T2	37 (13.8)	2 (20.0)	39 (14.0)	47 (17.1)		47 (17.1)	
Unknown	3 (1.1)	0 (0)	3 (1.1)	3 (1.1)		3 (1.1)	
Lymph node status of invasive cancers, No. (%)							
N0	203 (75.5)	6 (60.0)	209 (74.9)	195 (70.9)	NA	195 (70.9)	0.41
N+	47 (17.5)	2 (20.0)	49 (17.6)	51 (18.5)		51 (18.5)	
Unknown	19 (7.1)	2 (20.0)	21 (7.5)	29 (10.5)		29 (10.5)	
Receptor status of invasive cancers, No. (%)							
ER+, PR+/-, HER2-	224 (83.3)	8 (80.0)	232 (83.2)	236 (85.8)	NA	236 (85.8)	0.51
ER+/-, PR+/-, HER2+	16 (5.9)	0 (0)	16 (5.7)	17 (6.2)		17 (6.2)	
Triple negative	27 (10.0)	2 (20.0)	29 (10.4)	19 (6.9)		19 (6.9)	
Unknown	2 (0.7)	0 (0)	2 (0.7)	3 (1.1)		3 (1.1)	

[a] DCIS= Ductal carcinoma in situ; ER=Estrogen receptor; PR=Progesterone receptor; HER2=Human epidermal growth factor receptor 2; NST=no specific type.

[b] We compared the distributions in the two groups for the overall detected cancers

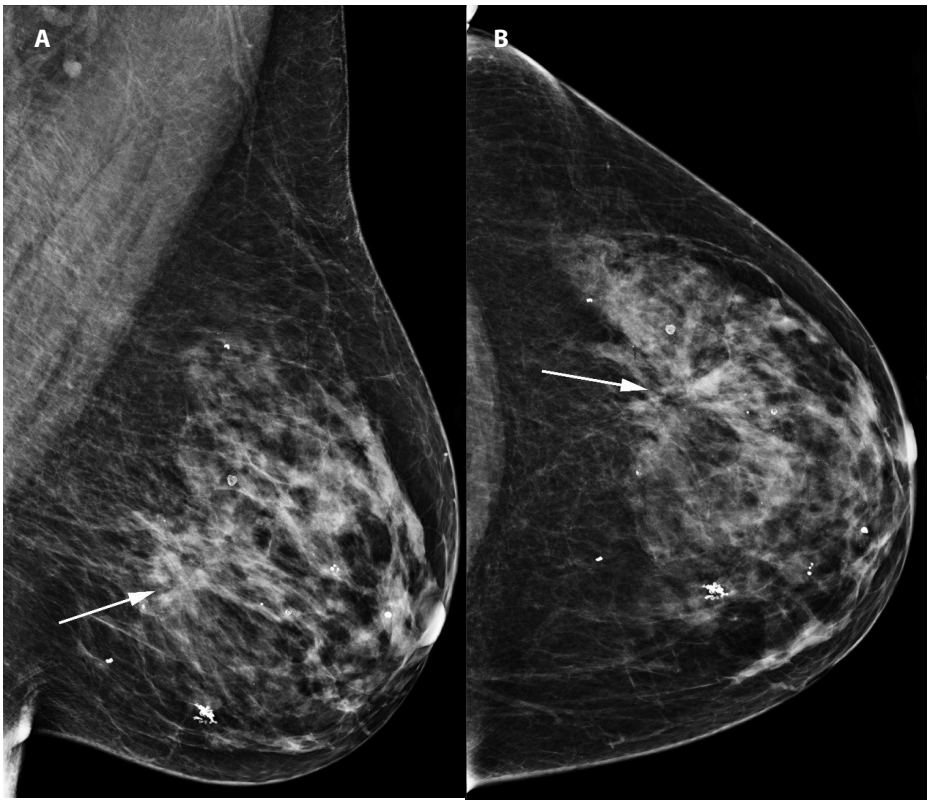


Figure 2: (A) Left mediolateral oblique (LMLO) view and (B) left craniocaudal (LCC) view in 74-year-old woman who underwent two-view screening mammography in 2018. This examination shows an architectural distortion (arrow), classified by the technologist as Breast Imaging Reporting and Data System category 4. Initially, the woman was not recalled after radiologist double reading, but was recalled at second instance after the quality assurance session. Ultrasound-guided core-needle biopsy revealed invasive lobular cancer (B&R grade 2, Estrogen and Progesterone receptor positive), without signs of axillary metastasis. Because of tumor type and discrepancy in size between mammography (ca. 3 cm) and echo (ca. 2 cm), an additional MRI (2.5 cm) was performed.

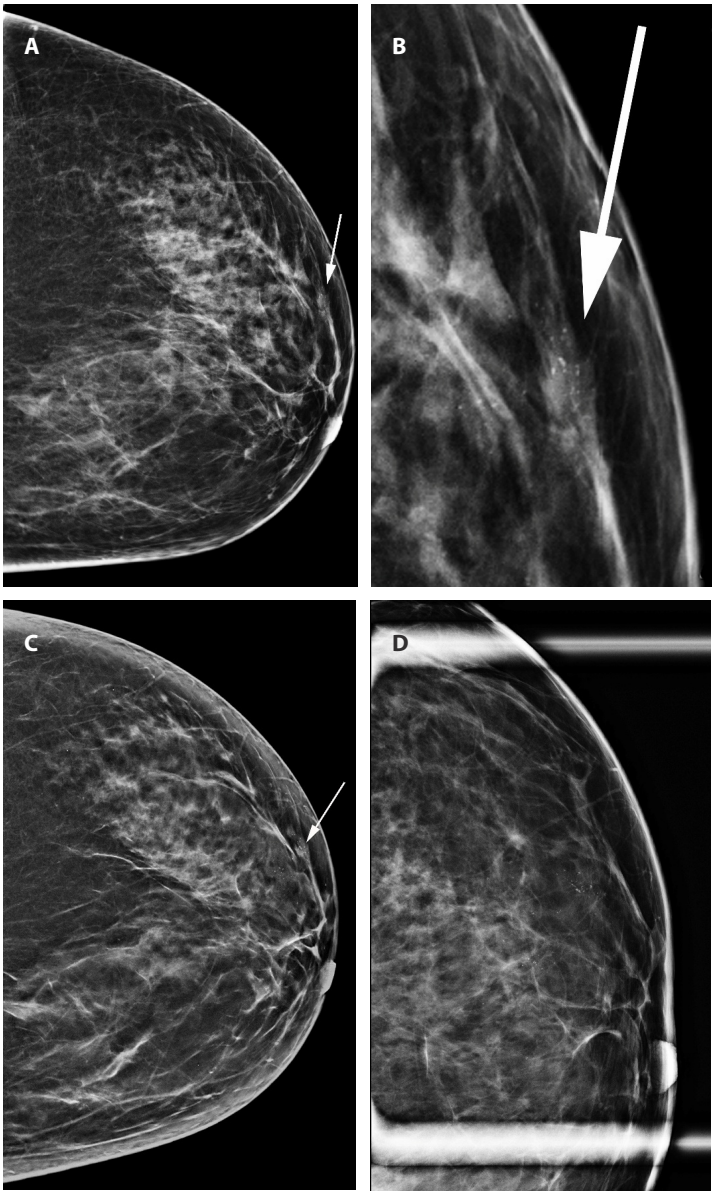


Figure 3: (A) Left craniocaudal (LCC) view in 58-year-old woman who underwent subsequent two-view screening mammography in 2018 shows new (compared with priors 2016) grouped, fine pleomorphic calcifications (arrow, [B] cropped image), classified by the technologist as Breast Imaging Reporting and Data System category 4. Initially, the woman was not recalled after radiologist double reading, but was recalled at second instance after the quality assurance session. (C) LCC view obtained in the hospital (2018), for assessment after recall, confirmed the calcifications. (D) The magnification view shows more details, with calcifications spread over an area of 40 mm. Stereotactic vacuum-assisted biopsy yielded grade 2 ductal carcinoma in situ (DCIS). Mastectomy was performed at the patient's request, with a final diagnosis of grade 2 DCIS sized 70 mm.

Table 4: Type of abnormality at mammography seen by the technologists during the pre-reading of screening examinations in the blinded group

			Secondary recalls		
	Warning signals	Total	Breast cancer	No breast cancer	PPV of recall
	(n=1339)	(n=51)	(n=19)	(n=32)	
Mammographic abnormality, No. (%)					
Mass	767 (57.3)	6 (11.8)	1 (5.3)	5 (15.6)	16.7%
Calcifications	237 (17.7)	18 (35.3)	11 (57.9)	7 (21.9)	61.1%
Mass with calcifications	37 (2.8)	4 (7.8)	0 (0)	4 (12.5)	0%
Asymmetry	153 (11.4)	18 (35.3)	4 (21.1)	14 (43.8)	22.2%
Architectural distortion	121 (9.0)	5 (9.8)	3 (15.8)	2 (6.3)	60.0%
Other	24 (1.8)	0 (0)	0 (0)	0 (0)	0%

PPV = positive predictive value

Discussion

In the Dutch breast cancer screening program, technologists pre-read mammograms to identify possible abnormalities, leading to warning signals for radiologists. The best moment to present these warning signals is unknown. This prospective study evaluated the effect that blinding of technologists' pre-reading has on radiologist early screening outcomes. It shows that blinding radiologists to this pre-reading resulted in a lower overall recall rate (blinded 2.1% vs. non-blinded 2.4%, P=0.001) and a higher positive predictive value (PPV) of recall (blinded 30.6% vs. non-blinded 26.2%, P=0.02) when compared to non-blinding, but we found no evidence of a difference in the cancer detection rate (blinded 6.5 per 1000 screens vs. non-blinded 6.4 per 1000 screens, P=0.75).

In a breast cancer screening program, one continuously strives to find an optimal balance between recall rate and cancer detection rate. The recall rate in the Netherlands is one of the lowest worldwide [17,18]. Low recall rates may result in more missed subtle cancers. However high recall rates result in unnecessary further assessment, patient anxiety, and increased costs. Different factors, such as national health policy issues (e.g., malpractice concerns) and characteristics of the screened population may strongly influence recall rates in different screening programs [19,20].

Our study shows that it is not effective to expose radiologists to technologists' warning signals during their initial interpretation of the screening mammograms.

In general, radiologists are getting more and more signals or markers to support them in detecting lesions on medical images. Due to the introduction of artificial intelligence, new computer-aided detection systems are becoming available for different modalities [21]; e.g., for mammograms or tomosynthesis in breast screening and for CT scans in lung screening. When implementing such tools in practice, careful evaluation is needed to decide whether these markers should be presented during or after reading images.

In a recent study performed within our screening region, Coolen et at. [22] found a limited role for quality assurance sessions in additional cancer detection. Less than 1% of the screen-detected cancers were additionally detected through these sessions. In that study, the radiologists where not blinded to the warning signals of the technologists. We found no evidence of a difference in cancer detection rate between the blinded and non-blinded group, but looking at the characteristics of the additional cancers, these appear to be clinically relevant findings, including DCIS of grade 2 or 3 and invasive cancers of grade 2 or 3.

More than half of the warning signals (767 of 1339 [57.3%]) given by the technologists concerned masses, resulting in only one additional cancer after secondary recall. Most cancers diagnosed through secondary recall showed suspicious calcifications (11 of 19 [57.9%]). The proportion of DCIS in these additional cancers was higher than that observed after primary recall, which is in accordance with previous studies [23,24]. Our results further show that warning signals concerning calcifications have the highest cancer probability, while masses have the lowest. These findings correspond with those reported by Coolen et al. [24] and Wivell et al. [25]. There is no obvious explanation, but we hypothesize that technologists may spend more time looking for calcifications and, in contrast to the radiologists, do not read mammograms in batches.

Our study has a few limitations. First, no data are available on interval cancers, i.e., cancers diagnosed between two consecutive screening rounds following a negative screening examination. Therefore, program sensitivity and specificity could not be calculated as outcome measures, and we did not have information on interval cancers for which the technologists had given a warning signal. Second, due to the very low number of cancers in a screening population, we could not identify small differences in cancer detection rate. However, we were mainly concerned with the effect on the recall rate. Third, when not blinded to warning signals, all radiologists made their own decisions on how to apply those signals. In contrast, in the blinded group, only two radiologists (J.N., W.S-P) saw the warning signals during the quality

assurance sessions, and decided whether secondary recall was necessary. Thus, the study design may have influenced the number of secondary recalls.

In conclusion, our results showed that radiologists interpreting screening mammograms for breast cancer who were blinded to technologists' warning signals had lower recall rates with higher PPVs than non-blinded radiologists, yet cancer detection rates seemed to remain unchanged. Thus, we suggest that radiologists should be blinded to warning signals of technologists while interpreting screening mammograms. They could be presented immediately after interpretation, to give the radiologists the opportunity to change their recall decision. Additional cancers may be detected through review at quality assurance sessions. Future research is necessary to investigate whether and to what extent such a new procedure increases the performance of radiologists and to ensure the effect of the quality assurance sessions. In addition, we will investigate technologists' additional views and whether these views prevent unnecessary recalls.

Acknowledgements

The authors thank the entire team of screening radiologists, technologists, data analysts, and secretarial support of our screening region.

Author contributions

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

Disclosures of Conflicts of Interest

T.D.G. No relevant relationships. W.S.P. No relevant relationships. D.v.d.W. No relevant relationships. J.N. No relevant relationships. B.K. No relevant relationships. E.T. No relevant relationships. R.M.P. No relevant relationships. M.J.M.B. No relevant relationships. L.E.M.D. No relevant relationships.

References

1. Verbeek ALM, Hendriks JHCL, Holland R, et al. Reduction of breast cancer mortality through mass screening with modern mammography. First results of the Nijmegen project, 1975-1981. Lancet. 1984; 1 (8388):1222–1224
2. Colette HJA, Day NE, Rombach JJ, et al. Evaluation of screening for breast cancer in a non-randomised study (the DOM Project) by means of a case-control study. Lancet. 1984; 1 (8388):1224–1226.
3. Dibden A, Offman J, Duffy SW, et al. Worldwide Review and Meta-Analysis of Cohort Studies Measuring the Effect of Mammography Screening Programs on Incidence-Based Breast Cancer Mortality. Cancers. 2020; 12(4):976-991.
4. Marmot MG, Altman DG, Cameron DA, et al. The benefits and harms of breast cancer screening: an independent review. Br J Cancer. 2013; 108(11): 1778–1786.
5. Tonita JM, Hillis JP, Lim, CH. Medical radiologic technologist review: Effects on a population-based breast cancer screening program. Radiology. 1999; 211(2): 529–533.
6. Torres-Mejía G., Smith RA, de la Luz Carranza-Flores M, et al. Technologists supporting radiologists in the interpretation of screening mammography: A viable strategy to meet the shortage in the number of radiologists. BMC Cancer. 2015; 15(1): 1–12.
7. Bennett RL, Sellars SJ, Blanks RG, et al. An observational study to evaluate the performance of units using two technologists to read screening mammograms. Clinical Radiology. 2012; 67(2): 114–121.
8. Moran S, Warren-Forward H. A retrospective study of the performance of technologists in interpreting screening mammograms. Radiography. 2011; 17(2): 126–131.
9. Duijm LEM, Groenewoud JH, Fracheboud J, et al. Introduction of additional double reading of mammograms by technologists: effects on a biennial screening program outcome. Eur J Cancer. 2008; 44(9): 1223–1228.
10. Sumkin JH, Klaman HM, Graham M, et al. Prescreening mammography by technologists: A preliminary assessment. Am J Roentgenol. 2003; 180(1): 253–256.
11. van den Biggelaar FJHM, Nelemans PJ, Flobbe K. Performance of technologists in mammogram interpretation: a systematic review. Breast. 2008; 17(1): 85–90.
12. Sankatsing VDV, van Ravesteyn NT, Heijnsdijk EAM, et al. The effect of population-based mammography screening in Dutch municipalities on breast cancer mortality: 20 years of follow-up. Int J Cancer. 2017; 141(4): 671–677.
13. Otten JDM, Fracheboud J, den Heeten GJ, et al. Likelihood of early detection of breast cancer in relation to false-positive risk in life-time mammographic screening: population-based cohort study. Ann Oncol. 2013; 24(10):2501-2506.
14. Sickles EA, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS® Mammography. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA, American College of Radiology; 2013.
15. Timmers JM, van Doorne-Nagtegaal HJ, Zonderland HM, et al. The Breast Imaging Reporting and Data System (BI-RADS) in the Dutch breast cancer screening program: its role as an assessment and stratification tool. Eur Radiol. 2012; 22(8): 1717–1723.
16. Amin MB, Edge S, Greene F, et al. (Eds.). AJCC Cancer Staging Manual (8th edition). Springer International Publishing: American Joint Commission on Cancer; 2017.
17. Otten JD, Karssemeijer N, Hendriks JH, et al. Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. J Natl Cancer Inst. 2005; 97(10), 748–754.

18. Ponti A, Anttila A, Ronco G, et al. Cancer screening in the European Union. Report on the implementation of the council recommendation on cancer screening. Second report. Brussels: European Commission, 2017.
19. Elmore JG, Nakano CY, Koepsell TD, et al. International variation in screening mammography interpretations in community-based programs. J Natl Cancer Inst. 2003; 95, 1384 – 93.
20. Le MT, Mothersill CE, Seymour CB, Mcneill, FE. Is the false-positive rate in mammography in North America too high? British Journal of Radiology. 2016; 89, 20160045.
21. Tariq A, Purkayastha S, Padmanaban GP, et al. Current Clinical Applications of Artificial Intelligence in Radiology and Their Best Supporting Evidence. J Am Coll Radiol. 2020; 17(11): 1371–1381.
22. Coolen AMP, Korte B, Tjan-Heijnen VCG, et al. Additional breast cancer detection at digital screening mammography through quality assurance sessions between technologists and radiologists. Radiology. 2020; 294(2): 509–517.
23. Duijm LEM, Groenewoud JH, Fracheboud, J, et al. Additional double reading of screening mammograms by radiologic technologists: Impact on screening performance parameters. J Natl Cancer Inst. 2007; 99(15): 1162–1170.
24. Coolen AMP, Lameijer JRC, Voogd AC, et al. Incorporation of the technologist's opinion for arbitration of discrepant assessments among radiologists at screening mammography. Breast Cancer Res Treat. 2018; 171(1): 143–149.
25. Wivell G, Denton ERE, Eve CB, et al. Can technologists read screening mammograms? Clinical Radiology. 2003; 58(1): 63–67.

Radboud

Chapter 5

Value of audits in breast cancer screening quality assurance programmes

Geertse TD, Holland R, Timmers JM, Paap E, Pijnappel RM, Broeders MJ, den Heeten GJ

European Radiology, 2015, 25(11):3338-47

Abstract

Objectives

To retrospectively evaluate the results of all audits performed in the past and to assess their value in the quality assurance of the Dutch breast cancer screening programme.

Methods

The audit team of the Dutch Reference Centre for Screening (LRCB) conducts triennial audits of all 17 reading units. During audits, screening outcomes like recall rates and detection rates are assessed and a radiological review is performed. This study investigates and compares the results of four audit series: 1996-2000, 2001-2005, 2003-2007, and 2010-2013.

Results

The analysis shows increased recall rates (from 0.66%, 1.07%, 1.22% to 1.58%), increased detection rates (from 3.3, 4.5, 4.8 to 5.4 per 1000) and increased sensitivity (from 64.5%, 68.7%, 70.5% to 71.6%), over the four audit series. The percentage "missed cancers" among interval cancers and advanced screen-detected cancers did not change (p=0.4).

Conclusions

Our audits not only provide an opportunity for assessing screening outcomes, but also provide moments of self-reflection with peers. For radiologists, an accurate understanding of their performance is essential to identify points of improvement. We therefore recommend a radiological review of screening examinations and immediate feedback as part of an audit.

Introduction

The aim of breast cancer screening is to prevent breast cancer related death by early detection and treatment [1,2]. Its benefits are well recognized, however negative 'side effects' have been debated since the start of the first screening programme [3]. The key issues are that asymptomatic women are invited for screening and therefore exposed to X-rays, and that false-positive and false-negative cases form an inevitable aspect of screening. Early on in breast cancer screening programmes, recommendations were made to introduce a comprehensive quality assurance (QA) scheme including quality control of equipment, training and accreditation of professionals, and evaluating screening outcomes with the aim of maintaining a favourable balance between the benefits and harms of screening [4,5]. Screening radiologists should, by repeated self-assessment, audit of practice and continuing education, strive to maintain and improve their skills to ensure that all women attending the screening receive an excellent service with minimal adverse effects. This is key to the success of any breast cancer screening programme.

From the start of the Dutch Breast Cancer Screening Programme (DBCSP) in 1989, a QA programme as part of the DBCSP was initiated, to achieve high positive predictive values (PPV) of around 50% and a restricted use of craniocaudal views in the subsequent examinations. An independent foundation, the Dutch Reference Centre for Screening (LRCB), is responsible for this QA programme. The LRCB has performed triennial audits at all reading units (RUs) since 1996, where screening outcomes such as recall rate, detection rate and PPV of recall are compared with targets and with mean national performance [6,7]. In addition, a radiological review of mammograms forms a substantial part of the audit. In this way, the Dutch programme distinguishes itself from other countries. In most QA programmes an audit consists only of an epidemiological surveillance [8].

The aim of this study was to retrospectively evaluate the results of 4 series of audits performed by the LRCB between 1996–2013 to investigate the value of these audits as a QA tool.

Materials and Methods

Screening setting

Details about the DBCSP have been described previously [9,10]. In brief, the DBCSP is centrally-organized with five regional screening organizations which operate

64 screening units (SUs) and 17 reading units (RUs). Every 2 years, all women aged 50-74 (approximately 2 million) receive a personal invitation. The attendance rate is approximately 80%. Non-blinded double reading was performed in the period of screen-film mammography (SFM). After the conversion to digital mammography (DM) in 2008-2010, blinded double reading was performed. Discrepant readings were solved by consensus between first and second reader or arbitration by a third reader. In the same period the Breast Imaging and Reporting Data System (BI-RADS) [11–13] was introduced. For women who are recalled, further assessment takes place at a hospital of their choice.

Audits

During a triennial audit, the performance of the team of radiologists of a RU is assessed (not individual performance). The audit follows a fixed protocol, divided into two parts: an evaluation of screening outcomes and a radiological review of mammograms.

Before every audit, data on screening outcomes (Table 1) were provided by the RU. The outcomes cover a period of 4 years, 1 or 2 years prior to the year of the audit. During the audit, the outcomes of the RU were compared to target values and to mean national figures [6,7,14]. The RUs collected assessment data from the hospitals on follow-up of all recalled women (98% coverage). Interval cancers should be obtained by linking the DBCSP data file to the file of the cancer registry. In the past, linkage was carried out at regional level. Between 2004-2013, linkage was not possible. Therefore, no complete data of interval cancers were available for the audits.

Until 2010, the radiological review included 80 interval cancers (ICs) and 40 screen-detected stage-II or more advanced cancers [15]. In 2010 the content of the radiological review has been adjusted and includes now 40 ICs, 40 stage-II cancers and 40 consecutive recalled cases.

For the radiological review, 40 ICs were collected by the RU, with help of the departments of radiation oncology and pathology at the hospitals, and the general practitioners. ICs were defined as cases of invasive breast cancer or ductal carcinoma in situ (DCIS) diagnosed within 24 months of a negative (no recall) screening examination and before a next screening examination [4]. For each case, the RU collected clinical diagnostic mammograms, clinical reports and the previous two screening mammograms. ICs after a woman's first screening examination were excluded. The two screening examinations before diagnosis were blindly reviewed by 2 of 6 experienced

Table 1: Screening outcomes for subsequent screening examinations in the four audit series

Outcome Measures period audit series period outcomes No. of included RUs	1st audit series 1996-2000 1990-1997 *16 RUs*	2nd audit series 2001-2005 1998-2003 *13 RUs*	3rd audit series 2003-2007 2001-2006 *14 RUs*	4th audit series 2010-2013 2006-2011 *17 RUs*	Trend[b]	P Value
Total number of screened women	966573	1913739	2053796	3106809		
Total number of recalled women	6525	20255	25801	47548		
Recall rate (%)	0.66 (0.5-1.0)	1.07 (0.7-1.5)	1.22 (0.7-1.9)	1.58 (1.0-2.2)	+0.29 (0.23, 0.35)	0.000
Total number of screen-detected cancers	3262	8450	9795	16762		
Detection rate (per 1000 screened)	3.3 (2.7-4.1)	4.5 (3.9-6.4)	4.8 (3.5-5.6)	5.4 (4.3-6.2)	+0.6 (0.5, 0.8)	0.000
PPVof recall (%)	51.9 (30.0-66.7)	43.5 (32.9-60.3)	41.5 (27.3-59.1)	35.5 (26.7-47.1)	-5.2 (-3.8, -6.6)	0.000
Proportion of DCIS (%)	11.7 (6.3-19.9)	14.4 (9.9-18.3)	13.3 (9.4-16.9)	16.4 (11.9-19.6)	+1.4 (0.8, 2.0)	0.000
Proportion of invasive cancers (%)	88.3 (80.1-93.7)	85.6 (81.7-90.1)	86.7 (83.1-90.6)	83.1 (78.2-88.1)	-1.5 (-1.0, -2.0)	0.000
Total number of detected invasive cancers	2837	7226	8477	13851		
Proportion of small tumors, T1a-c (%)	76.4 (68.4-83.5)	79.3 (74.9-83.7)	78.7 (73.7-81.9)	80.6 (76.9-84.8)	+1.3 (0.6, 1.9)	0.000
Proportion of advanced tumors, T2+ (%)	23.6 (16.5-31.6)	20.7 (16.3-25.1)	21.3 (18.1-26.3)	19.4 (15.2-23.1)	-1.3 (-0.6, -1.9)	0.000
Proportion N0[a] (%)	77.7 (74.8-82.6)	76.1 (71.4-79.9)	73.0 (67.4-76.2)	75.7 (68.6-81.9)	-0.4 (-1.4, 0.6)	0.412
Proportion N1 (%)	22.3 (17.4-25.2)	23.9 (20.1-28.6)	27.0 (23.8-32.6)	24.3 (18.1-31.4)	+0.4 (-0.6, 1.4)	0.412

Numbers in parentheses are the range, except for the column Trend

[a] Axillary lymph nodes were free of tumor

[b] Numbers in parentheses are the 95% confidence interval

RU=reading unit; PPV=positive predictive value; DCIS=ductal carcinoma in situ

radiologists of the audit team (more than 10 year screening experience). They have no prior knowledge of the clinical diagnostic mammograms, laterality and location of the IC. The cases were classified into 3 categories: true negative (category 1), minimal sign (category 2), and missed cancer (category 3). In case of a true negative, no abnormality is seen, and there is no reason for recall. A minimal sign indicates a possible subtle finding which did not justify recall. If an abnormality is clearly visible and there is a significant sign for recall, this is classified as a missed cancer. If the location of the IC does not correlate with the designated minimal sign or missed cancer, the case will be reclassified. The review takes place in the presence of RU radiologists. There is an open discussion where the attending radiologists exchange opinions. If the RU and audit team disagree on classification, the audit team decides.

Also 40 most recent consecutive stage-II cancers were collected by the RU from the DBCSP database. Stage-II cancers are defined as screen-detected cancers which have a positive lymph-node status (N+) and / or a size at diagnosis larger than 2 cm (T2+). For the audit RUs selected only stage-II cancers diagnosed after at least two negative screening examinations. The procedure is identical to that of the IC, where the screen-positive examination replaces the diagnostic mammograms.

For every audit held since 2010, 40 consecutive recalled cases were collected by the RU from the DBCSP database. For each case, the RU collects the screen-positive mammogram, the clinical reports of the assessment, and the previous mammogram (if available). The review procedure is similar to normal screening practice. The cases were classified as BI-RADS 0, 4 or 5. The cases where no recall was required by the audit team were recorded. The outcome of the diagnostic assessment is discussed after each case. Following the review, agreement between audit team and RU radiologists for assigning BI-RADS was determined using Cohen's kappa (κ) [16] as measurement of agreement [12]. In addition, the PPVs of the BI-RADS categories 0, 4 and 5 were calculated.

Directly following the audit, feedback was provided during a final meeting. A report was prepared, summarizing the results and providing recommendations for improvements.

Data collection

Data from 4 audit series: 1996-2000 (26 RUs), 2001-2005 (28 RUs), 2003-2007 (29 RUs) and 2010-2013 (17 RUs) were collected. In this period, around 8 million screens were performed, around 40,000 breast cancers were detected, and 10,000 cases were peer reviewed.

No complete data on ICs were available for the audits of the RUs, in the period 2004-2013. Recently a linkage between the DBCSP data file and the cancer registry at national level was successfully carried out, and complete data of ICs became available for the National Evaluation Team for Breast cancer screening (NETB). For interval cancer rate and sensitivity in this study, these recent published national data of the NETB are used [7].

Between 2008-2010, the RUs switched to soft-copy reading and the number of RUs was reduced from 29 to 17 (same number of radiologists, more readers per RU). During this period no audits were performed. To retain the relationships within RUs over time, for purposes of analysis, we aggregated the data of those RUs which were combined. As a result, for each audit series there were 17 RUs, but some RUs were excluded due to missing data. Table 1 shows the number of RUs for every audit series.

All data concern screens of women aged between 50-74, except for the first audit series, as until 1998 only women aged 50-69 were screened. In this study we focus only on subsequent screens as this represents the majority of the screening exams (85%) and will therefore be largely responsible for the impact of the DBCSP.

Analysis

For analysis we used SPSS (version 20.0.0 for Windows; SPSS, Chicago, III). Comparisons were made between the 4 audit series (or corresponding 4 periods for the data of NETB, Table 2), except for the review of 40 consecutive recalled cases where data were only available for the fourth series.

For screening outcomes, the means of each audit series were compared with regard to recall rate, detection rate, PPV of recall, interval cancer rate, sensitivity and radiological and pathological characteristics of the detected cancers. In 8% of the 32,391 invasive cancers, the radiological and pathological characteristics were incomplete (n=663 for Tx and n=1802 for Nx) and these cases were excluded. The trend over the 4 audit series was assessed by univariate linear regression for the observed rate/percentage, with the RU as fixed factor and the audit series number as continuous covariate. $P \leq .05$ was considered indicative of a statistically significant trend.

This analysis was also performed for the review of ICs and stage-II cancers, with regard to the means of the distribution of categories 1, 2 and 3. As we mainly focus on category 3 (missed cancers), for the trend analysis, we combined categories 1 and 2.

For the review of the consecutive recalled cases, we calculated the mean percentage agreement between the audit team and the RU. We also estimated the association between the "percentage of cases the audit team would not have recalled" and the mean recall rate of the RU. This estimation is conducted without 2 outliers, which were among the first reviews. It is likely that shortly after the introduction of this item, the audit team had to get used to this new part of the audit. Finally we calculated the mean PPV of the BI-RADS categories 0, 4 and 5.

Results

Recall rates increased from 0.66% in the first audit series to 1.07%, 1.22% and 1.58% respectively in the next three series (Table 1). In corresponding periods the data of NETB showed increased recall rates from 0.72%, 1.05%, 1.24% to 1.49% (Table 2). In both a positive trend is observed (+0.3%, $p<.05$). Also a positive trend in detection rates was observed in both tables (+0.6 per 1000 screened, $p<.01$), increasing from 3.3, 4.5, 4.8 to 5.4 per 1000 women screened in each audit series respectively (Table 1). Figure 1 shows the relation between recall rate, detection rate and audit series. A higher recall rate results in a higher detection rate. The relationship is not linear, but asymptotic. As a result, the PPV of recall decreased (negative trend of -5.2%, $p<.001$) from 51.9%, 43.5%, 41.5% to 35.5% respectively (Table 1). NETB data in Table 2, showed that the sensitivity increased from 64.6%, 68.7%, 70,5% to 71,6%, which corresponded to a positive trend (+2.3%, $p<.05$). The interval cancer rate was around 2 per 1000 women screened during the four periods, (Table 2). No trend was observed ($p=.051$).

For the proportion of DCIS of all screen-detected cancers, we noted a positive trend (+1.4%, $p<.001$), increasing from 11.7% in the first audit series to 16.4% in the fourth. Of all invasive detected cancers, the proportion of small tumours (T1a-c, size < 2 cm) increased (+1.3%, $p=.001$) from 76.4% in the first audit series to 80.6% in the fourth. In approximately 75% of screen-detected invasive cancers, there was no regional lymph node metastasis (N0). Over the 4 audit series, no trend was observed ($p=.412$) (Table 1).

For the review of the ICs, the percentage of missed cancers (category 3) varied between 20% and 30% in the four audit series (Table 3). No trend was observed over time ($p=.460$). We also did not observe a trend ($p=.323$) for the percentage of missed cancers for the review of the stage-II cancers, which again varied between 20% and 30%.

Table 2: Screening outcomes for subsequent screening examinations of the national evaluation of breast cancer screening

Outcome Measures period outcomes	1990-1997	1998-2003	2001-2006	2006-2009	Trend[a]	P Value
Total number of screened women	1376172	2851961	3667175	3054163		
Total number of recalled women	9977	30081	45392	45559		
Recall rate (%)	0.72	1.05	1.24	1.49	+0.25 (0.17, 0.33)	0.006
Total number of screen-detected cancers	4851	12643	17867	16537		
Detection rate (per 1000 screened)	3.5	4.4	4.9	5.4	+0.6 (0.3, 0.9)	0.012
Total number of interval cancers (< 24 months)	2656	5750	7481	6547		
Interval rate (per 1000 screened)	1.9	2.0	2.0	2.1	+0.1 (0, 0.1)	0.051
Sensitivity (%)	64.6	68.7	70.5	71.6	+2.3 (0.2, 4.4)	0.043

Data presented in this table are adopted from the Thirteenth evaluation report of the National Evaluation Team for Breast cancer screening [7]

[a] Numbers in parentheses are the 95% confidence interval

Table 3: Results of the radiological review of interval cancers and screen-detected stage II cancers in the four audit series

Results radiological review period audit series No. of included RUs	1st audit series 1996-2000 *17 RUs*	2nd audit series 2001-2005 ***17 RUs***	3rd audit series 2003-2007 ***14 RUs***	4th audit series 2010-2013 ***17 RUs***	Trend[a]	P Value
Total number of reviewed interval cancers	2034	1933	1544	664		
Category 1: true interval (%)	45.6 (35.9-56.4)	54.4 (44.9-63.5)	52.9 (40.0-69.1)	52.9 (41.0-65.0)		
Category 2: minimal sign (%)	26.0 (20.1-36.2)	23.7 (16.3-31.5)	22.6 (16.3-30.5)	21.3 (10.0-30.6)		
Category 3: missed cancer (%)	28.4 (17.2-39.1)	21.9 (13.8-31.3)	24.5 (11.8-38.8)	25.8 (12.8-35.9)	-0.5 (-1.7, 0.8)	0.460
Total number of reviewed StII cancers	849	1042	854	662		
Category 1: true interval (%)	45.3 (30.0-62.5)	51.3 (40.0-66.0)	46.7 (32.5-57.5)	51.4 (33.3-72.5)		
Category 2: minimal sign (%)	25.2 (12.5-34.5)	27.3 (12.2-38.8)	29.6 (17.5-40.0)	23.1 (7.9-35.9)		
Category 3: missed cancer (%)	29.6 (19.0-56.7)	21.4 (7.0-36.7)	23.5 (10.0-50.0)	25.4 (10.0-40.0)	-0.9 (-2.7, 0.9)	0.323

Numbers in parentheses are the range, except for the column Trend. For the trend analysis, categories 1 and 2 were combined.

[a] Numbers in parentheses are the 95% confidence interval

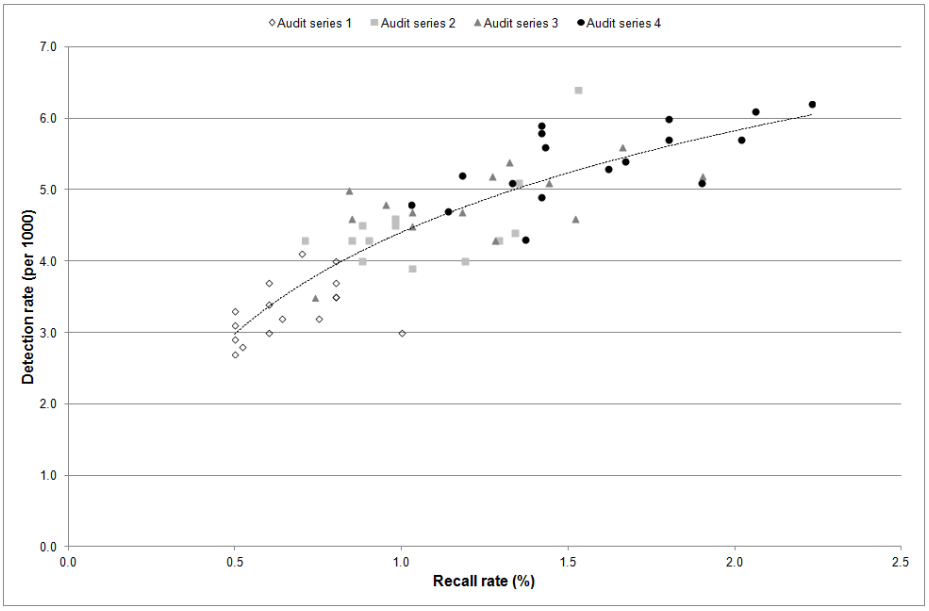


Figure 1: The relation between the recall rate (%) and the detection rate (per 1000 women screened) of the 17 RUs, determined during the 4 audit series; 1st audit series performed in 1996-2000, 2nd audit series performed in 2001-2005, 3rd audit series performed in 2003-2007 and 4th audit series performed in 2010-2013.

Table 4: Results of the radiological review of consecutive recalled cases in the 4th audit series

Reading Unit (RU)	No. of reviewed cases[a]	No. of agreements[b]	Agreements (%)	Kappa	Not recalled[d]
RU1	40	20	50.0	0.22	12 (30.0)
RU2	40	18	45.0	0.11	14 (35.0)
RU3	40	29	72.5	0.43	13 (32.5)
RU4	40	24	60.0	0.28	14 (35.0)
RU5	40	31	77.5	0.63	4 (10.0)
RU6	50	30	60.0	0.63	3 (6.0)
RU7	40	25	62.5	0.29	4 (10.0)
RU8	40	26	65.0	0.45	2 (5.0)
RU9	40	29	72.5	0.37	11 (27.5)
RU10	40	30	75.0	0.59	8 (20.0)
RU11	40	29	72.5	0.55	6 (15.0)
RU12	40	30	75.0	0.44	10 (25.0)
RU13	45	34	75.6	0.57	5 (11.1)
RU14	40	26	65.0	0.39	4 (10.0)
RU15	44	36	81.8	0.62	3 (6.8)
RU16	40	26	65.0	0.46	3 (7.5)
RU17	42	29	69.0	0.53	3 (7.1)
Average[c]			67.3 (45.0-81.8)	0.44 (0.11-0.63)	17.0 (5.0-35.0)

[a] The total number of reviewed cases during the review of consecutive recalled cases

[b] The total number of reviewed cases with an agreement in using BI-RADS among the radiologists of the RU and the radiologists of the audit team.

[c] Values in parentheses are the range

[d] The number of cases that were recalled in screening practice, but would not have been recalled by the audit team (values in parentheses are percentage). The average is the average percentage with the range in parentheses.

Table 5: Positive predictive value (PPV) for the BI-RADS categories 0, 4 and 5 for reading units in the 4th audit series

	BI-RADS 0			BI-RADS 4			BI-RADS 5		
	TP[a]	FP[b]	PPV[c]	TP[a]	FP[b]	PPV[c]	TP[a]	FP[b]	PPV[c]
RU1	0	4	0.0	7	27	20.6	2	0	100
RU2	2	4	33.3	15	18	45.5	1	0	100
RU3	1	30	3.2	2	7	22.2	0	0	-
RU4	0	14	0.0	2	22	8.3	0	0	-
RU5	1	24	4.0	5	5	50.0	3	2	60.0
RU6	3	10	23.1	12	12	50.0	13	0	100
RU7	1	14	6.7	7	17	29.2	1	0	100
RU8	5	14	26.3	8	10	44.4	3	0	100
RU9	0	25	0.0	5	10	33.3	0	0	-
RU10	2	15	11.8	11	9	55.0	2	0	100
RU11	4	17	19.0	7	8	46.7	3	1	75.0
RU12	0	9	0.0	6	22	21.4	2	1	66.7
RU13	12	19	38.7	6	3	66.7	5	0	100
RU14	2	10	16.7	8	16	33.3	4	0	100
RU15	1	15	6.3	12	11	52.2	5	0	100
RU16	4	15	21.1	14	4	77.8	3	0	100
RU17	6	15	28.6	10	6	62.5	4	1	80.0
Average			14.8			39.8			91.1

Data from the cases of the review of consecutive recalled cases (of the 4th audit series)

[a] The number of true-positive findings for the particular BI-RADS category

[b] The number of false-positive findings for the particular BI-RADS category

[c] The positive predictive value of recall for the particular BI-RADS category as percentage

In the audits after 2010, 701 consecutive recalled cases were reviewed. In 67.3% (range: 45.0%–81.8%), the audit team and the RU agreed on the BI-RADS score (Table 4). The inter-observer agreement of the RU with the audit team gives a $\kappa = 0.44$ (range: 0.11–0.63), which corresponds to moderate agreement [16, 17]. Overall, the PPV of BI-RADS 0 was 14.8% (range: 0–38.7%), PPV of BI-RADS 4 was 39.8% (range: 8.3-66.7%) and PPV of BI-RADS 5 was 91.1% (range: 60-100%) (Table 5). Finally, the number of cases recalled in screening practice but that would not have been recalled by the audit team were recorded. On average this happened in 17.0% (range: 5-35%) of the cases (Table 4). None of these 119 cases was a cancer (one year follow-up available). Figure 2 shows that the "percentage of cases the audit team would not have recalled" increased with increasing recall rate of the screening radiologists.

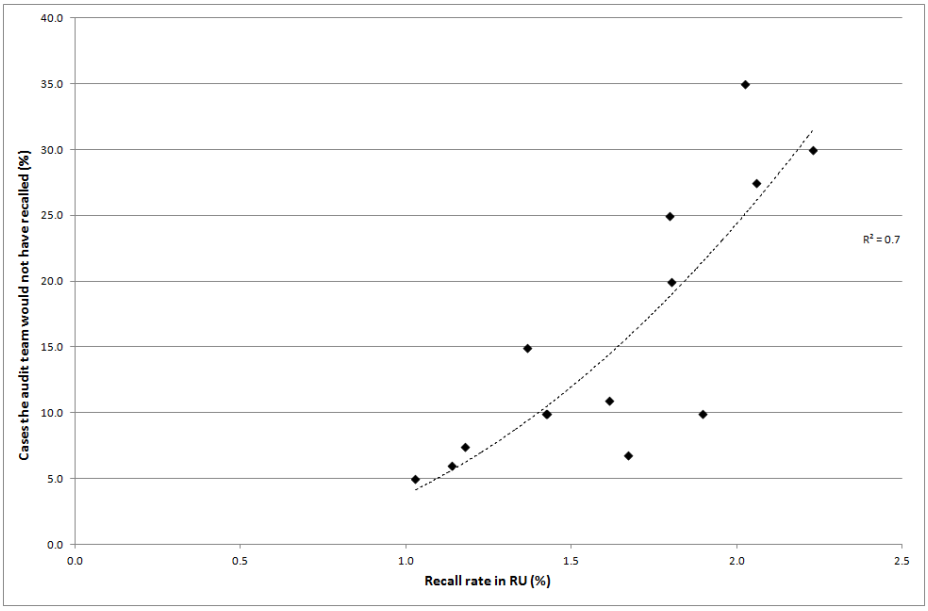


Figure 2: The association between the recall rate (%) and the "percentage of cases the audit team would not have recalled".

Discussion

Audits form a structural and obligatory part of the DBCSP quality assurance scheme. Since the start of the DBCSP, 4 audit series have been conducted, including 8 million screens, around 40,000 detected breast cancers and 10,000 peer reviewed cases. Our study was designed to investigate the value of those audits.

Over the 4 audit series we observed a positive trend in recall rate, detection rate and sensitivity, and a modest positive trend in the proportion of DCIS and of small tumours (T1a-c, size < 2 cm) detected. We observed no trend for the distribution of lymph-node status, the interval cancer rate and for the percentage missed cancers for reviews of ICs and stage-II cancers. Further we observed a moderate agreement in using BI-RADS among audit team vs. the RU. Finally we observed that the "percentage of cases the audit team would not have recalled" increased with a higher recall rate of the screening radiologists.

The recall rate in the DBCSP is one of the lowest worldwide. Low recall rates may result in more missed subtle cancers. However too high recall rates result in unnecessary further assessment, patient anxiety, and costs. Otten et al. [18]

reported that the extreme low recall rate noted in the early days of the DBCSP was not optimal, and recommended to lower the threshold for recall. During the audits, the LRCB advised all RU radiologists to lower the threshold for recall, which has contributed to the increased recall rates. As expected for this lower range of recall rates (<4%) these increased recall rates have resulted in increased detection rates, decreased PPV of recall and increased sensitivity [18,19].

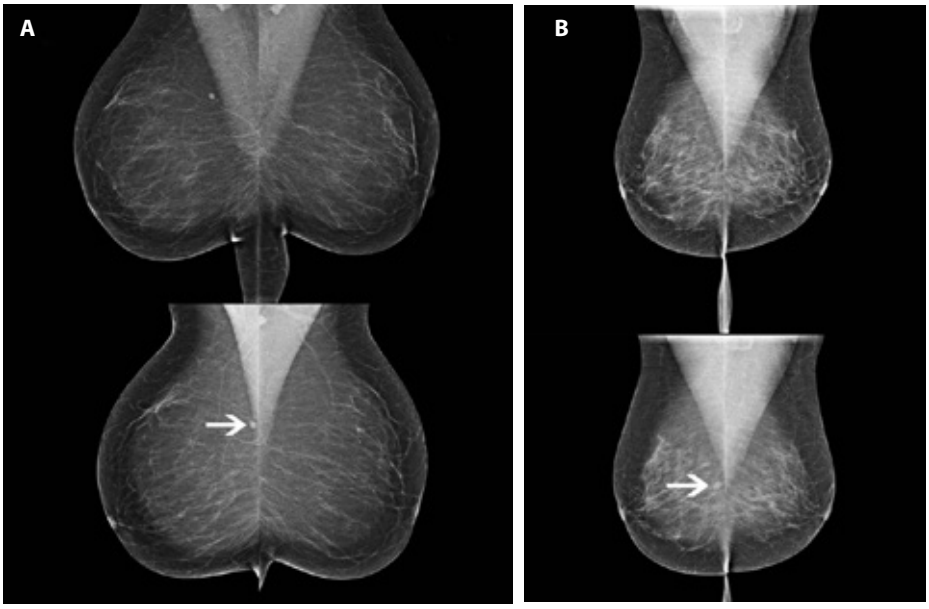


Figure 3: To avoid unnecessary recalls, all false positive cases should have clear signs that warrant recall. This figure gives two examples from the review of consecutive recalled cases. (A)This example is a mediolateral oblique screening mammogram of a woman recalled in screening practice with BI-RADS 0, because of a well-defined mass. Because of the shape and location of this mass, the experts of the audit team are convinced this concerns a lymph node. Since there was no change compared to the previous images, there are no clear signs for recall and according to the audit team it was not necessary to recall this woman. (B) This example is also a mediolateral oblique screening mammogram of a woman who was recalled in screening practice with BI-RADS 0, because of a well-defined mass. Although this mass also looks like a lymph node, it is new compared to the previous images. So, in contrast with the first example, this case has clear signs for recall.

The decreased PPV of recall implies a higher number of false positive cases. In these cases it is of interest to retrospectively determine which abnormality gave rise to the recall. During the reviews, on average 17% of the cases gave rise to discussion about the appropriateness of the recall. Figure 3 gives an example. These discussions encourage the RU radiologists to reflect on their recall behaviour, and should be aimed at avoiding unnecessary recalls in the future. Because diagnostic

work up is not included in the DBCSP, the RU radiologist is not necessarily the same as the radiologist engaged with the assessment. This is a disadvantage of the DBCSP. Still by far most RU radiologists work as breast radiologists and see most of the recalls from the screening in the multi disciplinary teams of their hospital.

Other discussions during the audits, led to criticizing the policy concerning single-view instead of two-view mammography, which resulted in changing this policy. Two-view mammography (MLO and CC) is now the standard. Because of the value of these discussions, we encourage as many RU radiologists to attend the audit (rewarded with CME points). After the audit the LRCB organises a two hour refresher course on site, to provide feedback to the radiologists who were not able to attend the audit.

Besides an increased detection rate and sensitivity, a modest positive trend in the proportion of DCIS and of small tumours (T1a-c, size < 2 cm) was observed. Like other studies [20–22], Van Luijt et al. [23] stated that this was related to better detection of microcalcifications using DM. In addition Van Luijt noted that the change from single-view to two-view mammography at subsequent examinations, which occurred together with the conversion to DM, may also have influenced the detection.

Also parallel to the conversion to DM, BI-RADS was introduced in the DBCSP. Our results showed a moderate agreement in using BI-RADS among audit team vs. the RU (κ=0.44), which matches results of Timmers et al. [12,13]. In our experience, the introduction of BI-RADS in screening resulted in an extra benchmark indicator, and assessment of PPV of BI-RADS 0, 4 and 5 offer new possibilities for influencing recall rates.

The percentage of missed cancers was between 20-30% on an aggregated level for both ICs and stage-II cancers. This also matches findings from other radiological review studies [24–29]. In a recent study, Pinto et al [30] report that the average error rate among radiologists is around 30%. The implementation of a peer review process is one method of measuring diagnostic errors. However, as noted by Pinto, it is important that radiologists are not 'blamed' for cases of missed cancers, and that these are treated as learning opportunities. In addition, it is possible that the proportion of missed cancers will decrease in the future. A study by Nederend et al [31], showed that a smaller proportion of interval cancers was missed at FFDM screening than at SFM. The ICs collected and reviewed during the audits in our study are however mostly diagnosed after a SFM examination (estimated 90%).

During the review, radiologists are unaware of the laterality and location of the IC, but know that all cases are ICs. As stated by Ciatto, "The more informed the review, the more likely that a screen will be classified as missed cancers." [25] This is a limitation of our review. A mix of ICs with proven normal cases could be investigated in the future. Another improvement would be to review only the subset of stage-II cancers that were larger than 2 cm at diagnosis (T2+), regardless of lymph-node status. After the introduction of the sentinel node procedure (SN), in the 40 reviewed stage-II cancers, more and more cases were included with a positive lymph-node status based on SN and a small tumour size. In our review these cases mostly turned out to be a true negative (category 1, no abnormality is seen). For a radiological review these cases are not the most interesting as they are not yet visible on the mammogram.

Our study has several limitations. The period of data collection for our study covers a long period of time, in which several changes took place in the screening programme. The conversion from SFM to DM changed the reading strategy from non-blinded to blinded double reading. In addition, two-view mammography became the standard procedure. Other studies showed that these modifications may effect for instance recall rate [20–23, 32, 33]. In our study, we were not able to investigate the influence of these changes on the outcome parameters. Another limitation of our study is the incompleteness (an estimated 20%) of the IC data during the audits. Given the fact that interval cancers were identified through different sources, we believe it is unlikely that selection bias was introduced in the review.

The European Guidelines [4] outline the optimal content of a QA scheme, including a review of ICs as a part of routine radiological audits. Implementation is not mandatory for the European countries. The content can be used to support and enhance local guidelines [34]. In practice, only a few countries implemented such a routine radiological review. There are countries which perform radiological review of ICs for evaluation studies or only as individual case-review. An audit not only provides an opportunity for assessing screening outcomes, but also provides moments of self-reflection with peers. We therefore recommend that in addition to benchmarking screening outcomes, a radiological review of screening examinations and immediate feedback should be part of an audit. This provides insights in recall behaviour and cancer characteristics that cannot be gathered from epidemiological surveillance. By reviewing cases where the mammograms show very subtle changes, radiologists will be able to improve their skills in

detecting small breast cancers. For radiologists, an accurate understanding of their performance is essential to know which points are most in need of improvement.

Acknowledgments

The authors thank the screening organisations (Foundation of Population Screening East, Foundation of Population Screening Mid-West, Foundation of Population Screening South, Foundation of Population Screening South-West) for providing the data, and all members of the audit team, for their help with the study.

Dr. P.G.M. Peer of the Department for Health Evidence of the Radboud University Medical Center, Nijmegen, The Netherlands kindly provided statistical advice for this manuscript.

References

1. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M (2013) The benefits and harms of breast cancer screening: an independent review. Br J Cancer 108:2205–2240
2. Kalager M, Zelen M, Langmark F, Adami H (2010) Effect of screening mammography on breast-cancer mortality in Norway. N Engl J Med 363:1203–1210
3. Gray JAM, Patnick J, Blanks RG (2013) Maximising benefit and minimising harm of screening. BMJ 336:480–483
4. Perry N, Broeders M, Wolf C De, Törnberg S, Karsa L Von (2006) European guidelines for quality assurance in breast cancer screening and diagnosis, 4th edn., European Commission, Luxembourg
5. Mammography Quality Standards: Final Rule (2007) Food and Drug Administration, Rockville
6. National Evaluation of breast cancer screening in the Netherlands, 1990-2007, the Twelth evaluation report (2009) LETB, Rotterdam
7. National Evaluation of breast cancer screening in the Netherlands, 1990-2011, the Thirteenth evaluation report (2014) LETB, Rotterdam
8. Houssami N, Irwig L, Ciatto S (2006) Radiological surveillance of interval breast cancers in screening programmes. Lancet Oncol 7:259–265
9. Fracheboud J, de Koning HJ, Beemsterboer PM, et al (1998) Nation-wide breast cancer screening in The Netherlands: results of initial and subsequent screening 1990-1995. National Evaluation Team for Breast Cancer Screening. Int J Cancer 75:694–698
10. Holland R, Rijken H, Hendriks J (2007) The Dutch Population-Based Mammography Screening: 30-Year Experience. Breast Care 2:12–18
11. Breast imaging reporting and data system (2003), 4th edn. American College of Radiology, Reston
12. Timmers JMH, van Doorne-Nagtegaal HJ, Verbeek a LM, den Heeten GJ, Broeders MJM (2012) A dedicated BI-RADS training programme: effect on the inter-observer variation among screening radiologists. Eur J Radiol 81:2184–2188
13. Timmers JM, van Doorne-Nagtegaal HJ, Zonderland HM, et al (2012) The Breast Imaging Reporting and Data System (BI-RADS) in the Dutch breast cancer screening programme: its role as an assessment and stratification tool. Eur Radiol 22:1717–1723
14. Interim Report 2011. Main results 2008-2009 breast cancer screening programme in the Netherlands (2011) LETB, Rotterdam
15. Van der Horst F, Hendriks JH, Rijken H (2003) Breast cancer screening in The Netherlands: audit and training of radiologists. Semin Breast Dis 6:114–122
16. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46.
17. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
18. Otten JD, Karssemeijer N, Hendriks JH, et al (2005) Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. J Natl Cancer Inst 97:748–754
19. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R (2001) Association of recall rates with sensitivity and positive predictive values of screening mammography. AJR Am J 177:543–549
20. Karssemeijer N, Bluekens AM, Beijerinck D, et al (2009) Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program. Radiology 253:353–358

21. De Gelder R, Fracheboud J, Heijnsdijk EAM, et al (2011) Digital mammography screening: weighing reduced mortality against increased overdiagnosis. Prev Med 53:134–140
22. Bluekens AMJ, Holland R, Karssemeijer N, Broeders MJM, den Heeten GJ (2012) Comparison of digital screening mammography and screen-film mammography in the early detection of clinically relevant cancers: a multicenter study. Radiology 265:707–714
23. Van Luijt PA, Fracheboud J, Heijnsdijk EAM, den Heeten GJ, de Koning HJ (2013) Nation-wide data on screening performance during the transition to digital mammography: Observations in 6 million screens. Eur J Cancer 31:1–9
24. Hoff SR, Abrahamsen AL, Samset JH, Vigeland E, Klepp O, Hofvind S (2012) Breast cancer: missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography-- results from a retrospective review. Radiology 264:378–386
25. Ciatto S, Catarzi S, Lamberini MP, et al (2007) Interval breast cancers in screening: the effect of mammography review method on classification. Breast 16:646–652
26. Hofvind S, Skaane P, Vitak B, et al (2005) Influence of review design on percentages of missed interval breast cancers: retrospective study of interval cancers in a population-based screening program. Radiology 237:437–443
27. Van Dijck JA, Verbeek AL, Hendriks JH, Holland R (1993) The current detectability of breast cancer in a mammographic screening program. A review of the previous mammograms of interval and screen-detected cancers. Cancer 72:1933–1938
28. Ciatto S, Bernardi D, Pellegrini M, et al (2012) Proportional incidence and radiological review of large (T2+) breast cancers as surrogate indicators of screening programme performance. Eur Radiol 22:1250–1254
29. Smallenburg VB, Setz-Pels W, Groenewoud JH, et al (2012) Malpractice claims following screening mammography in The Netherlands. Int J Cancer 131:1360–1366
30. Pinto A, Acampora C, Pinto F, Kourdioukova E, Romano L, Verstraete K (2011) Learning from diagnostic errors: a good way to improve education in radiology. Eur J Radiol 78:372–376
31. Nederend J, Duijm LEM, Louwman MWJ, et al. (2013) Impact of the transition from screen-film to digital screening mammography on interval cancer characteristics and treatment - A population based study from the Netherlands. Eur J Cancer 50: 31-39
32. Duijm LEM, Groenewoud JH, Hendriks JHCL, de Koning HJ (2004) Independent double reading of screening mammograms in The Netherlands: effect of arbitration following reader disagreements. Radiology 231:564–70
33. Klompenhouwer EG, Duijm LEM, Voogd AC, et al. (2014) Variations in screening outcome among pairs of screening radiologists at non-blinded double reading of screening mammograms: a population-based study. Eur Radiol 24: 1097-1104
34. Klabunde CN, Sancho-Garnier H, Broeders M, Thoresen S, Rodrigues VJL, Ballard-Barbash R (2001) Quality assurance for screening mammography data collection systems in 22 countries. Int J Technol Assess Health Care 17:528-541

Radboud

Chapter 6

Applying the "positive predictive value—recall diagram" to monitor performance and provide recommendations for screening radiologists

Geertse TD, Tetteroo E, Smid-Geirnaerdt MJA, Duijm LEM, Pijnappel RM, van der Waal D, Broeders MJM

European Radiology, September 2025, https://doi.org/10.1007/s00330-025-11978-3

Abstract

Objectives

To evaluate the suitability of "positive predictive value—recall" (PPV-recall) diagrams for monitoring performance and providing recommendations for groups of radiologists (RUs or reading units) in breast cancer screening.

Materials & Methods

This retrospective study used datasets from triennial quality assurance audits within the Dutch screening programme. The recall rate (RR), cancer detection rate (CDR), and PPV between 2010-2019 were plotted in PPV-recall diagrams separately for initial and subsequent screening. Using PPV-recall diagrams per year we compared variations in performance of the RUs within the screening programme. Each group's screening behaviour characteristics were evaluated over time with RU-specific PPV-recall diagrams and related audit recommendations.

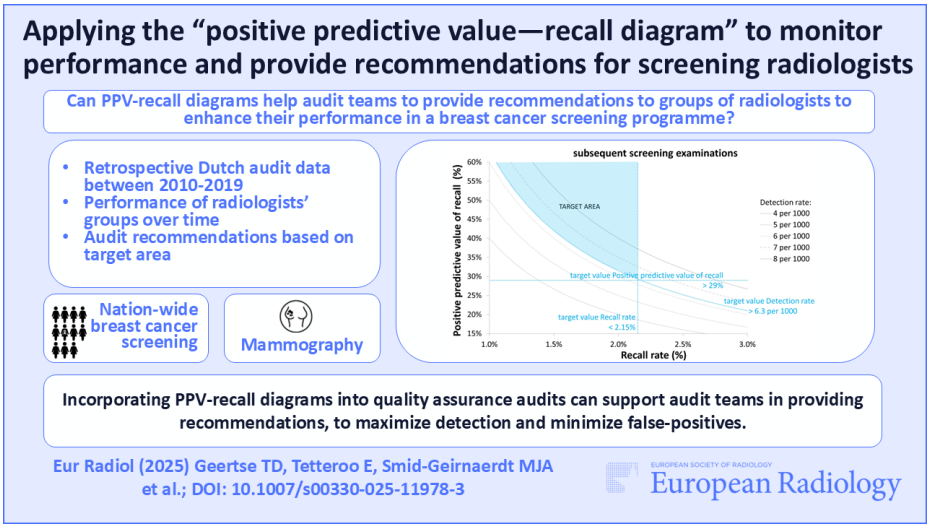
Results

The dataset comprised the aggregated results of 779,887 initial and 6,021,598 subsequent screenings read by 12 RUs between 2010-2019. The PPV-recall diagrams showed substantial variations in the individual RU performance over time, with PPVs ranging between 4.9-23.7% for initial and 21.2-54.3% for subsequent screening. Target values were less often met for initial (2010: 0 RUs; 2019: 5 RUs) than for subsequent screening (2010: 8 RUs; 2019: 10 RUs), resulting in more recommendations regarding initial screening (24 versus 13). All recommendations focused on adjusting RR, which often (17 out of 24) changed in the recommended direction, though not always sufficient to meet target values.

Conclusion

PPV-recall diagrams offer valuable insights into variations and interrelationships between screening outcomes, helping the audit team in providing recommendations for improvement. However, feedback based on these diagrams alone may not always be sufficient for individual radiologists to achieve these improvements.

Graphical abstract



Introduction

Breast cancer screening can prevent breast cancer-related death through early detection and treatment; however, it also inflicts harms such as overdiagnosis and false-positive screening results [1, 2]. Quality assurance (QA) programmes, including quality control of equipment, training and accreditation of professionals, and monitoring of screening outcomes, should ensure a favourable balance between the benefits and harms and guarantee a high level of quality within the screening programme [3-5].

For monitoring screening outcomes, an effective method is needed to provide insight into the reading performance of screening radiologists, to identify areas for improvement and, pinpoint (groups of) radiologists who may require additional training. Recall rate (RR), cancer detection rate (CDR), and positive predictive value of recall (PPV) are key parameters for this purpose, which are best monitored combined, as they are interrelated [6]. In 2001, Blanks introduced the PPV-recall diagram as a method for monitoring screening outcomes [7]. In a PPV-recall diagram the PPV is plotted against the RR, with the CDR shown as "isobars" in the graph, which visualises the interrelationship between these three screening outcomes [7]. The diagram may also include target values and/or a national average of these screening outcomes.

In the Netherlands, reading performance of the screening radiologists has been monitored at group level, with a group referred to as a reading unit [RU], during triennial audits. As part of the audit, the PPV-recall diagrams for the RUs have been reviewed for several years [8]. The effectiveness of using the PPV-recall diagram to assess quality, identify areas for improvement and formulate recommendations has not been fully established.

Therefore, the aim of this study was to retrospectively evaluate the suitability of PPV-recall diagrams for monitoring reading performance and providing recommendations for RUs in a breast cancer screening programme.

Materials and Methods

This retrospective descriptive study was performed under the national permit for breast cancer screening issued by the Dutch Ministry of Health, Welfare and Sport, which is equivalent to approval by a local institutional review board.

Screening procedure

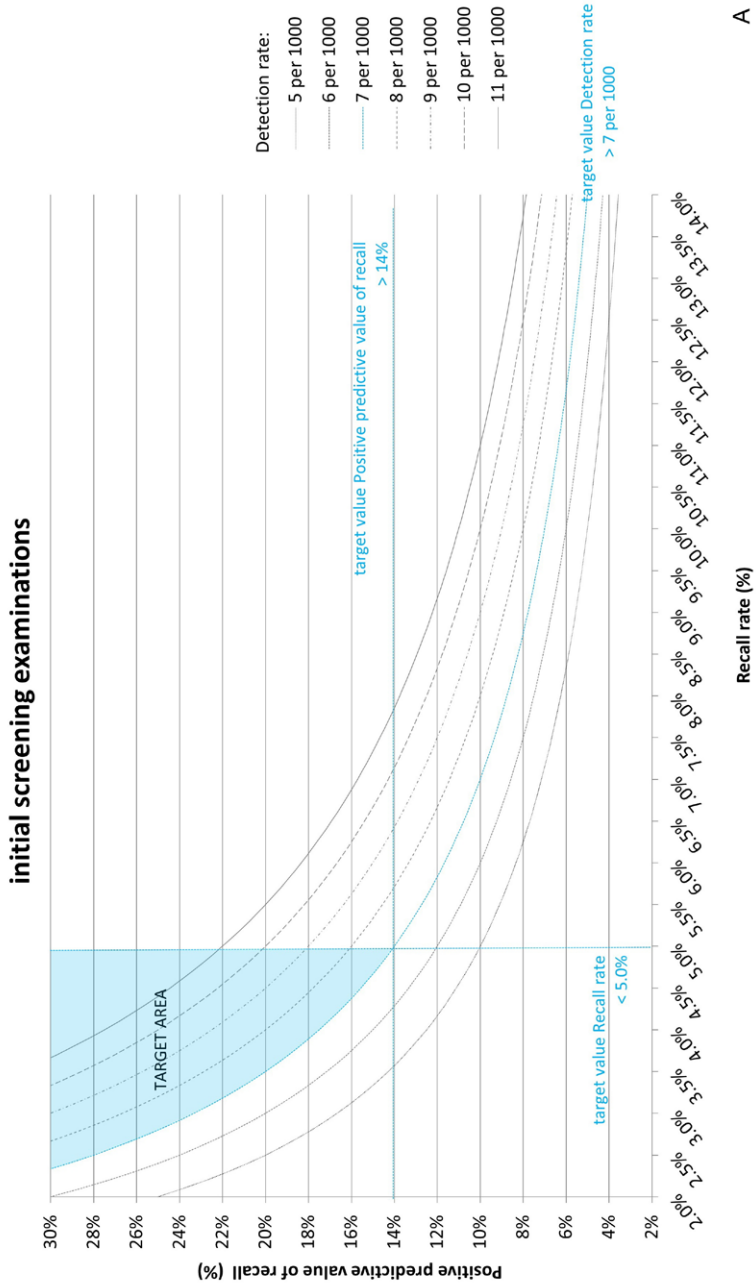
The Dutch breast cancer screening programme is centrally organised by the national screening organisation, which manages all screening units and RUs. Biennially, women aged 50-74 years are invited for a screening examination, performed by radiographers specialised in mammography. The standard examination is two-view mammography (mediolateral oblique view and craniocaudal view). Initially, screen-film mammography was used, with mammography systems from different manufacturers. In 2008, the transition to digital mammography (Lorad Selenia, Hologic) began and was completed by June 2010.

All mammograms are independently read by two certified screening radiologists. Discrepancies are solved by consensus or arbitration by a third reader and prior examinations are available for comparison in subsequent screening. Women with suspicious findings are recalled to a hospital for further assessment.

Audits

Triennial audits are conducted by the Dutch Expert Centre for Screening (LRCB) following a fixed protocol, described in detail in a previous study [8]. In summary, the screening organisation provides a dataset for each audit, divided into initial and subsequent screening outcomes, covering all examinations in a four-year period. The oldest year usually overlaps with the previous audit, as audits are conducted approximately every three years. For each audit, the mean screening outcomes over the four-year period are compared to target values (Table 1) and mean national values [9, 10], using tables and graphs, including the PPV-recall diagram. If the RR, PPV or CDR do not meet the target value, and thus the RU performs outside the "target area" in the PPV-recall diagram (see Figure 1), the audit team recommends improvements to the RU (audit recommendation).

After the transition to digital mammography in 2010, RRs increased considerably [11]. This led to the adjustment of target values in July 2016 (Table 1), based on an advisory report from the LRCB to the policymakers [12].


initial screening examinations
Positive predictive value of recall (%)
30%
28%
26%
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
TARGET AREA
target value Positive predictive value of recall
> 14%
target value Recall rate
< 5.0%
target value Detection rate
> 7 per 1000
Detection rate:
5 per 1000
6 per 1000
7 per 1000
8 per 1000
9 per 1000
10 per 1000
11 per 1000
2.0%
2.5%
3.0%
3.5%
4.0%
4.5%
5.0%
5.5%
6.0%
6.5%
7.0%
7.5%
8.0%
8.5%
9.0%
9.5%
10.0%
10.5%
11.0%
11.5%
12.0%
12.5%
13.0%
13.5%
14.0%
Recall rate (%)
A

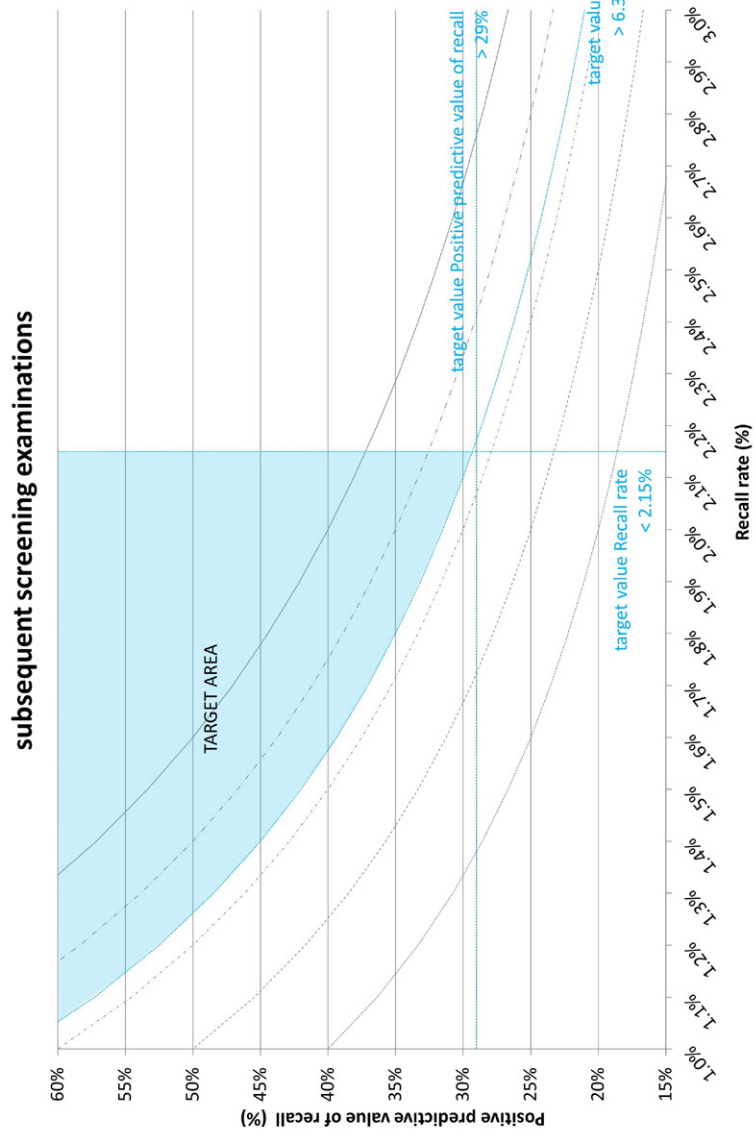


Figure 1: The PPV-recall diagram, showing positive predictive value of recall (PPV) versus recall rate (RR), with cancer detection rate (CDR) presented as isobars for initial screening (A) and subsequent screening (B). The "target area," where all target values are met, is shaded in blue. Screening outcomes outside this area deviate from the targets, and specific recommendations should be issued.

Table 1: Target values set for the Dutch breast cancer screening programme

Performance measure	Until July 2016	From July 2016
Initial screening:		
Recall rate (%)	2.5-3.5[a]	< 5.0
Detection rate (per 1000 screens)	> 5.5	> 7
Positive predictive value of recall (%)	> 16	> 14
Subsequent screening:		
Recall rate (%)	1.3-2.0[a]	< 2.15
Detection rate (per 1000 screens)	> 4.5	> 6.3
Positive predictive value of recall (%)	> 22.5	> 29

[a] Until July 2016 for the recall rate a target range was set instead of a target value

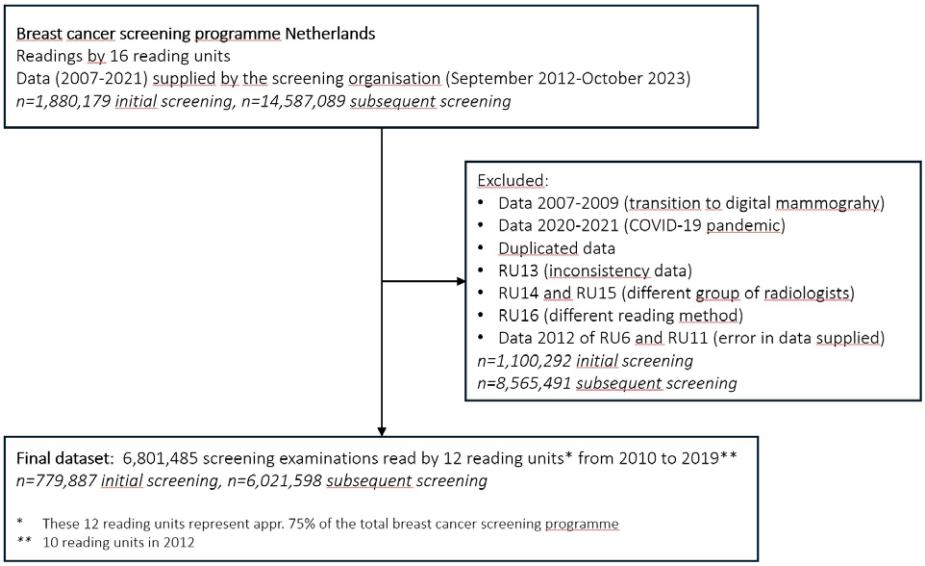


Figure 2: Flowchart to describe the selection of study data

Data collection

Data for this study were supplied by the screening organisation for audits conducted between September 2012 and October 2023, monitoring reading performance from 2007 to 2021 (Figure 2). Data were extracted on number of screenings, recalls, and screen-detected carcinomas (in-situ and invasive) per year, for both initial and subsequent screening. Only digital mammography data from 2010-2019 were used. Data from 2007-2009 were excluded due to the transition from analogue to digital mammography, and data from 2020-2021 were excluded due to the impact of the COVID-19 pandemic on the screening programme. In case of duplicate data, due to the triennial audit cycle and overlapping four-year periods (see "Audits"), only the most recent submissions were included. For inclusion, the data from a RU had to be complete and consistent, there could not have been a complete turnover of the radiologist team, and the RU had to adhere to the standard reading method. Otherwise the RU was excluded (see Figure 2). Audit reports were also available, describing audit outcomes and recommendations given to the RUs.

Data analysis

The screening outcomes RR, CDR, and PPV were calculated with 95% confidence intervals (95% CI), presented separately for initial and subsequent screening. The 95% CIs were calculated using the standard formula: 95% CI=$P \pm 1.96 * \sqrt{\frac{P*(1-P)}{N}}$, where P is the observed proportion and N is the number of screenings used to calculate the proportion. Furthermore, we calculated the median with interquartile range (IQR) for the RR, CDR, and PPV for each year.

The screening outcomes were graphically depicted in PPV-recall diagrams for both initial and subsequent screening, for each individual year within the study period (2010-2019). Each diagram contains data points for the 12 RUs, showing how they performed relative to each other. In addition, PPV-recall diagrams were plotted for each individual RU over a 10-year period. In these diagrams, audit years and recommendations were also indicated, enabling assessment of performance changes following audits, thus focusing more on the screening behaviour characteristics of each RU.

For all recommendations given by the audit team (based on the mean RR during the evaluation period), absolute changes in RR were assessed over the three years following the audit. We used the RR from the last year of the evaluation period as the 'baseline', for initial and subsequent screening. If the audit took place between 1 January and 30 June, the year in which the audit took place was also considered to be the first year after the audit. Changes in RR could not be assessed if the first year after the audit fell outside the study period (2020 or later).

Results

The dataset supplied by the screening organisation for the audits conducted between September 2012 and October 2023, consisted of the aggregated screening results of 16,467,268 mammography screenings (1,880,179 initial screenings and 14,587,089 subsequent screenings), read by 16 RUs. In total 9,665,783 screenings were excluded (1,100,292 initial screenings and 8,565,491 subsequent screenings). Our final dataset consisted of the aggregated screening results of 6,801,485 digital mammography screenings read by 12 RUs from 2010 to 2019 (Figure 2). The aggregated results per year are presented in Tables 2 and 3, for initial (n= 779,887) and subsequent screening (n=6,021,598), respectively.

Variations and trends in RU performance, for initial screening

For initial screening, the PPV-recall diagrams per year for 2010 to 2019 show substantial variations in performance among the individual RUs. The lowest PPV observed during the study period was 4.9% (2014, RU12) with a RR of 12.4% and a CDR of 6.1 per 1000 screenings. Conversely, the highest PPV was 23.7% (2014, RU2), based on a RR of 3.6% and a CDR of 8.6 per 1000 screenings. For overall performance across all units, the lowest median PPV was 11.8% (2013), and the highest median PPV was 15.3% (2012) (Table 2).

In 2010 (Figure 3A), following the transition to digital mammography, all RUs performed outside the "target area". The "target area" refers to the range where all three target values were met. Specifically, five RUs (RU8-RU12) had a PPV below the target value (16%), and all RUs had a RR exceeding the target value (3.5%). The median RR was 4.9%. All RUs had a CDR comparable to or above the target value (5.5 per 1000 screenings). In 2013, the median RR had increased to 6.3%, and all RUs still had a performance outside the "target area". Three RUs (RU4, RU10 and RU12) had a RR around or over 10% (Figure 3B). After new target values were introduced in mid-2016 (Table 1), most RUs (except RU2, RU8 and RU9) continued to perform outside the "target area" (Figure 3C). In the following three years, performance improved slightly. In 2019, five RUs (RU1, RU2, RU4, RU6 and RU7) performed on the border of the "target area" (Figure 3D) and no RRs exceeded 10%.

Variations and trends in RU performance, for subsequent screening

Similar to initial screening, the PPV-recall diagrams per year for 2010 to 2019 for subsequent screening show substantial variations in performance among the individual RUs. The lowest PPV observed during the study period was 21.2% (2014, RU12), based on a RR of 2.9% and a CDR of 6.0 per 1000 screenings. The highest PPV

Table 2: Aggregated results per year of all 12 reading units for recall rate, PPV of recall and detection rate for the initial screening examinations

Year (n=12)	Total initial screening examinations	Total recalls	Total screen-detected cancers	Recall rate %, median (IQR), range	PPV of recall %, median (IQR), range	Detection rate per 1000, median (IQR), range
2010	74399	3719	522	4.9 (4.1-6.1), 3.8-9.3	13.3 (10.1-17.6), 8.8-21.8	6.9 (5.9-7.8), 4.9-10.0
2011	80504	4103	586	5.3 (4.3-6.3), 3.7-7.3	14.3 (10.4-19.1), 7.4-21.5	7.9 (5.5-8.4), 5.2-8.8
2012[a]	74195	4234	597	5.1 (4.4-8.5), 4.1-9.5	15.3 (10.0-19.2), 7.9-20.5	8.1 (7.8-8.6), 6.8-9.5
2013	78939	5162	689	6.1 (4.9-9.4), 4.1-11.5	11.8 (9.5-18.7), 7.4-21.6	8.7 (7.8-10.0), 5.0-10.9
2014	81557	4921	634	5.5 (4.8-7.2), 3.6-12.4	13.1 (11.3-15.8), 4.9-23.7	7.8 (6.3-8.6), 5.1-10.6
2015	80455	4837	674	6.8 (5.2-7.2), 3.1-9.0	14.4 (12.0-16.7), 6.4-22.7	8.5 (7.0-10.1), 5.8-10.7
2016	80620	5187	743	7.0 (5.5-7.5), 3.2-9.3	13.2 (11.3-18.4), 10.4-23.2	9.3 (8.6-9.6), 7.3-10.6
2017	77698	4924	640	6.2 (5.3-7.7), 3.8-9.1	13.4 (10.6-15.4), 9.1-18.3	8.1 (7.1-9.1), 5.9-9.7
2018	76104	4509	620	5.6 (5.1-6.8), 4.7-7.4	14.3 (13.0-16.6), 9.0-18.0	8.5 (6.8-9.2), 6.7-9.4
2019	75416	4721	657	5.9 (5.2-6.9), 4.4-8.4	15.1 (11.6-16.7), 9.4-18.1	8.5 (7.8-9.4), 5.6-10.5

IQR interquartile range, *PPV* positive predictive value

[a] n=10 reading units

Table 3: Aggregated results per year of all 12 reading units for recall rate, PPV of recall and detection rate for the subsequent screening examinations

Year (n=12)	Total subsequent screening examinations	Total recalls	Total screen-detected cancers	Recall rate %, median (IQR), range	PPV of recall %, median (IQR), range	Detection rate per 1000, median (IQR), range
2010	543843	9220	3207	1.6 (1.4-2.1), 1.1-2.4	34.2 (28.5-40.6), 24.4-51.8	6.0 (5.2-6.3), 4.1-6.7
2011	607849	10534	3603	1.8 (1.5-1.9), 1.3-2.4	33.4 (32.4-38.5), 23.4-43.9	6.1 (5.7-6.3), 4.9-6.7
2012[a]	557550	10190	3396	1.9 (1.6-2.0), 1.4-2.5	33.4 (27.4-37.8), 24.4-41.0	6.1 (5.8-6.3), 5.1-7.0
2013	620961	12389	3956	1.9 (1.8-2.4), 1.5-2.8	31.1 (26.8-34.3), 24.3-42.6	6.4 (6.0-6.7), 5.6-7.9
2014	611087	11704	3926	1.9 (1.6-2.2), 1.1-2.9	34.7 (29.4-38.2), 21.2-46.7	6.2 (5.9-6.9), 5.3-7.5
2015	632228	11542	4135	1.9 (1.7-2.0), 1.1-2.2	34.7 (31.9-38.4), 30.4-54.3	6.6 (6.1-6.8), 5.8-7.4
2016	638783	12041	4276	1.9 (1.6-2.1), 1.0-2.5	35.0 (31.6-38.1), 28.5-54.0	6.8 (5.7-7.1), 5.3-8.0
2017	639283	11661	4112	1.8 (1.6-1.9), 1.0-2.5	34.6 (32.4-39.1), 28.4-51.1	6.5 (6.0-6.8), 5.2-7.2
2018	610791	10975	3925	1.7 (1.5-2.0), 1.3-2.4	36.6 (31.2-44.4), 28.3-45.6	6.3 (6.0-7.0), 5.3-7.1
2019	559223	10824	3759	1.8 (1.7-2.1), 1.4-2.7	37.2 (34.3-38.8), 26.7-40.1	6.8 (6.3-7.1), 5.7-7.5

IQR interquartile range, *PPV* positive predictive value

[a] n=10 reading units

was 54.3% (2015, RU2) with a RR of 1.1% and a CDR of 6.0 per 1000 screenings. For overall performance across all units, the lowest median PPV was 31.1% (2013), and the highest median PPV was 37.2% (2019) (Table 3).

Compared to initial screening, RUs more often met the target values. In 2010, 8 out of 12 RUs met all three target values (Figure 4A). Of the four RUs outside the "target area", three (RU8, RU9 and R11) had a RR above the target range of 1.3-2.0% and one (RU4) below. The median RR was 1.6% (Table 3).

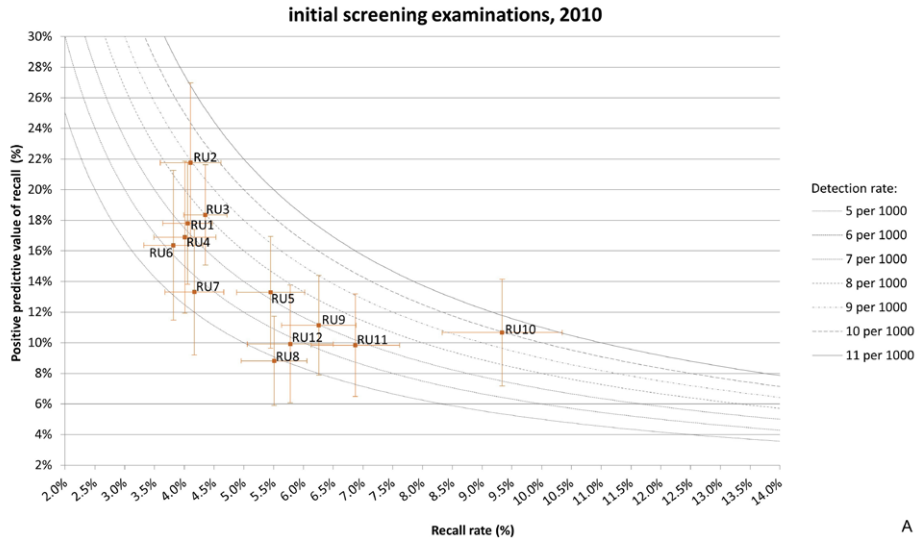
Similar to initial screening, the RR increased over the next three years. In 2013, the median RR was 1.9%, with five RUs (RU4, RU8, RU10-RU12) showing a RR above the target range (Figure 4B). Between 2013 and 2016, overall performance increased. The median PPV increased from 31.1% to 35.0% (Table 3). New target values were introduced in mid-2016 (Table 1). In 2016, all RUs except for one (RU7) met all three target values (Figure 4C). From 2016 to 2019, the overall performance increased slightly further. The median PPV increased to 37.2% in 2019 (Table 3). Now two RUs (RU3 and RU5) performed outside the "target area" (Figure 4D).

Screening behaviour characteristics of the individual reading units

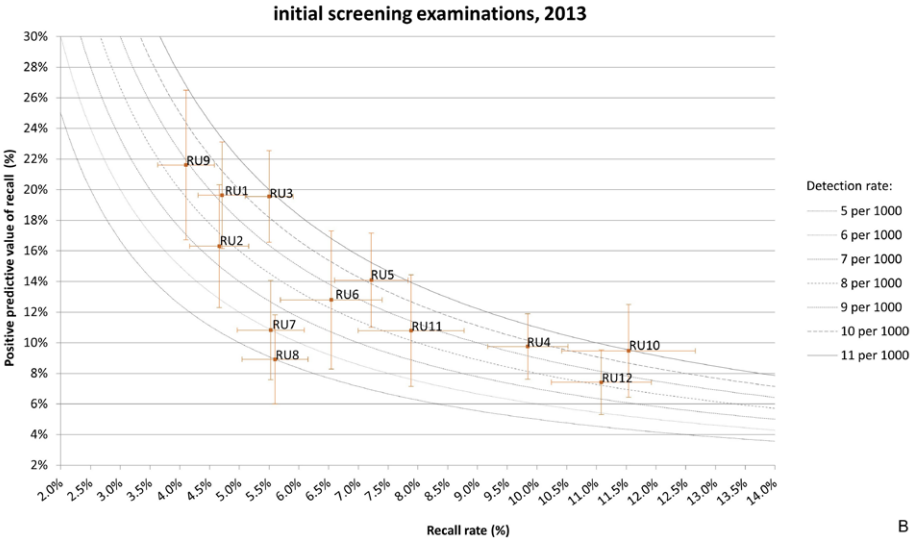
The PPV-recall diagrams per RU over all years provide insight into the screening behaviour characteristics of individual RUs. To illustrate this, we present two examples. We selected RU2 and RU12, because RU2 had the highest PPV observed during the study period for both initial (23.7%, 2014) and subsequent screening (54.3%, 2015) and RU12 had the lowest PPV for both initial (4.9%, 2014) and subsequent screening (21.2%, 2014).

For RU2, the PPV-recall diagrams (Figure 5) show a relatively low RR and high PPV. Four audits were conducted in the period 2010-2019. Based on the 2017 audit (data 2012-2015), RU2 was recommended to increase their RR, as the PPV-recall diagrams showed a high PPV alongside a declining RR for both initial and subsequent screening. Although the CDR met the target value, the audit team expected that the very low RR allowed for a slight increase, potentially resulting in an increased CDR. In the period after 2017 the RR increased while still meeting the target value, leading to a higher CDR in 2019.

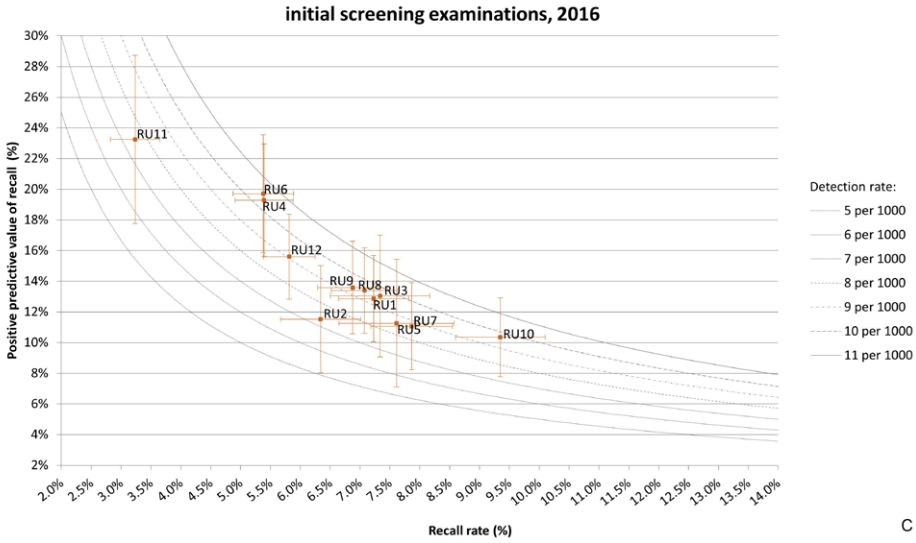
For RU12, the PPV-recall diagrams (Figure 6) show that this RU is characterized by a very high RR and low PPV in the period 2012-2015. Three audits were conducted during the study period, with recommendations provided during the audits in 2015 and 2018. Based on the 2015 audit (data 2010-2013), the radiologists were strongly advised to


initial screening examinations, 2010
Positive predictive value of recall (%)
Recall rate (%)
RU1
RU2
RU3
RU4
RU5
RU6
RU7
RU8
RU9
RU10
RU11
RU12
Detection rate:
5 per 1000
6 per 1000
7 per 1000
8 per 1000
9 per 1000
10 per 1000
11 per 1000

A


initial screening examinations, 2013
Positive predictive value of recall (%)
Recall rate (%)
RU1
RU2
RU3
RU4
RU5
RU6
RU7
RU8
RU9
RU10
RU11
RU12
Detection rate:
5 per 1000
6 per 1000
7 per 1000
8 per 1000
9 per 1000
10 per 1000
11 per 1000

B



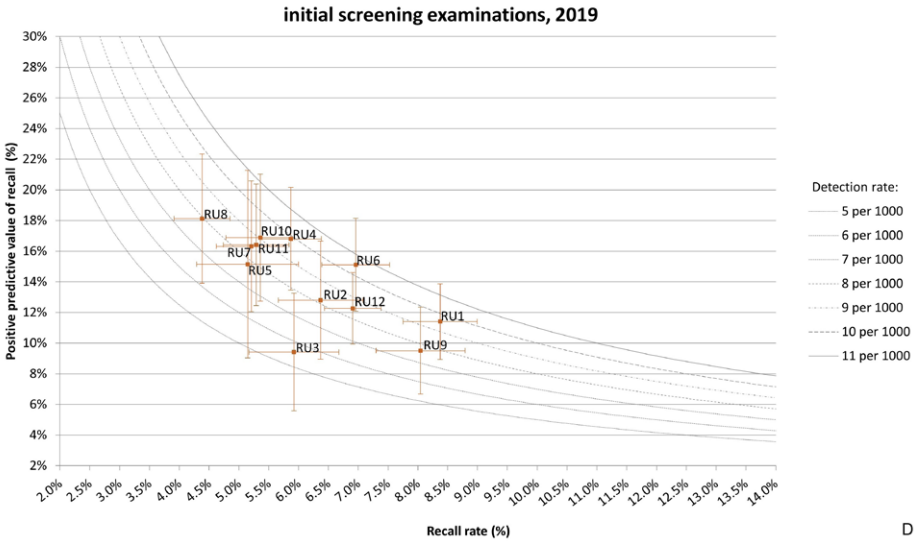
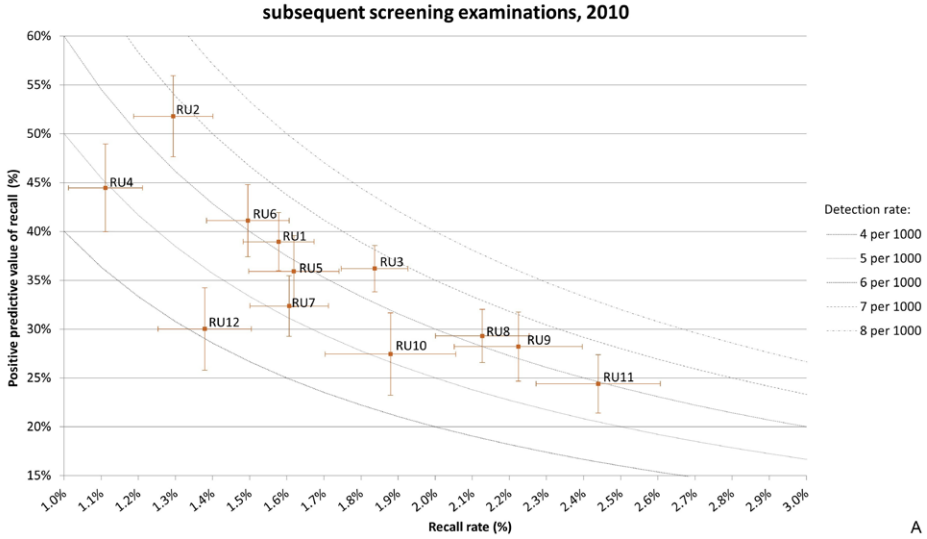
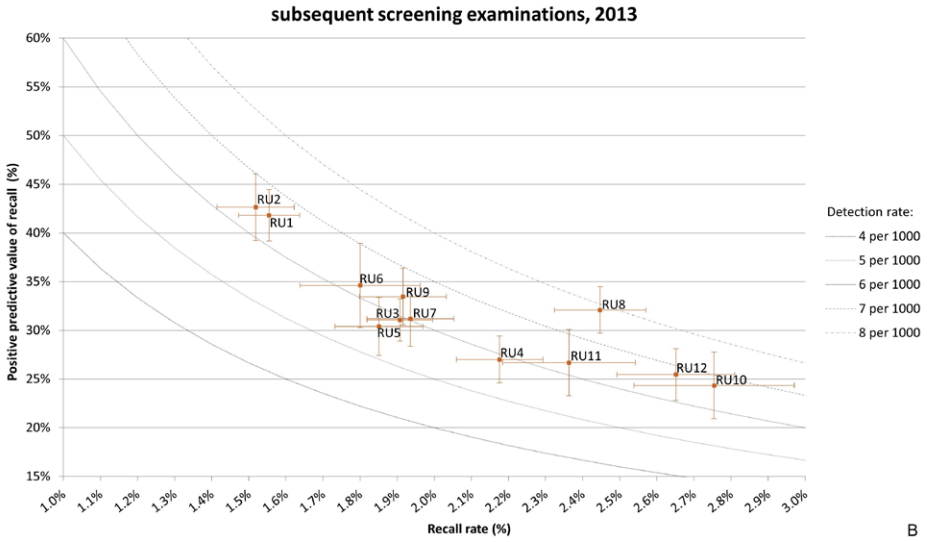


Figure 3: PPV-recall diagrams, with cancer detection rate (CDR) presented as isobars, for initial screening examinations for 2010 (A), 2013 (B), 2016 (C) and 2019 (D), showing how the 12 RUs performed relative to each other.


subsequent screening examinations, 2010
Positive predictive value of recall (%)
Recall rate (%)
Detection rate:
4 per 1000
5 per 1000
6 per 1000
7 per 1000
8 per 1000
RU1
RU2
RU3
RU4
RU5
RU6
RU7
RU8
RU9
RU10
RU11
RU12

A


subsequent screening examinations, 2013
Positive predictive value of recall (%)
Recall rate (%)
Detection rate:
4 per 1000
5 per 1000
6 per 1000
7 per 1000
8 per 1000
RU1
RU2
RU3
RU4
RU5
RU6
RU7
RU8
RU9
RU10
RU11
RU12

B

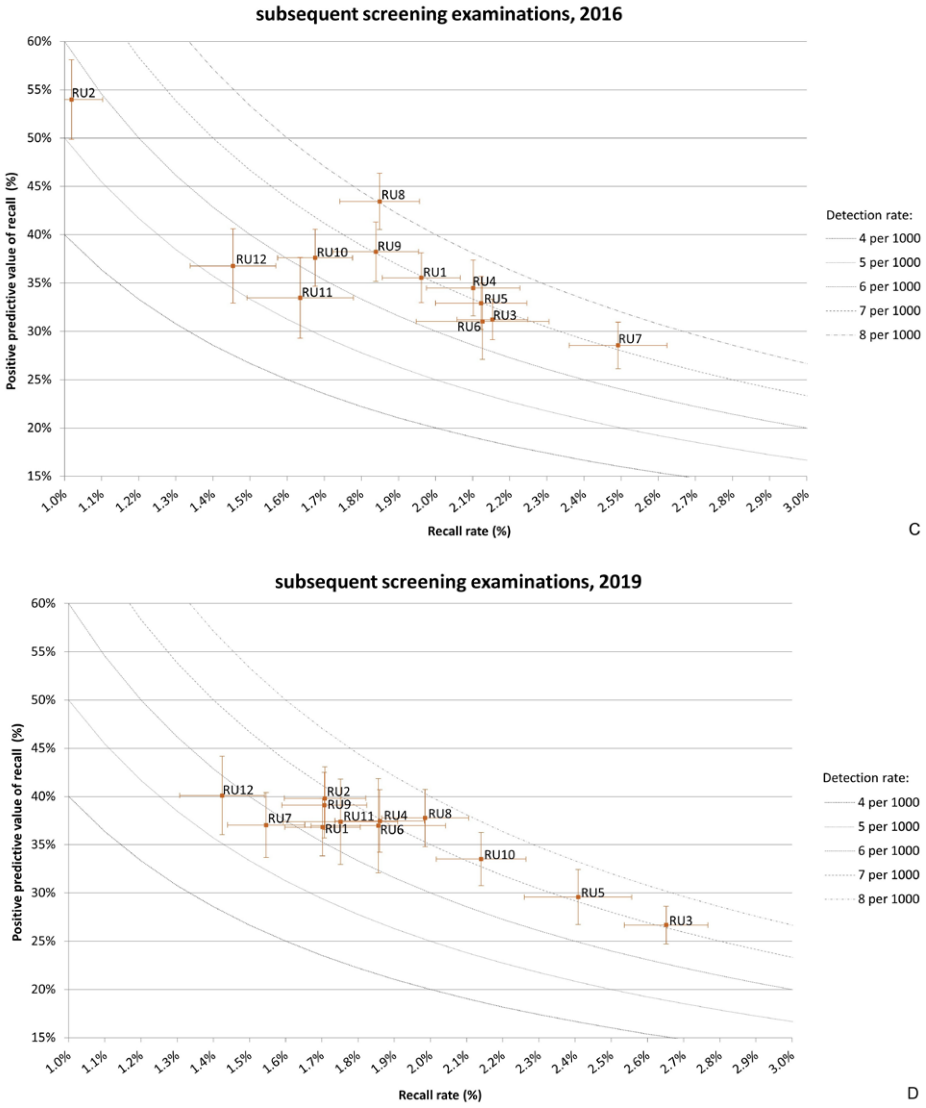
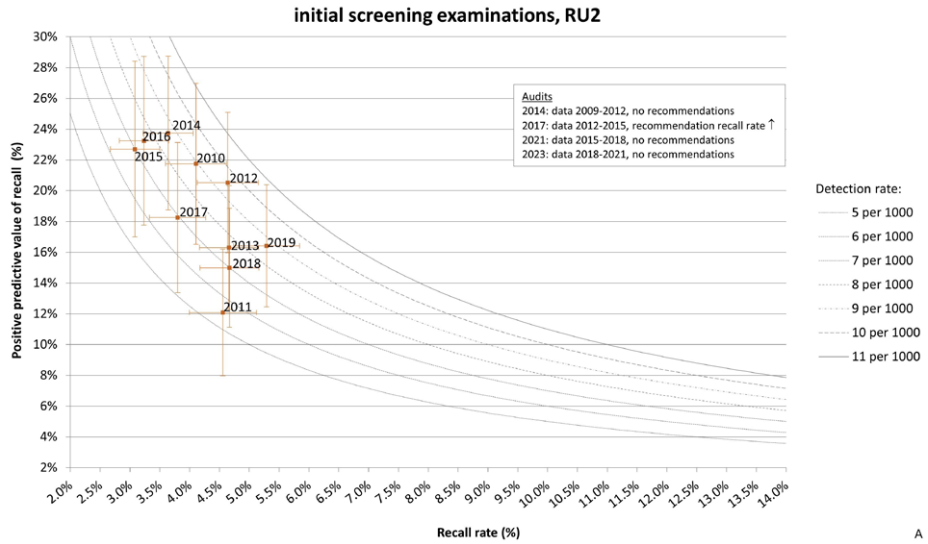


Figure 4: PPV-recall diagrams, with cancer detection rate (CDR) presented as isobars, for subsequent screening examinations for 2010 (A), 2013 (B), 2016 (C) and 2019 (D), showing how the 12 RUs performed relative to each other.



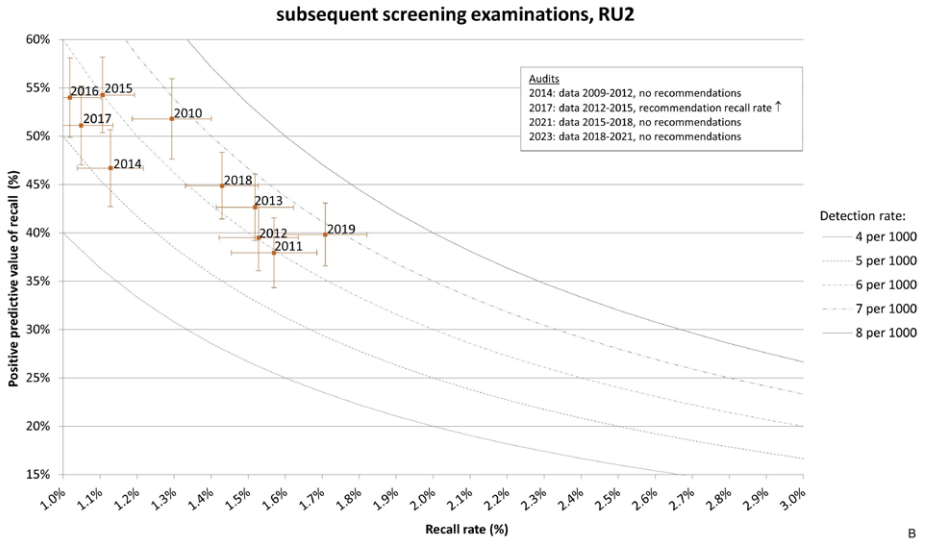


Figure 5: PPV-recall diagrams, with cancer detection rate (CDR) presented as isobars, for RU2 for initial (A) and subsequent screening examination (B), both over the period 2010-2019. The audit year, period of the audit data and audit recommendations are indicated in the frame.

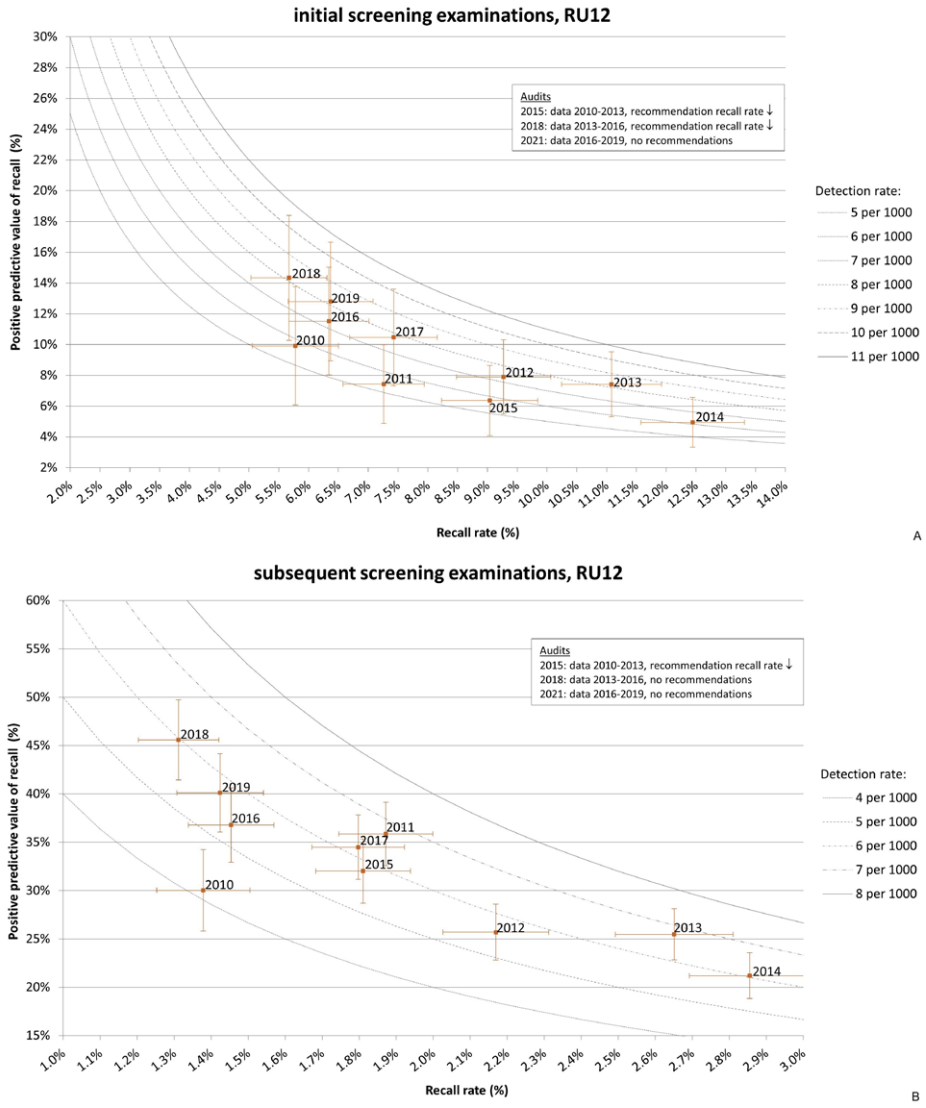


Figure 6: PPV-recall diagrams, with cancer detection rate (CDR) presented as isobars, for RU12 for initial (A) and subsequent screening examination (B), both over the period 2010-2019. The audit year, period of the audit data and audit recommendations are indicated in the frame.

reduce their RR, as their PPV was low and the RR had rapidly increased for both initial and subsequent screening. After 2015, the PPV-recall diagrams showed a considerable reduction in RR, meeting the target value for subsequent screening. However, for initial screening, the RR remained higher than the target value of 5.0%. During the 2018 audit (data 2013-2016), the radiologists were once again recommended to lower their RR, this time only for initial screening. In the period after 2018, the RR did not decrease.

Audit recommendations and performance changes

A total of 42 audits were conducted across the 12 RUs, resulting in 37 recommendations: 24 for initial and 13 for subsequent screening. All recommendations concerned lowering, increasing or stabilising RR. The latter was regarding subsequent screening, combined with a recommendation to lower the RR only for initial screening.

For initial screening (Table 4), the RR changed in the desired direction for nine recommendations out of 24, with five RUs meeting the target value after one year. For six recommendations the RR remained unchanged or changed in the opposite direction, whereas for nine recommendations changes in RR could not be assessed (were outside study period).

For subsequent screening (Table 5), the RR changed in the desired direction for eight recommendations, with six RUs meeting the target value after one year. For one RU, the RR increased while the recommendation was to stabilise, and for four RUs changes in RR could not be assessed.

Discussion

The PPV-recall diagrams demonstrated substantial variations in individual RU performance over a time period of 10 years, with PPVs ranging from 4.9-23.7% for initial and 21.2-54.3% for subsequent screening. Target values were less often met for initial (2010: 0 RUs; 2019: 5 RUs) than for subsequent screening (2010: 8 RUs; 2019: 10 RUs), resulting in more audit recommendations for initial screening (24 versus 13). All recommendations focused on adjusting RR, which often (9 out of 15 for initial and 8 out of 9 for subsequent screening) changed in the recommended direction, though not always sufficiently to meet target values (5 out of 9 for initial and 6 out of 8 for subsequent screening).

In 2001, Blanks et al. from the UK described how they use the PPV-recall diagram as a tool to gain insight into the quality of screening programmes and how this can

generate suggestions to improve screening quality [7]. Similar to our study, they found a substantial variation in reading performance between radiologist groups. However, they did not report whether providing recommendations to these screening groups resulted in improved screening outcomes.

Other papers related to the QA procedures of the UK breast cancer screening programme have also presented PPV-recall diagrams [13-15]. These studies highlighted the correlation between RR, CDR, and PPV, and demonstrated that the PPV-recall diagram serves as a graphical tool to illustrate this relationship. However, similar to Blanks et al. [7], these studies did not report recommendations to support quality improvement. A recent study on the Flemish breast cancer screening programme also uses the PPV-recall diagram for illustrative purposes only [16].

In addition, the study of Miglioretti et al. [6] used the PPV-recall diagram to update performance criteria (target values) for screening radiologists. This study also focused on the interrelationship between RR, CDR, and PPV. Specifically, the authors suggested to define acceptable ranges for RR and PPV conditional on the CDR. In the PPV-recall diagram, these different zones were shaded using distinct colours. Based on our target values, we also defined an area in the PPV-recall diagram that indicated where all target values are met (see Figure 1). Identifying the target area in the diagram visually demonstrates where screening outcomes deviate and specific recommendations should be issued.

It should be noted though that the target values specified for the Dutch breast cancer screening programme (see table 1) differ considerably from those established in e.g., the UK screening programme [7, 13-15] and the US [6]. Target values are inherently dependent on national policy, incidence of breast cancer, screening interval and age range of the screened population. Nonetheless, our approach within the QA programme could be generalised. The performance of a RU should meet the combination of established target values for RR, CDR, and PPV. If performance falls outside these thresholds, the RU receives a recommendation. The extent to which performance deviates from the target values determines the severity of the recommendation.

Our results suggest that the audit recommendations may not always be sufficient to ensure that RUs consistently meet the target values in subsequent years. There are several explanations for this finding. First, the three-year audit frequency may be too infrequent. Several studies suggest that feedback should be based on recent performance to optimise effectiveness [17-21]. Second, feedback at group level may be too indirect. The audit data are provided at RU level. While optimal reading

performance is a shared responsibility, improvement often depends on actions taken by individual radiologists. It is therefore plausible that the effectiveness of the audit process could increase if group performance feedback were combined with feedback on individual performance [20]. Third, the PPV-recall diagram may not be intuitive for radiologists and might require further explanation. Bowles and Geller concluded that radiologists generally prefer graphic displays over tables in audit feedback; however, their findings also revealed that radiologists found a graph similar to the PPV-recall diagram to be too complex [22]. Fourth, the recommendations lack clear guidance on how to achieve the goals. Our study showed that recommendations always focused on adjusting the RR, but did not suggest specific methods for achieving this. Feedback is likely to be more effective when accompanied by a plan with specific actions to reach targets [17, 20]. Radiologists might benefit from additional methods to optimise their recall decision, such as reviewing test sets with specific mammographic abnormalities, in order to identify which types of abnormalities were unnecessarily recalled in false-positive screenings and which false-negative screenings could have been detected in the previous screening round.

An alternative explanation why the provided feedback may not always lead to achieving the target values, is that these target values might be unrealistic and require adjustment. In mid-2016, the target values for the RR were increased from 3.5% to 5.0% for initial screening and from 2.0% to 2.15% for subsequent screening. Still, RUs generally have a too high recall rate for initial screening. All target values were mainly based on the study of Otten et al.[23], which was conducted with analogue screening examinations and in a laboratory setting, using test sets. This highlights the need for a new study based on digital mammography in daily screening practice, to determine optimal target values for the RR. With this goal, the Recall and detection Of breast Cancer in Screening (ROCS) study was started within the Dutch breast cancer screening programme in 2019. Sechopoulos et al. [24] published the study design in 2022 and the analysis for this study is ongoing.

Our study has certain limitations. First, we were unable to investigate whether there was a causal relationship between a recommendation and a change in performance after an audit, as it is not possible to separate the impact of an audit from other factors. Second, due to incomplete data regarding interval cancers during the audits, we could not compare the results of the PPV-recall diagrams with the sensitivity and specificity of the RUs.

In conclusion, our results show that the PPV-recall diagram is a suitable tool for monitoring reading performance of groups of screening radiologists within a

breast cancer screening programme. For the audit team, it provides insight into the interrelationship between RR, CDR, and PPV, and into performance variations, which can help in formulating recommendations for improvement. However, for individual radiologists, audit feedback based on the PPV-recall diagram on group level alone may not always be sufficient to achieve these improvements. To better support radiologists, feedback on overall group performance could be combined with feedback on individual performance, specific action plans for improvement could be provided, and feedback could be presented in a clearer and more intuitive way.

Acknowledgements

We acknowledge the use of ChatGPT in translating some parts of the text from Dutch to English, for checking grammatical errors throughout the text and reducing the number of words. The auteurs thank the Dutch screening organisation (Bevolkingsonderzoek Nederland) for providing the data.

Conflict of Interest

L.D. is a member of the Scientific Editorial Board of European Radiology (section: Breast) and, as such, did not participate in the selection or review processes for this article. The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

References

1. Dibden A, Offman J, Duffy SW, Gabe R (2020) Worldwide Review and Meta-Analysis of Cohort Studies Measuring the Effect of Mammography Screening Programmes on Incidence-Based Breast Cancer Mortality. Cancers 12(4): 976
2. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M (2013) The benefits and harms of breast cancer screening: an independent review. British journal of cancer 108(11): 2205–2240
3. Perry N, Broeders M, DeWolf C, Törnberg S, Holland R, Von Karsa L (2006) European guidelines for quality assurance in breast cancer screening and diagnosis – Fourth edition. Office for Official Publications of the European Communities, Luxembourg
4. European Commission: Joint Research Centre, European Quality Assurance Scheme for Breast Cancer Services, Publications Office of the European Union (2025) Available via https://data.europa.eu/doi/10.2760/7067385. Accessed 14 April 2025
5. U.S. Food and Drug Administration, Mammography Quality Standards (2023) Available via https://www.fda.gov/radiation-emitting-products/mammography-quality-standards-act-and-program. Accessed 26 November 2024
6. Miglioretti DL, Ichikaw L, Smith RA et al (2015) Criteria for identifying radiologists with acceptable screening mammography interpretive performance on basis of multiple performance measures. AJR American journal of roentgenology 204(4): W486–W491
7. Blanks RG, Moss SM, Wallis MG (2001) Monitoring and evaluating the UK National Health Service Breast Screening Programme: evaluating the variation in radiological performance between individual programmes using PPV-referral diagrams. Journal of medical screening 8(1): 24–28
8. Geertse TD, Holland R, Timmers JM et al (2015) Value of audits in breast cancer screening quality assurance programmes. European radiology 25(11): 3338–3347
9. Dutch Expert Centre for Screening, Audit protocol breast cancer screening (2022, in Dutch) Available via https://lrcb.nl/wp-content/uploads/2022/04/Visitatieprotocol-borstkankerscreening-versie-2022.pdf. Accessed 26 November 2024
10. IKNL, National evaluation of breast cancer screening in the Netherlands (2018/2019) Available via https://iknl.nl/getmedia/cc1b7da5-28aa-43e6-b70a-b3dadf08ab89/monitor-bevolkingsonderzoek-borstkanker-2018-2019.pdf. Accessed 26 November 2024
11. Bluekens AM, Holland R, Karssemeijer N, Broeders MJ, den Heeten GJ (2012) Comparison of digital screening mammography and screen-film mammography in the early detection of clinically relevant cancers: a multicenter study. Radiology 265(3): 707–714
12. Dutch Expert Centre for Screening, Advies en voorstel voor streefwaarde m.b.t. hoogte verwijscijfer [Advice and proposal for target value with regard to the level of the recall rate] (2015), Netherlands
13. Perry NM (2003) Interpretive skills in the National Health Service Breast Screening Programme: performance indicators and remedial measures. Seminars in Breast Disease 6(3): 108–113
14. Bennett RL, Blanks RG (2007) Should a standard be defined for the Positive Predictive Value (PPV) of recall in the UK NHS Breast Screening Programme? Breast (Edinburgh, Scotland) 16(1): 55–59
15. Cohen SL, Blanks RG, Jenkins J, Kearins O (2018) Role of performance metrics in breast screening imaging - where are we and where should we be? Clinical radiology 73(4): 381–388
16. Goossens M, De Brabander I, De Grève J et al (2019) Flemish breast cancer screening programme: 15 years of key performance indicators (2002-2016). BMC Cancer. 19(1):1012

17. Ivers N, Jamtvedt G, Flottorp S et al (2012) Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database of Systematic Reviews 2012, Issue 6. Art. No.: CD000259
18. Ivers N, Yogasingam S, Lacroix M et al (2025) Audit and feedback: effects on professional practice. Cochrane Database of Systematic Reviews 2025, Issue 3. Art. No.: CD000259
19. Brehaut JC, Colquhoun HL, Eva KW et al (2016) Practice Feedback Interventions: 15 Suggestions for Optimizing Effectiveness. Annals of internal medicine 164(6): 435–441
20. Brown B, Gude WT, Blakeman T et al (2019) Clinical Performance Feedback Intervention Theory (CP-FIT): A new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. Implementation Science 14(1): 1–25
21. Hofvind S, Bennett RL, Brisson J et al (2016) Audit feedback on reading performance of screening mammograms: An international comparison. Journal of Medical Screening 23(3): 150–159
22. Bowles EJA, Geller BM (2009) Best Ways to Provide Feedback to Radiologists on Mammography Performance. AJR Am J Roentgenol 193(1): 157–164
23. Otten JD, Karssemeijer, N, Hendriks JH et al (2005) Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. Journal of the National Cancer Institute 97(10): 748–754
24. Sechopoulos I, Abbey CK, van der Waal D et al (2022) Evaluation of reader performance during interpretation of breast cancer screening: the Recall and detection Of breast Cancer in Screening (ROCS) trial study design. European radiology 32(11): 7463–7469

Supplementary materials

The supplementary materials are online available at:
https://static-content.springer.com/esm/art%3A10.1007%2Fs00330-025-11978-3/MediaObjects/330_2025_11978_MOESM1_ESM.pdf

Radboud

Chapter 7

General Discussion

The tasks of breast cancer screening radiologists are complex and challenging. Their performance plays a critical role in the effectiveness of a breast cancer screening programme. In this thesis, efforts have been made to generate deeper insights into the value of training and audits of screening radiologists. It also explores new ways to better support screening radiologists in improving their performance. This final chapter presents a reflection on the main findings of the conducted studies, outlines ideas for future research, and provides suggestions for potential adjustments to the reading procedures used by screening radiologists.

Training

Due to the low breast cancer prevalence and the large number of examinations, reading screening mammograms requires a completely different mindset compared to reading clinical mammograms. It is not a given that an excellent clinical radiologist is also an excellent screening radiologist [1]. Therefore, the need for specific training for screening radiologists is widely acknowledged and also recommended by the European Commission Initiative on Breast Cancer (ECIBC) [2, 3].

In the Netherlands, new screening radiologists are required to complete an initial training. This training involves reading test sets followed by feedback discussions with expert screening radiologists, to develop a screening-oriented mindset. Our literature review in **chapter 2** showed that training is effective in improving the performance of radiologists in breast cancer screening. However, the optimal training method remains unclear, as high-quality studies directly comparing different approaches are lacking.

Reading test sets seems to be a successful method for learning how to read screening mammograms or for improving the performance of screening radiologists, and it is used in several countries worldwide [4–9]. Surveyed radiologists indicated that they consider this approach beneficial [10]; 65% considered the feedback obtained from reading test sets, combined with feedback on their performance in real-world screening practice, to be the most important factor in improving their skills. In the literature, test sets mostly involve self-study. However, radiologists may find it more valuable to receive face-to-face feedback on their performance from an expert screening radiologist, allowing for direct interaction and discussion with peers. Only the study by Lee et al. [5] used such a face-to-face approach; however, the added value of this method compared to self-study has not been studied and remains unknown.

To determine whether a radiologist who performs well on a test set also performs well in screening practice, a strong correlation between both settings is required. The literature suggests that test set outcomes correlate only weakly to moderately with screening practice [11–15]. Consequently, there is insufficient evidence to support test set performance as a reliable predictor of performance in real-world screening practice. However, for new screening radiologists, test set performance is the only available indicator of competence. Chen et al. [14] showed that outlier readers in test set performance (defined as outcomes more than 1.5 times the interquartile range below the first quartile) had significantly lower performance in real-world screening practice in terms of cancer detection rate and positive predictive value of recall (PPV) compared with the non-outlier readers. This suggests that test set outcomes could potentially help prevent poor performance in real-world screening, for example by providing additional training to the outlier readers.

Reading screening mammograms

When reading screening mammograms, radiologists aim to detect abnormalities. Comparing with previous mammograms is considered helpful in distinguishing benign from suspicious abnormalities that may indicate breast cancer [16, 17]. **Chapter 3**, demonstrated that priors are particularly important in the case of common well-circumscribed masses. Although these are almost always cysts [18], they often lack typical benign characteristics on a mammogram, which may lead to uncertainty about whether to recall. To better distinguish between a cyst and a potential cancer, it is important to determine whether the well-circumscribed mass is new or has grown. This can only be done with a prior mammogram.

If no prior mammogram is available, there are generally two main options: recall all cases for further assessment or wait until the next screening mammogram. The observations in **chapter 3** – based on a limited exploration – suggest that the waiting approach could be feasible, as the recalled cases involving a well-circumscribed mass were rarely associated with malignancy. Moreover, the few breast cancers that were detected were generally not aggressive in nature. This waiting approach could help prevent many false-positive recalls. However, further research is needed to confirm its feasibility.

This waiting approach is not widely supported in the literature. Most studies describe the approach of recall followed by short-interval imaging surveillance, with follow-up examinations at 6, 12, 24 months, and optional at 36-months [19-23].

These studies emphasize that women's compliance decreases with each subsequent follow-up examination [20-23]. In addition, Berg et al. [22] showed that the majority of the few diagnosed cancers during follow-up were diagnosed no later than 6 months. They therefore suggested only a short follow-up period instead of 2 years.

The determination of the best approach to managing well-circumscribed masses depends on what is considered the optimal trade-off between detecting cancers and avoiding false-positives. This is influenced by multiple factors, mainly policy choices, which are also culturally determined. In the Netherlands, there is a strong emphasis on avoiding false-positives. As a result, the waiting approach is more likely to be regarded as the best approach compared to screening programmes where the balance shifts more towards maximising cancer detection. Furthermore, it is important to consider the extent to which radiologists experience fear of being sued for malpractice, if a malignancy is detected during the screening interval. Moreover, the preferences of the participants must be taken into account. For example, in the Netherlands, additional diagnostic work-up after recall is not fully covered by the national health insurance system. Instead, the costs fall under the compulsory deductible, meaning that individuals must pay these expenses out of pocket up to a certain annual limit. This financial burden may contribute to a more negative attitude towards false-positives.

A well-circumscribed mass is a type of abnormality on a mammogram that is challenging to interpret in terms of its suspicion for malignancy, but is commonly clearly detectable and therefore less likely to be overlooked. In contrast, other abnormalities such as calcifications are typically very small and subtle in appearance, and therefore more likely to be missed. This is particularly relevant in the context of breast cancer screening, where the low prevalence of breast cancers and the high volumes of examinations make the detection of abnormalities challenging in general. Therefore, double reading – where two radiologists independently read each mammogram – is standard practice in many breast cancer screening programmes [3, 24]. Studies have demonstrated that double reading significantly increases the cancer detection rate [25-27].

Based on the assumption that more pairs of eyes lead to more cancers being detected – as is the case with double reading – we investigated in **chapter 4** whether involving radiographers in prereading of the mammograms could also be helpful in detecting more cancers. However, it turned out that the results did not improve but became less favourable. Presenting the radiographers' warning signals to the radiologists resulted in a higher recall rate and lower PPV, while cancer

detection rate appeared to remain unchanged. This may potentially be related to the timing of presenting the warning signals, which was immediately upon opening the screening examination.

Studies on double reading by Klompenhouwer et al. [28] and on Artificial Intelligence (AI) support by Fogliato et al. [29], have demonstrated that the timing of presenting additional information – such as the decision of the first reader or AI recommendation – to the radiologist is crucial. Klompenhouwer et al. showed that blind double reading (where the second reader has no information about the first reading) is preferable. Fogliato et al. found that when the radiologist was already informed about the AI recommendation during the reading process, this introduced an anchoring bias, in which radiologists are unconsciously influenced by the information (the 'anchor') they receive when making decisions. This bias led to false-positive recalls and its impact was directly related to the accuracy of the AI system. When the AI recommendation was incorrect, it resulted in poorer performance by the radiologist. Notably, this anchoring bias did not occur when the radiologist was only provided with the information after the reading process. In that case, the final result improved as consensus was sought in the event of conflicting decisions. As a follow-up to our study, in which we investigated whether radiographers' warning signals could help detect more cancers, we could examine the effect of presenting the radiographers' warning signals only after the radiologist has completed the reading process.

Audits

An audit is a widely used method to monitor the performance of professionals, as part of a quality assurance (QA) programme. During an audit the performance of screening radiologists is independently evaluated. This means that the assessment is conducted by auditors who are not involved in the daily screening work of the radiologists being audited. The audit assesses compliance with established requirements and target values for screening performance and identifies opportunities for improvement. Screening radiologists must find the operating point that maximises cancer detection (true-positives) while minimising false-positives. The recall rate, cancer detection rate, and PPV are the key quality indicators. Within screening programmes, target values have been established for these quality indicators [8, 30-34]. These target values vary between countries, partly because they depend on national policy issues – such as screening interval and age range of the screened population –, but also because of contextual factors

like the underlying incidence of breast cancer and cultural or legal considerations, for example the fear of being sued for malpractice.

In the Netherlands, radiologists are assessed at group level in triennial audits. **Chapters 5 and 6** demonstrate that, overall, groups of radiologists have improved their performance over the years, although variation between groups remains. Audits may have contributed to this improvement, but their specific impact cannot be isolated, as multiple factors influence radiologists' performance [35]. For instance in **chapter 5**, the transition from analogue to digital mammography, which took place in the Netherlands between 2008 and 2010, has also had an impact on radiologist' performance.

Chapter 6 highlights that, in audits, the "positive predictive value – recall diagram" (PPV-recall diagram) is a useful visual method to provide insight into the interrelationship between the three quality indicators (recall rate, cancer detection rate, and PPV). This insight is essential for determining how radiologists can move towards the optimal target area. Previous studies have shown that radiologists generally prefer graphical representations over tables [36], and that they favour comparisons against target values and the performance of their peers [36, 37]. In this regard, the use of the PPV–recall diagram aligns well with their informational needs. However, it is also important to acknowledge the fact that radiologists often struggle to interpret performance data correctly and tend to overestimate their performance [37, 38]. Therefore, a PPV-recall diagram should be made as intuitive as possible. The diagram could be enhanced with graphical cues – such as a shaded target area and directional arrows indicating desired improvement – to address potential misinterpretations.

Chapter 6 further reveals that, despite the fact that over the years groups of radiologists have improved their performance, most groups of radiologists operate outside the target area for initial screening, and some groups also for subsequent screening. As improvements are ultimately driven by changes made by individual radiologists, it is crucial for each radiologist to gain insight into their own performance in relation to both the target values and the group average [39]. This raises the question whether radiologists could be better supported in achieving targets by incorporating individual performance feedback alongside group-level evaluations during audits. This approach has already been successfully implemented in the UK breast cancer screening programme [8].

It should be taken into account that target values are generally best regarded as reference values rather than strict requirements, as variations in performance are to

be expected. Especially in individual performance, these fluctuations play a major role due to the smaller numbers. This complicates the assessment of individual performance and can make it difficult to determine whether a radiologist truly deviates from a target value. In addition, contextual factors may limit the extent to which target values can be met. For example, in the Netherlands, longer screening intervals due to a shortage of radiographers have affected performance indicators in the last years. This type of change, but also the potential implementation of tomosynthesis in screening or the future introduction of AI, may provide grounds for reconsidering the established target values.

We observe that professionals experience audits as stressful, partly due to the emphasis on performance assessment. Audits do serve to check whether performance aligns with established target values, and radiologists may be required to improve if their performance is lacking. This focus on compliance can create pressure and shift attention away from learning and development. Nevertheless, the primary goal of the audits should be to provide constructive feedback that facilitates further performance improvement. In this sense, audits should be framed and experienced as opportunities for reflection and professional growth, rather than solely as compliance checks. This highlights the importance of presenting and communicating audits in a way that sufficiently emphasises their role as learning opportunities. **Chapter 5** shows that peer reviews, as part of an audit, provide valuable learning opportunities for screening radiologists. Interval cancers are particularly relevant cases for peer review. The European Guidelines [40] recommend incorporating peer review into the audit process, by categorising the screening mammogram prior to the interval cancer into three groups: category 1 (normal), category 2 (minimal signs, uncertain), and category 3 (suspicious findings/missed diagnosis). By reviewing these cases with peers, radiologists can discuss whether the screening mammograms classified as category 3 – where the abnormality was either not detected or misinterpreted as benign – reveal underlying patterns, and how such pitfalls might be better recognised.

Literature indicates that peer review is not standard practice. A survey of international breast cancer screening programmes found that 10 out of 17 respondents conducted peer reviews of interval cancers as part of their audits [41]. This survey also showed that the classification into categories 1, 2 and 3 is generally not used as a formal quality indicator in these audits. A possible explanation for this limited use is the concern about the reliability of the classification. Fitzpatrick et al. [41] emphasise that the reproducibility of this classification is limited due to hindsight bias. This bias refers to the fact that, once it is known that a woman has developed an interval

cancer, it becomes almost impossible to review prior mammograms objectively, as this knowledge (often unconsciously) influences the reader's interpretation and the classification. For training purposes, this limitation is not problematic. However, QA-tools must be capable of measuring performance objectively. This supports the statement that peer review should primarily be regarded as a training instrument, rather than as an objective QA-tool.

Introduction of Artificial Intelligence (AI) in breast cancer screening

In the near future, we expect AI systems to be introduced in breast cancer screening. Currently, several prospective studies are underway to investigate the implementation of AI systems [42-45]. The outcomes of these studies will provide greater clarity on the optimal strategy of using AI in breast cancer screening. The decisions that will be made will undoubtedly depend on a range of factors, such as the availability of radiologists, regulatory frameworks, and the attitudes of both participants and radiologists within a given screening programme or country regarding the use of AI. Nevertheless, it is expected that strategies combining radiologists with AI – whether as a triage tool, as decision-support for radiologists, or by replacing one radiologist with AI – will offer the most effective approach to integrating AI into breast cancer screening [46]. This highlights the expectation that radiologists and their performance will continue to play a crucial role in breast cancer screening following AI implementation. Consequently, the findings presented in this thesis retain their relevance. Ultimately, the combination of highly skilled radiologists and increasingly sophisticated AI algorithms will drive further advancements in breast cancer screening outcomes.

Future directions

Drawing upon the findings presented in this thesis, several opportunities for future research and practical adjustments in the reading process of screening radiologists can be identified:

- High-quality studies are needed to determine the most effective training method for screening radiologists. This should include the added value of face-to-face feedback compared to self-study.

- For the accreditation of new screening radiologists, further research is necessary to investigate whether the correlation between performance in test sets and real-world screening practice is strong enough to be predictive of their future screening performance. So far, studies have generally demonstrated a weak correlation. Nevertheless, test sets appear to be useful in identifying negative outlier readers who also demonstrate poor performance in real-world screening practice. Future research could specifically focus on this subgroup.
- A study with a large sample size is necessary to provide sufficient evidence on whether it is justified to opt for recalling well-circumscribed masses at subsequent screening – if they have grown – rather than immediately at initial screening. If the mammograms from these cases can also be made available for research purposes, it would be possible to investigate whether AI can improve the differentiation between benign and malignant well-circumscribed masses.
- More research should be conducted on the optimal timing for presenting additional information, such as warning signals from radiographers or markings from an AI system, during the reading process. Our suggestion is that these should be presented only after the radiologist has made an initial decision.
- Regarding audits, further research is desirable on the benefits of feedback on individual performance, in addition to group-level feedback. This could help to increase personal responsibility and engagement in quality improvement among screening radiologists.
- If there are changes in contextual factors which may influence the performance of screening radiologists – such as extended screening intervals or the introduction of new technologies like tomosynthesis or AI – target values should be evaluated and adjusted to maintain their relevance for performance assessment.

References

1. Elmore JG, Cook AJ, Bogart A, et al. Radiologists' interpretive skills in screening vs. diagnostic mammography: are they related? *Clin Imaging*. 2016;40(6):1096–1103.
2. Hofvind S, Bennett RL, Brisson J, et al. Audit feedback on reading performance of screening mammograms: an international comparison. *J Med Screen*. 2016;23(3):150–159.
3. European Commission, Joint Research Centre. European Commission Initiative on Breast Cancer – European Quality Assurance Scheme for Breast Cancer Services. Luxembourg: Publications Office of the European Union; 2025. Available from: https://publications.jrc.ec.europa.eu/repository/handle/JRC140425. Accessed 11 Sep 2025
4. Geller BM, Bogart A, Carney PA, et al. Educational interventions to improve screening mammography interpretation: a randomized controlled trial. *AJR Am J Roentgenol*. 2014;202(6):W586–W596.
5. Lee EH, Jun JK, Jung SE, Kim YM, Choi N. The efficacy of mammography boot camp to improve the performance of radiologists. *Korean J Radiol*. 2014;15(5):578–585.
6. Timmers JM, Verbeek AL, Pijnappel RM, Broeders MJ, den Heeten GJ. Experiences with a self-test for Dutch breast screening radiologists: lessons learnt. *Eur Radiol*. 2014;24(2):294–304.
7. Suleiman WI, Rawashdeh MA, Lewis SJ, et al. Impact of Breast Reader Assessment Strategy on mammographic radiologists' test reading performance. *J Med Imaging Radiat Oncol*. 2016;60(3):352–358.
8. Cohen SL, Blanks RG, Jenkins J, Kearins O. Role of performance metrics in breast screening imaging – where are we and where should we be? *Clin Radiol*. 2018;73(4):381–388.
9. Trieu PDY, Lewis SJ, Li T, et al. Improving radiologist's ability in identifying particular abnormal lesions on mammograms through training test set with immediate feedback. *Sci Rep*. 2021;11(1):9899.
10. Michalopoulou E, Clauser P, Gilbert FJ, et al. A survey by the European Society of Breast Imaging on radiologists' preferences regarding quality assurance measures of image interpretation in screening and diagnostic mammography. *Eur Radiol*. 2023;33(11):8103–8111.
11. Rutter CM, Taplin S. Assessing mammographers' accuracy: a comparison of clinical and test performance. *J Clin Epidemiol*. 2000;53(5):443–450.
12. Soh BP, Lee WB, Mello-Thoms C, et al. Certain performance values arising from mammographic test set readings correlate well with clinical audit. *J Med Imaging Radiat Oncol*. 2015;59(4):403–410.
13. Miglioretti DL, Ichikawa L, Smith RA, et al. Correlation between screening mammography interpretive performance on a test set and performance in clinical practice. *Acad Radiol*. 2017;24(10):1256–1264.
14. Chen Y, James JJ, Cornford EJ, Jenkins J. The relationship between mammography readers' real-life performance and performance in a test set-based assessment scheme in a national breast screening program. *Radiol Imaging Cancer*. 2020;2(5):e200016.
15. Qenam BA, Li T, Alshabibi A, Frazer H, Ekpo E, Brennan P. Test-set results can predict participants' development in breast-screen cancer detection: an observational cohort study. *Health Sci Rep*. 2024;7(6):e2161.
16. Callaway MP, Boggis CR, Astley SA, Hutt I. The influence of previous films on screening mammographic interpretation and detection of breast carcinoma. *Clin Radiol*. 1997;52(7):527–529.
17. Akwo JD, Trieu P, Lewis S. Does the availability of prior mammograms improve radiologists' observer performance? A scoping review. *BJR Open*. 2023;5(1):20230038.

18. Berment H, Becette V, Mohallem M, Ferreira F, Chérel P. Masses in mammography: what are the underlying anatomopathological lesions? *Diagn Interv Imaging*. 2014;95(2):124–133.
19. Sickles EA. Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. *Radiology*. 1991;179(2):463–468.
20. Helvie MA, Pennes DR, Rebner M, Adler DD. Mammographic follow-up of low-suspicion lesions: compliance rate and diagnostic yield. *Radiology*. 1991;178(1):155–158.
21. Chung CS, Giess CS, Gombos EC, et al. Patient compliance and diagnostic yield of 18-month unilateral follow-up in surveillance of probably benign mammographic lesions. *AJR Am J Roentgenol*. 2014;202(4):922–927.
22. Berg WA, Berg JM, Sickles EA, et al. Cancer yield and patterns of follow-up for BI-RADS category 3 after screening mammography recall in the National Mammography Database. *Radiology*. 2020;296(1):32–41.
23. Common J, Abdullah P, Alabousi A. A single-center audit of BI-RADS 3 assessment category utilization in mammography and breast ultrasound. *Can Assoc Radiol J*. 2023;74(1):69–77.
24. Taylor-Phillips S, Stinton C. Double reading in breast cancer screening: considerations for policy-making. *Br J Radiol*. 2020;93(1106):20190610.
25. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. *Radiology*. 1994;191(1):241–244.
26. Brown J, Bryan S, Warren R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. *BMJ*. 1996;312(7034):809–812.
27. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. *AJR Am J Roentgenol*. 2003;180(5):1461–1467.
28. Klompenhouwer EG, Voogd AC, den Heeten GJ, et al. Blinded double reading yields a higher programme sensitivity than non-blinded double reading at digital screening mammography: a prospected population-based study in the south of The Netherlands. *Eur J Cancer*. 2015;51(3):391–399.
29. Fogliato R, Chappidi S, Lungren M, et al. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. *ACM Int Conf Proc Ser*. 2022:1362–1374.
30. Armaroli P, Riggi E, Basu P, et al. Performance indicators in breast cancer screening in the European Union: a comparison across countries of screen positivity and detection rates. *Int J Cancer*. 2020;147(7):1855–1863.
31. Taylor K, Parashar D, Bouverat G, et al. Mammographic image quality in relation to positioning of the breast: a multicentre international evaluation of the assessment systems currently used, to provide an evidence base for establishing a standardised method of assessment. *Radiography*. 2017;23(4):343–349.
32. Lee CS, Parise C, Burleson J, Seidenwurm D. Assessing the recall rate for screening mammography: comparing the Medicare Hospital Compare dataset with the National Mammography Database. *AJR Am J Roentgenol*. 2018;211(1):127–132.
33. Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. *Radiology*. 2010;255(2):354–361.
34. National Institute for Public Health and the Environment. Indicatoren voor Bevolkingsonderzoek naar Borstkanker, Versie 2.2 [Indicators for Breast Cancer Screening, Version 2.2]. Netherlands; 2017.

35. Clerkin N, Ski CF, Brennan PC, Strudwick R. Identification of factors associated with diagnostic performance variation in reporting of mammograms: a review. *Radiography*. 2023;29(2):340–346.

36. Bowles EJA, Geller BM. Best ways to provide feedback to radiologists on mammography performance. *AJR Am J Roentgenol*. 2009;193(1):157–164.

37. Funaro K, Ataya D, Niell B. Understanding the mammography audit. *Radiol Clin North Am*. 2021;59(1):41–55.

38. Cook AJ, Elmore JG, Zhu W, et al. Mammographic interpretation: radiologists' ability to accurately estimate their performance and compare it with that of their peers. *AJR Am J Roentgenol*. 2012;199(3):695–702.

39. Brown B, Gude WT, Blakeman T, et al. Clinical Performance Feedback Intervention Theory (CP-FIT): a new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. *Implement Sci*. 2019;14(1):1–25.

40. Perry N, Broeders M, de Wolf C, et al. European guidelines for quality assurance in breast cancer screening and diagnosis. 4th ed. Luxembourg: European Commission; Office for Official Publications of the European Communities; 2006.

41. Fitzpatrick P, Byrne H, Flanagan F, et al. Interval cancer audit and disclosure in breast screening programmes: an international survey. *J Med Screen*. 2023;30(1):36–41.

42. Lång K, Josefsson V, Larsson AM, 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(8):936–944.

43. 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(10):e703–e711.

44. Ng AY, Oberije CJG, Ambrózay É, et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nat Med*. 2023;29(12):3044–3049.

45. Hernström V, Josefsson V, Sartor H, 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(3):e175–e183.

46. Raya-Povedano JL. AI in breast cancer screening: a critical overview of what we know. *Eur Radiol*. 2024;34(7):4774–4775.

Radboud

Chapter 8

Summary
Samenvatting

Summary

Breast cancer is the leading cause of cancer-related mortality among women worldwide. Screening enables early detection, which increases the likelihood of successful treatment and helps reduce breast cancer mortality. Mammography is still the most effective screening method. However, breast cancer screening also presents disadvantages and challenges. These include false-positive recalls, which can cause anxiety and unnecessary follow-up procedures, and overdiagnosis, which may result in overtreatment. Both contribute to increased healthcare costs. The screening performance of radiologists – who determine whether a woman should be recalled for further diagnostic assessment – plays an important role in balancing the benefits and harms of screening.

In this thesis, we aimed to identify factors that help screening radiologists in improving their reading performance. To this end, we examined various quality assurance tools and screening strategies, including radiologist training, recall strategies, warning signals from radiographers to radiologists, and audits and feedback.

In chapter 2, we reviewed the existing literature to evaluate whether training improves radiologists' performance in breast cancer screening. A systematic search identified 18 relevant studies, of which 17 showed performance improvement in at least one training activity. Two measurement approaches were found: performance on test sets and performance in actual screening practice. Sensitivity and specificity were the most frequently reported outcomes in test set studies, whereas recall rate was most commonly used to assess performance in actual screening practice. Despite methodological heterogeneity and generally moderate study quality – often due to small sample sizes and lack of control groups – the conclusion of the study was that training helps radiologists to improve their screening performance.

Well-circumscribed masses are a very common type of abnormality seen on mammograms in a screening population. In the Dutch breast cancer screening programme, these masses should be recalled for further assessment, if they are solitary, new or growing. This results in a substantial number of false-positive recalls, especially at initial screening. **In chapter 3**, we aimed to find a potential solution to reduce the number of false-positive recalls. We examined characteristics of well-circumscribed masses at mammography by combining findings from a narrative literature review with an exploration of screening practice. The 15 articles identified in the literature search showed that well-circumscribed masses have a low positive predictive value (PPV) of 0–2%, while solitary, new, or enlarging well-

circumscribed masses have a higher PPV of 10–12%. In our exploration of screening practice, we found that 25% of all recalls were based on a well-circumscribed mass, with a PPV of 2.0% at initial and 10.6% at subsequent screenings. Most screen-detected malignant well-circumscribed masses had a favourable prognosis. Based on these findings, we identified a potential improvement in the recall strategy. The absence of prior mammograms at initial screening limits radiologists' ability to identify new or growing well-circumscribed masses, contributing to a low PPV. Access to prior clinical imaging could improve this PPV. If no prior mammograms are available, screening radiologists may consider recalling these lesions at subsequent screening – if they have grown – rather than at initial screening. However, further research is needed to provide sufficient evidence on whether this adjusted recall strategy is justified in practice.

In the Dutch breast cancer screening programme, radiographers (in chapter 4 referred to as technologists) pre-read the mammograms to identify possible abnormalities, leading to warning signals for radiologists. **In chapter 4**, we examined the impact of these warning signals on early screening outcomes. We conducted a prospective study in which radiologists were alternately blinded or nonblinded to these warning signals on a monthly basis. The blinded group demonstrated a significantly lower overall recall rate (2.1% vs. 2.4%, P=0.001) and a higher positive predictive value (30.6% vs. 26.2%, P=0.02), while there was no evidence of cancer detection rate differences between the groups (6.5 vs. 6.4 per 1000 screens, P=0.75). Radiographers most frequently flagged masses, while these warning signals had the lowest cancer probability (PPV 0% and 17% respectively for masses with or without calcifications. Warning signals concerning calcifications had the highest cancer probability (PPV 61%). The findings of this study suggest that directly presenting warning signals to radiologist is not effective. It may be more beneficial to present the signal after the radiologist's initial recall decision. Future research is necessary to confirm this.

For radiologists, an accurate understanding of their performance is essential for identifying areas of improvement. Since the start of the Dutch breast cancer screening programme, a comprehensive quality assurance system has been an integral part of its structure. This includes triennial audits in which screening outcomes are evaluated and cancer cases are peer reviewed. The radiologists are audited at group level (RUs or reading units). **In chapter 5**, we assessed the value of these audits as a quality assurance tool. We retrospectively evaluated audit data from four audit series (1996-2000, 2001-2005, 2003-2007 and 2010-2013). The audits showed a consistent increase in detection rate (from 3.3 to 5.4 per 1000) and sensitivity (from 64.5% to

71.6%). For radiological reviews of interval cancers and advanced screen-detected cancers, the percentage of missed cancers remained stable (20–30%). In the review of consecutive recalled cases, on average 17% of recalled cases led to discussion about the appropriateness of the recall. These discussions offer radiologists a valuable opportunity for self-reflection with peers. Therefore, radiological review and immediate feedback should be a standard component of audit procedures, complementing epidemiological surveillance.

To effectively monitor radiologists' screening performance at group level, recall rate, cancer detection rate, and positive predictive value of recall are key parameters in an audit. They are best monitored combined, as they are interrelated. For this purpose the PPV-recall diagram can be used. **In chapter 6**, we retrospectively evaluated the suitability of this PPV-recall diagrams as a tool for monitoring and providing recommendations to RUs. Audit data from 12 RUs between 2010–2019 was used. PPV-recall diagrams showed substantial variations in the individual RU performance over time, with PPVs ranging between 4.9-23.7% for initial and 21.2-54.3% for subsequent screening. Target values were less often met for initial (2010: 0 RUs; 2019: 5 RUs) than for subsequent screening (2010: 8 RUs; 2019: 10 RUs), resulting in more recommendations regarding initial screening (24 versus 13). All recommendations focused on adjusting recall rate, which often (17 out of 24) changed in the recommended direction, though not always sufficient to meet target values. The study highlights that PPV-recall diagrams are valuable for audit teams to identify performance issues and formulate recommendations. However, feedback at group level may be insufficient for individual radiologists to achieve sufficient improvement. To enhance effectiveness, the feedback on group-level could be combined with feedback on individual performance data.

This thesis highlights the complexity of interpreting mammograms in breast cancer screening and underscores the critical role of radiologists in ensuring the programme's effectiveness. Based on the findings, several practical recommendations have been proposed to improve the radiologists' reading process and to guide future research. In the near future, AI systems are expected to be integrated into breast cancer screening, in a way that complements rather than replaces radiologists. As a result, radiologists' performance will remain central to the success of the screening programme, and the insights from this thesis will continue to be highly relevant.

Samenvatting

Borstkanker is wereldwijd de belangrijkste oorzaak van sterfte aan kanker onder vrouwen. Screening maakt vroege opsporing mogelijk. Dit vergroot de kans op een succesvolle behandeling en helpt het aantal sterfgevallen aan borstkanker te verminderen. Mammografie is nog altijd de meest effectieve methode. Borstkankerscreening kent echter ook nadelen en uitdagingen. Zoals fout-positieve verwijzingen, die kunnen leiden tot angst en onnodig vervolgonderzoek, en overdiagnose, wat kan resulteren in overbehandeling. Beide dragen bij aan hogere zorgkosten. De screeningsprestaties van radiologen – die bepalen of een vrouw moet worden doorverwezen voor verdere diagnostiek – spelen een belangrijke rol in het optimaliseren van de balans tussen deze voor- en nadelen.

In dit proefschrift hebben we onderzocht welke factoren radiologen kunnen helpen om hun beoordelingsprestaties te verbeteren. Daartoe hebben we verschillende kwaliteitsborgingsinstrumenten en strategieën voor screening onderzocht, waaronder training van radiologen, verwijsstrategieën, signalering door laboranten aan radiologen, en audits en feedback.

In hoofdstuk 2 hebben we de bestaande literatuur bestudeerd om te evalueren of training de prestaties van radiologen bij borstkankerscreening verbetert. Een systematische zoekopdracht identificeerde 18 relevante studies, waarvan er 17 een verbetering lieten zien door ten minste één trainingsactiviteit. Er werden twee meetmethoden gevonden: resultaten op testsets en resultaten in de daadwerkelijke screeningspraktijk. Sensitiviteit en specificiteit waren de meest gerapporteerde uitkomstmaten bij testsets als meetmethode, terwijl het verwijscijfer de meest gerapporteerde uitkomstmaat was in studies die de resultaten in de daadwerkelijke screeningspraktijk als meetmethode gebruikten. Ondanks methodologische heterogeniteit en de over het algemeen matige studiekwaliteit – vaak als gevolg van kleine steekproefgroottes en een gebrek aan controlegroepen – was de conclusie van onze literatuurstudie dat training de screeningsprestaties van radiologen verbetert.

Scherp begrensde massa's zijn een veelvoorkomend type afwijking op mammogrammen in een screeningspopulatie. In de Nederlandse borstkankerscreening moeten radiologen deze massa's verwijzen voor verdere diagnostiek als ze solitair, nieuw of groeiend zijn. Dit leidt echter tot een aanzienlijk aantal fout-positieve verwijzingen, vooral bij de eerste screeningsonderzoeken. **In hoofdstuk 3** hebben we gezocht naar een mogelijke oplossing om het aantal fout-positieve verwijzingen te kunnen verminderen. We onderzochten de kenmerken van scherp begrensde

massa's op mammogrammen, waarin we de resultaten van een literatuuronderzoek met een verkenning van de screeningspraktijk combineerden. De 15 artikelen die uit de literatuurzoekopdracht naar voren kwamen, toonden aan dat scherp begrensde massa's een lage positief voorspellende waarde (PVW) hebben van 0–2%. De subgroep van solitaire, nieuwe of groeiende scherp begrensde massa's hebben een hogere PVW van 10–12%. In onze verkenning van de screeningspraktijk bleek dat 25% van alle verwijzingen werd gedaan op basis van een scherp begrensde massa, met een PVW van 2,0% bij de eerste screeningsonderzoeken en een PVW van 10,6% bij vervolgscreeningsonderzoeken. De meeste bij screening ontdekte maligne scherp begrensde massa's waren niet-agressief van karakter. Op basis van deze gegevens konden we een mogelijke verbetering in de verwijsstrategie voor deze scherp begrensde massa's identificeren. Het ontbreken van een voorgaand mammogram bij het eerste screeningsonderzoek maakt het voor de radioloog onmogelijk om te bepalen of de scherp begrensde massa nieuw of gegroeid is, wat bijdraagt aan een lage PVW. Het beschikbaar maken van eerdere klinische mammogrammen zou deze PVW kunnen verbeteren. Bij het ontbreken van een voorgaand mammogram zou de screeningsradioloog ervoor kunnen kiezen om pas bij een volgend screeningsonderzoek te verwijzen indien gegroeid, in plaats van bij het eerste screeningsonderzoek. Meer onderzoek is echter noodzakelijk om voldoende bewijs te leveren dat deze aangepaste verwijsstrategie veilig in de praktijk geïmplementeerd kan worden.

In de Nederlandse borstkankerscreening hebben de laboranten de mogelijkheid om een mammogram ook te beoordelen op eventuele afwijkingen. Als zij iets verdachts zien, geven ze een waarschuwing aan de screeningsradiologen die de mammogrammen beoordelen. **In hoofdstuk 4** onderzochten we de impact van deze waarschuwingssignalen op vroege screeningsuitkomsten. We voerden een prospectieve studie uit, waarbij de radiologen tijdens het beoordelen van de mammogrammen afwisselend per maand afgeschermd of niet afgeschermd werden voor deze signalen. De afgeschermde groep liet een significant lager verwijscijfer zien (2,1% versus 2,4%, P=0,001) en een hogere PVW (30,6% versus 26,2%, P=0,02), terwijl tussen de twee groepen geen verschil werd gevonden in het detectiecijfer (6,5 versus 6,4 per 1000 gescreende vrouwen, P=0,75). De signalen van laboranten gingen meestal over massa's, terwijl die juist de laagste kans hebben om kanker te zijn (PVW: 0% bij massa's met calcificaties en 17% voor massa's zonder). Signalering met betrekking tot calcificaties heeft de hoogste kans om kanker te zijn (PVW 61%). De resultaten van deze studie laten zien dat het niet effectief is om waarschuwingssignalen direct aan radiologen te tonen. Mogelijk dat het tonen van het signaal ná de beoordeling door de radioloog het verwijscijfer wel

verbetert zonder het detectiecijfer te beïnvloeden. Toekomstig onderzoek is nodig om dit te bevestigen.

Voor radiologen is een nauwkeurig inzicht in hun prestaties essentieel om te weten op welke punten verbetering nodig is. Sinds de start van de Nederlandse borstkankerscreening is een uitgebreid kwaliteitsborgingssysteem onderdeel van het programma. Driejaarlijkse audits maken deel hiervan uit, waarbij screeningsuitkomsten worden geëvalueerd en kankercasussen worden herbeoordeeld samen met collega radiologen. Elk team van screeningsradiologen wordt als beoordelingseenheid (BE) geaudit. **In hoofdstuk 5** onderzochten we de waarde van deze audits als instrument voor kwaliteitsborging. We evalueerden retrospectief de auditgegevens van vier auditseries (1996–2000, 2001–2005, 2003–2007 en 2010–2013). De audits toonden een consistente stijging in het detectiecijfer (van 3,3 naar 5,4 per 1000) en sensitiviteit (van 64,5% naar 71,6%). Bij radiologische herbeoordelingen van intervalkankers en 'advanced screen-detected' kankers bleef het percentage gemiste kankers stabiel (20–30%). Bij de herbeoordeling van een set opeenvolgend verwezen cliënten leidde gemiddeld 17% van de verwijzingen tot discussie over de noodzakelijkheid van de verwijzing. Deze discussies bieden een waardevolle gelegenheid tot zelfreflectie met collega's. Daarom zouden radiologische herbeoordelingen en directe feedback een standaard onderdeel moeten zijn in audits, als aanvulling op epidemiologische monitoring.

Om de screeningsprestaties van radiologen tijdens een audit effectief te kunnen monitoren, zijn het verwijscijfer, detectiecijfer en de PVW belangrijke parameters. Deze kunnen het beste gecombineerd worden gemonitord, aangezien ze onderling samenhangen. Hiervoor kan het "PVW-verwijsdiagram" worden gebruikt. **In hoofdstuk 6** evalueerden we retrospectief de geschiktheid van dit PVW-verwijsdiagram als hulpmiddel bij het monitoren van screeningsprestaties en het geven van aanbevelingen aan BE's. Hiervoor werd gebruikgemaakt van auditgegevens van 12 BE's uit de periode 2010–2019. PVW-verwijsdiagrammen toonden aanzienlijke variatie in de prestaties tussen BE's over de tijd, met PVW variërend van 4,9–23,7% bij eerste screeningsonderzoeken en 21,2–54,3% bij vervolgscreeningsonderzoeken. De streefwaarden werden minder vaak gehaald bij eerste screeningsonderzoeken (2010: 0 BE's; 2019: 5 BE's) dan bij vervolgscreeningsonderzoeken (2010: 8 BE's; 2019: 10 BE's), wat leidde tot meer aanbevelingen voor eerste screeningsonderzoeken (24 versus 13). Alle aanbevelingen richtten zich op het aanpassen van het verwijscijfer, wat in veel gevallen (17 van de 24) in de gewenste richting veranderde, maar niet altijd voldoende om de streefwaarden te halen. Deze studie benadrukt dat PVW-verwijs-

diagrammen waardevol zijn voor auditteams om onderprestatie te identificeren en aanbevelingen te formuleren. Feedback op BE-niveau blijkt echter niet altijd voldoende voor individuele radiologen om de beoogde verbetering te bereiken. Om de effectiviteit te vergroten, kan groepsfeedback gecombineerd worden met feedback op individueel prestatieniveau.

Dit proefschrift benadrukt de complexiteit van het beoordelen van mammogrammen binnen de borstkankerscreening en de belangrijke rol van radiologen in het waarborgen van de effectiviteit ervan. Op basis van de bevindingen zijn verschillende praktische aanbevelingen gedaan om het beoordelingsproces van radiologen te verbeteren en richting te geven aan toekomstig onderzoek. In de nabije toekomst wordt verwacht dat AI-systemen zullen worden geïntegreerd in de borstkankerscreening, op een manier die radiologen aanvult in plaats van vervangt. Dit betekent dat de prestaties van radiologen ook na de implementatie van AI een belangrijke rol blijven spelen binnen het screeningsprogramma en dat de inzichten uit dit proefschrift relevant blijven.

8

Radboud

Appendices

Research Data Management

PhD Portfolio

List of publications

Curriculum vitae

Dankwoord

Research Data Management

All studies in this thesis, with the exception of the literature studies described in chapter 2 and 3, involved the collection of data from participants of the Dutch breast cancer screening programme. The client records of participating women are owned and managed by the Dutch screening organisation. These studies were conducted under the national permit for breast cancer screening issued by the Dutch Ministry of Health, Welfare and Sport and fall outside the scope of the Dutch Medical Research Involving Human Subjects act. This means that no formal approval of an Institutional Review Board is required. By participating in the screening programme, women consent to the use of their data available for evaluation and research purposes, unless they explicitly choose to opt out. The privacy of the participants in these studies was warranted by the use of fully anonymised data.

Chapter 4 describes a prospective study conducted within the Dutch breast cancer screening programme in a region covered by the Foundation of Population Screening South. For this study, fully anonymised client records of participating women were shared with the researchers. For chapters 5 and 6, screening data was provided by the Dutch screening organisation in the context of the audits conducted by the Dutch expert centre for screening (LRCB), as part of the quality assurance programme. These aggregated and anonymised data are stored in the archives of the LRCB, on a local LRCB network drive. The data were reused for the purposes of these studies.

All raw study data are stored on a local LRCB network drive, accessible only to researchers involved in the studies. Screening performance data of radiologists were calculated from aggregated data of participating women. The processed data and the results of analyses are also stored on the same local LRCB network drive.

The data from all studies in this thesis will be retained for 15 years after termination of the study. The opt-out procedure used in the Dutch breast cancer screening programme does not allow for consent to be given for public data sharing. In addition, the data are not owned by Radboudumc. Therefore, data cannot be made F&A compliant according to the Radboud RDM policy. Questions about the data can be addressed to t.geertse@lrcb.nl.

The chapters 3, 4 and 6 have been published open access.

PhD portfolio of Tanya Geertse

Department: **Medical Imaging**
PhD period: **01/07/2015 – 01/10/2025**
PhD Supervisor(s): **Prof. dr. M.J.M. Broeders, Prof. dr. R.M. Pijnappel**
PhD Co-supervisor(s): **Dr. D. van der Waal, Dr. L.E.M. Duijm**

Training activities	Hours	ECTS
Courses		
• Literature Review for your PhD: how to search & where to publish (2021)	4.5	0.15
• RIHS - Introduction course for PhD candidates (2021)	15.00	0.5
• E-learning Biostatistics for medical sciences, nursing science, health sciences and epidemiology (2021)	40.00	1.4
• MED-BMS14 Design and analysis of experiments (2021)	84	3.00
• MED-BMS84 Longitudinal and multilevel analysis (2022)	28	1.00
• Radboudumc - Scientific integrity (2022)	20.00	0.7
• Radboudumc - eBROK course (for Radboudumc researchers working with human subjects) (2022)	26.00	0.95
• Radboudumc – eBROK re-registration (October 2025)	3	0.1
Seminars		
• Webinar Research Integrity Round #1, September 2021, 16:00-17:30 (2021)	1.5	0.05
• Webinar Research Integrity Round #2, December 1, 16:00-17:30 (2021)	1.5	0.05
• Meet the expert: Research Data Storage (2022)	1.5	0.05
• Webinar Research Integrity Round, September 2022,, 16:00-17:30 (2022)	1.5	0.05
• Webinar Research Integrity Round 22 March 2023, 16:00-17:30 (2023)	1.5	0.05
• Webinar Research Integrity Round 13 December 2023 (2023)	1.5	0.05
• Webinar Research Integrity Round 13 March 2024 (2024)	1.5	0.05
Conferences		
• ECR (poster presentation) (2023)	32.5	1.15
• ICSN meeting 2025, Aarhus Denmark (oral presentation) (2025)	28	1.00
Other		
• Workshop: Prepare your defence (2022)	1.5	0.05
• LRCB Onderzoekersmiddag 2023 (2023)	3	0.1
• LRCB Onderzoekersmiddag 2024 (2024)	3	0.1
• LRCB Onderzoeksmiddag 2025 (oral presentation) (2025)	4.5	0.15
• Quarterly LRCB Research meeting (2025)	30.00	1.1
Total	**330.5**	**11.7**

List of publications

Papers in international journals

Geertse TD, Tetteroo E, Smid-Geirnaerdt MJA, et al. Applying the "positive predictive value–recall diagram" to monitor performance and provide recommendations for screening radiologists. Eur Radiol. 2025 Sep 4. DOI: 10.1007/s00330-025-11978-3.

Geertse TD, van der Waal D, Vreuls W, et al. The dilemma of recalling well-circumscribed masses in a screening population: A narrative literature review and exploration of Dutch screening practice. Breast. 2023;69:431-440

Sechopoulos I, Abbey CK, van der Waal D, **Geertse TD**, Tetteroo E, Pijnappel RM, Broeders MJM. Evaluation of reader performance during interpretation of breast cancer screening: the Recall and detection Of breast Cancer in Screening (ROCS) trial study design. Eur Radiol. 2022;32(11):7463-7469

Geertse TD, Setz-Pels W, van der Waal D, et al. Added Value of Prereading Screening Mammograms for Breast Cancer by Radiologic Technologists on Early Screening Outcomes. Radiology. 2022;302(2):276-283

Geertse TD, Paap E, van der Waal D, Duijm LEM, Pijnappel RM, Broeders MJM. Utility of Supplemental Training to Improve Radiologist Performance in Breast Cancer Screening: A Literature Review. J Am Coll Radiol. 2019;16(11):1528-1546

Geertse TD, Holland R, Timmers JM, et al. Value of audits in breast cancer screening quality assurance programmes. Eur Radiol. 2015;25(11):3338-3347

Papers in conference proceedings

Craig K. Abbey, Michael A. Webster, **Tanya Geertse**, Danielle van der Waal, Eric Tetteroo, Ruud Pijnappel, Mireille J. M. Broeders, Ioannis Sechopoulos, "Sequential reading effects in Dutch screening mammography," Proc. SPIE 11316, Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, 113160G (16 March 2020)

Kenneth C. Young, Abdulaziz Alsager, Jennifer M. Oduko, Hilde Bosmans, Beatrijs Verbrugge, **Tanya Geertse**, Ruben van Engen, "Evaluation of software for reading images of the CDMAM test object to assess digital mammography systems," Proc. SPIE 6913, Medical Imaging 2008: Physics of Medical Imaging, 69131C (18 March 2008)

van Engen, R.E., Swinkels, M.M.J., Oostveen, L.J., **Geertse, T.D.**, Visser, R. (2006). Using a Homogeneity Test as Weekly Quality Control on Digital Mammography Units. In: Astley, S.M., Brady, M., Rose, C., Zwiggelaar, R. (eds) Digital Mammography. IWDM 2006. Lecture Notes in Computer Science, vol 4046. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11783237_36

Geertse, T.D., van Engen, R. E., Oostveen, L. J., Thijssen, M. A. O., & Karssemeijer, N. (2004). Spectrum optimization for a selenium digital mammography system, Chapel Hill, NC, University of North Carolina at Chapel Hill. In Proceedings of the IWDM (Vol. 18, pp. 116-122).

Abstracts in conference proceedings

T.D. Geertse, E. Tetteroo, M.J.A. Smid-Geirnaerdt, L.E.M. Duijm, R.M. Pijnappel, D. van der Waal, M.J.M. Broeders. "Utilizing the 'positive predictive value – recall diagram' for monitoring key performance indicators of groups of radiologists reading screening mammograms." International Cancer Screening Network (ICSN), 2025

D. van der Waal, C. Abbey, E. Tetteroo, **T. Geertse**, M. Smid-Geirnaerdt, I. Sechopoulos, M. Broeders. "Finding the optimal recall rate in breast cancer screening: First results from the Recall and detection Of breast Cancer in Screening (ROCS) study" International Cancer Screening Network (ICSN), 2025

T.D. Geertse, D. van der Waal, R.M. Pijnappel, E. Tetteroo, L.E.M. Duijm, W. Vreuls, M.J.M. Broeders. "The dilemma of recalling well-circumscribed masses in a screening population: a literature review and exploration of screening practice in a Dutch screening region." European Congress of Radiology (ECR), 2023

T.D. Geertse, R. Holland, C.G.C.M. van Landsveld-Verhoeven, J.M.H. Timmers, K.H. Schuur, R. Pijnappel, F. Jansen, M.J.M. Broeders and G.J. den Heeten. "Audits as part of quality assurance in the Dutch breast cancer screening programme." European Congress of Radiology (ECR), 2013

T. Geertse, K. Young, R. Bouman, R. van Engen. "Specifying Minimum Standards for Image Quality of Mammography Systems Based on Automatic Reading of CDMAM Images". Radiological Society of North America (RSNA), 2006

T.D. Geertse, R.E. van Engen, L.J. Oostveen, R. Visser, and M.A.O. Thijssen. "Image Quality of Commercially Available Digital Mammography Systems Compared by Contrast-Detail Analysis." European Congress of Radiology (ECR), 2004

T.D. Geertse, K.R. Bijkerk, R.E. van Engen, L.J. Oostveen, M.M.J. Swinkels, and R.D. ter Wee. "Revision of the Dutch Protocol for Acceptance Inspections of Breast Cancer Screening Units." European Radiology 12(1, Suppl.):375, 2002

Curriculum vitae

Tanya Geertse werd op 16 april 1972 geboren in Kruiningen (Reimerswaal). In 1990 behaalde zij haar VWO-diploma aan het Buys Ballot College in Goes. In datzelfde jaar is ze begonnen met de studie Technische Natuurkunde aan de Hogeschool Eindhoven, waarvan ze in 1994 het diploma behaalde. Van 1994 tot en met 1998 heeft ze gewerkt bij Organon Teknika in Boxtel (Pharma Division AKZO) als Natuurkundig Laboratorium Assistent op de onderzoeksafdeling. Hier is haar belangstelling voor het uitvoeren van onderzoek ontstaan, maar ook de wens om dit te doen in een academische wereld in plaats van in een zeer gesloten research omgeving van een commercieel bedrijf.

In 1998 is ze gaan werken bij het Landelijk Referentie Centrum voor bevolkingsonderzoek op Borstkanker (LRCB) als Fysisch Technisch medewerker. Naast de kwaliteitscontroles van de mammografen, ontwikkelmachines en lichtkasten van het bevolkingsonderzoek, voerde ze samen met haar collega's van de Fysische Groep ook kleine onderzoeken uit naar het verbeteren van de bestaande kwaliteitsmetingen. Daarnaast werkte ze aan het ontwikkelen van nieuwe kwaliteitsmetingen op nieuwe technologieën, zoals de digitale mammografie.

In 2009 zei ze de techniek vaarwel en werd coördinator visitaties en opleidingen, eveneens bij het LRCB. Naast de typische coördinatortaken ging zij ook screeningsradiologen ondersteunen bij het maken van lesmateriaal, zoals testsets voor de radiologenopleiding bij het LRCB. En bij de visitaties werd ze al snel betrokken bij de inhoud van de radiologische onderdelen, zoals kwaliteitsindicatoren voor het monitoren van hun beoordelingsprestaties. Na goed ingewerkt te zijn op deze inhoud, is ze in 2012 begonnen met het uitvoeren van onderzoek naar mogelijke verbeteringen in de prestaties van de screeningsradiologen. Dit heeft uiteindelijk geleid tot dit proefschrift.

Ook na het afronden van dit promotietraject zal ze bij het LRCB werkzaam blijven en zich inzetten voor het verbeteren van de kwaliteit van het bevolkingsonderzoek op borstkanker.

Dankwoord

Jarenlang heb ik aan dit proefschrift gewerkt. Nu is het af en daar ben ik ontzettend trots op. Het was een intensief traject, soms uitdagend, maar vooral leerzaam en verrijkend. Natuurlijk heb ik dit niet alleen gedaan. Er zijn vele mensen om mij heen, (oud-)collega's, vrienden en familie, die ik dank verschuldigd ben. Jullie betrokkenheid, aanmoediging en hulp hebben een wereld van verschil gemaakt. Daarvoor wil ik jullie hartelijk bedanken. Een aantal personen wil ik graag in het bijzonder noemen.

Eigenlijk begon dit hele avontuur in 2013 met een presentatie over de resultaten van de visitaties van 1996-2013 bij de ECR in Wenen. Roland Holland had al meerdere jaren het idee dat al deze data eens bij elkaar gezet moest worden en dat dit wel eens een mooie paper kon opleveren. Deze paper kwam er in 2015. Roland, ik wil je graag bedanken voor deze inspiratie en dat je dit (mede) mogelijk hebt gemaakt.

In mijn jaargesprek in 2015 heb ik aangegeven dat het schrijven van deze eerste paper moeilijk was, maar dat ik wel graag verdere mogelijkheden zou krijgen in het uitvoeren van onderzoek. Ik zag in het schrijven van een tweede artikel de mogelijkheid om uitgedaagd te worden en me verder te kunnen ontwikkelen. Ruud en Piet, ik wil jullie graag bedanken dat jullie dit allemaal prima vonden en mij deze kans hebben geboden. Ruud, ik wil jou ook bedanken voor het vertrouwen in mij, dat je in 2017 voor mij een toegang tot promotie bij het UMC Utrecht hebt aangevraagd. Zij hadden minder vertrouwen in mij, maar jij bent tot de dag van vandaag in mij blijven geloven.

Mireille, ook jou wil ik graag bedanken voor al je vertrouwen. Toen je in 2020 hoogleraar werd bij het Radboudumc opende dit nieuwe deuren. Als mijn kamergenoot kende je mijn dromen en kwam met een nieuw plan. Je bood me aan voor mij een toegang tot promotie aan te vragen bij het Radboudumc en Ruud zou dan mijn 2e promotor worden. En dat lukte! Nu kon het avontuur echt beginnen. Ik ben je ook dankbaar voor je begeleiding. Als ik je nodig had, dan was je er. Ik heb veel van je mogen leren, o.a. dat je als een goede onderzoeker heel kritisch moet zijn. Dat ben jij als geen ander en dat was soms best lastig voor mij. Uiteindelijk kom ik daarna altijd tot inzicht dat je helemaal gelijkt hebt en ligt er nu een heel mooi resultaat. Ik bewonder je juist ook vanwege deze eigenschap. Daarnaast hebben we ook heel veel lol gehad samen, bijvoorbeeld tijdens congressen en onze tripjes naar Brazilië.

Daniëlle, ik wil jou graag bedanken dat je mijn co-promotor wilde zijn. Sinds 2018 zijn we collega's en trekken we veel met elkaar op. Niet alleen op het werk, maar ook privé. We hebben o.a. veel plezier gehad tijdens de spelletjesavonden, samen met Ruben en Danielle. Op het gebied van onderzoek heb ik veel van je geleerd. Ik hoop dat we nog lang samen mogen werken.

Lucien, amigo, het blijft altijd een feest om met jou samen te mogen werken. Of het nu voor de radiologentrainingen is of als mijn co-promotor, je bent altijd enthousiast en motiverend. Regelmatig wandelde je even binnen om te vragen hoe het met me was en om me een hart onder de riem te steken als dat nodig was. Je bent altijd bereikbaar voor vragen en razend snel met het beantwoorden van e-mails. En nu gaan we eindelijk het feestje vieren.

Alle overige co-auteurs, Janine, Ellen, Ard, Wikke, Joost, Eric, Willem en Maartje, ik wil jullie graag bedanken voor de fijne samenwerking en de input en feedback die ik van jullie heb mogen ontvangen.

Mijn oprechte dank gaat uit naar de leden van de manuscriptcommissie: prof. dr. Pim Assendelft, prof. dr. Chantal van Ongeval en prof. dr. Sabine Siesling. Ik waardeer het zeer dat jullie de tijd hebben genomen om mijn proefschrift zorgvuldig te beoordelen en bereid zijn om hierover met mij in gesprek te gaan tijdens de verdediging. Daarnaast wil ik prof. dr. Marjolein van de Pol, dr. Monique Dorrius en dr. Patricia Hugen hartelijk bedanken voor hun bereidheid om als opponenten op te treden.

Klaas en Frits, bedankt dat jullie mijn paranimfen willen zijn. Toen ik in 2009 de techniek vaarwel zei en coördinator visitaties en opleiding werd, waren jullie het die mij enthousiast maakten voor het werk van een screeningsradioloog. Ik heb van jullie mogen leren hoe ik naar een mammogram moet kijken en hoe complex jullie vak is. Dit wakkerde bij mij de wens aan om iets voor de screeningsradiologen te kunnen betekenen. Velen visitaties en trainingen hebben we met elkaar uitgevoerd, niet alleen in Nijmegen maar ook in het buitenland. Klaas, ik zal nooit vergeten hoe we samen in een Caddy, volgeladen met werkstations, naar Luxemburg reden om daar een paar dagen training te verzorgen. Wat heb ik daar een plezier aan beleefd. En Frits, ook al ben je met pensioen, afgelopen jaar ging je nog gezellig mee naar de Antillen voor een visitatie.

Collega's, en oud collega's, jullie wil ik ook graag bedanken voor de interesse die jullie hebben getoond, voor de ontspannen lunches en voor het geduld. Er zijn perioden geweest dat ik mijn normale werkzaamheden misschien niet de aandacht

heb gegeven die had gemoeten. Hierover hebben jullie niet geklaagd, maar juist begrip getoond. Ruben, ik wil jou in het bijzonder bedanken. We werken al 27 jaar samen en als 'kantoorburen' lopen we regelmatig bij elkaar binnen voor een praatje. Je bent zeer betrokken geweest en hebt me waardevolle feedback gegeven.

Al mijn leuke en lieve (schoon)familie en vrienden, jullie wil ik ook graag bedanken voor jullie interesse in mijn bezigheden. En zeker ook voor de vele gezellige momenten. Verjaardagen, familie- en vriendenweekenden en festivals, deze ontspannende momenten waren tijdens dit traject extra belangrijk voor mij. Jullie zijn mij allen zeer dierbaar, toch wil ik Mayke, Ankie, Corine en Antoinette speciaal bedanken voor de sauna-uitjes, de wandelingen, de kaartjes die ik kreeg en de traan naast de lach.

Lieve Else, wat ben ik gelukkig met jou als zus! Jij bent de liefste zus die ik maar kan wensen. Bedankt voor jouw onvoorwaardelijke liefde en steun. Ik weet dat je super trots op me bent. Ik vond het wel even wennen toen je terug verhuisde naar Zeeland. Maar de logeerpartijtjes van tegenwoordig zijn me ook wel zeer dierbaar.

Pap en mam, wat ben ik dankbaar dat jullie dit moment nog mogen meemaken. Bedankt voor de stevige basis die jullie me hebben gegeven. Jullie hebben nooit druk op mij gelegd om te presteren, dat deed ik zelf wel. Maar jullie hebben me wel altijd ondersteund in wat ik graag wilde bereiken en zijn altijd trots op mij. Ik hoop dat we nog heel veel mooie momenten samen mogen meemaken.

Niek, hoe kan ik jou bedanken. Jij betekent alles voor mij. Jij was het die mij ervan overtuigde dat ik dit kon en was het ook die mij er doorheen trok als ik helemaal stuk was. Zonder jou had ik echt de eindstreep nooit gehaald. Laten we proosten en het leven vieren! En uh, sorry voor het ontbreken van die stellingen…

Radboud

Addendum to Chapter 6

Applying the “positive predictive value – recall diagram” to monitor performance and provide recommendations for screening radiologists

Geertse TD, Tetteroo E, Smid-Geirnaerdt MJA, Duijm LEM, Pijnappel RM, van der Waal D, Broeders MJM.

European Radiology, September 2025, https://doi.org/10.1007/s00330-025-11978-3

This addendum contains Tables 4 and 5, which were unintentionally omitted from Chapter 6 of the thesis.

Table 4 Summary of recommendations during audits conducted during the study period and changes in recall rate after the audit, regarding initial screening

Reading unit	Audit[a]	Evaluation period	Mean RR evaluation period[b]	Recommendation[c]	RR in last year of evaluation period[d]	Difference in RR between last year of the evaluation period and:			
						1st year after audit [e]	2nd year after audit [e]	3rd year after audit [e]	target value is met after 1st year [f]
RU1	January 2019	2014-2017	5.9%	RR↓	6.5%	-2.1%	NA	NA	✓
RU2	April 2017	2012-2015	4.0%	RR↑	3.1%	+0.7%	+1.6%	+2.2%	✓
RU4	April 2015	2010-2013	7.5%	RR↓	9.8%	-1.4%	-0.5%	-1.6%	X
RU4	June 2018	2013-2016	9.0%	RR↓	9.3%	-4.1%	-3.9%	NA	✓
RU5	June 2015	2010-2013	5.9%	RR↓	7.2%	-0.2%	-0.3%	-1.0%	X
RU5	June 2018	2013-2016	6.6%	RR↓	6.9%	=0.0%	+1.2%	NA	X
RU6	March 2017	2012-2015	6.4%	RR↓	7.2%	-2.3%	-2.1%	-2.0%	✓
RU7	January 2016	2011-2014	4.9%	RR↓	6.0%	+1.8%	+3.1%	+1.4%	X
RU7	December 2018	2014-2017	7.4%	RR↓	9.1%	-3.9%	NA	NA	✓
RU10	December 2013	2010-2012	7.7%	RR↓	8.2%	-0.8%	-1.0%	-1.0%	X
RU10	November 2016	2012-2015	8.6%	RR↓	7.2%	+0.5%	+0.1%	+1.2%	X
RU11	June 2013	2010-2011	6.7%	RR↓	6.6%	+1.3%	-0.1%	+0.5%	X
RU11	May 2016	2011-2014	7.0%	RR↓	6.5%	+0.8%	-0.2%	-1.0%	X
RU12	February 2015	2010-2013	8.3%	RR↓	11.1%	-2.0%	-4.7%	-3.7%	X
RU12	March 2018	2013-2016	9.7%	RR↓	6.3%	-0.7%	=0.0%	NA	X

RU Reading unit, *RR* recall rate, *NA* not available

[a] For nine audits the changes in RR could not be assessed as the first year after the audit fell outside the study period (2020 or later), and are therefore not included in this table;

[b] At the time of the audit, all recommendations given by the audit team were based on the mean RR of the evaluation period;

[c] RR↓ = recommendation to decrease recall rate; RR↑ = recommendation to increase recall rate;

[d] For this study, the RR in the last year of the evaluation period was used as the 'baseline', against which changes in RR after the audit were compared;

[e] If the audit took place between 1 January and 30 June, the year in which the audit took place was also considered to be the first year after the audit; NA = not available, data outside study period;

[f] ✓= the target value is met; X = the target value is not met.

Table 5 Summary of recommendations during audits conducted during the study period and changes in recall rate after the audit, regarding subsequent screening

Reading unit	Audit[a]	Evaluation period	Mean RR evaluation period[b]	Recommendation[c]	RR in last year of evaluation period[d]	Difference in RR between last year of the evaluation period and:			
						1st year after audit [e]	2nd year after audit [e]	3rd year after audit [e]	target value is met after 1st year [f]
RU2	April 2017	2012-2015	1.3%	RR↑	1.1%	-0.1%	+0.3%	+0.6%	X
RU5	June 2015	2010-2013	1.6%	RR=	1.9%	+0.1%	+0.3%	-0.1%	✓
RU5	June 2018	2013-2016	1.9%	RR=	2.1%	=0.0%	+0.3%	NA	✓
RU7	January 2016	2011-2014	1.8%	RR=	2.1%	+0.4%	+0.4%	-0.1%	X
RU7	December 2018	2014-2017	2.3%	RR↓	2.5%	-1.0%	NA	NA	✓
RU10	December 2013	2010-2012	2.1%	RR↓	2.5%	-0.4%	-0.7%	-0.8%	✓
RU11	June 2013	2010-2011	2.4%	RR↓	2.4%	-0.1%	-0.2%	-0.4%	X
RU11	May 2016	2011-2014	2.3%	RR↓	2.2%	-0.6%	-0.8%	-0.7%	✓
RU12	February 2015	2010-2013	2.0%	RR↓	2.7%	-0.8%	-1.2%	-0.9%	✓

RU Reading unit, *RR* recall rate, *NA* not available

[a] For four audits the changes in RR could not be assessed as the first year after the audit fell outside the study period (2020 or later), and are therefore not included in this table;

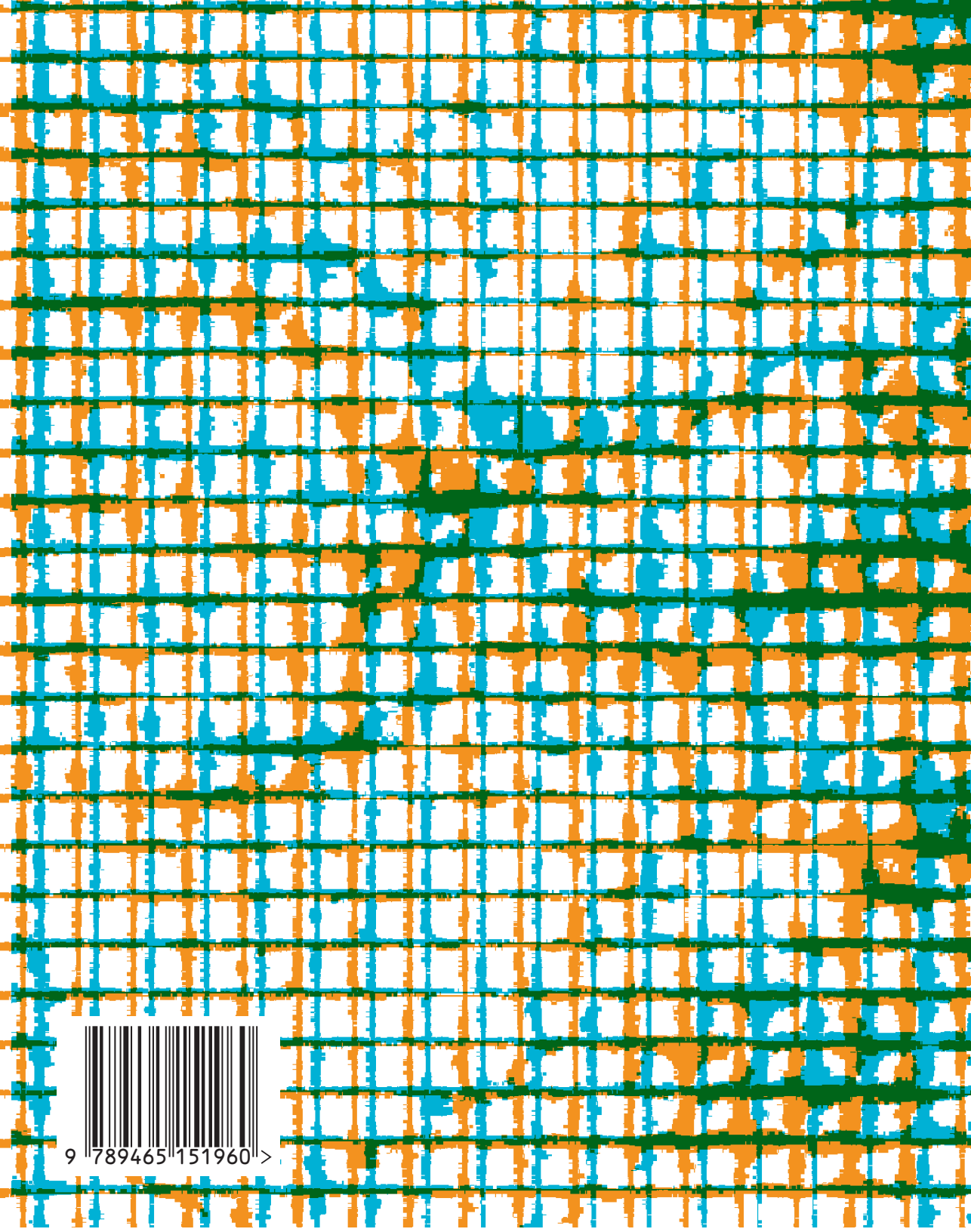
[b] At the time of the audit, all recommendations given by the audit team were based on the mean RR of the evaluation period;

[c] RR↓ = recommendation to decrease recall rate; RR↑ = recommendation to increase recall rate; RR= = recommendation to stabilise recall rate;

[d] For this study, the RR in the last year of the evaluation period was used as the 'baseline', against which changes in RR after the audit were compared;

[e] If the audit took place between 1 January and 30 June, the year in which the audit took place was also considered to be the first year after the audit; NA = not available, data outside study period;

[f] ✓= the target value is met; X = the target value is not met.



9 789465 151960 >

Radboudumc university medical center

IN DEI NOMINE FELICITER

Radboud University